Towards Understanding Practical Randomness Beyond Noise: Differential Privacy and Mixup

Hanshen Xiao and Srinivas Devadas

MIT, {hsxiao,devadas}@mit.edu

Abstract. Information-theoretical privacy relies on randomness. Representatively, Differential Privacy (DP) has emerged as the gold standard to quantify the individual privacy preservation provided by given randomness. However, almost all randomness in existing differentially private optimization and learning algorithms is restricted to noise perturbation. In this paper, we set out to provide a privacy analysis framework to understand the privacy guarantee produced by other randomness commonly used in optimization and learning algorithms (e.g., parameter randomness). We take *mixup*: a random linear aggregation of inputs, as a concrete example. Our contributions are twofold. First, we develop a rigorous analysis on the privacy amplification provided by *mixup* either on samples or updates, where we find the hybrid structure of *mixup* and the Laplace Mechanism produces a new type of DP guarantee lying between Pure DP and Approximate DP. Such an average-case privacy amplification can produce tighter composition bounds. Second, both empirically and theoretically, we show that proper *mixup* comes almost free of utility compromise.

1 Introduction

Differential privacy (DP) has emerged as a standard measure of the individual-level privacy risk during an aggregate analysis on a dataset. Informally, a differentially private algorithm maps any two close datasets to similar probability distributions over outputs and thus, from outputs observed, it is hard to distinguish the participation of an individual. In Dwork et al.'s pioneering work [DMNS06], such indistinguishability is parameterized by a positive real number ϵ in a multiplicative manner:

Definition 1 (Pure ϵ-DP). *A randomized algorithm $A : \mathcal{X} \to \mathcal{O}$, achieves ϵ-DP if for any adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ in $\mathcal{X}$, and any set S in the output domain $\mathcal{O}$ of $A(\cdot)$,*

$$\Pr[A(\mathcal{D}) \in S] \leq e^{\epsilon} \Pr[A(\mathcal{D}') \in S]. \tag{1}$$

Here, we call two datasets $\mathcal{D}$ and $\mathcal{D}'$ adjacent if $\mathcal{D}$ and $\mathcal{D}'$ only differ in one data point, denoted by $\mathcal{D} \sim \mathcal{D}'$ in the following. Stemming from (1), there is a long line of works to relax the original metric to measure the difference between the distributions of $A(\mathcal{D})$ and $A(\mathcal{D}')$ in Definition 1, for example, approximate (ϵ, δ)-DP [DKM^{+}06], where under the same setup a failure probability of (1) at most δ is admitted:

Definition 2 (Approximate (ϵ, δ)-DP). *A randomized algorithm $A : \mathcal{X} \to \mathcal{O}$, achieves (ϵ, δ)-DP if for any adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ in $\mathcal{X}$, and any set S in the output domain $\mathcal{O}$ of $A(\cdot)$,*

$$\Pr[A(\mathcal{D}) \in S] \leq e^{\epsilon} \Pr[A(\mathcal{D}') \in S] + \delta. \tag{2}$$

Other variants include concentrated DP [DR16, BS16, BDRS18], Renyi DP [Mir17] and the recently proposed f-DP [DRS19, BDLS19]. Those relaxations provide versatile frameworks to analyze a larger class of randomized algorithms with tighter bounds when handling composition, i.e., the cumulative privacy risk under repetition of mechanisms on one dataset. However, compared to sharpened composition control, another issue that usually gets overlooked is *how to introduce randomness for privacy preservation?*

Randomness beyond Noise: The simplest way to randomize an algorithm is perturbation. For example, to make a deterministic algorithm A satisfy ϵ-DP, one can add Laplace noise in a scale of the sensitivity, i.e., $\max_{\mathcal{D},\mathcal{D}'} \|A(\mathcal{D}) - A(\mathcal{D}')\|$, to the output [DMNS06]. In general, DP does not come for free: lower bounds on utility loss in many tasks are known, for example, (strongly) convex optimization [BST14, TTZ15] and Principal Components Analysis (PCA) [CSS12, DTTZ14], etc.

Though privacy is usually not free of utility loss, it does **not** mean randomness will always come with a performance compromise. The purposes to introduce randomness in optimization and learning are far more than privacy preservation, for example, stochastic gradient Langevin dynamics (SGLD) [MWZZ18, LCLC19] for nonconvex optimization and uniform noise perturbed gradient descent to escape saddle points [JGN$^+$17]. Randomness can even strengthen the training performance, e.g., random dropout [SHK$^+$14] and data augmentation [SK19]. Generally speaking, data augmentation represents a large class of methods to improve robustness and reduce memorization (instead of generalization), especially in computer version: Training is conducted on similar but different virtual examples compared to the raw data through random cropping [KSH12], erasing [ZZK$^+$20] and mixup [ZCDLP18], etc. However, compared to simple noise perturbation, algorithm-oriented randomness has been rarely formally studied from a privacy-preservation viewpoint.
Restricted Randomness: Indeed, typically algorithm-oriented randomness does not produce reasonable DP guarantees, or a controllable (high-probability) worst-case distinguishability, as described in Definition 1 and 2. The reason is twofold: the random operators are usually localized and data dependent. For example, consider random dropout (pruning) [SHK$^+$14], where a node in a neural network is ignored independently with some fixed rate, or random erasing [ZZK$^+$20], where a rectangle region of an image is randomly erased and replaced with random values. In a network or an image of privacy concern processed by the above mechanisms, the random transformation is multiplicative over the private input while the output is restricted to a bounded domain determined by the specific input processed. As a result, given proper sensitivity restriction, say two images differing in one pixel, when random cropping is applied, one can still successfully distinguish the two images with a constant probability, if the distinct pixel is not erased. A similar argument holds for the saddle point escaping algorithm [JGN$^+$17], where the gradient is perturbed by a bounded noise uniformly selected from a sphere.

Though practical randomness may not necessarily lead to DP, greater randomness potentially implies better privacy. As a first step to formalize the privacy gain and utility of practical or heuristic randomness, we will use DP as the primitive. To this end, we consider a natural hybrid structure of both kinds of randomness, for example, Laplace noise and *mixup*.
Mixup: In this paper, we use *mixup* to denote a simple aggregation structure with random weights. Given N inputs $x_1, x_2, ..., x_N$, *mixup* outputs $\sum_{i=1}^{N} \omega_i x_i$, with random $\omega_i \in (0,1)$ and $\sum_{i=1}^{N} \omega_i = 1$. One successful example of *mixup* is [ZCDLP18], where a surprisingly efficient data augmentation is described: Given the raw data $(\boldsymbol{x}_i, y_i)$, $i = 1, 2, ..., N$, where $\boldsymbol{x}_i$ is the feature and y_i is the associated label, a virtual training sample $(\tilde{\boldsymbol{x}}, \tilde{y})$ is constructed by:

$$\tilde{\boldsymbol{x}} = \lambda \boldsymbol{x}_{i_1} + (1-\lambda)\boldsymbol{x}_{i_2}, \tilde{y} = \lambda y_{i_1} + (1-\lambda) y_{i_2}. \tag{3}$$

Here, $(\boldsymbol{x}_{i_1}, y_{i_1})$ and $(\boldsymbol{x}_{i_2}, y_{i_2})$ are randomly drawn while $\lambda \in (0,1)$ is a random variable selected. *Mixup* based data augmentation has been widely studied and shown to be very powerful empirically in thorough experiments and subsequent works [BCG$^+$19], [TCB$^+$19], [PXZ19]. This raises two interesting questions: *On privacy, what kind of privacy amplification is provided by mixup? On utility, are there other applications of mixup but with theoretical performance guarantees?* In this work, we set out to answer the two questions.
Existing Privacy Preservation with Mixup: The shuffled and randomly composite samples in *mixup* seem to provide some natural privacy protection. However, it is noted that the mixed samples generated using (3) are restricted within the convex hull of original samples. Similarly, *mixup* itself is not sufficient to produce reasonable DP guarantees, though its potential privacy preservation via some other measurements such as computational hardness is still an open question. Based on *mixup*, there are several appealing proposals such as *Instahide* [HSLA20], for private training, and *Datamix* [LWG$^+$20], for private inference. The authors of Instahide conjecture that breaking sign-flipping equipped *mixup* can be reduced to that of the subset sum problem. Unfortunately, as far as we know, existing heuristic *mixup* related data privacy preservations have not been studied under any systematic privacy notions. Consequently, Carlini et al. [CDG$^+$20] and Chen et al. [CSZ20] point out several vulnerabilities and potential attacks on Instahide.

1.1 A High-level Picture of Methodology and Contribution

One of our main results is that the hybrid of mixup (or the other algorithmic randomizations listed above) and the regular Laplace mechanism produces a similar worst-case guarantee compared to the pure Laplace mechanism, but generally improves the average-case privacy loss, which leads to a tighter composition bound. We introduce the following alternative definition of the ϵ privacy loss defined in (1) to give a more refined, formal illustration.

Definition 3 **([LNR$^+$17]).** *The* ex-post *privacy loss $\epsilon(o)$ for an algorithm A on an output o is defined as*

$$\epsilon(o) = \sup_{\mathcal{D},\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(o) = \sup_{\mathcal{D},\mathcal{D}'} \log(\frac{\mathbb{P}(A(\mathcal{D}) = o)}{\mathbb{P}(A(\mathcal{D}') = o)}),$$

for arbitrary $\mathcal{D} \sim \mathcal{D}'$.

Definition 3 provides a point-wise measurement on the privacy loss where $\epsilon_{\mathcal{D},\mathcal{D}'}(o)$ captures, for any specific output o, the likelihood ratio between two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$. By taking a supremum over all pairs $(\mathcal{D}, \mathcal{D}')$, $\epsilon(o)$ provides a worst-case ϵ loss at output o. It is not hard to observe that under the Laplace Mechanism, where a pure ϵ-DP is produced, $\epsilon(o) = \epsilon$ uniformly. In contrast, under a Gaussian Mechanism with an approximate (ϵ, δ)-DP, $\sup_o \epsilon(o) = \infty$ is unbounded, and we have to resort to the failure probability δ.[1]

In general, the output of the hybrid structure can be viewed as a mixture of some restricted randomness and noise. We find that the corresponding privacy guarantee is between pure DP and approximate DP. To be specific, the hybrid structure of *mixup* and the Laplace mechanism provides a bounded worst-case privacy guarantee $\sup_o \epsilon(o) \leq \epsilon_0$, for some constant ϵ_0, whereas $\epsilon(o)$ does not uniformly equal to ϵ_o, where there exists o such that $\epsilon(o) < \epsilon_0$.

Another way to characterize such average-case amplification is to view the privacy loss under a relaxed divergence metric. For example, if we measure the privacy loss through KL-divergence,

$$D(A(\mathcal{D})||A(\mathcal{D}')) = \mathbb{E}_{o \leftarrow A(\mathcal{D})} \log \big[\frac{\mathbb{P}(A(\mathcal{D})(o)}{\mathbb{P}(A(\mathcal{D}')(o)}\big],$$

the hybrid structure is strictly better than that of a pure Laplace, which implies a sharper (advanced) composition bound when applying concentration inequalities [DRV10]. In the rest of the paper, we set out to quantify the privacy amplification factor and understand the impact of algorithmic randomness on utility.

Contributions and Organization: In this paper, we have two main contributions. First, we systematically study the privacy amplification from the convolution of random mixing and regular noise (Laplace) mechanisms. The privacy framework presented here can be used to study the privacy gain from *restricted and localized randomness* which captures many empirical training improvements such as the data augmentation techniques listed earlier. Second, we propose a new application of *mixup* to mix the immediate updates during (decentralized) optimization. To understand the utility and privacy tradeoff, we provide results regarding the effect of *mixup* on the convergence rate.

The remainder of the paper can be summarized as follows. In Section 2, we provide an analysis of sample mixing as described by (3) and show that sample mixing produces a privacy amplification factor in a simple compositional setting of iterative optimization. In Section 3, we describe another model of the hybrid mixup and noise architecture, namely, update mixing during optimization. In Section 4, we show how to incorporate the proposed update mixing architecture into more complicated decentralized algorithms. Two concrete examples, Modified private ADMM (Algorithm 1) and (Decentralized) SGD (Algorithm 2), are proposed, and we further study the local differential privacy (LDP) amplification. On the utility side, in Section 5, we theoretically prove that proper *mixup* on immediate updates during optimization comes almost free of utility compromise.

2 Hybrid Architecture of Mixup and Noise: Sample Mixing

Differentially private (Stochastic) Gradient Descent ((S)GD) and its variants have been extensively studied [BST14, LNR$^+$17, CMS11, JT13, WGX18, WYX17, WX19]. A common strategy is to perturb the gradient in each iteration with well-scaled noise to keep track of the cumulative privacy loss. In this section, we provide an analysis of sample mixing as described by (3) in the context of private optimization.

Imagine we run SGD to minimize the empirical loss over samples $\mathcal{S} = \{\boldsymbol{s}_i, i = 1, 2, ..., n\}$ with *mixup*, where we use $\boldsymbol{s}_i$ to denote $(\boldsymbol{x}_i, y_i)$ and the objective loss function of the parameter $\boldsymbol{\theta}$ is defined as $\sum_{i=1}^n f(\boldsymbol{\theta}, \boldsymbol{s}_i)$. Across each iteration, we assume two samples $\boldsymbol{s}_{i_1}$ and $\boldsymbol{s}_{i_2}$ are randomly selected from $\mathcal{S}$ and mixed as $\tilde{\boldsymbol{s}} = \lambda \boldsymbol{s}_{i_1} + (1 - \lambda)\boldsymbol{s}_{i_2}$ with

[1] It is not hard to verify that $\epsilon(o) \leq \epsilon$ with probability at least $1 - \delta$, also termed as probabilistic DP [GMW$^+$11], is stronger than the (ϵ, δ) approximate DP notion defined in (2).

some random weight $\lambda \in (0,1)$. Thus, in general, an SGD protocol to privately update $\boldsymbol{\theta}$ with mixed samples can be described as follows:

$$\boldsymbol{\theta}^k = \boldsymbol{\theta}^{k-1} - \eta_k \underline{\nabla f(\boldsymbol{\theta}^{k-1}, \lambda_k \boldsymbol{s}_{i_{k,1}} + (1-\lambda_k)\boldsymbol{s}_{i_{k,2}})} + \boldsymbol{\Delta}^k. \tag{4}$$

Here, $i_{k,1}$ and $i_{k,2}$ are two different random indexes independently sampled from $[1:n]$ and λ_k is randomly drawn from $(0,1)$ in iteration k. $\boldsymbol{\Delta}^k$ is the noise added, of which each dimension is assumed to be a Laplace, $\mathrm{Lap}(0,\beta)$, with probability density $\mathbb{P}(z) = \beta/2 \cdot e^{-\beta|z|}$. For quantification, we assume the l_∞ sensitivity of the gradient, i.e., $\sup_{\boldsymbol{\theta},\boldsymbol{s},\boldsymbol{s}'} \|\nabla f(\boldsymbol{\theta},\boldsymbol{s}) - \nabla f(\boldsymbol{\theta},\boldsymbol{s}')\|_\infty \leq \mathcal{B}$, for two arbitrary sample candidates $\boldsymbol{s}$ and $\boldsymbol{s}'$.

Example 1: Assume that the gradient $\nabla f(\boldsymbol{\theta},\boldsymbol{s})$ is linear to the sample $\boldsymbol{s}$ (for example $f(\cdot)$ is a ridge regression), and samples vary between $[0,1]$. Ignoring the constants $\boldsymbol{\theta}^{k-1}$ and η_k in (4), two resultant mixture distributions of $\boldsymbol{\theta}^k$ can be: Case 1, imagine two samples $s_{i_{k,1}} = 0$ and $s_{i_{k,2}} = 1$ are selected and mixed, where for simplicity we assume $\nabla f(\boldsymbol{\theta}^{k-1}, 0) = 0$ and $\nabla f(\boldsymbol{\theta}^{k-1}, 1) = 1$. Then, the distribution is equivalent to a mixture of a Laplace and a uniform distribution between 0 and 1, written as $Lap(0,\beta) * U[0,1]$, where $*$ denotes convolution of distributions and $U[p,q]$ represents the uniform distribution between $[p,q]$; Case 2, if selected samples are identical, say $s_{i_{k,1}} = s_{i_{k,2}} = 1$, then the mixed sample is still 1 and the corresponding distribution becomes a pure Laplace $Lap(1,\beta)$.

The above example captures the underlying key problem we study here: what is the additional privacy gain from the convolution of a (sample-dependent) randomness, whose support set is bounded, and the Laplace Mechanism compared to the pure Laplace? Intuitively, the worst-case privacy loss is still preserved by the Laplace Mechanism but the additional randomness from mixing *smoothens* the divergence on average, which would be helpful especially when handling the composition. Indeed, our following analysis matches this intuition, where we prove that sample mixing as used in (4) produces an amplification factor determined by $\tau = \beta\mathcal{B}$.

Without loss of generality, we assume the step size $\eta_k = 1$ and only consider the one-dimensional case; the multi-dimensional analysis is a straightforward composition. To capture the privacy analysis of the updating procedure (4), for simplicity, we assume the gradient of randomly-mixed samples $\nabla f(\boldsymbol{\theta}^{k-1}, \lambda_k \boldsymbol{s}_{i_{k,1}} + (1-\lambda_k)\boldsymbol{s}_{i_{k,2}})$ (underlined in (4)) is uniformly distributed between $\boldsymbol{s}_{i_{k,1}}$ and $\boldsymbol{s}_{i_{k,2}}$[2]. Thus, the l_∞ sensitivity bound $\mathcal{B}$ on the gradient is equivalent to that of the samples. With the above setup, the distribution of $\boldsymbol{\theta}^k$ is equivalent to $Lap(0,\beta) * U[\boldsymbol{s}_{i_{k,1}}, \boldsymbol{s}_{i_{k,2}}]$, where we ignore the constant $\boldsymbol{\theta}^{k-1}$, the earlier update from iteration $(k-1)$, behaving as a uniform shift on the distribution.

The following theorem states that the privacy loss of (4) satisfies a bounded $\sup_{\boldsymbol{\theta}^k} \epsilon(\boldsymbol{\theta}^k) = \frac{2}{n} \cdot \log(\frac{e^{\mathcal{B}\beta}-1}{\mathcal{B}\beta})$. In comparison, without mixup, a pure Laplace mechanism will produce $\epsilon(\boldsymbol{\theta}^k) = \frac{\mathcal{B}\beta}{n}$ [DMNS06] uniformly for any $\boldsymbol{\theta}^k$. Moreover, we apply (ϵ,δ) measurement, defined in Definition 2 as a high probability bound on ϵ, to quantify the privacy amplification, shown below.

Theorem 1 *Let $\tau = \beta\mathcal{B}$. With the above setup and assumption on the data mixing, sensitivity and noise, the hybrid data mixing and Laplace Mechanism shown in (4) satisfies $\sup_{\boldsymbol{\theta}^k} \epsilon(\boldsymbol{\theta}^k) = \frac{2}{n} \cdot \log(\frac{e^\tau - 1}{\tau})$. In an approximate DP view, if $\delta > (1-e^{-\tau})/(2\tau)$, it produces an (ϵ,δ)-DP such that*

$$\epsilon = \frac{2}{n} \log\big(\max\{\frac{2e^{\tau-\beta\psi(\delta)} - e^{\tau-2\beta\psi(\delta)} - 1}{\mathcal{B}}, \frac{\mathcal{B} - 1/(2\delta)}{1/(2\delta)} \cdot \frac{e^{\beta/(2\delta)}(1-e^{-\beta(\mathcal{B}-1/(2\delta))})}{1-e^{-\beta(\mathcal{B}-1/(2\delta))}}\}\big) \tag{5}$$

where $\psi(\delta) = \frac{\mathcal{B}}{1-e^{-\tau/2}} \cdot (\delta - \frac{1-e^{-\tau}}{2\tau})$.

Proof. Without loss of generality (w.l.o.g.), we assume that samples are within $[0,\mathcal{B}]$. For two adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$, each of n elements, we assume the differing elements are s and s' in $\mathcal{D}$ and $\mathcal{D}'$, respectively. Recall the subsampling of two samples defined in the SGD protocol (4). It is of probability $\frac{2}{n}$ that one may select the s from $\mathcal{D}$ or s' from $\mathcal{D}$, which determines the subsampling factor in the $\epsilon(\boldsymbol{\theta}^k)$ bound [BBG18]. In the following, we only need to consider the case where the differing sample is selected. Let s_0 denote the other element selected from the intersection of $\mathcal{D} \cap \mathcal{D}'$. Therefore, to derive an upper bound of $\epsilon(\boldsymbol{\theta}^k)$, we need to compare the two distributions $U[s,s_0] * Lap(0,\beta)$ and $U[s',s_0] * Lap(0,\beta)$. In the following, we use x to denote the output $\boldsymbol{\theta}^k$.

[2] Here, we indeed capture the form $\nabla f(\theta^{k-1}, \lambda_k \boldsymbol{s}_{i_{k,1}} + (1-\lambda_k)\boldsymbol{s}_{i_{k,2}} + \boldsymbol{\Delta}^k)$ with perturbation directly on samples. Thus, the gradient function $\nabla f(\cdot)$ becomes a postprocessing and we only need to focus on the term $\lambda_k \boldsymbol{s}_{i_{k,1}} + (1-\lambda_k)\boldsymbol{s}_{i_{k,2}} + \boldsymbol{\Delta}^k$.

First, we consider an extreme case, if $s_0 = s'$ or $s_0 = s$, where w.l.o.g. we assume $s_0 = s' = 0$. Thus, $U[s', s_0] * Lap(0, \beta)$ is reduced to $Lap(0, \beta)$ and $U[s, s_0] * Lap(0, \beta)$ becomes $U[0, s_0] * Lap(0, \beta)$. The probability density of $U[0, s_0] * Lap(0, \beta)$ can be described as

$$\mathbb{P}(x) = \begin{cases} \frac{e^{\beta x}(1-e^{-\beta s_0})}{2s_0} & x < 0 \\ \frac{2-e^{-\beta x}-e^{-\beta(s_0-x)}}{2s_0} & 0 \leq x < s_0 \\ \frac{e^{-\beta(x-s_0)}(1-e^{-\beta s_0})}{2s_0} & x \geq s_0. \end{cases} \tag{6}$$

When $x < 0$, it is straightforward to see that

$$e^{\epsilon} \leq \frac{Lap(0, \beta)(x)}{U[0, s_0] * Lap(0, \beta)(x)} \leq \frac{\beta e^{x\beta}}{\frac{1}{\mathcal{B}} \cdot e^{x\beta}(1 - e^{-\beta \mathcal{B}})} = \frac{\beta \mathcal{B}}{1 - e^{-\beta \mathcal{B}}},$$

where the equality is achieved when $s_0 = \mathcal{B}$. On the other hand, if $x \geq 0$, the maximal of ϵ is achieved when $x \geq \mathcal{B}$ and $s = \mathcal{B}$, where $e^{\epsilon} \leq \frac{e^{\beta \mathcal{B}}-1}{\beta \mathcal{B}}$. Moreover, it is easy to verify that:

$$\frac{e^{\beta \mathcal{B}} - 1}{\beta \mathcal{B}} \geq \frac{\beta \mathcal{B}}{1 - e^{-\beta \mathcal{B}}},$$

which is equivalent to $2 + (\beta \mathcal{B})^2 \leq e^{-\beta \mathcal{B}} + e^{\beta \mathcal{B}}$, for arbitrary β and $\mathcal{B}$. Thus, we have $\epsilon \leq \log(\frac{e^{\beta \mathcal{B}}-1}{\beta \mathcal{B}})$.

Second, if $s \neq s_0$ and $s' \neq s_0$, we have to consider the following two scenarios: (a). $s < s_0 \leq s'$, where w.l.o.g. we set $s = 0$. (b). $s_0 \leq s < s'$, where w.l.o.g. we set $s_0 = 0$.

In (a), for the upper bound of e^{ϵ}, we have the following fact: for $x < 0$, the ratio e^{ϵ}, which equals

$$\frac{\frac{1}{s_0} \cdot (1 - e^{-\beta s_0})}{\frac{1}{s'-s_0} e^{-\beta s_0}(1 - e^{-\beta(s'-s_0)})}, \tag{7}$$

is a non-decreasing function with respect to s_0. Therefore, for $s_0 \in [0, s']$, the upper bound is achieved when $s_0 = s'$. Here, we use the following fact that $\lim_{x \to 0} \frac{1-e^{-\alpha x}}{x} = \alpha$. Furthermore, it is noted that the function $\frac{e^{\beta s'}-1}{\beta s'}$ increases with increasing s'. To conclude, the extreme case of ϵ when $x < 0$ happens when $s_0 = s' = \mathcal{B}$. Similarly, for $x \in [0, s_0]$, we have the following observation on the probability density ratio:

$$\frac{s' - s_0}{s_0} \frac{2 - e^{-\beta x} - e^{-\beta(s_0-x)}}{e^{-\beta(s_0-x)}(1 - e^{-\beta(s'-s_0)})} \leq \frac{s' - s_0}{s_0} \frac{1 - e^{-\beta s_0}}{e^{-\beta s_0}(1 - e^{-\beta(s'-s_0)})},$$

which is equivalent to $(e^{-\beta x} - 1)^2 \geq 0$. Due to the symmetry, therefore, the upper bound of e^{ϵ} in (a) also happens when $s_0 = s' = \mathcal{B}$, which is reduced to the first case. With a similar reasoning, we can show that the extreme case of (b) also happens when $s_0 = s = 0$ while $s' = \mathcal{B}$. Thus, we have proved that $\epsilon \leq \log(\frac{e^{\beta \mathcal{B}}-1}{\beta \mathcal{B}})$.

In the following, we turn to characterize such privacy amplification with (ϵ, δ)-DP language, i.e., a high probability bound of ϵ.

We need to consider two scenarios (a). $0 = s \leq s_0 \leq s' \leq \mathcal{B}$; (b). $0 = s_0 \leq s \leq s' \leq \mathcal{B}$. First, we still consider the case when $s = s_0$. For $x \sim Lap(0, \beta)$, $\Pr(x \leq 0) = \frac{1}{2}$ and for $x \leq 0$, the probability density ratio between $Lap(0, \beta)$ and $U[0, s'] * Lap(0, \beta)$ is a constant equaling $\frac{\beta \mathcal{B}}{1-e^{-\beta \mathcal{B}}}$. Now, when $s \neq s_0$, back to (a), for $x \sim U[0, s_0] * Lap(0, \beta)$, it is easy to verify that $\Pr(x \leq \frac{s_0}{2}) = \frac{1}{2}$ and for $x \geq \frac{s_0}{2}$, the density ratio of $U[0, s_0] * Lap(0, \beta)(x)$ over $U[s_0, s'] * Lap(0, \beta)(x)$ is decreasing as x gets larger. From the earlier analysis, we know that when $x < 0$, where $\Pr(x < 0) = \frac{1-e^{-\beta \mathcal{B}}}{2\beta \mathcal{B}}$, the density ratio is a constant upper bounded by $\frac{e^{\beta \mathcal{B}}-1}{\beta \mathcal{B}}$. For $t \in [0, \frac{s_0}{2}]$, $\Pr(x \leq t) = \frac{1-e^{-\beta \mathcal{B}}}{(2-e^{-\beta x}-e^{-\beta(s_0-x)})\beta \mathcal{B}} + \frac{t}{s_0} - \frac{1}{2\beta \mathcal{B}}(1 - e^{-\beta t} + e^{-\beta(s_0-t)} - e^{-\beta s_0})$.

Therefore, provided a failure probability $\delta > \frac{1-e^{-\beta s_0}}{2\beta s_0}$, since $2 - e^{-\beta x} - e^{-\beta(s_0-x)}$, for $x \in (0, s_0)$, is upper bounded by $2(1 - e^{-\beta s_0/2})$ and thus $\mathbb{P}(x) = \frac{2-e^{-\beta x}-e^{-\beta(s_0-x)}}{2s_0} \leq \frac{1-e^{-\beta s_0/2}}{s_0}$ for $0 \leq x < s_0$, we have

$$\Pr\left(x \leq \frac{s_0}{1 - e^{-\beta s_0/2}}(\delta - \frac{1 - e^{-\beta s_0}}{2\beta s_0})\right) \leq \delta.$$

Let $\psi(\delta) = \frac{s_0}{1-e^{-\beta s_0/2}}(\delta - \frac{1-e^{-\beta s_0}}{2\beta s_0})$ and substitute the above into the expression of e^{ϵ}, then we have the relationship of (ϵ, δ) that if $\psi(\delta) > 0$,

$$\begin{aligned} e^{\epsilon} &= \frac{\frac{1}{s_0}(2 - e^{-\beta x} - e^{-\beta(s_0 - x)})}{\frac{1}{s'-s_0}e^{-\beta(s_0-x)}(1 - e^{-\beta(s'-s_0)})} \\ &\leq \frac{(2 - e^{-\beta\psi(\delta)} - e^{-\beta(s_0 - \psi(\delta))})(\mathcal{B} - s_0)}{1 - e^{-\beta(\mathcal{B}-s_0)}} \cdot \frac{e^{\beta(s_0 - (\delta - \frac{1-e^{-\beta\mathcal{B}}}{2\beta\mathcal{B}})/s_0)}}{s_0} \\ &\leq \frac{(2 - e^{-\beta\psi(\delta)} - e^{-\beta(\mathcal{B} - \psi(\delta))})e^{\beta(\mathcal{B}-\psi(\delta))}}{\mathcal{B}}, \end{aligned} \tag{8}$$

where the third inequality is because the product term on the right hand side of the second row of (8) is non-decreasing in s_0 and so we set $s_0 = \mathcal{B}$.

If $\psi(\delta) < 0$, we have that the extreme case happens when s_0 is such that $\frac{1-e^{-\beta\mathcal{B}}}{2\mathcal{B}} < \delta = \frac{1-e^{-\beta s_0}}{2s_0}$, where $s_0 < \frac{1}{2\delta}$ and

$$e^{\epsilon} \leq \frac{\frac{1-e^{-\beta s_0}}{2s_0}}{\frac{e^{-\beta s_0}(1-e^{-\beta(\mathcal{B}-s_0)})}{2(\mathcal{B}-s_0)}} \leq \frac{\mathcal{B} - 1/(2\delta)}{1/(2\delta)} \cdot \frac{e^{\beta/(2\delta)}(1 - e^{-\beta(\mathcal{B}-1/(2\delta))})}{1 - e^{-\beta(\mathcal{B}-1/(2\delta))}}.$$

With similar reasoning, in (b), if $0 = s_0 \leq s \leq s' \leq \mathcal{B}$, for $x \sim U[0, s] * Lap(0, \beta)$, once $\delta > \frac{1-e^{-\beta\mathcal{B}}}{2\beta\mathcal{B}}$, e^{ϵ} can be upper bounded by $\frac{\beta\mathcal{B}}{1-e^{-\beta\mathcal{B}}}$. On the other hand, if $x \sim U[0, s'] * Lap(0, \beta)$, it is not hard to observe that for either $x < 0$ or $x \geq 0$, the bound of ϵ is strictly controlled by that derived in (a). Thus, the claim holds.

In Theorem 1, we characterize the privacy amplification in a (ϵ, δ) form. In the following, we propose another metric to quantify the additional gain, i.e., in the worst case, how much privacy loss can be saved in expectation compared to the pure Laplace Mechanism. Stemming from Definition 3, in (4), under a pure Laplace Mechanism without sample mixing, for arbitrary $\mathcal{D}$ and output $\boldsymbol{\theta}^k$, $\sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}^k)$ uniformly equals $\mathcal{B}\beta/n$. Thus, under the hybrid sample mixing and Laplace structure, we define the ratio

$$\gamma = \frac{\sup_{\mathcal{D}} \mathbb{E}_{\mathcal{D}} \ \sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}^k)}{\beta\mathcal{B}/n}, \tag{9}$$

where $\mathbb{E}_{\mathcal{D}}$ denotes that $\boldsymbol{\theta}^k$ is produced based on dataset $\mathcal{D}$, to capture the expected ex-post loss savings in the worst case. The following theorem upper bounds the ratio. We calculate the expected ϵ-loss in the worst case.

Theorem 2 *With the same setup, the expected privacy loss $\mathbb{E}_{\boldsymbol{\theta}^k}\epsilon(\boldsymbol{\theta}^k)$ in (4) over the distribution of $\boldsymbol{\theta}^k$ in the worst case satisfies that for any $\mathcal{D} \sim \mathcal{D}'$,*

$$\begin{aligned} \sup_{\mathcal{D}} \mathbb{E}_{\mathcal{D}} \sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}^k) \leq \frac{2}{n} \cdot \max\big\{ \log(\frac{\beta\mathcal{B}}{2(1-e^{\beta\mathcal{B}/2})}), \max_{s \leq s_0 \in [0,\mathcal{B}]} \big[\log(\frac{s_0(1-e^{-\beta(s_0-s)})}{(s_0-s)(1-e^{-\beta s_0})})\big\} \\ + \log\big(\frac{(\mathcal{B}-s_0)(7e^{\beta(s_0-s)} - (12\beta(s_0-s)+3))}{6\beta(1-e^{-\beta(\mathcal{B}-s_0)})(s_0-s)^2} + \frac{e^{\beta(s_0-s)}(\mathcal{B}-s_0)(1-e^{\beta(s_0-s)})^2}{2\beta(s_0-s)^2(1-e^{-(\mathcal{B}-s_0)})}\big)\big], \end{aligned} \tag{10}$$

where $\mathbb{E}_{\mathcal{D}}$ denotes that $\boldsymbol{\theta}^k$ is produced based on dataset $\mathcal{D}$.

Proof. With a similar reasoning as shown in the proof of Theorem 1, we only need to consider two adjacent datasets in a form $\mathcal{D} = \{s, s_0\}$ and $\mathcal{D}' = \{s', s_0\}$, where $s, s', s_0 \in [0, \mathcal{B}]$.

First, we consider an extreme case if $s_0 = s$, which produces a pure Laplace distribution. To determine the $\sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(x)$, when $x < s$, to maximize the ratio, s' should equal $\mathcal{B}$ and

$$e^{\sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(x)} = \frac{\beta e^{-\beta(s-x)}}{\frac{1}{\mathcal{B}-s}e^{-\beta(s-x)}(1 - e^{-\beta(\mathcal{B}-s)})} = \frac{\beta(\mathcal{B}-s)}{1 - e^{-\beta(\mathcal{B}-s)}}.$$

Symmetrically, when $x \geq s$, s' should be set as 0 and $e^{\sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(x)} = \frac{\beta e^{-\beta x}}{\frac{1}{s} e^{-\beta x}(1-e^{-\beta s})} = \frac{\beta s}{1-e^{-\beta s}}$. Thus, since when $x \sim Lap(s, \beta)$, $\Pr(x < s) = \Pr(x \geq s) = \frac{1}{2}$,

$$\mathbb{E}_{x \sim Lap(s,\beta)} \sup_{s'} \epsilon_{s,s'}(x) = \quad \frac{1}{2}\big(\log(\frac{\beta(\mathcal{B}-s)}{1-e^{-\beta(\mathcal{B}-s)}}) + \log(\frac{\beta s}{1-e^{-\beta s}})\big). \tag{11}$$

With some calculation, since $s(\mathcal{B}-s)$ achieves the maximal when $s = \frac{\mathcal{B}}{2}$, we have $\sup_s \mathbb{E}_{x \sim Lap(s,\beta)} \sup_{s'} \epsilon_{s,s'}(x) = \log(\frac{\beta \mathcal{B}}{2(1-e^{\beta \mathcal{B}/2})})$.

Now, we turn to the more generic case, where we assume $s \neq s_0$. There are four cases regarding s', s, s_0, which can be (a). $s' \leq s_0 < s$; (a)'. $s' \leq s < s_0$; (b). $s < s_0 \leq s'$; (b)'. $s_0 < s \leq s'$. First, to simplify the study on the upper bound of the ratio, we show there is no need to consider cases (a)' and (b)'. To show this, we need the following trivial fact: for two positive numbers z_1 and z_2 and an arbitrary weight $w \in [0,1]$, if $wz_1 + (1-w)z_2 \leq z_2$, then clearly $z_2 \geq z_1$. Back to our cases, recall the convolutional density function $U[0,s] * Lap(0,\beta)(x) = \int_0^s \frac{\beta}{2s} e^{-\beta|x-t|} dt$, which can be viewed as an average of the Laplace density $\frac{\beta}{2} e^{-\beta|x-t|}$ over the interval $[0,s]$. Thus, we take (a)' as an example. Let I_1 and I_2 be the integral of $e^{-\beta|x-t|}$ over the range $[s', s]$ and $[s, s_0]$, respectively. Clearly, the distribution density of x produced by $\mathcal{D} = \{s, s_0\}$ is $\frac{I_2}{s_0-s}$, while that of $\mathcal{D}' = \{s', s_0\}$ is $\frac{I_2+I_1}{s_0-s'} = \frac{I_2}{s_0-s} \cdot \frac{s_0-s}{s_0-s'} + \frac{I_1}{s-s'} \cdot \frac{s-s'}{s_0-s'}$. Thus, if $\frac{I_2}{s_0-s} \geq \frac{I_2+I_1}{s_0-s'}$, then $\frac{I_2}{s_0-s} \geq \frac{I_1}{s-s'}$. Therefore, in such a case, the ratio achieved by case (a) where $s' \leq s < s_0$ should be larger than that of case (a)'. We can make a similar argument for (b) and (b)'. Therefore, we have that the supremum of $\epsilon_{\mathcal{D},D'}$ is achieved either in $s' = 0$ or $s' = \mathcal{B}$.

With the above preparation, w.l.o.g., we assume $s < s_0$. We first have the following simple observation: when $x < \frac{s}{2}$, s' should be set to be $\mathcal{B}$ and

$$\sup_{s'} e^{\epsilon_{s,s'}(x)} = \frac{\frac{e^{\beta(x-s)}}{2(s_0-s)}(1-e^{-\beta(s_0-s))}}{\frac{e^{\beta(x-s_0)}}{2(\mathcal{B}-s_0)}\big(1-e^{-\beta(\mathcal{B}-s_0)}\big)} \quad = \frac{\mathcal{B}-s_0}{s_0-s} \cdot \frac{e^{s_0-s}(1-e^{-\beta(s_0-s)})}{1-e^{-\beta(\mathcal{B}-s_0)}}. \tag{12}$$

Similarly, when $x \geq \frac{\mathcal{B}+s_0}{2}$, s' should be 0 and

$$\sup_{s'} \epsilon_{s,s'}(x) = \frac{s_0}{s_0-s} \cdot \frac{1-e^{-\beta(s_0-s)}}{1-e^{-\beta s_0}}.$$

Now, we consider the other case when $x \in [\frac{s}{2}, \frac{\mathcal{B}+s_0}{2}]$. To derive an upper bound, we simply bound $\sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'} \leq \epsilon_{\mathcal{D},s'=0} + \epsilon_{\mathcal{D},s'=1}$. We first calculate the expectation of $\epsilon_{\mathcal{D},s'=1}$ when x is restricted to $[s, s_0]$. Due to the concavity of $\log(\cdot)$, with Jensen's inequality, this quantity can be upper bounded by

$$\mathbb{E}_{x \in [s,s_0]} \epsilon_{\mathcal{D},s'=1} \leq \quad \Pr(x \in [s,s_0]) \log\big[\frac{1}{\Pr(x \in [s,s_0])} \int_s^{s_0} \frac{(\frac{2-e^{-\beta(x-s)}-e^{-\beta(s_0-x)}}{2(s_0-s)})^2}{\frac{e^{-\beta(s_0-x)}(1-e^{-\beta(\mathcal{B}-s_0)})}{2(\mathcal{B}-s_0)}} dx\big]. \tag{13}$$

The integral quantity in (13) can be expressed as

$$\int_s^{s_0} \frac{(\frac{2-e^{-\beta(x-s)}-e^{-\beta(s_0-x)}}{2(s_0-s)})^2}{\frac{e^{-\beta(s_0-s)}(1-e^{-\beta(\mathcal{B}-s_0)})}{2(\mathcal{B}-s_0)}} dx \quad \leq \frac{\mathcal{B}-s_0}{2(1-e^{-\beta(\mathcal{B}-s_0)})(s_0-s)^2} \frac{7e^{\beta(s_0-s))} - (12\beta(s_0-s)+3)}{3\beta}. \tag{14}$$

Similarly, for $\epsilon_{\mathcal{D},s'=0}(x)$, it is not hard to find that

$$\epsilon_{\mathcal{D},s'=0} \leq \log(\frac{\frac{1-e^{-\beta(s_0-s)}}{2(s_0-s)}}{\frac{1-e^{-\beta s_0}}{2s_0}}) = \log(\frac{s_0(1-e^{-\beta(s_0-s)})}{(s_0-s)(1-e^{-\beta s_0})}),$$

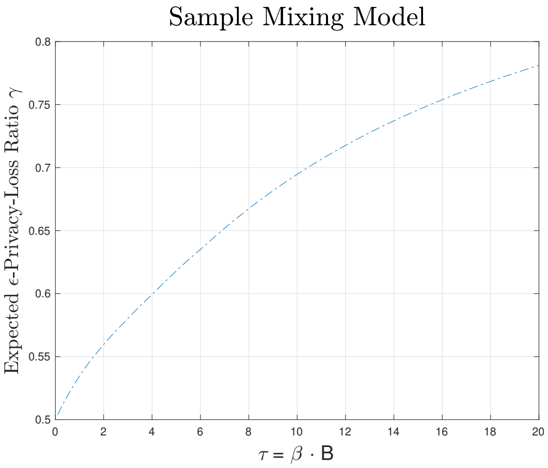
where equality is achieved when $x \geq s_0$. Thus, putting things together, we have

$$\begin{aligned} \sup \mathbb{E}_{\mathcal{D}} \sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}^k) \leq \max_{s \leq s_0 \in [0,\mathcal{B}]} \big[\log(\frac{s_0(1-e^{-\beta(s_0-s)})}{(s_0-s)(1-e^{-\beta s_0})}) \\ + \log\big(\frac{(\mathcal{B}-s_0)(7e^{\beta(s_0-s)} - (12\beta(s_0-s)+3))}{6\beta(1-e^{-\beta(\mathcal{B}-s_0)})(s_0-s)^2} + \frac{e^{\beta(s_0-s)}(\mathcal{B}-s_0)(1-e^{\beta(s_0-s)})^2}{2\beta(s_0-s)^2(1-e^{-(\mathcal{B}-s_0)})})\big]. \end{aligned} \tag{15}$$

Finally, multiplying by the subsampling factor $\frac{2}{n}$, we have the claim.

In Fig. 1, we present the simulation of the ratio γ defined in (9). From Fig. 1, the privacy amplification ratio, within $[0.5, 1]$, is determined by $\tau = \beta\mathcal{B}$, where a smaller τ produces a better privacy amplification. Now, we make two remarks on further generalization.

Fig. 1: Privacy Amplification in the Hybrid of Sample Mixing and Laplace Mechanism



Public Dataset and Subsampling: There is an interesting variant of the sample mixing model with the existence of a public dataset. Recall the subsampling across n private samples in the SGD protocol (4), which contributes a factor $1/n$ and $2/n$ to the ϵ-loss in the case without and with sample mixing, respectively. Imagine if the entire dataset is formed by a public set of m samples and a private set of n samples. Without mixup, the vanilla subsampling contributes a factor $1/(m+n)$ to the ϵ privacy leakage. In comparison, if samples are randomly mixed between the private and public sets, the factor becomes $1/n$ in contrast to $2/n$ when we only mix private samples. Thus, γ can be further improved by $\frac{m+n}{2n}$, as large as an additional 50% reduction for $m \ll n$.

(Advanced) Composition: It is well known that relaxed to approximate DP, a K-fold composition of (ϵ, δ)-DP mechanisms can produce a $(\tilde{\epsilon}, K\delta + \tilde{\delta})$-DP, with $\tilde{\epsilon} = K\epsilon^2 + \sqrt{-2K\epsilon^2 \log \tilde{\delta}}$ [KOV17], where $\tilde{\delta}$ is a free parameter. We consider the composition of $\epsilon(\boldsymbol{\theta}^k)$ across multiple iterations. Following the idea in [KOV17], since $\epsilon(\boldsymbol{\theta}^k)$ is also bounded from Theorem 1, by $\frac{2}{n} \cdot \log(\frac{e^{\mathcal{B}\beta}-1}{\mathcal{B}\beta})$, we may also apply Azuma's Inequality to derive a high-probability bound. However, compared to the pure Laplace Mechanism, as shown in Theorem 2 and Fig. 1, sample mixing enjoys better average loss, which can produce a sharpened composition bound. It is noted that the metric $\sup_{\mathcal{D}} \mathbb{E}_{\mathcal{D}} \sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}^k)$ proposed, which is defined on the point-wise worst-case ex-post ϵ-loss, is stronger than the KL-divergence between the two output distributions from $\mathcal{D}$ and $\mathcal{D}'$. Thus, γ presented above also provides a generic lower bound on the privacy amplification, i.e., how much privacy loss is saved, over the composition of multiple dimensions and iterations. In this paper we only consider the hybrid between *mixup* and the Laplace Mechanism, but it can be easily generalized to a Gaussian Mechanism with a similar reasoning.

3 Hybrid Architecture of Mixup and Noise: Update Mixing

Besides sample aggregation, local update aggregation is also a building block of distributed optimization. In the following, we shift our focus to a novel hybrid structure of update mixing and a Laplace Mechanism. Consider a more complicated case, which is a distributed optimization amongst N agents and the goal is to collaboratively minimize the sum of their loss functions $\sum_{i=1}^{N} f_i(\boldsymbol{\theta}_i)$ under a consensus restriction $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = ... = \boldsymbol{\theta}_N$, where $f_i(\cdot)$ denotes the local loss function held by agent i. For the i-th agent at the k-th iteration, a generic updating protocol of a distributed GD can be described as follows,

$$\boldsymbol{\theta}_i^k = \underline{\sum_{j=1}^{N} w_{ij}^k \boldsymbol{\theta}_j^{k-1}} - \eta_k \nabla f_i(\boldsymbol{\theta}_i^{k-1}) + \boldsymbol{\Delta}_i^k. \tag{16}$$

Here, $w_{ij}^k \in [0,1]$ is the weight assigned to $\boldsymbol{\theta}_j^k$ at iteration k such that $\sum_{j=1}^N w_{ij}^k = 1$. We still assume the noise follows a Laplace distribution. Regular distributed GD fixes the weight w_{ij}^k in (16) to be the constant $1/N$. In this case, $\boldsymbol{\theta}_i^k$ is in a pure Laplace distribution provided the earlier updates. Now, we imagine that w_{ij}^k is randomly generated across iterations but still with $\sum_{j=1}^N w_{ij}^k = 1$. Then, the underlined quantity in (16) is randomly distributed in a convex hull of earlier immediate updates $\boldsymbol{\theta}_i^{k-1}$. Thus, the distribution of $\boldsymbol{\theta}_i^k$ is indeed a mixture of a Laplace noise and a bounded random variable.

Still, w.l.o.g., we only consider the one-dimensional case in the following. The multi-dimensional case is a straightforward composition. To capture the distribution of $\boldsymbol{\theta}_i^k$ in (16), for simplicity we assume $N = 2$, i.e., the underlying quantity in (16) is a 2-mix, and the difference between $\boldsymbol{\theta}_1^{k-1}$ and $\boldsymbol{\theta}_2^{k-1}$ is ω. Thus, equivalently, we only need to consider the following mixture distribution: the sum of Lap$(0,\beta)$ and an independent uniform distribution within interval $[0,\omega]$ denoted by $U[0,\omega]$, whose probability density is a convolution Lap$(0,\beta) * U[0,\omega]$. We assume the l_∞ sensitivity of ∇f_i to be $\mathcal{B}$ and $\eta_k = 1$.

Example 2: Imagine that in (16), the two earlier updates are $\boldsymbol{\theta}_1^{k-1} = 0$ and $\boldsymbol{\theta}_2^{k-1} = 1$, i.e., $\omega = 1$. Two extreme cases can be: Case 1, $\eta_k \nabla f_1(\boldsymbol{\theta}_1^{k-1}) = 0$ and $\boldsymbol{\theta}_1^k$ produced follows $Lap(0,\beta) * U[0,1]$; Case 2, with the sensitivity assumption, $\eta_k \nabla f_1(\boldsymbol{\theta}_1^{k-1})$ can also be as large as $\mathcal{B}$, and correspondingly the distribution of $\boldsymbol{\theta}_1^k$ is $Lap(0,\beta) * U[\mathcal{B}, \mathcal{B}+1]$.

Theorem 3 *With the above setup and assumptions on the update mixing, sensitivity and noise, the hybrid structure shown in (16) satisfies*

$$\epsilon(\boldsymbol{\theta}_i^k) \le \max_{t=\pm\mathcal{B}} \left| \log\left[\int_0^{\omega} e^{-\beta|\boldsymbol{\theta}_i^k - c|} dc \Big/ \int_t^{t+\omega} e^{-\beta|\boldsymbol{\theta}_i^k - c|} dc \right] \right| \quad \le \beta\mathcal{B}.$$

In particular, let $\tau' = \omega\beta$, in an approximate (ϵ,δ)-DP view, if $\delta > (1-e^{-\tau'})/(2\tau')$, then

$$\epsilon \le \log\big(\max\{\frac{2e^{\beta(\mathcal{B}-\psi(\delta))}}{1-e^{-\tau'}}, \frac{2}{1-e^{-\tau'}}\}\big),$$

where $\psi(\delta) = \omega/(1-e^{-\tau'}) \cdot (\delta - (1-e^{-\tau'})/(2\tau'))$.

The proof of Theorem 3 can be found in Appendix B.

Theorem 3 states that, without mixing in (16), i.e., weights ω_{ij}^k are fixed, if the pure Laplace mechanism produces ϵ_0-DP, then with the same setup, the hybrid of update mixing and the Laplace mechanism satisfies the same worst case, i.e., $\sup_{\boldsymbol{\theta}_i^k} \epsilon(\boldsymbol{\theta}_i^k) = \epsilon_0$. On the other hand, similar to the argument of sample mixup, update mixing also strengthens the average-case privacy loss, captured by the high-probability ϵ bound given in Theorem 3. Analogous to Theorem 2, we also quantify the expected privacy loss savings in the worst case for the update mixing model.

Theorem 4 *Let $\mathit{\Phi}(t,z) = \frac{\beta}{2\omega} e^{-\beta|t-z|}$, in the hybrid of update mixing and the Laplace model, when $\omega > \mathcal{B}$,*

$$\sup_{\mathcal{D}} \mathbb{E}_{\mathcal{D}} \sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}_i^k) \le \log\left\{ 2\int_0^{\frac{\omega-\mathcal{B}}{2}} \int_{-\mathcal{B}}^{\omega-\mathcal{B}} \mathit{\Phi}(t,z)dtdz \;\;+\;\; e^{\beta\mathcal{B}}\left[1 - 2\int_0^{\frac{\omega-\mathcal{B}}{2}} \int_0^{\omega} \mathit{\Phi}(t,z)dtdz \right] \right\},$$

which is $O(1/\omega + 1/(\omega\beta))$ if we take $\mathcal{B}$ as a constant.

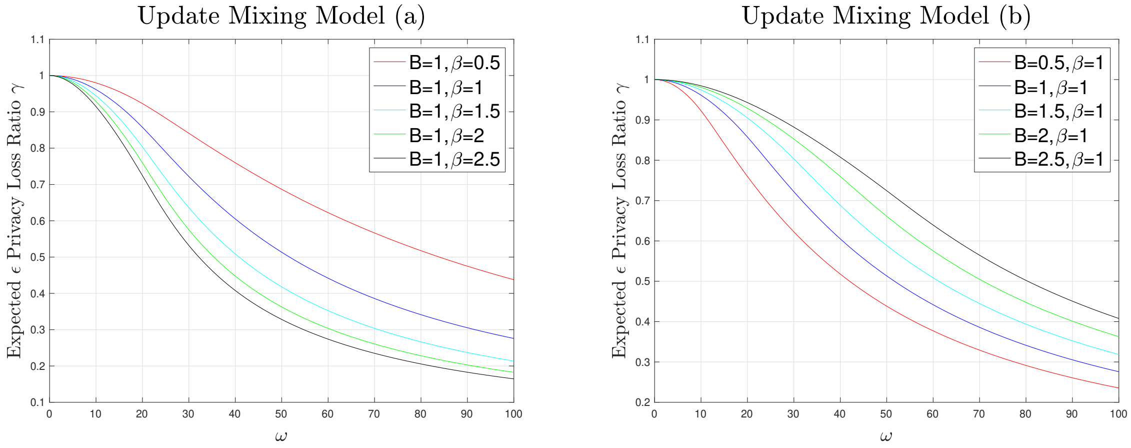
The proof of Theorem 4 is given in Appendix C.

In Fig. 2, we present the simulation results on the ratio

$$\gamma = \frac{\sup_{\mathcal{D}} \mathbb{E}_{\mathcal{D}} \ \sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}_i^k)}{\beta\mathcal{B}},$$

which quantifies the privacy gain compared to pure Laplace in the update mixing model. We show the effect of β and $\mathcal{B}$ on γ in Fig. 2(a) and (b), respectively. Clearly, larger ω (larger update difference) and β (smaller noise) with a smaller sensitivity $\mathcal{B}$ produce stronger privacy amplification (smaller γ), which coincides with our analysis.

Fig. 2: Privacy Amplification in the Hybrid of Update Mixing and Laplace Mechanism



4 Decentralized Locally Private Optimization with Update Mixing

We now proceed to consider a stronger privacy guarantee, Local Differential Privacy (LDP), and construct concrete decentralized (distributed) algorithms that incorporate such update mixing structure.

In practice, data is usually stored across multiple agents and they have to collaborate on a distributed optimization. Without loss of generality, consider a decentralized optimization problem across N agents in a connected network. The network is modeled by an undirected graph $\mathcal{G}(N, E)$. Nodes are indexed as $N = \{1, ..., N\}$ and when two nodes i and j are neighbors that can communicate, $(i, j) \in E$. Each node i has a function $f(\boldsymbol{s}_i, \boldsymbol{\theta}_i)$ that we regard as a loss function determined by the sample $\boldsymbol{s}_i$ held locally with the parameter $\boldsymbol{\theta}_i$ to be optimized. In this paper, we always assume that $f(\boldsymbol{s}_i, \cdot)$ is a differentiable convex function $\mathcal{C} \to \mathbb{R}$ and $\boldsymbol{\theta}_i \in \mathcal{C} \subset \mathbb{R}^d$. $\mathcal{C}$ can be viewed as the constraint, assumed to be a closed convex set. We express the objective function to be minimized as

$$\min_{\boldsymbol{\theta}_{[1:N]}} \sum_{i=1}^{N} f(s_i, \boldsymbol{\theta}_i), \quad s.t. \sum_{i=1}^{N} A_i \boldsymbol{\theta}_i = \boldsymbol{c}. \tag{17}$$

In many learning problems, $\boldsymbol{\theta}_{[1:N]}$ stands for one parameter to be collaboratively optimized. Here, $[1 : N]$ is the compact form of $\{1, 2, ..., N\}$. We term the problem as consensus optimization if the constraint requires that all $\boldsymbol{\theta}_i$ be equal, which can still be enforced by a linear constraint $\sum_{i=1}^{N} A_i \boldsymbol{\theta}_i = \mathbf{0}$ [MO17]. In such a scenario, agents may not trust each other. To this end, a stronger notion is Local Differential Privacy (LDP) [BLR13, DJW13, KOV14], where each agent runs a local randomization procedure to privatize its local dataset before release. If each local randomization procedure satisfies ϵ-DP with respect to (w.r.t.) its own local dataset, we then say the whole protocol achieves ϵ-LDP. Formally, in the context of decentralized optimization,

Definition 4 **(*ex-post* local privacy loss).** *The local privacy loss $\epsilon_i(o)$ for agent i in a decentralized optimization algorithm A on an output o is defined as*

$$\Pr(A(\mathcal{D}_i) = o) \leq e^{\epsilon_i(o)} \Pr(A(\mathcal{D}'_i) = o). \tag{18}$$

Here, $\mathcal{D}_i = (\hat{\boldsymbol{s}}_1, ..., \hat{\boldsymbol{s}}_i, ..., \hat{\boldsymbol{s}}_N)$ and $\mathcal{D}'_i = (\hat{\boldsymbol{s}}_1, ..., \hat{\boldsymbol{s}}'_i, ..., \hat{s}_N)$ are two arbitrary candidate sets of samples, where $\hat{\boldsymbol{s}}_i$ denotes the sample agent i holds, i.e., $\mathcal{D}_i$ and $\mathcal{D}'_i$ differ at the samples agent i holds.

Algorithm 1 Modified Private ADMM with First-order Approximation

```
Input: Local functions $f_{[1:N]}$, step penalty $\zeta$.
Initialize $\boldsymbol{\theta}^0_{[1:N]}$ randomly, $\boldsymbol{\lambda}^0_{[1:N]} = \mathbf{0}$. Each agent selects a private constant $H_i$.
for $k = 0, 1, 2, ...K-1$ do
    Agents $i = 1$ to $N$ do in parallel:
    Randomly pick two positive diagonal matrices $\bar{\boldsymbol{\rho}}^{k+1}_i$ and $\boldsymbol{\Gamma}^{k+1}_i$ such that $(N-1)\bar{\boldsymbol{\rho}}^{k+1}_i + \boldsymbol{\Gamma}^{k+1}_i = H_i \cdot \boldsymbol{I}_d$ and update $\boldsymbol{\theta}^{k+1}_i$ in parallel, where $\boldsymbol{\Delta}^{k+1}_i \sim Lap(\mathbf{0}, \beta_{k+1})$:
        $$\begin{aligned}\boldsymbol{\theta}^{k+1}_i := & \underline{\frac{\boldsymbol{\Gamma}^{k+1}_i}{H_i}\boldsymbol{\theta}^k_i + \frac{(N-1)\bar{\boldsymbol{\rho}}^{k+1}_i}{H_i}\frac{\sum_{j\neq i}\boldsymbol{\theta}^k_j}{N-1}}_A - \underline{H_i^{-1}\nabla f_i(\boldsymbol{\theta}^k_i)}_B \\ & + H_i^{-1}\boldsymbol{\lambda}^k_i + \boldsymbol{\Delta}^{k+1}_i.\end{aligned} \quad (21)$$
    Exchange $\boldsymbol{\theta}^{k+1}_i$ and then update $\boldsymbol{\lambda}^{k+1}_i := \boldsymbol{\lambda}^k_i - \zeta\sum_{j\neq i}(\boldsymbol{\theta}^{k+1}_j - \boldsymbol{\theta}^{k+1}_i)$.
end for
```

LDP described in Definition 4 states that even all other nodes are colluding against node i, from the output, it is still hard to distinguish agent i's private data. Similarly, the above point-wise privacy loss easily produces the classic ϵ-LDP by setting $\epsilon = \max_i \sup_o \epsilon_i(o)$ [LNR$^+$17, WCHRP17]. For an iterative optimization, o in Definition 4 includes all (immediate) outputs across iterations. For simplicity, we use $f_i(\boldsymbol{\theta}_i)$ to denote $f(\boldsymbol{s}_i, \boldsymbol{\theta}_i)$ in the following.

Algorithm Construction: Generally speaking, there are two types of decentralized optimization. One is (sub)gradient based, such as decentralized (stochastic) gradient descent (D(S)GD) methods [NO09], [LZZ$^+$17], and EXTRA [SLWY15]. The latter relies on solving a constrained problem with dual variables to minimize some Lagrangian function, such as Alternating Direction Method of Multipliers (ADMM) [WO12]. However, the key idea is the same where agents average out updates from neighbors and perform broadcasts, and collaboratively and iteratively approach the global optimum. In the previous section, we briefly described how to incorporate *mixup* into GD. In the following, we consider the more complicated case of ADMM.

The updating rule of ADMM with the dual method for (17) relies on solving an optimization problem:

$$\boldsymbol{\theta}^{k+1}_i := \arg\min_{\boldsymbol{\theta}_i} \mathcal{L}(\boldsymbol{\theta}^k_1, ..., \boldsymbol{\theta}_i, ..., \boldsymbol{\theta}^k_N, \boldsymbol{\lambda}^k) + \frac{\Gamma}{2}\left\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^k_i\right\|^2 + \frac{\rho}{2}\left\|A_i\boldsymbol{\theta}_i + \sum_{j\neq i}^{N} A_j\boldsymbol{\theta}^k_j - \boldsymbol{c}\right\|^2 + \boldsymbol{\Delta}^{k+1}_i, \quad (19)$$

where the Lagrangian function is defined as $\mathcal{L}(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N, \boldsymbol{\lambda}) = \sum_{i=1}^N f_i(\boldsymbol{\theta}_i) - \boldsymbol{\lambda}^T(\sum_{i=1}^N A_i\boldsymbol{\theta}_i - \boldsymbol{c})$. Here, $\|\cdot\|_q$ denotes the l_q norm. For brevity, $\|\cdot\|$ denotes the standard l_2 norm in the following. The Lagrangian multiplier $\boldsymbol{\lambda}^{k+1}$ is updated through $\boldsymbol{\lambda}^{k+1} := \boldsymbol{\lambda}^k - \zeta(\sum_{i=1}^N A_i\boldsymbol{\theta}^{k+1}_i - \boldsymbol{c})$. If we allow each agent to independently select random penalties ρ and Γ across iterations, the $\boldsymbol{\theta}^{k+1}_i$ produced similarly follows a mixture Laplace distribution with a random mean.

However, we have to stress that closed-form optima of (19) may not exist. To reduce the computational complexity, in contrast to previous ADMM protocols [WO12, ZZ17, ZKL18], we consider a first-order approximation for each f_i:

$$f_i(\boldsymbol{\theta}_i) \approx f_i(\boldsymbol{\theta}^k_i) + \nabla f_i(\boldsymbol{\theta}^k_i)(\boldsymbol{\theta}_i - \boldsymbol{\theta}^k_i). \quad (20)$$

Through randomizing ρ and Γ in (19) accompanied with the approximation (20), a modified private ADMM with *mixup* is accordingly derived as Algorithm 1. Meanwhile, stemming from (16), we formally describe a Distributed SGD with *mixup* as Algorithm 2 [3].

It is clear that Algorithms 1 and 2 share a very similar structure except for the dual variable $\boldsymbol{\lambda}^k_i$ in ADMM. Intuitively, Term A in either (21) or (22) behaves as *mixup* to merge the updates from the previous iteration and Term B corresponds to the effect from the function f_i on updating $\boldsymbol{\theta}^{k+1}_i$ followed by a Laplace noise. Thus, with fixed updates from the last iteration, in each dimension, the produced distribution of (21) and (22) is equivalent to the update mixing model

[3] Here, we omit the projection step if $\mathcal{C} \subsetneq \mathbb{R}^d$ since we are interested in arbitrary constraints. 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 the general case in Theorem 5.

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, $S_{[1:2K]}$.
Initialize $\boldsymbol{y}_1^0$ and $\boldsymbol{y}_2^0$.
**for** $k = 0, 1, 2, ..., K-1$ **do**
    **Agents** $i$ **in** $S_{2k+1}$ **and** $S_{2k+2}$ do in parallel :
    Randomly pick a positive diagonal matrix $\boldsymbol{w}_i$ of which the non-zero elements are within $(0, 1)$. Then, with $\boldsymbol{\Delta}_i^{k+1} \sim Lap(\mathbf{0}, \beta_{k+1})$, update $\boldsymbol{\theta}_i$ as:
        $$\boldsymbol{\theta}_i := \underbrace{\boldsymbol{w}_i \boldsymbol{y}_1^k + (\boldsymbol{I}_d - \boldsymbol{w}_i)\boldsymbol{y}_2^k}_{A} - \underbrace{\eta_{k+1} N \nabla f_i\Big(\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2}\Big)}_{B} + \boldsymbol{\Delta}_i, \tag{22}$$
    Broadcast $\boldsymbol{\theta}_i$ to agents in $S_{2k+3}$ and $S_{2k+4}$ where $\boldsymbol{y}_1^{k+1} = \frac{1}{|S_{2k+1}|}\sum_{i \in S_{2k+1}} \boldsymbol{\theta}_i$ and $\boldsymbol{y}_2^{k+1} = \frac{1}{|S_{2k+2}|}\sum_{i \in S_{2k+2}} \boldsymbol{\theta}_i$.
**end for**
```

described in Section 3. We can still use $\mathcal{B}$ to denote the l_∞ sensitivity where $\max_i \sup_{\boldsymbol{\theta}} \|\nabla f(\hat{s}_i, \boldsymbol{\theta}) - \nabla f(\hat{s}_i', \boldsymbol{\theta})\|_\infty \leq \mathcal{B}$ for two arbitrary candidates $\hat{\boldsymbol{s}}_i$ and $\hat{\boldsymbol{s}}_i'$ of the sample from agent i. What remains to complete the privacy analysis on Algorithms 1 and 2 is simply a composition of the privacy loss on each dimension of immediate update $\boldsymbol{\theta}_i$ across each iteration and is omitted.

5 Analysis of LDP-Utility Tradeoff

Based on fundamental analysis of non-private ADMM and decentralized (stochastic) GD [MO17, OHTG13, SLY$^+$14, CHW15], a systematic study of decentralized optimization with composite incorporation of random aggregation and noise perturbation is developed throughout this section. It is worth mentioning that, with properly selected parameters, *the theorems and framework of proofs provided are invariant to whether update mixing is incorporated or not.* In other words, upper bounds shown in Theorems 5-6 also match the best-known existing privacy-utility tradeoff without *mixup*. We focus on the $\epsilon(, \delta)$ privacy guarantee of LDP in the following.

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

A function $f(\boldsymbol{\theta}) : \mathcal{C} \to \mathbb{R}$ is L-Lipschitz continuous if for any $\boldsymbol{\theta}, \boldsymbol{y} \in \mathcal{C}$, $|f(\boldsymbol{\theta}) - f(\boldsymbol{y})| \leq L\|\boldsymbol{\theta} - \boldsymbol{y}\|$;

A function $f(\boldsymbol{\theta}) : \mathcal{C} \to \mathbb{R}$ is M-smooth if $\nabla f(\cdot)$ is M-Lipschitz continuous: for any $\boldsymbol{\theta}, \boldsymbol{y} \in \mathcal{C}$, $\|\nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{y})\| \leq M\|\boldsymbol{\theta} - \boldsymbol{y}\|$.

$\boldsymbol{I}$ is the $d \times d$ identity matrix. We use $\|\boldsymbol{z}\|_G^2$ to denote $\boldsymbol{z}^T G \boldsymbol{z}$ in the following.

Analysis of Algorithm 1: Let $G_x = diag\{\boldsymbol{H}_1, ..., \boldsymbol{H}_N, 1/\zeta\}$ denote a diagonal matrix, where $\boldsymbol{H}_i = H_i \cdot \boldsymbol{I} = \boldsymbol{\Gamma}_i^{k+1} + A_i^T \boldsymbol{\rho}_i^{k+1} A_i$ is positive definite.

Theorem 5 *If $f_i(\cdot), i = 1, 2, .., N$, are convex and M-smooth, and for any $i, j \in [1:N]$*

$$\begin{cases} \frac{H_j - M}{N^2}(\underline{\rho}_i - \frac{\zeta}{2}) \geq \sigma_{\max,j}^2 \bar{\rho}_i^2 \\ \frac{1}{(N-1)^2}(\underline{\rho}_i - \frac{\zeta}{2})(\underline{\rho}_j - \frac{\zeta}{2}) \geq (\bar{\rho}_i - \underline{\rho}_i)^2, \end{cases} \tag{23}$$

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

$$|\mathbb{E}[F(\bar{\boldsymbol{\theta}}^K)] - F(\boldsymbol{\theta}_*)| \leq \frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{G_x}^2 + \frac{4}{\zeta}\|\boldsymbol{\lambda}^*\|^2}{2K} + \frac{\sum_{k=1}^K \sum_{i=1}^N (M + H_i + \frac{1}{\zeta})\mathbb{E}[\|\boldsymbol{\Delta}_i^k\|^2]}{K}, \tag{24}$$

where $\bar{\boldsymbol{\theta}}^K = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\theta}_k$ and $\boldsymbol{\lambda}^*$ is the state of $\boldsymbol{\lambda}^k$ when $\boldsymbol{\theta}_k$ reaches the optimum $\boldsymbol{\theta}_* = (\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, ..., \boldsymbol{\theta}_N^*)$ of (17). Under (ϵ, δ)-LDP, the utility loss is $O(\frac{\sqrt{N} d\mathcal{B}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{G_x}}{\epsilon})$.

Here, $\mathcal{B}$ still stands for the l_∞ sensitivity. Theorem 5 characterizes the convergence rate of the decentralized optimization in an arbitrary communication graph. The takeaway from the above theorem is that, with or without update mixing, regular ADMM and the proposed Algorithm 1 enjoy the same convergence rate and privacy-utility tradeoff in an asymptotic view.

The high-level idea of the proof (see Appendix D) can be summarized as two steps: First, we derive the convergence rate of regular ADMM with the *mixup* structure (Lemma 1 and 2); Second, we study the first-order approximation error (Lemma 3), and then combine both to complete the proof.

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 $\boldsymbol{\theta}_* = (\boldsymbol{\theta}^*, \boldsymbol{\theta}^*, ..., \boldsymbol{\theta}^*)$ is the optimum solution to (17).

Theorem 6 *Assume that $f_i(\boldsymbol{\theta})$ is convex and $\|\nabla f_i(\boldsymbol{\theta})\|^2$ is bounded by G^2 for each i. Moreover, assume that*

$$\max_{k,j\in\{1,2\}} \mathbb{E}[(\boldsymbol{\Delta}_j^k/\eta_k)^2] \leq V^2.$$

When we select the step size $\eta_k = \frac{1}{c\sqrt{k}}$ for some constant c, then $\frac{1}{N}\mathbb{E}[\sum_{i=1}^N f_i\big(\frac{\sum_{k=0}^{K-1}(\boldsymbol{y}_1^k+\boldsymbol{y}_2^k)}{2K}\big) - F(\boldsymbol{\theta}_*)]$ is upper bounded by

$$\frac{c(\|\boldsymbol{y}_1^0 - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^0 - \boldsymbol{\theta}^*\|^2)}{4\sqrt{K}} + \frac{(\log K + 2)\sqrt{K+1}(G^2+V^2)}{2Kc}. \tag{25}$$

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

The proof of Theorem 6 can be found in Appendix E. (25) matches the lower bound of [KOV14] and is essentially a more generalized form of the utility guarantees derived in [STU17]. In addition, we provide a non-asymptotic view of the convergence rate through a study on the random stochastic matrix, included in Appendix F.

In Appendix A, we provide two sets of experiments to support all the theoretical results obtained so far. In Appendix A.2, we compare Algorithm 1 and 2 with existing private decentralized optimization algorithms [ZKL18, ZZ17, DGPH19]. In Appendix A.3, we move our focus to study the impact of update mixing in the hybrid structure on utility, compared to the pure Laplace mechanism without mixing (by fixing all weights). We test both practical and synthetic datasets.

As a conclusion, either from a theoretically asymptotic convergence rate view, shown in Theorem 5 and 6, which matches the information-theoretically optimal lower bound, or from an empirical view, through experiments shown in Appendix A, random mixing does not hurt performance but produces a straightforward privacy amplification, captured by the ratio γ described in Theorem 2 and 4.

6 Other Related Works and Conclusion

As a powerful paradigm, randomness can be used to strengthen various aspects of performance, such as learning capacity and computational complexity, and meet different requirements. In this paper, we set out to bridge empirical randomization commonly used in machine learning and the classic noise (Laplace/Gaussian) mechanism commonly used in Differential Privacy with a hybrid structure. This provides a systematic framework to understand and make use of optimization or learning-utility-oriented randomness to further strengthen privacy preservation.

In this work, we take *mixup* as an example and study privacy amplification in the convolution of input mixing and the Laplace mechanism. Another well-known amplification technique in DP is the mixture of subsampling and regular noise mechanisms. Balle et al. in [BBG18] propose applying coupling and α-divergence to derive a tighter privacy loss bound, which we believe can also be generalized to our case. Moreover, we observe that public dataset can further strengthen the privacy amplification during sample mixing, which is discussed in Section 2. This is different from the approach mentioned in [FMTT18].

Inspired by sample mixing, we propose a novel update mixing structure and show its applications in decentralized optimization. As far as we know, the utility-LDP tradeoff presented here is the most generic result so far, where existing private ADMM or DGD works [ZZ17, ZKL18, HMV15, HHG$^+$19, HTP17] either assume strong convexity or central aggregators.

Bibliography

[BBG18] 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.

[BCG$^+$19] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5050–5060, 2019.

[BDLS19] Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. *arXiv preprint arXiv:1911.11607*, 2019.

[BDRS18] Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86, 2018.

[BLR13] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[BST14] 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.

[CDG$^+$20] Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. An attack on instahide: Is private learning possible with instance encoding? *arXiv preprint arXiv:2011.05315*, 2020.

[CHW15] 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.

[CMS11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(Mar):1069–1109, 2011.

[CSS12] Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pages 989–997, 2012.

[CSZ20] Sitan Chen, Zhao Song, and Danyang Zhuo. On instahide, phase retrieval, and sparse matrix factorization. *arXiv preprint arXiv:2011.11181*, 2020.

[DGPH19] Jiahao Ding, Yanmin Gong, Miao Pan, and Zhu Han. Optimal differentially private admm for distributed machine learning. *arXiv preprint arXiv:1901.02094*, 2019.

[DJW13] 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.

[DKM$^+$06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DR16] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[DRS19] Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[DRV10] Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60. IEEE, 2010.

[DTTZ14] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

[FMTT18] 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.

[GMW$^+$11] Michaela Gotz, Ashwin Machanavajjhala, Guozhang Wang, Xiaokui Xiao, and Johannes Gehrke. Publishing search logs—a comparative study of privacy guarantees. *IEEE Transactions on Knowledge and Data Engineering*, 24(3):520–532, 2011.

[HHG$^+$19] Zonghao Huang, Rui Hu, Yuanxiong Guo, Eric Chan-Tin, and Yanmin Gong. Dp-admm: Admm-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 2019.

[HMV15] 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.

[HSLA20] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. Instahide: Instance-hiding schemes for private distributed learning. In *International Conference on Machine Learning*, pages 4507–4518. PMLR, 2020.

[HTP17] S. Han, U. Topcu, and G. J. Pappas. Differentially private distributed constrained optimization. *IEEE Transactions on Automatic Control*, 62(1):50–64, 2017.

[JGN$^+$17] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1724–1732. JMLR. org, 2017.

[JT13] Prateek Jain and Abhradeep Thakurta. Differentially private learning with kernels. *Journal of Machine Learning Research*, 2013.

[KOV14] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In *Advances in neural information processing systems*, pages 2879–2887, 2014.

[KOV17] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. *IEEE Transactions on Information Theory*, 63(6):4037–4049, 2017.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[LCLC19] Bai Li, Changyou Chen, Hao Liu, and Lawrence Carin. On connecting stochastic gradient mcmc and differential privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 557–566, 2019.

[LNR$^+$17] 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.

[LWG$^+$20] Zhijian Liu, Zhanghao Wu, Chuang Gan, Ligeng Zhu, and Song Han. Datamix: Efficient privacy-preserving edge-cloud inference. In *Computer Vision – ECCV 2020*, pages 578–595. Springer International Publishing, 2020.

[LZZ$^+$17] 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.

[Mir17] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE, 2017.

[MO17] Ali Makhdoumi and Asuman Ozdaglar. Convergence rate of distributed admm over networks. *IEEE Transactions on Automatic Control*, 62(10):5082–5095, 2017.

[MWZZ18] Wenlong Mou, Liwei Wang, Xiyu Zhai, and Kai Zheng. Generalization bounds of sgld for non-convex learning: Two theoretical viewpoints. *Proceedings of Machine Learning Research vol*, 75:1–34, 2018.

[NO09] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48, 2009.

[OHTG13] Hua Ouyang, Niao He, Long Tran, and Alexander Gray. Stochastic alternating direction method of multipliers. In *International Conference on Machine Learning*, pages 80–88, 2013.

[PXZ19] Tianyu Pang, Kun Xu, and Jun Zhu. Mixup inference: Better exploiting mixup to defend adversarial attacks. *arXiv preprint arXiv:1909.11515*, 2019.

[SHK$^+$14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[SK19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[SLWY15] 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.

[SLY$^+$14] 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.

[STU17] 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.

[TCB$^+$19] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32:13888–13899, 2019.

[TTZ15] Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly optimal private lasso. In *Advances in Neural Information Processing Systems*, pages 3025–3033, 2015.

[WCHRP17] Daniel Winograd-Cort, Andreas Haeberlen, Aaron Roth, and Benjamin C Pierce. A framework for adaptive differential privacy. *Proceedings of the ACM on Programming Languages*, 1(ICFP):1–29, 2017.

[WGX18] 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.

[WO12] 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.

[WX19] 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.

[WYX17] 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.

[ZCDLP18] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[ZKL18] 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.

[ZZ17] 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.

[ZZK$^+$20] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

A Simulation Results

A.1 Theoretical privacy amplification

First, in Fig. 3 (a), we show the relationship between γ, ω and β in Theorem 4, where $\mathcal{B}$ is fixed to 0.001. With ω ranging from 0.1 to 1 and β from 2 to 10, clearly larger ω and β, corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification.

A.2 Comparison with existing works

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 the features while converting labels $\{\geq 50k, < 50k\}$ to $\{-1, 1\}$. The training samples are denoted by $\{\boldsymbol{x}^i_j \in \mathbb{R}^{14}, L^i_j \in \{-1,1\} | i = 1, \cdots, N, j = 1, \cdots, n_i\}$. Consistent with [ZKL18], [ZZ17], we select the loss function $L(x) = \log(1 + \exp(-x))$. Thus, N agents are collaboratively solving the following logistic regression:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} f_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \log(1 + \exp(-L^i_j \boldsymbol{\theta}^T \boldsymbol{x}^i_j)) + \frac{1}{2}\|\boldsymbol{\theta}\|^2 \right).$$

Tests are run with different parameter settings. 100 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 E.

In Fig. 3 (b) and (c), we uniformly assume that $H_i = D = 10$ and $\zeta = 0.5$ for Algorithm 1. In the case of private ADMM, previous works all assume fixed parameters in the optimization protocol. In [ZZ17], the Lagrangian multiplier at the beginning of each iteration is perturbed, while [DGPH19] considers the output perturbation at the end of each iteration. Further, in [ZKL18], the authors introduce a sequence of increasing step penalty, which can bring better utility-privacy tradeoff empirically. For [ZZ17], [DGPH19] with constant fixed penalty, we assume $\varGamma_i = 0.5D$ and $\rho_i = \frac{0.5D}{|N_i|}$, corresponding to the expectation of the penalty terms in Algorithm 1. Here N_i denotes the neighbors of agent i. As for [ZKL18], we follow their setting that $\varGamma^k_i = 0.5 \times 1.02^k |N_i|$ and $\rho^k_i = 0.5 \times 1.02^k$. [4]

In the privacy part, with the same assumption in [ZKL18], we assume f_i and $\hat{f}_i$ may only differ in one sample and thus, due to the normalization, $\mathcal{B} = \frac{1}{n_i} = 0.01$, and $\boldsymbol{J} = \frac{2.8}{Dn_i}$, the Jacobian constant required by [ZKL18] in their privacy analysis. It is noted that ∇L is within $(-1, 0]$, while the privacy analysis of [DGPH19] takes the globally upper bound on ∇L as the sensitivity. This makes their bound too loose and we omit their privacy loss bound in our simulation. The results are illustrated in Fig. 3 (b) and (c), where $N = 10$, $E = 20$. The accuracy logarithm is defined by $\log\|(\boldsymbol{\theta}^k_i - \boldsymbol{\theta}^*_i)/d\|$, across 100 iterations averaged across 100 runs. The difference between the best and the worst accuracy over 100 runs is also marked.

A.3 Impact from *Mixup* on Performance

In this subsection, we provide more details on the impact of random aggregation on the utility.

In Fig.4, we test Algorithm 1 and 2 with their corresponding variants without update mixing but a pure Laplace Mechanism: the parameters in Algorithm 1 and 2 are all fixed to be constants. The task here is still the logistic regression of the *Adult* dataset defined above. For simplicity, in the following, (F) denotes the latter case of fixed parameters (without *mixup*). In the experiment, communication graphs are randomly generated across 100 trials, where N and E are the number of vertices and edges, respectively. In addition, we assume each agent holds a dataset of size 1000 and 200 in Algorithm 1 and 2, respectively. One of the key observations is, with the same setup, the hybrid randomization

[4] 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.

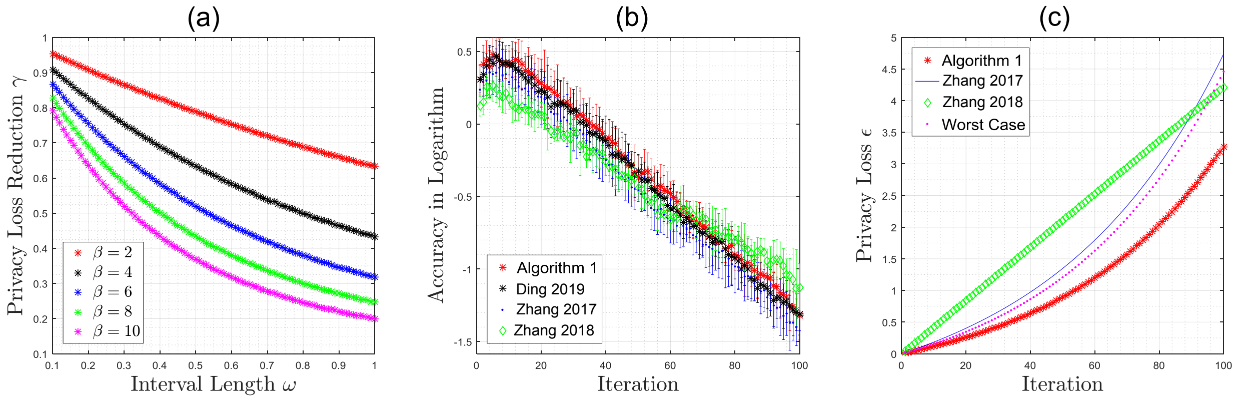


Fig. 3: Simulation on γ and comparison with [ZKL18] [ZZ17], [DGPH19] on UCI *Adult* dataset over graphs $N = 10$, $E = 20$

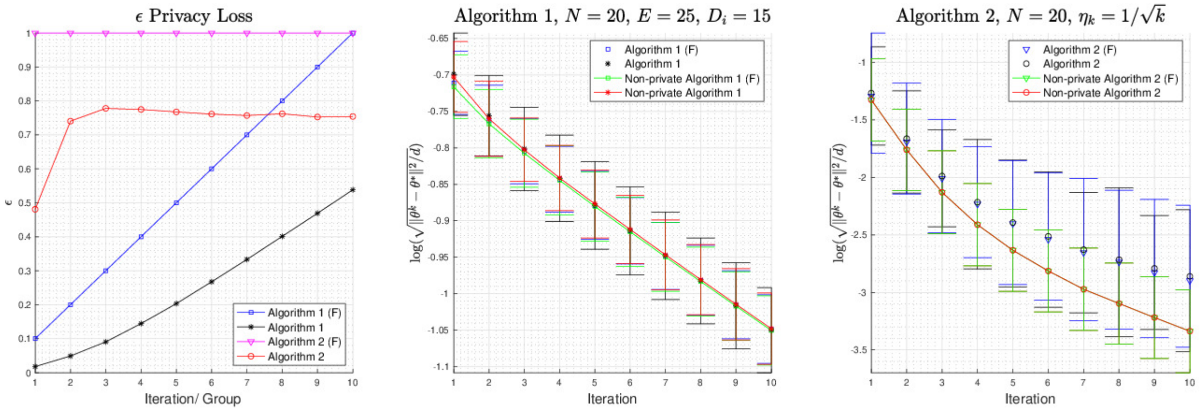


Fig. 4: Comparison between Algorithm 1 & 2 (with update mixing), Algorithm 1 & 2 (F) (without update mixing) and their non-private versions on logistic regression over *Adult*.

achieves almost the same utility loss in optimization accuracy as that of the regular Laplace Mechanism. For the privacy side, the same noise is applied in both cases where we fix the ϵ privacy budget to be 1. Update mixing renders a sharpened privacy amplification, where the privacy loss is reduced empirically ranging from 30% to 50%, and performs even better with a sparser graph or a larger privacy budget. Also, earlier iterations enjoy better privacy amplification since the divergence amongst $\boldsymbol{\theta}^k_{[1:N]}$ (or $\boldsymbol{y}^k_{[1:2]}$ in Algorithm 2) is larger. This is consistent with Theorem 4 where a larger interval length ω renders smaller γ.

We further consider synthetic datasets where each data point (z_i, y_i) is i.i.d. generated by the model $y_i = \text{sign}[\frac{1}{1+e^{\langle \boldsymbol{\theta}^*, z_i\rangle + e_i}} - \frac{1}{2}]$, where $z_i \in \mathbb{R}^{20}$ and $y_i \in \{-1, 1\}$. Here, in the first example shown in Fig. 5, we select e_i to be i.i.d. Gaussian noise $\mathcal{N}(\mathbf{0}, 0.5^2)$; in the second example shown in Fig. 6, we consider heavy-tailed noise and select e_i to be i.i.d. Lognormal noise of parameter $\mu = 1, \sigma = 1$. The probability density of a Lognormal noise of parameter (μ, σ) is $\mathbb{P}(e_i = x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$. In this case, we further normalize e_i such that the average of them equals 0. Accordingly, we implement Algorithm 1 and Algorithm 2 with corresponding fixed parameter and non-private versions on the logistic regression over the two synthetic datasets generated. Here, we assume in Algorithm 1, each agent holds 500 samples while in in Algorithm 2, each agent holds 200 samples.

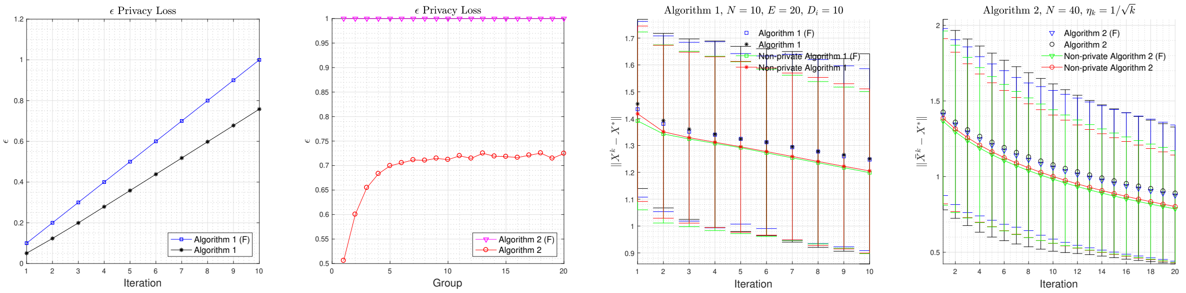


Fig. 5: Algorithm 1 & 2 and their fixed-parameter and non-private versions on logistic regression over the synthetic dataset in Gaussian distribution

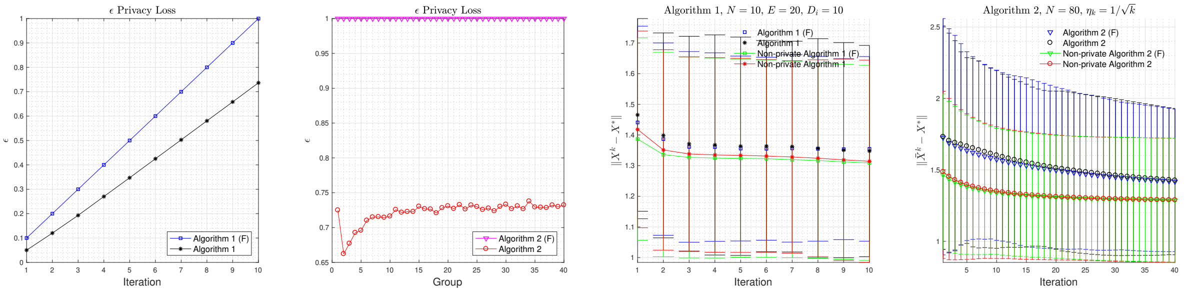


Fig. 6: Algorithm 1 & 2 and their fixed-parameter and non-private versions on logistic regression over the synthetic dataset in Lognormal distribution

B Proof of Theorem 3

Proof. W.l.o.g., we assume $\boldsymbol{\theta}_1^{k-1} = 0$ and $\boldsymbol{\theta}_2^{k-1} = \omega$. We use x to denote $\boldsymbol{\theta}_i^k$ in the following and the upper bound of ϵ can be reformulated as follows. For $x \in \mathbb{R}$, we consider

$$\max_{|t| \le \mathcal{B}} \left| \log \frac{\int_0^\omega \beta e^{-\beta|x-c|} dc}{\int_t^{t+\omega} \beta e^{-\beta|x-c|} dc} \right|, \tag{26}$$

for some positive numbers ω, β and $\mathcal{B}$, which correspond to the length of the interval, Laplace noise factor and sensitivity, respectively.

For a fixed t, $|t| \le \mathcal{B}$, if $x \notin [0, \omega] \cup [t, \omega + t]$, then

$$\left| \log \frac{\int_0^\omega \beta e^{-\beta|x-c|dc}}{\int_t^{t+\omega} \beta e^{-\beta|x-c|} dc} \right| = \left| \log \frac{\int_0^\omega e^{-\beta|x-c|} dc}{e^{\beta t} \int_0^\omega e^{-\beta|x-c|} dc} \right| = |\beta t| \le \beta \mathcal{B}. \tag{27}$$

Thus, in the following, we only need to consider the rest cases. When $x \in [0, \omega]$, then $\int_0^\omega \beta e^{-\beta|x-c|} dc = 2 - e^{-\beta x} - e^{-\beta(\omega - x)}$. In addition, if $x \in [t, \omega + t]$, then $\int_t^{\omega+t} \beta e^{-\beta|x-c|} dc = 2 - e^{-\beta(x-t)} - e^{-\beta(\omega+t-x)}$. To show

$$e^{-\beta|t|} \le \frac{2 - e^{-\beta x} - e^{-\beta(\omega-x)}}{2 - e^{-\beta(x-t)} - e^{-\beta(\omega+t-x)}} \le e^{\beta|t|},$$

it is equivalent to showing

$$\begin{cases} 2e^{\beta|t|} - e^{-\beta x + \beta|t|} - e^{-\beta(\omega-x)+\beta|t|} \ge 2 - e^{-\beta(x-t)} - e^{-\beta(\omega+t-x)}, \\ 2 - e^{-\beta x} - e^{-\beta(\omega-x)} \le 2e^{\beta|t|} - e^{-\beta(x-t)+\beta|t|} - e^{-\beta(\omega+t-x)+\beta|t|}. \end{cases} \tag{28}$$

Due to the symmetry, we merely prove the case when $t \geq 0$, where (28) can be rewritten as,

$$\begin{cases} \quad 2e^{\beta t} - e^{-\beta(x-t)} - e^{-\beta(\omega-x-t)} \geq 2 - e^{-\beta(x-t)} - e^{-\beta(\omega+t-x)}, \\ 2 - e^{-\beta x} - e^{-\beta(\omega-x)} \leq 2e^{\beta t} - e^{-\beta(x-2t)} - e^{-\beta(\omega-x)}. \end{cases} \tag{29}$$

Clearly, for the first inequality, it suffices to show

$$2(e^{\beta t} - 1) \geq (e^{2\beta t} - 1)e^{-\beta(\omega+t-x)}, \tag{30}$$

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

$$2e^{\beta x} \geq e^{\beta t} + 1, \tag{31}$$

which holds since $x > t$.

At last, we consider $x \notin [t, t+\omega]$. Again, due to the symmetry, we can assume $t \geq 0$ and $x \leq t$. Then, it is equivalent to showing:

$$\begin{cases} 2e^{\beta t} - e^{-\beta(x-t)} - e^{-\beta(\omega-x)+\beta t} \geq e^{-\beta(t-x)} - e^{-\beta(\omega+t-x)}, \\ \quad 2 - e^{-\beta x} - e^{-\beta(\omega-x)} \leq e^{\beta x} - e^{-\beta(\omega-x)}. \end{cases} \tag{32}$$

As for the first inequality, assume that $g(t) = 2e^{\beta t} - e^{-\beta(x-t)} - e^{-\beta(t-x)} - e^{-\beta(\omega-x)+\beta t} + e^{-\beta(\omega+t-x)}$. It is noted that when $t = 0$, x should be also be 0 since $x \leq t$ and $x \in [0, \omega]$ as assumed, and $g(0) = 0$. On the other hand,

$$\frac{dg}{dt} = \beta\big(2e^{\beta t} - e^{-\beta(x-t)} + e^{-\beta(t-x)} - e^{-\beta(\omega-x)+\beta t} - e^{-\beta(\omega+t-x)}\big). \tag{33}$$

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

$$2e^{\beta t} - e^{-\beta(x-t)} + e^{-\beta(t-x)} - e^{-\beta(t-x)+\beta t} - e^{-\beta(t+t-x)} \geq 0. \tag{34}$$

It is clear that $e^{\beta t} \geq e^{-\beta(x-t)}$ and $e^{-\beta(t-x)} \geq e^{-\beta(2t-x)}$ as both x and t are non-negative. Furthermore, $e^{\beta t} \geq e^{-\beta(t-x)+\beta t} = e^{\beta x}$ since $t \geq x$. Therefore, (33) is non-negative. The second inequality of (32) is exactly the AM-GM inequality that

$$2 \leq e^{-\beta x} + e^{\beta x}.$$

In a nutshell, we have proven that (26) is upper bounded by $\max_{|t| \leq \mathcal{B}} |t\beta| = \beta\mathcal{B}$. Moreover, when x belongs to the intersection of the two intervals, $(0, \omega)$ and $(t, \omega + t)$, the above inequalities are strict, i.e., (26) is strictly smaller than $\beta\mathcal{B}$, which corresponds to the pure Laplace mechanism case where we fix all parameters to be constants.

Finally, we prove the approximate (ϵ, δ)-DP argument. It is noted that the density function of $U[0, \omega] * Lap(0, \beta)(x)$ is increasing when $x < \frac{\omega}{2}$, and decreasing when $x > \frac{\omega}{2}$. Therefore, when $x \in [0, \omega]$, since $\frac{2 - e^{-\beta x} - e^{-\beta(\omega-x)}}{2\omega} < \frac{1 - e^{-\beta\omega/2}}{\omega}$, it is easy to verify that $\Pr(x < \psi(\delta)) \leq \delta$, where

$$\psi(\delta) = \frac{\delta - \frac{1 - e^{-\beta\omega}}{2\omega}}{\frac{1 - e^{-\beta\omega/2}}{\omega}}.$$

Due to the symmetry, w.l.o.g., we consider the other extreme case where x is generated by $U[\mathcal{B}, \omega + \mathcal{B}] * Lap(0, \beta)$ and compare the density ratio. When $x \in [0, \omega] \cap [\mathcal{B}, \omega + \mathcal{B}]$, the ratio becomes

$$\frac{2 - e^{-\beta x} - e^{-\beta(\omega-x)}}{2 - e^{-\beta(x-\mathcal{B})} - e^{-\beta(\omega+\mathcal{B}-x)}},$$

which is upper bounded by $\frac{2}{1 - e^{-\beta\omega}}$. On the other hand, if $x \notin [\mathcal{B}, \omega + \mathcal{B}]$ but within $[0, \omega]$, the ratio becomes

$$\frac{2 - e^{-\beta x} - e^{-\beta(\omega-x)}}{e^{\beta(x-\mathcal{B})}(1 - e^{-\beta\omega})},$$

which is upper bounded by $\frac{2e^{\beta(\mathcal{B}-x)}}{1 - e^{-\beta\omega}}$. Taking the maximum of both, the claim holds.

C Proof of Theorem 4

Proof. Following the normalization in the proof of Theorem 3, we still assume $x = \boldsymbol{\theta}_i^k$ is a Laplace distribution whose mean is uniformly distributed in $[0, \omega]$, conditional on all prior intermediate outputs. As a corollary of Theorem 3,

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

where the maximization is achieved when t either equals to $\mathcal{B}$ or $-\mathcal{B}$. To quantify $\sup_{\mathcal{D}} \mathbb{E}_{\mathcal{D}} \sup_{\mathcal{D}'} \epsilon_{\mathcal{D},\mathcal{D}'}(\boldsymbol{\theta}_i^k)$, it suffices to calculate

$$\int_{-\infty}^{\infty} \int_0^{\omega} \mathit{\Theta}(x) \frac{\beta}{2\omega} e^{-\beta|x-y|} dy dx, \tag{35}$$

since the probability density function of x is $\int_0^{\omega} \frac{\beta}{2\omega} e^{-\beta|x-y|} dy$. With the concavity of $\log(\cdot)$, (35) is upper bounded by

$$\log \int_{-\infty}^{\infty} \int_0^{\omega} \frac{\beta}{2\omega} e^{-\beta|x-y|} \cdot \quad \max_{t=\pm\mathcal{B}} \left\{ \frac{\int_0^{\omega} e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^{\omega} e^{-\beta|x-z|} dz} \right\} dx dy. \tag{36}$$

Now, we take a closer look into $\max_{t=\pm\mathcal{B}} \{ \frac{\int_0^{\omega} e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^{\omega} e^{-\beta|x-z|} dz} \}$. In the proof of Theorem 3, once $x \notin [0, \omega]$, $\max_{t=\pm\mathcal{B}} \{ \frac{\int_0^{\omega} e^{-\beta|x-z|} dz}{\int_t^{\omega+t} e^{-\beta|x-z|} dz}, \frac{\int_t^{\omega+t} e^{-\beta|x-z|} dz}{\int_0^{\omega} e^{-\beta|x-z|} dz} \} = e^{\beta\mathcal{B}}$.

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

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

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

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

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

Thus, fortunately, we can avoid the complicated integral at least in $x \in [0, \frac{\omega}{2} - \mathcal{B}]$, or $x \in [\frac{\omega}{2} + \mathcal{B}, \omega]$ where it is simplified to $O(\int_t^{\omega+t} e^{-\beta|x-y|} dy)$. Now, we can split $\mathbb{R}$ into three parts, $(-\infty, 0) \cup (\omega, \infty)$, $[0, \frac{\omega-\mathcal{B}}{2}] \cup [\frac{\omega+\mathcal{B}}{2}, \omega]$ and $(\frac{\omega-\mathcal{B}}{2}, \frac{\omega+\mathcal{B}}{2})$. To avoid the tedious term when $x \in (\frac{\omega-\mathcal{B}}{2}, \frac{\omega+\mathcal{B}}{2})$, here, we simply substitute the global upper bound to derive a closed-form expression but one may obtain the expression of γ exactly with the same reasoning. Note the symmetry on $t = \pm\mathcal{B}$, (36) is upper bounded by

$$\log \left\{ e^{\omega\mathcal{B}} \left[1 - 2 \int_0^{(\omega-\mathcal{B})/2} \int_0^{\omega} \frac{\beta}{2\omega} e^{-\beta|x-y|} dy dx \right] \quad + 2 \int_0^{(\omega-\mathcal{B})/2} \int_{-\mathcal{B}}^{\omega-\mathcal{B}} \frac{\beta}{2\omega} e^{-\beta|x-y|} dy dx \right\}. \tag{37}$$

D Proof of Theorem 5

We describe the construction of Algorithm 1 in two steps to solve (17). First, we recall the conventional ADMM (19) without first-order approximation but incorporating random penalties across iterations. The updating procedure of node i at the $(k+1)$th iteration becomes,

$$\begin{cases} \tilde{\boldsymbol{\theta}}_i^{k+1} := \arg\min_{\boldsymbol{\theta}_i} f_i(\boldsymbol{\theta}_i) - \boldsymbol{\lambda}^{kT} \left(A_i \boldsymbol{\theta}_i + \sum_{j \neq i} A_j \boldsymbol{\theta}_j^k - \boldsymbol{c} \right) \\ \qquad + \frac{1}{2} \left\| A_i \boldsymbol{\theta}_i + \sum_{j \neq i} A_j \boldsymbol{\theta}_j^k - \boldsymbol{c} \right\|_{\boldsymbol{\rho}_i^{k+1}}^2 + \frac{1}{2} \left\| \boldsymbol{\theta}_i - \boldsymbol{\theta}_i^k \right\|_{\boldsymbol{\Gamma}_i^{k+1}}^2 ; \\ \boldsymbol{\theta}_i^{k+1} = \tilde{\boldsymbol{\theta}}_i^{k+1} + \boldsymbol{\Delta}_i^{k+1}, \end{cases} \tag{38}$$

and the Lagrangian multiplier is updated accordingly as $\tilde{\boldsymbol{\lambda}}^{k+1} := \boldsymbol{\lambda}^k - \boldsymbol{\gamma}_i^{k+1}\boldsymbol{\rho}_i^{k+1}(\sum_{i=1}^N A_i\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{c})$; and $\boldsymbol{\lambda}^{k+1} := \boldsymbol{\lambda}^k - \boldsymbol{\gamma}_i^{k+1}\boldsymbol{\rho}_i^{k+1}(\sum_{i=1}^N A_i\boldsymbol{\theta}_i^{k+1} - \boldsymbol{c})$. $\boldsymbol{\gamma}_i^{k+1}\boldsymbol{\rho}_i^{k+1} = \zeta \cdot \boldsymbol{I}$ is a global constant set up at the beginning. Let $\boldsymbol{u}^{k+1} = [\boldsymbol{\theta}_{[1:N]}^{k+1}, \boldsymbol{\lambda}^{k+1}]$ and $\boldsymbol{u}^* = [\boldsymbol{\theta}_{[1:N]}^*, \boldsymbol{\lambda}]$, where $\boldsymbol{\theta}_{[1:N]}^*$ stand for the optimum to (17) and $\boldsymbol{\lambda}$ is an arbitrary point in $\mathbb{R}^d$. In Algorithm 1, we further apply first-order approximation in (38) as

$$\boldsymbol{\theta}_i^{k+1} := \boldsymbol{H}_i^{-1}\big[A_i^T\big(\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1}\big(\sum_{j \neq i} A_j\boldsymbol{\theta}_j^k - \boldsymbol{c}\big)\big) \quad + \boldsymbol{\Gamma}_i^{k+1}\boldsymbol{\theta}_i^k - \nabla f_i(\boldsymbol{\theta}_i^k)\big] + \boldsymbol{\Delta}_i^{k+1}. \tag{39}$$

The utility analysis of Algorithm 1 is thus developed in two steps: first we derive the utility loss of (38) under *mixup* structure; then we bound the loss from the approximation in Algorithm 1. Combine both, we then complete the proof.

Lemma 1. *If $f_{[1:N]}$ are all M-smooth convex functions, following (38),*

$$\begin{aligned} &2\mathbb{E}[F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}_*) - \boldsymbol{\lambda}^T A(\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*)] \\ &\leq -\mathbb{E}[h^{k+1}] + \sum_{i=1}^N (M + H_i + \frac{1}{\zeta})\mathbb{E}[\|\boldsymbol{\Delta}_i^{k+1}\|^2] + \mathbb{E}[\|\boldsymbol{u}^k - \boldsymbol{u}^*\|_G^2 - \|\boldsymbol{u}^{k+1} - \boldsymbol{u}^*\|_G^2], \end{aligned} \tag{40}$$

where $G = diag\{\boldsymbol{H}_1, ..., \boldsymbol{H}_N, 1/\zeta\}$. $diag\{...\}$ denotes a diagonal matrix. Here, $\boldsymbol{H}_i = H_i \cdot \boldsymbol{I} = \boldsymbol{\Gamma}_i^{k+1} + A_i^T\boldsymbol{\rho}_i^{k+1}A_i$ is positive definite. h^{k+1} is some remainder term which is non-negative if for any $i, j \in [1:N]$

$$\begin{cases} \frac{H_j}{N^2}(\underline{\rho}_i - \frac{\zeta}{2}) \geq \sigma_{\max,j}^2\bar{\rho}_i^2 \\ \frac{1}{(N-1)^2}(\underline{\rho}_i - \frac{\zeta}{2})(\underline{\rho}_j - \frac{\zeta}{2}) \geq (\bar{\rho}_i - \underline{\rho}_i)^2 \end{cases} \tag{41}$$

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

Proof. Since f_i is convex,

$$\langle \nabla f_i(\tilde{\boldsymbol{\theta}}_i^{k+1}), \tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^* \rangle \geq f_i(\tilde{\boldsymbol{\theta}}_i^{k+1}) - f_i(\boldsymbol{\theta}_i^*). \tag{42}$$

Due to the optimality condition satisfied, we have

$$\nabla f_i(\tilde{\boldsymbol{\theta}}_i^{k+1}) = A_i^T(\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1}(A_i\tilde{\boldsymbol{\theta}}_i^{k+1} + \sum_{j \neq i} A_j\boldsymbol{\theta}_j^k - \boldsymbol{c})) + \boldsymbol{\Gamma}_i^{k+1}(\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1}). \tag{43}$$

Also from the KKT condition, for the optimal states $\boldsymbol{\theta}_{[1:N]}^*$, $\sum_{i=1}^N A_i\boldsymbol{\theta}_i^* = \boldsymbol{c}$. Substitute the above equations into (42), we have

$$\begin{aligned} &\langle A_i^T(\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1}(A\tilde{\boldsymbol{\theta}}^k - \boldsymbol{c}) + \boldsymbol{\rho}_i^{k+1}A_i(\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1})), \tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^* \rangle + (\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*)^T\boldsymbol{\Gamma}_i^{k+1}(\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1}) \\ &\quad \geq f_i(\tilde{\boldsymbol{\theta}}_i^{k+1}) - f_i(\boldsymbol{\theta}_i^*). \end{aligned} \tag{44}$$

Here, $A\boldsymbol{\theta}^k = \sum_{i=1}^N A_i\boldsymbol{\theta}_i^k$. Let $\boldsymbol{\lambda}^k = \boldsymbol{\lambda}^k - \boldsymbol{\lambda} + \boldsymbol{\lambda}$, we can rewrite (44) as

$$\begin{aligned} &\langle \boldsymbol{\lambda}^k - \boldsymbol{\lambda}, A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*) \rangle + (\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*)^T(\boldsymbol{\Gamma}_i^{k+1} + A_i^T\boldsymbol{\rho}_i^{k+1}A_i)(\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1}) - \langle A\tilde{\boldsymbol{\theta}}^k - \boldsymbol{c}, \boldsymbol{\rho}_i^{k+1}A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*) \rangle) \\ &\qquad \geq f_i(\tilde{\boldsymbol{\theta}}_i^{k+1}) - f_i(\boldsymbol{\theta}_i^*) - \boldsymbol{\lambda}^T A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*). \end{aligned} \tag{45}$$

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

$$\begin{aligned} &\frac{1}{\zeta}(\langle \boldsymbol{\lambda}^k - \tilde{\boldsymbol{\lambda}}^{k+1}, \boldsymbol{\lambda}^k - \tilde{\boldsymbol{\lambda}}^{k+1} \rangle + \langle \tilde{\boldsymbol{\lambda}}^{k+1} - \boldsymbol{\lambda}, \boldsymbol{\lambda}^k - \tilde{\boldsymbol{\lambda}}^{k+1} \rangle) + \sum_{i=1}^N (\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*)^T(\boldsymbol{\Gamma}_i^{k+1} + A_i^T\boldsymbol{\rho}_i^{k+1}A_i)(\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1}) \\ &- \langle A\tilde{\boldsymbol{\theta}}^k - \boldsymbol{c}, \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1}A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*) \rangle \geq F(\tilde{\boldsymbol{\theta}}^{k+1}) - F(\boldsymbol{\theta}_i^*) - \boldsymbol{\lambda}^T A(\tilde{\boldsymbol{\theta}}^{k+1} - \boldsymbol{\theta}_*). \end{aligned} \tag{46}$$

Let the matrix $G = diag\{\boldsymbol{H}_1, ..., \boldsymbol{H}_N, \frac{1}{\zeta}\}$, where $\boldsymbol{H}_i = H_i \cdot \boldsymbol{I}$. With the identity: $\|\boldsymbol{u}^k - \boldsymbol{u}^*\|_G^2 - \|\tilde{\boldsymbol{u}}^{k+1} - \boldsymbol{u}^*\|_G^2 = 2(\tilde{\boldsymbol{u}}^{k+1} - \boldsymbol{u}^*)^T G(\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^{k+1}) + \|\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^{k+1}\|_G^2$, we have

$$\begin{aligned} &\|\boldsymbol{u}^k - \boldsymbol{u}^*\|_G^2 - \|\tilde{\boldsymbol{u}}^{k+1} - \boldsymbol{u}^*\|_G^2 - \|\boldsymbol{u}^k - \tilde{\boldsymbol{u}}^{k+1}\|_G^2 + \frac{2}{\zeta}\|\boldsymbol{\lambda}^k - \tilde{\boldsymbol{\lambda}}^{k+1}\|^2 \\ &\quad - 2\langle A\boldsymbol{\theta}^k - \boldsymbol{c}, \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*)\rangle \geq 2[F(\tilde{\boldsymbol{\theta}}^{k+1}) - F(\boldsymbol{\theta}_i^*) - \boldsymbol{\lambda}^T A(\tilde{\boldsymbol{\theta}}^{k+1} - \boldsymbol{\theta}_*)]. \end{aligned} \tag{47}$$

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

$$\begin{aligned} h^{k+1} =& \sum_{i=1}^N H_i \|\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1}\|^2 - \frac{1}{\zeta}\|\boldsymbol{\lambda}^k - \tilde{\boldsymbol{\lambda}}^{k+1}\|^2 + 2\langle \sum_{i=1}^N A_i(\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1} + \tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*), \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*)\rangle \\ =& \sum_{i=1}^N H_i \|\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1}\|^2 - \zeta\|A(\tilde{\boldsymbol{\theta}}^{k+1} - \boldsymbol{\theta}_*)\|^2 + 2\langle \sum_{i=1}^N A_i(\boldsymbol{\theta}_i^k - \tilde{\boldsymbol{\theta}}_i^{k+1}), \sum_{i=1}^N \boldsymbol{\rho}_i^{k+1} A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*)\rangle \\ &+ 2\rho_0\|A(\tilde{\boldsymbol{\theta}}^{k+1} - \boldsymbol{\theta}_*)\|^2 + 2\langle \sum_{i=1}^N A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*), \sum_{i=1}^N (\boldsymbol{\rho}_i^{k+1} - \rho_0 \boldsymbol{I}) A_i(\tilde{\boldsymbol{\theta}}_i^{k+1} - \boldsymbol{\theta}_i^*)\rangle. \end{aligned} \tag{48}$$

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

$$\begin{cases} \frac{H_j}{N^2}(\underline{\rho}_i - \frac{\zeta}{2}) \geq \sigma_{\max,j}^2 \bar{\rho}_i^2, \\ \frac{1}{(N-1)^2}(\underline{\rho}_i - \frac{\zeta}{2})(\underline{\rho}_j - \frac{\zeta}{2}) \geq (\bar{\rho}_i - \underline{\rho}_i)^2. \end{cases} \tag{49}$$

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

$$f_i(\boldsymbol{\theta} + \varDelta_i^{k+1}) \leq f_i(\boldsymbol{\theta}) + \langle \varDelta_i^{k+1}, \nabla f_i(\boldsymbol{\theta})\rangle + \frac{M}{2}\|\varDelta_i^{k+1}\|^2.$$

In addition, $\mathbb{E}[\|\tilde{\boldsymbol{u}}^{k+1} - \boldsymbol{u}^{k+1}\|_G^2]$ w.r.t. the randomness of $\boldsymbol{\varDelta}^{k+1}$ is $\|\tilde{\boldsymbol{u}}^{k+1} - \boldsymbol{u}^*\|_G^2 + \mathbb{E}[\|\boldsymbol{u}^{k+1} - \boldsymbol{u}^*\|_G^2]$, since the mean of noise added is zero. Putting things together, the claim follows.

With Lemma 1, we can show the convergence and utility loss w.r.t. the norm $\|\cdot\|_G$.

Lemma 2. *Under the same condition as in Lemma 1 such that $h^k \geq 0$, let $\bar{\boldsymbol{\theta}}^K = \frac{1}{K}\sum_{k=1}^K \boldsymbol{\theta}_k$, then*

$$|\mathbb{E}[F(\bar{\boldsymbol{\theta}}^K)] - F(\boldsymbol{\theta}_*)| \leq \frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{G_x}^2 + \frac{4}{\zeta}\|\boldsymbol{\lambda}^*\|^2}{2K} + \frac{\sum_{k=1}^K \sum_{i=1}^N (M + H_i + \frac{1}{\zeta})\mathbb{E}[\|\boldsymbol{\varDelta}_i^k\|^2]}{K}, \tag{50}$$

where $\boldsymbol{\lambda}^*$ is the state of $\boldsymbol{\lambda}^k$ when $\boldsymbol{\theta}_k$ reaches the optimum. Under (ϵ, δ)-LDP, [5] the utility loss is $O(\frac{\sqrt{N}d\mathcal{B}\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{G_x}}{\epsilon})$. Here, $G_x = diag\{\boldsymbol{H}_1, ..., \boldsymbol{H}_N\}$.

Proof. Summing up (40) for $k = 0, 1, ..., K-1$ and applying the Jensen inequality, we have

$$\begin{aligned} \mathbb{E}[K\big(F(\bar{\boldsymbol{\theta}}^K) - F(\boldsymbol{\theta}_*)\big) - \boldsymbol{\lambda}^T A \sum_{k=0}^K (\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*)] &\leq \sum_{k=0}^{K-1} \mathbb{E}[F(\boldsymbol{\theta}_{k+1}) - F(\boldsymbol{\theta}_*) + \lambda^T A\boldsymbol{\theta}_{k+1}] \\ &\leq \frac{1}{2}\sum_{k=0}^{K-1}\sum_{i=1}^N (M + H_i + \frac{1}{\zeta})\mathbb{E}[\|\varDelta_i^{k+1}\|^2] + \mathbb{E}[\|\boldsymbol{u}^0 - \boldsymbol{u}^*\|_G^2]. \end{aligned} \tag{51}$$

[5] K-fold composition of (ϵ, δ)-(L)DP mechanisms can produce a $(\tilde{\epsilon}, K\delta + \tilde{\delta})$-DP, where $\tilde{\epsilon} = K\epsilon^2 + \epsilon\sqrt{-2K\log\tilde{\delta}}$.

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

$$\mathbb{E}[F(\bar{\boldsymbol{\theta}}^K) - F(\boldsymbol{\theta}_*)] \leq \frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{G_x}^2}{2K} + \frac{\sum_{k=1}^K \sum_{i=1}^N (\sqrt{M} + H_i + \frac{1}{\zeta}) \mathbb{E}[\|\boldsymbol{\Delta}_i^k\|^2]}{K}. \tag{52}$$

On the other hand, with KKT condition, $-\lambda^{*T} A(\bar{\boldsymbol{\theta}}^K - \boldsymbol{\theta}_*) = -\nabla F(\boldsymbol{\theta}_*)(\bar{\boldsymbol{\theta}}^K - \boldsymbol{\theta}_*)$. Applying convexity of $F(\cdot)$, we have

$$F(\boldsymbol{\theta}^*) \leq F(\bar{\boldsymbol{\theta}}^K) - \nabla F(\boldsymbol{\theta}_*)^T (\bar{\boldsymbol{\theta}}^K - \boldsymbol{\theta}_*). \tag{53}$$

Thus, by letting $\boldsymbol{\lambda} = 2\boldsymbol{\lambda}^*$ in (51) and adding $-\lambda^{*T} A(\bar{\boldsymbol{\theta}}^K - \boldsymbol{\theta}_*)$ on both sides of (53), we have

$$\begin{aligned} &-\lambda^{*T} A \sum_{k=0}^{K} (\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*) \leq F(\bar{\boldsymbol{\theta}}^K) - F(\boldsymbol{\theta}^*) - 2\boldsymbol{\lambda}^{*T} A \sum_{k=0}^{K} (\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*) \\ &\leq \frac{\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_*\|_{G_x}^2 + \frac{4}{\zeta}\|\boldsymbol{\lambda}^*\|^2}{2K} + \frac{\sum_{k=1}^K \sum_{i=1}^N (M + H_i + \frac{1}{\zeta}) \mathbb{E}[\|\boldsymbol{\Delta}_i^k\|^2]}{K}. \end{aligned} \tag{54}$$

Following [BST14], $\mathbb{E}[\|\boldsymbol{\Delta}_i^k\|^2 = O(\frac{d^2 K \mathcal{B}_\infty^2}{\epsilon^2})$ under (ϵ, δ)-LDP guarantee. Putting the upper and lower bounds of $F(\bar{\boldsymbol{\theta}}^K) - F(\boldsymbol{\theta}^*)$ together, let $K = O(\frac{\sqrt{N} d \mathcal{B}}{\epsilon})$, the claim follows.

Finally, we can bridge the above utility analysis and the case where we further apply first-order approximation in (38) as

$$\boldsymbol{\theta}_i^{k+1} := \boldsymbol{H}_i^{-1} \big[A_i^T \big(\boldsymbol{\lambda}^k - \boldsymbol{\rho}_i^{k+1} \big(\sum_{j \neq i} A_j \boldsymbol{\theta}_j^k - \boldsymbol{c} \big) \big) + \boldsymbol{\Gamma}_i^{k+1} \boldsymbol{\theta}_i^k - \nabla f_i(\boldsymbol{\theta}_i^k) \big] + \boldsymbol{\Delta}_i^{k+1}. \tag{55}$$

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

Lemma 3. *Under the modified updating procedure above, Lemma 1 holds with the same setup and h^{k+1} is non-negative if for any $i, j \in [1:N]$:*

$$\begin{cases} \frac{H_j - M}{N^2} (\underline{\rho}_i - \frac{\zeta}{2}) \geq \sigma_{\max,j}^2 \bar{\rho}_i^2 \\ \frac{1}{(N-1)^2} (\underline{\rho}_i - \frac{\zeta}{2})(\underline{\rho}_j - \frac{\zeta}{2}) \geq (\bar{\rho}_i - \underline{\rho}_i)^2. \end{cases} \tag{56}$$

Proof. With the smooth assumptions on ∇f_i, i.e., for any $\boldsymbol{\theta}$ and $\boldsymbol{y}$,

$$\|\nabla f_i(\boldsymbol{\theta}) - \nabla f_i(\boldsymbol{y})\| \leq M \|\boldsymbol{\theta} - \boldsymbol{y}\|,$$

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

$$f_i(\boldsymbol{\theta}) - f_i(\boldsymbol{y}) \leq \nabla f_i^T(\boldsymbol{z})(\boldsymbol{\theta} - \boldsymbol{y}) + \frac{M}{2} \|\boldsymbol{\theta} - \boldsymbol{z}\|^2.$$

Therefore, by replacing $\boldsymbol{\theta}_i^{k+1}$ with $\boldsymbol{\theta}_i^k$ in (42), all the deductions in Theorem 1 stay the same except that the term $\sum_{i=1}^N H_i \|\boldsymbol{\theta}_i^{k+1} - \boldsymbol{\theta}_i^k\|^2$ in the expression of h^{k+1} in (48) becomes $\sum_{i=1}^N (H_i - M) \|\boldsymbol{\theta}_i^{k+1} - \boldsymbol{\theta}_i^k\|^2$. Therefore, by replacing H_i in (41) with $H_i - M$, the claim follows.

Once h^k is non-negative as guaranteed by the above lemma, Lemma 2 still holds under the first-order approximation based updating rule, which completes the proof.

E Proof of Theorem 6

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

$$\boldsymbol{\theta}_i := \boldsymbol{w}_i \boldsymbol{y}_1^k + (\boldsymbol{I}_d - \boldsymbol{w}_i) \boldsymbol{y}_2^k - \eta_{k+1} \nabla f_i \big(\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2} \big) + \boldsymbol{\Delta}_i. \tag{57}$$

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

$$\begin{aligned}
\|\boldsymbol{y}_1^{k+1} - \boldsymbol{\theta}^*\|^2 &= \|\frac{1}{|S_{2k+1}|}\sum_{i\in S_{2k+1}}\big(\boldsymbol{w}_i\boldsymbol{y}_1^k + (\boldsymbol{I}_d - \boldsymbol{w}_i)\boldsymbol{y}_2^k - \boldsymbol{\theta}^* - \zeta_{k+1}\nabla f_i(\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2})\big)\|^2 \\
&= \|\frac{1}{|S_{2k+1}|}\sum_{i\in S_{2k+1}}\big(\boldsymbol{w}_i(\boldsymbol{y}_1^k - \boldsymbol{\theta}^*) + (\boldsymbol{I}_d - \boldsymbol{w}_i)(\boldsymbol{y}_2^k - \boldsymbol{\theta}^*)\big)\|^2 \\
&- 2\zeta_{k+1}\langle\frac{1}{|S_{2k+1}|}\sum_{i\in S_{2k+1}}(\boldsymbol{w}_i\boldsymbol{y}_1^k + (\boldsymbol{I}_d - \boldsymbol{w}_i)\boldsymbol{y}_2^k) - \boldsymbol{\theta}^*, \sum_{i\in S_{2k+1}}\frac{\nabla f_i(\frac{\boldsymbol{y}_1^k+\boldsymbol{y}_2^k}{2}) + \zeta_{k+1}^{-1}\boldsymbol{\Delta}_i}{|S_{2k+1}|}\rangle \\
&+ \|\sum_{i\in S_{2k+1}}\frac{1}{|S_{2k+1}|}\big(\zeta_{k+1}\nabla f_i(\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2}) + \boldsymbol{\Delta}_i\big)\|^2.
\end{aligned} \tag{58}$$

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,

$$(\sum_{i=1}^N \omega_i r_i)^2 \le \sum_{i=1}^N \omega_i r_i^2. \tag{59}$$

It is noted that in (58), the sum of weights of $(\boldsymbol{y}_1^k - \boldsymbol{\theta}^*)$ and $(\boldsymbol{y}_2^k - \boldsymbol{\theta}^*)$ is always the identity. With (59), by taking expectation on both sides of (58), we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{y}_1^{k+1} - \boldsymbol{\theta}^*\|^2] \le \mathbb{E}[\frac{\|\boldsymbol{y}_1^k - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^k - \boldsymbol{\theta}^*\|^2}{2}] &+ (\eta_{k+1}^2 G^2 + \mathbb{E}[\|\bar{\boldsymbol{\Delta}}_1^k\|^2]) \\
&- 2\eta_{k+1}\langle\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2}, \nabla F(\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2})\rangle,
\end{aligned} \tag{60}$$

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

$$\mathbb{E}[\frac{1}{|S_i|}\sum_{i\in S_i}\nabla f_i(\boldsymbol{\theta})] = \nabla F(\boldsymbol{\theta})$$

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

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

$$\mathbb{E}[\|\boldsymbol{y}_2^{k+1} - \boldsymbol{\theta}^*\|^2] \le \mathbb{E}[\frac{\|\boldsymbol{y}_1^k - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^k - \boldsymbol{\theta}^*\|^2}{2}] + (\eta_{k+1}^2 G^2 + \mathbb{E}[\|\bar{\boldsymbol{\Delta}}_2^k\|^2]) - 2\eta_{k+1}\langle\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2}, \nabla F(\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2})\rangle, \tag{61}$$

where $\bar{\boldsymbol{\Delta}}_2^k = \frac{1}{|S_{2k+2}|}\sum_{i\in S_{2k+1}}\|\Delta\|_i$. Applying the fact that $\langle\frac{\boldsymbol{y}_1^k+\boldsymbol{y}_2^k}{2} - \boldsymbol{\theta}^*, \nabla F(\frac{\boldsymbol{y}_1^k+\boldsymbol{y}_2^k}{2})\rangle \ge F(\frac{\boldsymbol{y}_1^k+\boldsymbol{y}_2^k}{2}) - F(\boldsymbol{\theta}^*)$ and taking the average on both sides of (60) and (61), we can bound $F(\frac{\boldsymbol{y}_1^k+\boldsymbol{y}_2^k}{2}) - F(x^*)$ as,

$$\begin{aligned}
&F(\frac{\boldsymbol{y}_1^k + \boldsymbol{y}_2^k}{2}) - F(\boldsymbol{\theta}^*) \le \frac{\eta_{k+1}^{-1}}{4}(\|\boldsymbol{y}_1^k - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^k - \boldsymbol{\theta}^*\|^2 - \|\boldsymbol{y}_1^{k+1} - \boldsymbol{\theta}^*\|^2 - \|\boldsymbol{y}_2^{k+1} - \boldsymbol{\theta}^*\|^2) \\
&+ (\frac{\eta_{k+1}}{2}G^2 + \frac{\mathbb{E}[\|\bar{\boldsymbol{\Delta}}_1^k\|^2] + \mathbb{E}[\|\bar{\boldsymbol{\Delta}}_2^k)\|^2]}{4\eta_{k+1}}).
\end{aligned} \tag{62}$$

Before we can derive a global convergence analysis, we need to give an upper bound on $\|\boldsymbol{y}_1^k - \boldsymbol{\theta}^*\|$ and $\|\boldsymbol{y}_2^k - \boldsymbol{\theta}^*\|$ with the initial divergence $\|\boldsymbol{y}_1^0 - \boldsymbol{\theta}^*\|$, $\|\boldsymbol{y}_2^0 - \boldsymbol{\theta}^*\|$ and the noise $\mathbb{E}[\|\bar{\boldsymbol{\Delta}}_1^k\|^2]$ and $\mathbb{E}[\|\bar{\boldsymbol{\Delta}}_2^k\|^2]$. It is noted that, with rearrangement on (62) and the fact $F(\boldsymbol{\theta}) - F(\boldsymbol{\theta}^*) \ge 0$,

$$\mathbb{E}[\|\boldsymbol{y}_1^{k+1} - \boldsymbol{\theta}^*\|^2] \le \mathbb{E}[\frac{\|\boldsymbol{y}_1^k - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^k - \boldsymbol{\theta}^*\|^2}{2}] + \eta_{k+1}^2(G^2 + \mathbb{E}[\|\eta_{k+1}^{-1}\boldsymbol{\Delta}_1^{k+1}\|^2)]. \tag{63}$$

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

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

with the assumption $\max_{k,j \in \{1,2\}} \mathbb{E}[(\bar{\boldsymbol{\Delta}}_j^k/\eta_k)^2] \leq V^2$. Thus, (62) can be further formulated as

$$\begin{aligned}
&\sum_{k=0}^{K-1} \frac{\mathbb{E}[F(\frac{\boldsymbol{y}_1^k+\boldsymbol{y}_2^k}{2}) - F(\boldsymbol{\theta}^*)]}{K} \\
&\leq \frac{\sum_{k=1}^{K}(\eta_{k+1}^{-1} - \eta_k^{-1})(\|\boldsymbol{y}_1^k - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^k - \boldsymbol{\theta}^*\|^2) + \eta_1^{-1}(\|\boldsymbol{y}_1^0 - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^0 - \boldsymbol{\theta}^*\|^2) + 2\sum_{k=0}^{K-1} \eta_{k+1}(G^2 + V^2)}{4K} \\
&= O(\frac{c\sqrt{K}(\|\boldsymbol{y}_1^0 - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^0 - \boldsymbol{\theta}^*\|^2) + c^{-1}(\log K + 2)\sqrt{K+1}(G^2 + V^2)}{K}),
\end{aligned} \tag{64}$$

and $\mathbb{E}[\sum_{i=1}^{N} f_i(\sum_{k=0}^{K-1}\sum_{i=1}^{N} \boldsymbol{\theta}_i^k/NK) - f_i(\boldsymbol{\theta}^*)] \leq \sum_{k=0}^{K-1}\sum_{i=1}^{N} \frac{\mathbb{E}[f_i(\bar{\boldsymbol{\theta}}^k)]-f_i(x^*)}{K}$. Here, we use the trick of SGD proof of selecting such a sequence of decreasing step size. To finally disclose the utility-privacy tradeoff, we specify the parameter of noise. In pure ϵ-LDP setting, since the sensitivity is bounded by $\mathcal{B}$ in l_∞, on each dimension we may add a noise following Lap$(0, \frac{\epsilon}{d\eta_k\mathcal{B}})$ to produce a total ϵ loss from d dimensions. Under the relaxed (ϵ, δ)-DP setting, with the strong composition theorem [KOV17], the variance $\mathbb{E}[(\boldsymbol{\Delta}_i^k/\eta_k)^2]$ can be reduced to $\tilde{O}(\frac{K}{N}d(\frac{\sqrt{d}\mathcal{B}}{\epsilon})^2)$. Substituting those into (64), we complete the proof of the Theorem that

$$\begin{cases}
\text{pure } \epsilon - LDP : \tilde{O}\Big(\frac{\sqrt{\|\boldsymbol{y}_1^0 - \boldsymbol{\theta}^*\|^2 + \|\boldsymbol{y}_2^0 - \boldsymbol{\theta}^*\|^2}(G + \frac{\sqrt{K}}{\sqrt{N}}\frac{d^{3/2}\mathcal{B}}{\epsilon})}{\sqrt{K}}\Big) = \tilde{O}(\frac{d^{3/2}\mathcal{B}}{\sqrt{N}\epsilon}) \\
\text{relaxed } (\epsilon, \delta) - LDP : \tilde{O}(\frac{d\mathcal{B}}{\sqrt{N}\epsilon}).
\end{cases} \tag{65}$$

F Non-asymptotic Convergence Rate Analysis

Theorems 5 and 6 have provided a unified upper bound of privacy-utility tradeoff, which is also asymptotically tight [KOV14]. In the following, we aim to capture the gap between the performance of fixed parameters and update mixing in a *non-asymptotic view*. For simplicity, we still consider the consensus case. In general, the framework of the proposed private DGD can be expressed as

$$\boldsymbol{\theta}_{k+1} = W_{k+1}\boldsymbol{\theta}_k - \xi_{k+1}\nabla F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k) + \boldsymbol{\Delta}^{k+1} \tag{66}$$

where, recalling w_{ij} in (16), $W_{k+1}(i,j) = w_{ij} \cdot \boldsymbol{I}$ is a random stochastic matrix determined by the weights selected by each agent. Here, $\boldsymbol{I}$ is the $d \times d$ identity matrix and $W(i,j)$ denotes the element of W at the crossing of the i^{th} row and j^{th} column (in a block sense). We call a non-negative matrix W stochastic if the sum of entries in each row is 1. W is doubly stochastic if W^T is stochastic as well. Let $W(i,:)$ and $W(:,j)$ denote the i^{th} row and j^{th} column, respectively. When $\mathbb{E}[W_k]$ is doubly stochastic, a similar upper bound of utility-privacy tradeoff for (66) can be derived as follows,

Theorem 7 *Selecting $\xi_k = O(\frac{1}{\sqrt{k}})$, when $f_{[1:N]}(\cdot)$ are L-Lipschitz, and $\mathbb{E}[W_k]$ is doubly stochastic,*

$$|F(\frac{\sum_{k=0}^{K-1}\boldsymbol{\theta}_k}{K}) - F(\boldsymbol{\theta}_*)| \leq \frac{\bar{\mathcal{R}}}{\sqrt{K}} + L\sum_{i=1}^{N}\mathbb{E}[\mathcal{T}_i^K], \tag{67}$$

where $W_0 = \boldsymbol{I}$, $\bar{\mathcal{R}}$ is a term invariant to the randomness of W_k, specified in the proof, and

$$\mathcal{T}_i^K = \|\frac{\sum_{k=0}^{K-1}\sum_{l=1}^{N}\mathbb{E}[W_{k+1}(l,:)]^T\boldsymbol{\theta}_k}{KN} - \frac{\sum_{k=0}^{K-1}\mathbb{E}[W_{k+1}(i,:)]^T\boldsymbol{\theta}_k}{K}\| + \|\frac{\sum_{k=0}^{K-1}(\boldsymbol{\theta}_i^k - \mathbb{E}[W_{k+1}(i,:)]^T\boldsymbol{\theta}_k)}{K}\|.$$

Proof. Applying the convexity of $F(\cdot)$ that $\langle \nabla F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k), \mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k - \boldsymbol{\theta}_* \rangle \geq F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}_*)$, we can derive the following:

$$\begin{aligned} &\mathbb{E}[F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}_*)] \\ &\leq \frac{1}{2}(\zeta^{k+1})^{-1}\big(\mathbb{E}\big[\|W_{k+1}\boldsymbol{\theta}_k - \boldsymbol{\theta}_*\|^2 + \|\xi^{k+1}\nabla F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k) - \boldsymbol{\Delta}^{k+1}\|^2 - \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*\|\big]\big) \\ &\leq \frac{1}{2}(\zeta^{k+1})^{-1}\big(\mathbb{E}\big[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_*\|^2 - \|\boldsymbol{\theta}_{k+1} - \boldsymbol{\theta}_*\|^2 + \|\xi^{k+1}\nabla F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k)\|^2\big] + \mathbb{E}[\|\boldsymbol{\Delta}^{k+1}\|^2]\big). \end{aligned} \tag{68}$$

Following the proof of Theorem 6, we may sum up and average out both sides of (68) for $k = 0, 1, ..., K-1$, which turns to be,

$$\frac{1}{K}\sum_{k=0}^{k-1} F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}_*) \geq F(\frac{\sum_{k=0}^{K-1}\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k}{K}) - F(\boldsymbol{\theta}_*).$$

However, such a bound cannot be straightforwardly used for measuring utility since $\boldsymbol{\theta}_k$ is not necessarily in consensus, i.e., $\boldsymbol{\theta}^k_i = \boldsymbol{\theta}^k_j$ for any $i, j \in [1:N]$, and thus $F(\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k) - F(\boldsymbol{\theta}_*) \geq 0$ may not hold. However, when we assume $f_i(\cdot)$ is L-Lipschitz, such a gap can be fixed via (69):

$$\begin{aligned} &\sum_{i=1}^{N} f_i(\frac{\sum_{k=0}^{K-1}\sum_{l=1}^{N}\mathbb{E}[W_{k+1}(l,:)^T]\boldsymbol{\theta}_k}{KN}) - F(\boldsymbol{\theta}_*) \\ &\leq F(\frac{\sum_{k=0}^{K-1}\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k}{K}) - F(\boldsymbol{\theta}_*) + L\sum_{i=1}^{N}\|\frac{\sum_{k=0}^{K-1}\sum_{l=1}^{N}\mathbb{E}[W_{k+1}(l,:)^T]\boldsymbol{\theta}_k}{KN} - \frac{\sum_{k=0}^{K-1}\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k}{K}\|, \end{aligned} \tag{69}$$

where $W_{k+1}(i,:)$ denotes the i^{th} row of W_{k+1}. It is noted that $\sum_{i=1}^{N} f_i(\frac{\sum_{k=0}^{K-1}\sum_{i=1}^{N}\mathbb{E}[W_{k+1}(i,:)^T]\boldsymbol{\theta}^k_i}{KN}) - F(\boldsymbol{\theta}_*) \geq 0$ since $\boldsymbol{\theta}_*$ is the optimum under consensus restrain. Therefore,

$$|F(\frac{\sum_{k=0}^{K-1}\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k}{K}) - F(\boldsymbol{\theta}_*)| \leq \frac{\bar{\mathcal{R}}}{\sqrt{K}} + L\sum_{i=1}^{N}\|\frac{\sum_{k=0}^{K-1}\sum_{l=1}^{N}\mathbb{E}[W_{k+1}(l,:)]^T\boldsymbol{\theta}_k}{KN} - \frac{\sum_{k=0}^{K-1}\mathbb{E}[W_{k+1}(i,:)]^T\boldsymbol{\theta}_k}{K}\| \tag{70}$$

where $\frac{\bar{\mathcal{R}}}{\sqrt{K}}$ corresponds to the average of the sum of the right hand of (68) following the proof of Theorem 6, which is invariant to the randomness of W_k. Moreover, applying L-Lipschitz continuity again,

$$|F(\frac{\sum_{k=0}^{K-1}\boldsymbol{\theta}_k}{K}) - F(\frac{\sum_{k=0}^{K-1}\mathbb{E}[W_{k+1}]\boldsymbol{\theta}_k}{K})| \leq L\sum_{i=1}^{N}\|\frac{\sum_{k=0}^{K-1}(\boldsymbol{\theta}^k_i - \mathbb{E}[W_{k+1}(i,:)]^T\boldsymbol{\theta}_k)}{K}\|. \tag{71}$$

Putting things together, the claim follows.

Theorem 7 indicates that, to study the utility loss, it suffices to consider *the rate of $\boldsymbol{\theta}_k$ towards the consensus*, i.e., $\boldsymbol{\theta}^k_i = \boldsymbol{\theta}^k_j$ for $i, j \in [1:N]$, where the deviation amongst $\boldsymbol{\theta}_k$ controls $\mathcal{T}^K_i$ on the right hand of (67). This is consistent with intuition. In the non-private case, where the convergence proofs in [MO17, SLY+14] guarantee $\boldsymbol{\theta}_k$ approaches the unique consensus optima $\boldsymbol{\theta}_*$, the excess loss is then proportional to the divergence among $\boldsymbol{\theta}_k$ in expectation. For quantification, we introduce the following metric $\phi(\boldsymbol{\theta})$, which denotes the largest deviation between any two elements of $\boldsymbol{\theta}$ in l_2 norm. For example, $\phi(\boldsymbol{\theta}_k) = \max_{i,j}\|\boldsymbol{\theta}^k_i - \boldsymbol{\theta}^k_j\|$. With the above understanding, we move our focus to $\phi(\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K)$. To proceed, we rewrite $\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K$ in the following form,

$$\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K = \big(\sum_{k=0}^{K-1}\big(\prod_{j=1}^{k}W_j\boldsymbol{\theta}_0 + \sum_{j=1}^{k}\prod_{l=j+1}^{k}W_l R_j\big)\big)/K \tag{72}$$

where for simplicity we rewrite $\boldsymbol{\theta}_{k+1} = W_{k+1}\boldsymbol{\theta}_k + R_{k+1}$ for some remainder term R_{k+1} and $\prod_{l=j}^{k}W_k = \boldsymbol{I}$ if $j > k$. Now, we measure the impact of random aggregation, i.e., random stochastic W_k applied in (66), on the consensus rate of $\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K$, compared to the fixed $W_k = W$ case. From (72), we consider the extreme scenario where the communication graph is fully connected and $W(i,j) = \frac{1}{N}\cdot\boldsymbol{I}$. The justification of this choice is as follows. Such fixed weight matrix W with identical rows has the property that for any $\boldsymbol{\theta}$, elements in $W\boldsymbol{\theta}$ are identical, i.e., $\phi(W\boldsymbol{\theta}) = 0$. Let $W_k = W$ in (72), then all the terms that are a multiple of W reach consensus and thus $\phi(\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K) = \phi((\boldsymbol{\theta}_0 + \sum_{k=1}^{K-1}R_k)/K)$. In contrast, for randomized W_k, those terms of multiples of W_k, such as $\prod_k W_k\boldsymbol{\theta}_0$ in (72), do not necessarily reach consensus, which produces the gap between the random and fixed cases. For simplicity, we assume N is even and consider the following way to randomize W_k: $W_k(i,j) = r_i^{k+1}\times\frac{2}{N}\cdot\boldsymbol{I}$ if $j\leq\frac{N}{2}$, otherwise $W(i,j) = (1-r_i^{k+1})\times\frac{2}{N}\cdot\boldsymbol{I}$. r_i^k are i.i.d. random variables in $(0,1)$. Clearly, $\mathbb{E}[W_{k+1}] = W$.

Lemma 4. *For two matrices W_1 and W_2 independently following the distribution described above,*

$$\mathbb{E}[\phi\big((W_1W_2)(:,j)\big)] \leq \frac{1}{2}\mathbb{E}[\phi(W_2(:,j))]. \tag{73}$$

Proof. For simplicity, we omit the $\boldsymbol{I}$ in the elements of W_k. 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 $W_2(1,j) = \min_l\{W_2(l,j)\}$. Thus,

$$\begin{aligned}\big|\tilde{W}(i_1,j)-\tilde{W}(i_2,j)\big| &= \big|\sum_{l=1}^N(W_1(i_1,l)-W_1(i_2,l))W_2(l,j)\big| \\ &= \big|\big(1-\sum_{l=2}^N W_1(i_1,l)-1+\sum_{l=2}^N W_1(i_2,l)\big)W_2(1,j)+\sum_{l=2}^N(W_1(i_1,l)-W_1(i_2,l))W_2(l,j)\big| \\ &= \big|\sum_{l=2}^N(W_1(i_1,l)-W_1(i_2,l))(W_2(l,j)-W_2(1,j))\big|.\end{aligned} \tag{74}$$

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 (74),

$$\begin{aligned}\mathbb{E}[|\tilde{W}(i_1,j)-\tilde{W}(i_2,j)|] &\leq \frac{N}{2}\mathbb{E}[\max_l W_1(i_1,l)-\min_l W_2(i_2,l)]\mathbb{E}[\phi(W_2(:,j)] \\ &= (\frac{3}{4}-\frac{1}{4})\mathbb{E}[\phi(W_2(:,j)] = \frac{1}{2}\mathbb{E}[\phi(W_2(:,j)].\end{aligned} \tag{75}$$

Lemma 4 indicates that even with randomness in W_k where $\phi(W_k\boldsymbol{\theta}) = 0$ does not necessarily hold, the product of W_k converges to a matrix of identical rows with an expected exponential rate. It is noted that for fixed $\boldsymbol{\theta}$, $\phi(\prod_k W_k\boldsymbol{\theta}) = O(\max_j\phi((\prod_k W_k)(:,j))$. Then, the gap can be bounded theoretically as follows.

Theorem 8 *Under (ϵ,δ)-LDP and the setup claimed above, for W_k fixed to W, $\phi(\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K) = O(\frac{1}{\sqrt{K}}+\frac{1}{K}+\frac{d\mathcal{B}_\infty}{\epsilon})$; as for randomized W_k generated as Lemma 4, $\phi(\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K) = O(\frac{1}{\sqrt{K}}+\frac{1}{K}+(1+\frac{1}{\sqrt{N}})\frac{d\mathcal{B}_\infty}{\epsilon})$.*

Proof. First, we rewrite the expression of R_k. Recalling (66), $R_k = \big(-\xi_k\nabla f_1(W\boldsymbol{\theta}_{k-1})+\boldsymbol{\Delta}_1^k, -\xi_k\nabla f_2(W\boldsymbol{\theta}_{k-1})+\boldsymbol{\Delta}_2^k, ..., -\xi_k\nabla f_N(W\boldsymbol{\theta}_{k-1})+\boldsymbol{\Delta}_N^k\big)$. It is noted that, though $\nabla F(\boldsymbol{\theta}^*) = \sum_{i=1}^N\nabla f_i(\boldsymbol{\theta}^*) = \boldsymbol{0}$, $\nabla f_i(\boldsymbol{\theta}^*)$ does not necessarily equal to 0. On the other hand, when we select $\xi_k = O(1/\sqrt{k})$, the sensitivity of $\xi_k\nabla f_i(\boldsymbol{\theta}^*)$ is $O(1/\sqrt{k})$, where accordingly the noise $\|\boldsymbol{\Delta}_i^k\|$ is also scaled in $O(\frac{1}{\sqrt{k}}\cdot\frac{d\sqrt{K}\mathcal{B}_\infty}{\epsilon})$ for (ϵ,δ) privacy guarantee when we ignore other terms. Thus, for fixed $W_k = W$,

$$\phi(\frac{\sum_{k=0}^{K-1}\boldsymbol{\theta}_k}{K}) = \phi(\frac{\boldsymbol{\theta}_0+\sum_{k=1}^{K-1}R_k}{K}) = O(\frac{1}{\sqrt{K}}+\frac{1}{K}+\frac{d\mathcal{B}_\infty}{\epsilon}).$$

Here, we use the fact $\sum_{k=1}^K\frac{1}{\sqrt{k}} = O(\sqrt{K})$.

In comparison, for randomized W_k, applying Lemma 4 on $\mathbb{E}[\phi(W_{k+1}\nabla F(\mathbb{E}[W_k]\boldsymbol{\theta}))]$, we have

$$\mathbb{E}[\phi(W_{k+1}\nabla F(\mathbb{E}[W_k]\boldsymbol{\theta}))] = \mathbb{E}[\phi(W_k \nabla F(W\boldsymbol{\theta}))] \leq \frac{1}{2}\phi(\nabla F(W\boldsymbol{\theta}))).$$

Moreover, since Δ_i^k are i.i.d. noise, due to the construction of W_k in the setup, $\mathbb{E}[\phi(W_{k+1}\boldsymbol{\Delta}^k)] = O(\frac{1}{\sqrt{N}}\mathbb{E}[\|\boldsymbol{\Delta}_i^k\|])$, where $\boldsymbol{\Delta}^k = (\boldsymbol{\Delta}_1^k, \boldsymbol{\Delta}_2^k, ..., \boldsymbol{\Delta}_N^k)$. Similarly, applying Lemma 4, the sum $\mathbb{E}[\sum_{K=1}^{\infty}\phi(\prod_{k=1}^{K} W_k(:,j))]$, which is a geometric decaying sequence, is bounded by 2. Therefore, $\phi(\frac{\sum_{k=0}^{K-1}\boldsymbol{\theta}_k}{K}) = O(\frac{1}{\sqrt{K}} + \frac{2}{K} + (1 + \frac{1}{\sqrt{N}}) \cdot \frac{d\mathcal{B}_\infty}{\epsilon})$, which asymptotically enjoys $O(\frac{d\mathcal{B}_\infty}{\epsilon})$ as N and K increase.

Theorem 8 shows that, with either fixed and randomized W_k, the consensus distance in the average $\sum_{k=0}^{K-1}\boldsymbol{\theta}_k/K$ measured with ϕ gradually approaches $d\mathcal{B}_\infty/\epsilon$ as N and k increase. This explains why update mixing comes almost free of utility loss.