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 $\epsilon$ in a multiplicative manner:

**Definition 1 (Pure $\epsilon$-DP).** A randomized algorithm $A : \mathcal{X} \to \mathcal{O}$, achieves $\epsilon$-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 $A(\cdot)$,

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

(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 $(\epsilon, \delta)$-DP [DKM06], where under the same setup a failure probability of $\delta$ is admitted:

**Definition 2 (Approximate $(\epsilon, \delta)$-DP).** A randomized algorithm $A : \mathcal{X} \to \mathcal{O}$, achieves $(\epsilon, \delta)$-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 $A(\cdot)$,

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

(2)

Other variants include concentrated DP [DR16 BST16 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 $\epsilon$-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 non-convex optimization and uniform noise perturbed gradient descent to escape saddle points [IGN+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 vision. 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 [IGN+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 $x_1, x_2, \ldots, x_N$, 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 $(x_i, y_i)$, $i = 1, 2, \ldots, N$, where $x_i$ is the feature and $y_i$ is the associated label, a virtual training sample $(\tilde{x}, \tilde{y})$ is constructed by:

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

Here, $(x_{i_1}, y_{i_1})$ and $(x_{i_2}, y_{i_2})$ 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, JCB+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 preserving 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 $\epsilon$ privacy loss defined in (1) to give a more refined, formal illustration.
We find that the corresponding privacy guarantee is between pure DP and approximate DP. To be specific, the hybrid architecture of Mixup and Noise: Sample Mixing and (Decentralized) SGD (Algorithm 2), are proposed, and we further study the local differential privacy (LDP) into more complicated decentralized algorithms. Two concrete examples, Modified private ADMM (Algorithm 1) update mixing during optimization. In Section 4, we show how to incorporate the proposed update mixing architecture of iterative optimization. In Section 3, we describe another model of the hybrid mixup and noise architecture, namely, the utility and privacy tradeoff, we provide results regarding the effect of Proposition 8 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 how 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, the hybrid structure is strictly better than that of a pure Laplace, which implies a sharper (advanced) composition bound when applying concentration inequalities \cite{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 Definition 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 (SGD) and its variants have been extensively studied \cite{BST14,LNR+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 Definition 3 in the context of private optimization.

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

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

for arbitrary $D \sim D'$.

**Definition 3 \cite{LNR+CMS11+JT13}.** The ex-post privacy loss $\epsilon(o)$ for an algorithm $A$ on an output $o$ is defined as

\[ \epsilon(o) = \sup_{D, D'} \epsilon_{D, D'}(o) = \sup_{D, D'} \log \frac{P(A(D) = o)}{P(A(D') = o)}, \]

for arbitrary $D \sim D'$.
some random weight $\lambda \in (0, 1)$. Thus, in general, an SGD protocol to privately update $\theta$ with mixed samples can be described as follows:

$$\theta^k = \theta^{k-1} - \eta_k \nabla f(\theta^{k-1}, \lambda k s_{i_{k,1}} + (1 - \lambda k) s_{i_{k,2}}) + \Delta^k.$$  \hfill (4)

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

**Example 1:** Assume that the gradient $\nabla f(\theta, s)$ is linear to the sample $s$ (for example $f(\cdot)$ is a ridge regression), and samples vary between $[0, 1]$. Ignoring the constants $\theta^{k-1}$ and $\eta_k$ in (4), two resultant mixture distributions of $\theta^k$ 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 $\nabla f(\theta^{k-1}, 0) = 0$ and $\nabla f(\theta^{k-1}, 1) = 1$. 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 $\nabla f(\theta^{k-1}, \lambda k s_{i_{k,1}} + (1 - \lambda k) s_{i_{k,2}})$ (underlined in (4)) is uniformly distributed between $s_{i_{k,1}}$ and $s_{i_{k,2}}$. Thus, the $l_{\infty}$ sensitivity bound $\mathcal{B}$ on the gradient is equivalent to that of the samples. With the above setup, the distribution of $\theta^k$ is equivalent to $Lap(0, \beta) * U[s_{i_{k,1}}, s_{i_{k,2}}]$, where we ignore the constant $\theta^{k-1}$, 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 $\sup_{\theta^k} \epsilon(\theta^k) = 2 \cdot \log(\frac{e^{\beta} - 1}{\beta})$. In comparison, without mixup, a pure Laplace mechanism will produce $\epsilon(\theta^k) = \frac{B^2}{n}$ uniformly for any $\theta^k$. Moreover, we apply $(\epsilon, \delta)$ measurement, defined in Definition 2 as a high probability bound on $\epsilon$, 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 $\sup_{\theta^k} \epsilon(\theta^k) = \frac{B^2}{n} \cdot \log(\frac{e^{\beta} - 1}{\beta})$. In an approximate $DP$ view, if $\delta > (1 - e^{-\tau})/(2\tau)$, it produces an $(\epsilon, \delta)$-DP such that

$$\epsilon = \frac{B}{n} \cdot \log \left( \frac{2e^{\tau - \beta \psi(\delta)} - e^{\tau - 2\beta \psi(\delta)} - 1}{B} - \frac{e^{\beta/2} - 1}{1/e^{\beta/2}} \cdot \frac{B - 1/(2\delta)}{1/(2\delta)} \cdot \frac{1 - e^{-\beta(B-1/(2\delta))}}{1 - e^{-1/B}} \right) \hfill (5)$$

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

**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 $\frac{\varepsilon}{2}$ that one may select the $s$ from $\mathcal{D}$ or $s'$ from $\mathcal{D}'$, which determines the subsampling factor in the $\epsilon(\theta^k)$ bound [BB#18]. 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 $\epsilon(\theta^k)$, 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 $\theta^k$.

---

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

4
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, \( U[s', s_0] * \text{Lap}(0, \beta) \) is reduced to \( \text{Lap}(0, \beta) \) and \( U[s, s_0] * \text{Lap}(0, \beta) \) becomes \( U[0, s_0] * \text{Lap}(0, \beta) \). The probability density of \( U[0, s_0] * \text{Lap}(0, \beta) \) can be described as

\[
\mathbb{P}(x) = \begin{cases} 
\frac{e^{\beta x}}{2 - e^{-\beta x} - e^{-\beta(s_0-x)}} & x < 0 \\
\frac{2 e^{-\beta(s_0-x)} - e^{-\beta x}}{2s_0} & 0 \leq x < s_0 \\
e^{-\beta s_0} & x \geq s_0
\end{cases}
\]  

(6)

When \( x < 0 \), it is straightforward to see that

\[
e^{\varepsilon} \leq \frac{\text{Lap}(0, \beta)(x)}{U[0, s_0] * \text{Lap}(0, \beta)(x)} \leq \frac{\beta e^{\varepsilon \beta}}{1 - e^{-\beta B}} = \frac{\beta B}{1 - e^{-\beta B}},
\]

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

\[
\frac{e^{\beta B} - 1}{\beta B} \geq \frac{\varepsilon}{1 - e^{-\beta B}},
\]

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

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 \( e^{\varepsilon} \), we have the following fact: for \( x < 0 \), the ratio \( e^{\varepsilon} \), which equals

\[
\frac{1}{s_0 \cdot (1 - e^{-\beta s_0})} \cdot e^{-\beta s_0} \left(1 - e^{-\beta(s'-s_0)}\right)^{-1},
\]

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 \( \lim_{x \to 0} \frac{1 - e^{-\beta x}}{x} = \alpha \). Furthermore, it is noted that the function \( \frac{e^{\beta s'} - 1}{\beta s'} \) increases with increasing \( s' \). To conclude, the extreme case of \( \varepsilon \) when \( x < 0 \) happens when \( s_0 = s' = B \). Similarly, for \( x \in [0, s_0] \), we have the following observation on the probability density ratio:

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

which is equivalent to \( (e^{-\beta x} - 1)^2 \geq 0 \). Due to the symmetry, therefore, the upper bound of \( e^{\varepsilon} \) in (a) also happens when \( s_0 = s' = B \), 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 \( s' = B \). Thus, we have proved that \( \varepsilon \leq \log\left(\frac{e^{\beta B} - 1}{\beta B}\right) \).

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

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

\[
\text{Pr}(x \leq t) = \frac{1 - e^{-\beta s_0}}{(2e^{\beta s_0} - e^{-\beta(s_0-t)})\beta B} + \frac{1}{2} - \frac{1}{2e^{\beta t}}(1 - e^{-\beta t}) \cdot \frac{e^{-\beta s_0}}{1 - e^{-\beta(s_0-t)} - e^{-\beta s_0}}.
\]

Therefore, provided a failure probability \( \delta > \frac{1 - e^{-\beta s_0}}{(2e^{\beta s_0} - e^{-\beta(s_0-t)})\beta B} \), since \( 2 - e^{-\beta t} - e^{-\beta(s_0-t)} \), 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

\[
\text{Pr}(x \leq \frac{s_0}{1 - e^{-\beta s_0/2}}(\delta - \frac{1 - e^{-\beta s_0}}{2\beta s_0})) \leq \delta.
\]
Let $\psi(\delta) = \frac{s_0}{1-e^{-\beta s_0}} \left( \delta - \frac{1-e^{-\beta s_0}}{2s_0} \right)$ and substitute the above into the expression of $e^\epsilon$, then we have the relationship of $(\epsilon, \delta)$ that if $\psi(\delta) > 0$,

$$
e^\epsilon = \frac{\frac{1}{s_0} \left( 2 - e^{-\beta x} - e^{-\beta (s_0 - x)} \right)}{e^{-\beta (s_0 - \psi(\delta))} (1 - e^{-\beta (x - s_0)})} \leq \frac{\left( 2 - e^{-\beta \psi(\delta)} - e^{-\beta (s_0 - \psi(\delta))} \right) (B - s_0)}{1 - e^{-\beta (B - s_0)}} \cdot \frac{e^{\beta (s_0 - \delta ) / s_0}}{s_0} \leq \frac{\left( 2 - e^{-\beta \psi(\delta)} - e^{-\beta (B - \psi(\delta))} \right) e^{\beta (B - \psi(\delta))}}{B},$$

(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 $s_0 = B$.

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

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

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

In Theorem 1, we characterize the privacy amplification in a $(\epsilon, \delta)$ 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 $D$ and output $\theta^k$, $\sup_{D'} \epsilon_{D, D'}(\theta^k)$ uniformly equals $B\beta / n$. Thus, under the hybrid sample mixing and Laplace structure, we define the ratio

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

(9)

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

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

$$\sup_{D} \mathbb{E}_{D} \sup_{D'} \epsilon_{D, D'}(\theta^k) \leq \frac{2}{n} \max \left\{ \log \left( \frac{\beta B}{2(1 - e^{-\beta B/2})} \right) , \max_{s \leq s_0 \in [0, B]} \left[ \log \left( \frac{s_0 (1 - e^{-\beta (s_0 - s)})}{(s_0 - s)(1 - e^{-\beta s_0})} \right) \right] + \log \left( \frac{(B - s_0)(7e^{\beta (s_0 - s)} - (12\beta (s_0 - s) + 3))}{6\beta (1 - e^{-\beta (B - s_0)})(s_0 - s)^2} + \frac{e^{\beta (s_0 - s)} (B - s_0)(1 - e^{-\beta (s_0 - s)})^2}{2\beta (s_0 - s)^2 (1 - e^{-\beta (s_0 - s)})} \right) \right\},$$

(10)

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

**Proof.** With a similar reasoning as shown in the proof of Theorem 1 we only need to consider two adjacent datasets in a form $D = \{s, s_0\}$ and $D' = \{s', s_0\}$, where $s, s', s_0 \in [0, B]$. First, we consider an extreme case if $s_0 = s$, which produces a pure Laplace distribution. To determine the $\sup_{D'} \epsilon_{D, D'}(x)$, when $x < s$, to maximize the ratio, $s'$ should equal $B$ and

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

\[
\mathbb{E}_{x \sim \text{Lap}(s,\beta)} \sup_{s'} \epsilon_{s,s'}(x) = \frac{1}{2} \left( \log(\frac{\beta(\mathcal{B} - s)}{1 - e^{-\beta(\mathcal{B} - s)}}) + \log(\frac{\beta s}{1 - e^{-\beta s}}) \right).
\] (11)

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

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). \( s' \leq s_0 \leq s \); (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 \( w z_1 + (1 - w) z_2 \leq z_2 \), then clearly \( z_2 \geq z_1 \). Back to our cases, recall the convolutional density function \( U[0,s] \ast \text{Lap}(0,\beta)(x) = \int_{0}^{s} \frac{\beta}{2 \pi} e^{-\beta|x-t|} dt \), 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 \( D = \{ s, s_0 \} \) is \( \frac{I_1 + I_2}{s_0 - s} \), while that of \( D' = \{ s', s_0 \} \) is \( \frac{I_1 + I_2}{s_0 - s} + \frac{I_2}{s'} \) \( \frac{s_0 - s}{s'} \). Thus, if \( \frac{I_1 + I_2}{s_0 - s} \geq \frac{I_1}{s_0 - s} \), then \( \frac{I_1 + I_2}{s_0 - s} \geq \frac{I_1}{s_0 - s} \). 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 \( \epsilon_{D,D'} \) 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

\[
\sup_{s'} \epsilon_{s,s'}(x) = \frac{e^\beta(x-x)}{2(s_0-s)} \frac{1-e^{-\beta(s_0-s)}}{1-e^{-\beta(s_0-s)}} = \frac{B-s_0}{s_0-s} \frac{1-e^{-\beta(s_0-s)}}{1-e^{-\beta(s_0-s)}}.
\] (12)

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

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

Now, we consider the other case when \( x \in \left[ \frac{s}{2}, \frac{\mathcal{B}+s_0}{2} \right] \). To derive an upper bound, we simply bound \( \sup_{D'} \epsilon_{D,D'} \leq \epsilon_{D,s'=0} + \epsilon_{D,s'=1} \). We first calculate the expectation of \( \epsilon_{D,s'=1} \) 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

\[
\mathbb{E}_{x \in [s, s_0]} \epsilon_{D,s'=1} \leq \Pr(x \in [s, s_0]) \log \left[ \frac{1}{\Pr(x \in [s, s_0])} \int_{s}^{s_0} \frac{(2 - e^{-\beta(x-s)} - e^{-\beta(s_0-x)})^2}{e^{-\beta(s_0-x)}(1-e^{-\beta(s_0-x)})} dx \right].
\] (13)

The integral quantity in (13) can be expressed as

\[
\int_{s}^{s_0} \frac{(2 - e^{-\beta(x-s)} - e^{-\beta(s_0-x)})^2}{e^{-\beta(s_0-x)}(1-e^{-\beta(s_0-x)})} dx \leq \frac{B-s_0}{2(1-e^{-\beta(s_0-s)})} \frac{7e^\beta(s_0-s)}{(s_0-s)^2} - \frac{12\beta(s_0-s) + 3}{3\beta}.
\] (14)

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

\[
\epsilon_{D,s'=0} \leq \log \left( \frac{1-e^{-\beta(s_0-s)}}{1-e^{-\beta s_0}} \right) = \log \left( \frac{s_0(1-e^{-\beta(s_0-s)})}{(s_0-s)(1-e^{-\beta s_0})} \right),
\]

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

\[
\sup_{D} \sup_{\Theta'} \mathbb{E}_{D'}(\theta^k) \leq \max_{s,s_0 \in [0,B]} \left[ \log \left( \frac{s_0(1-e^{-\beta(s_0-s)})}{(s_0-s)(1-e^{-\beta s_0})} \right) + \log \left( \frac{(B-s_0)(7e^\beta(s_0-s) - 12(s_0-s) + 3)}{6\beta(1-e^{-\beta(s_0-s)})(s_0-s)^2} + \frac{e^\beta(s_0-s)(B-s_0)(1-e^{-\beta(s_0-s)})^2}{2\beta(s_0-s)^2(1-e^{-\beta(s_0-s)})^2} \right) \right].
\] (15)
Finally, multiplying by the subsampling factor $\frac{2}{n}$, we have the claim.

In Fig. 1, we present the simulation of the ratio $\gamma$ defined in (9). From Fig. 1, the privacy amplification ratio, within $[0.5, 1]$, is determined by $\tau = \beta B$, where a smaller $\tau$ 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 $\epsilon$-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 $\epsilon$ 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, $\gamma$ 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 $(\epsilon, \delta)$-DP mechanisms can produce a $(\tilde{\epsilon}, K\delta + \tilde{\delta})$-DP, with $\tilde{\epsilon} = K\epsilon^2 + \sqrt{-2K\epsilon^2 \log \delta}$ [KOV17], where $\tilde{\delta}$ is a free parameter. We consider the composition of $\epsilon(\theta^k)$ across multiple iterations. Following the idea in [KOV17], since $\epsilon(\theta^k)$ is also bounded from Theorem 1 by $\frac{\delta}{n} \cdot \log \left(\frac{n-1}{B\epsilon^2}\right)$, 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 $\sup_{D_D} \sup_{D_D'} \epsilon_D, \epsilon_{D_D'}(\theta^k)$ proposed, which is defined on the point-wise worst-case ex-post $\epsilon$-loss, is stronger than the KL-divergence between the two output distributions from $D$ and $D'$. Thus, $\gamma$ 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 $\sum_{i=1}^{N} f_i(\theta_i)$ under a consensus restriction $\theta_1 = \theta_2 = \ldots = \theta_N$, 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,

$$\theta_i^k = \sum_{j=1}^N w_{ij}^k \theta_j^{k-1} - \eta_k \nabla f_i(\theta_i^{k-1}) + \Delta_i^k. \quad (16)$$

Here, $w_{ij}^k \in [0, 1]$ is the weight assigned to $\theta_j$ 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, $\theta_i^k$ 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 $\theta_i^{k-1}$. Thus, the distribution of $\theta_i^k$ 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 $\theta_i^k$ in (16), for simplicity we assume $N = 2$, i.e., the underlying quantity in (16) is a 2-mix, and the difference between $\theta_i^{k-1}$ and $\theta_j^{k-1}$ is $\omega$. Thus, equivalently, we only need to consider the following mixture distribution: the sum of $\text{Lap}(0, \beta)$ and an independent uniform distribution within interval $[0, \omega]$ denoted by $U[0, \omega]$, whose probability density is a convolution $\text{Lap}(0, \beta) * U[0, \omega]$. We assume the $l_\infty$ sensitivity of $\nabla f_i$ to be $\mathcal{B}$ and $\eta_k = 1$.

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

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

$$\epsilon(\theta_i^k) \leq \max_{\tau \in [0, \mathcal{B}]} \left| \log \left[ \frac{\int_0^\omega e^{-\beta|\theta_i^k-\epsilon|} d\epsilon}{\int_t^{t+\omega} e^{-\beta|\theta_i^k-\epsilon|} d\epsilon} \right] \right| \leq \beta \mathcal{B}.$$  

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

$$\epsilon \leq \log \left( \max \left( \frac{2e^{3\beta(1-\tau')}}{1-e^{-\tau'}}, \frac{2}{1-e^{-\tau'}} \right) \right),$$

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 $w_{ij}^k$ are fixed, if the pure Laplace mechanism produces $\epsilon_0$-DP, then with the same setup, the hybrid of update mixing and the Laplace mechanism satisfies the same worst case, i.e., $\sup_{t, z} \epsilon(\theta_i^k) = \epsilon_0$. 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 $\epsilon$ bound given in Theorem 3. Analogous to Theorem 3, we also quantify the expected privacy loss savings in the worst case for the update mixing model.

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

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

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

$$\gamma = \frac{\sup_{t, z} \mathbb{E}_{\mathcal{D}} \sup_{\mathcal{D}, \mathcal{D}'} \epsilon_{\mathcal{D}, \mathcal{D}'}(\theta_i^k)}{\beta \mathcal{B}},$$

9
which quantifies the privacy gain compared to pure Laplace in the update mixing model. We show the effect of $\beta$ and $B$ on $\gamma$ in Fig. 2(a) and (b), respectively. Clearly, larger $\omega$ (larger update difference) and $\beta$ (smaller noise) with a smaller sensitivity $B$ produce stronger privacy amplification (smaller $\gamma$), 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 $G(N, E)$. Nodes are indexed as $N = \{1, \ldots, N\}$ and when two nodes $i$ and $j$ are neighbors that can communicate, $(i, j) \in E$. Each node $i$ has a function $f(s_i, \theta_i)$ that we regard as a loss function determined by the sample $s_i$ held locally with the parameter $\theta_i$ to be optimized. In this paper, we always assume that $f(s_i, \cdot)$ is a differentiable convex function $C \rightarrow \mathbb{R}$ and $\theta_i \in C \subset \mathbb{R}^d$. $C$ can be viewed as the constraint, assumed to be a closed convex set. We express the objective function to be minimized as

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

In many learning problems, $\theta[1:N]$ stands for one parameter to be collaboratively optimized. Here, $[1 : N]$ is the compact form of $\{1, 2, \ldots, N\}$. We term the problem as consensus optimization if the constraint requires that all $\theta_i$ be equal, which can still be enforced by a linear constraint $\sum_{i=1}^{N} A_i \theta_i = 0$ [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 local randomization procedure satisfies $\epsilon$-DP with respect to (w.r.t.) its own local dataset. If each local randomization procedure satisfies $\epsilon$-DP with respect to (w.r.t.) its own local dataset, we then say the whole protocol achieves $\epsilon$-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

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

Here, $D_i = (\hat{s}_1, \ldots, \hat{s}_i, \ldots, \hat{s}_N)$ and $D'_i = (\hat{s}_1, \ldots, \hat{s'}_i, \ldots, \hat{s}_N)$ are two arbitrary candidate sets of samples, where $\hat{s}_i$ denotes the sample agent $i$ holds, i.e., $D_i$ and $D'_i$ differ at the samples agent $i$ holds.
where the Lagrangian function is defined as we consider the more complicated case of ADMM.

Algorithm 1 Modified Private ADMM with First-order Approximation

Input: Local functions \( f_i(\cdot) \), step penalty \( \zeta \).

1. Initialize \( \theta_i^0 \) randomly, \( \lambda_i^0 = 0 \). Each agent selects a private constant \( H_i \).
2. for \( k = 0, 1, 2, ..., K - 1 \) do
   - Agents \( i = 1 \) to \( N \) do in parallel:
     - Randomly pick two positive diagonal matrices \( \bar{\rho}_i^{k+1} \) and \( \Gamma_i^{k+1} \) such that \( (N - 1)\bar{\rho}_i^{k+1} + \Gamma_i^{k+1} = H_i \cdot I_d \) and update \( \theta_i^{k+1} \) in parallel, where \( \Delta_i^{k+1} \sim \text{Lap}(0, \beta_i^{k+1}) \):
       \[
       \theta_i^{k+1} := \frac{\bar{\rho}_i^{k+1}}{H_i} \theta_i^k + \frac{(N - 1)\bar{\rho}_i^{k+1} \sum_{j \neq i} \theta_j^k}{H_i} \frac{\lambda_i^k}{N - 1} - H_i^{-1} \nabla f_i(\theta_i^k)_B + H_i^{-1} \lambda_i^k + \Delta_i^{k+1}.
       \]
     - Exchange \( \theta_i^{k+1} \) and then update \( \lambda_i^{k+1} := \lambda_i^k - \zeta \sum_{j \neq i} (\theta_j^{k+1} - \theta_i^{k+1}) \).
   - end for

LDP described in Definition \[ \text{LVR}_1 \] 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 \( \epsilon \)-LDP by setting \( \epsilon = \max_i \sup_{\theta_i} \epsilon_i(o) \) \[ \text{LVR}_1 \] \[ \text{WCHRP}_{17} \]. For an iterative optimization, \( o \) in Definition \[ \text{LVR}_1 \] includes all (immediate) outputs across iterations. For simplicity, we use \( f_i(\theta_j) \) to denote \( f(s_i, \theta_i) \) 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 \[ \text{NO09} \], \[ \text{LZZ}_{17} \], and EXTRA \[ \text{SLWY}_{15} \]. The latter relies on solving a constrained problem with dual variables to minimize some Lagrangian function, such as Alternating Direction Method of Multipliers (ADMM) \[ \text{WO12} \]. 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 \text{mixup} into GD. In the following, we consider the more complicated case of ADMM.

The updating rule of ADMM with the dual method for \[ \text{LVR}_1 \] relies on solving an optimization problem:

\[
\theta_i^{k+1} := \arg \min_{\theta_i} \mathcal{L}(\theta_i^{k}, \ldots, \theta_i^{k}, \theta_N^{k}, \lambda) + \frac{\Gamma}{2} \| \theta_i - \theta_i^{k} \|_q^2 + \frac{\rho}{2} \| A_i \theta_i + \sum_{j \neq i} A_j \theta_j^{k} - c \|_q^2 + \Delta_i^{k+1},
\]

where the Lagrangian function is defined as \( \mathcal{L}(\theta_1, \ldots, \theta_N, \lambda) = \sum_{i=1}^N f_i(\theta_i) - \lambda^T (\sum_{i=1}^N A_i \theta_i - c) \). Here, \( \| \|_q \) denotes the \( l_q \) norm. For brevity, \( \| \| \) denotes the standard \( l_2 \) norm in the following. The Lagrangian multiplier \( \lambda^{k+1} \) is updated through \( \lambda^{k+1} := \lambda^k - \zeta (\sum_{i=1}^N A_i \theta_i^{k+1} - c) \). If we allow each agent to independently select random penalties \( \rho \) and \( \Gamma \) across iterations, the \( \theta_i^{k+1} \) produced similarly follows a mixture Laplace distribution with a random mean.

However, we have to stress that closed-form optima of \[ \text{LVR}_1 \] may not exist. To reduce the computational complexity, in contrast to previous ADMM protocols \[ \text{WO12} \] \[ \text{LZZ}_{17} \] \[ \text{ZKL}_{18} \], we consider a first-order approximation for each \( f_i \):

\[
f_i(\theta_i) \approx f_i(\theta_i^{k}) + \nabla f_i(\theta_i^{k})(\theta_i - \theta_i^{k}).
\]

Through randomizing \( \rho \) and \( \Gamma \) in \[ \text{LVR}_1 \] accompanied with the approximation \[ \text{LVR}_2 \], a modified private ADMM with \text{mixup} is accordingly derived as Algorithm \[ \text{LVR}_3 \]. Meanwhile, stemming from \[ \text{LVR}_3 \], we formally describe a Distributed SGD with \text{mixup} as Algorithm \[ \text{LVR}_4 \].

It is clear that Algorithms \[ \text{LVR}_1 \] and \[ \text{LVR}_2 \] share a very similar structure except for the dual variable \( \lambda_i^k \) in ADMM. Intuitively, Term \( A \) in either \[ \text{LVR}_1 \] or \[ \text{LVR}_2 \] behaves as \text{mixup} to merge the updates from the previous iteration and Term \( B \) corresponds to the effect from the function \( f_i \) on updating \( \theta_i^{k+1} \) followed by a Laplace noise. Thus, with fixed updates from the last iteration, in each dimension, the produced distribution of \[ \text{LVR}_1 \] and \[ \text{LVR}_2 \] is equivalent to the update mixing model

\[ \text{Here, we omit the projection step if } C \subseteq \mathbb{R}^d \text{ since we are interested in arbitrary constraints. For simplicity, in this section, we assume } G \text{ is fully connected and only consider the consensus problem temporarily, while we provide a convergence proof for the general case in Theorem } 5 \].

11
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 $y^0_i$ and $\bar{y}^0_i$.

for $k = 0, 1, 2, \ldots, K - 1$ do
  Agents in $S_{2k+1}$ and $S_{2k+2}$ do in parallel:
  Randomly pick a positive diagonal matrix $w_i$ of which the non-zero elements are within $(0,1)$. Then, with $\Delta_i^{k+1} \sim \text{Lap}(0, \beta_{k+1})$, update $\theta_i$ as:
  \[
  \theta_i := w_i y_i^k + (I_i - w_i) y_{\bar{i}}^k - \eta_{k+1} N \nabla f_i \left( \frac{y_i^k + y_{\bar{i}}^k}{2} \right) + \Delta_i,
  \]

  Broadcast $\theta_i$ to agents in $S_{2k+3}$ and $S_{2k+4}$ where $y_i^{k+1} = \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} \theta_i$ and $y_{\bar{i}}^{k+1} = \frac{1}{|S_{2k+2}|} \sum_{i \in S_{2k+2}} \theta_i$.
end for

The analysis of the LDP-utility tradeoff is described in Section 5. We can still use $B$ to denote the $l_\infty$ sensitivity where $\max_i \sup_{\theta} \| \nabla f_i(s_i, \theta) - \nabla f_i(s_i', \theta) \|_\infty \leq B$ for two arbitrary candidates $s_i$ and $s_i'$ 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 $\theta_i$ 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 and 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 $\theta_k = (\theta_1^k, \theta_2^k, \ldots, \theta_N^k)$ and $F(\theta_k) = \sum_{i=1}^N f_i(\theta_i^k)$, and accordingly $\nabla F(\theta_k) = (\nabla f_1(\theta_1^k), \ldots, \nabla f_N(\theta_N^k))$. We first recall some commonly used notions in convex optimization analysis:

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

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

Analysis of Algorithm 1 Let $G_x = \text{diag}(H_1, \ldots, H_N, 1/\zeta)$ denote a diagonal matrix, where $H_i = H_i \cdot I = H_i^{k+1} + A_i^T \rho_i^{k+1} A_i$ is positive definite.

Theorem 5 If $f_i(\cdot), i = 1, 2, \ldots, N$, are convex and $M$-smooth, and for any $i, j \in [1 : N]$
\[
\left\{ \begin{array}{l}
\frac{H_j - M}{2N^2} (\rho_i - \bar{\rho}_i)^2 \geq \sigma_{\text{max},j}^2 \rho_i^2 \\
\frac{1}{N - 1^2} (\rho_i - \bar{\rho}_i) (\rho_j - \bar{\rho}_j) \geq (\bar{\rho}_i - \bar{\rho}_j)^2,
\end{array} \right.
\]
(23)

where $\rho_i \cdot I \leq \rho_i^{k+1} \leq \bar{\rho}_i \cdot I$ for some constants $0 < \rho_i < \bar{\rho}_i$ and $\sigma_{\text{max},j}$ is the largest singular value of $A_i$, then for Algorithm 1
\[
\|E[F(\hat{\theta}^K)] - F(\theta_*)\| \leq \frac{\|\theta_0 - \theta_*\|_{G_x}^2 + 4 \|\lambda^*\|^2}{2K} + \sum_{k=1}^K \sum_{i=1}^N (M + H_i + \frac{1}{\zeta}) E[\|\Delta_i^k\|^2],
\]
(24)

where $\hat{\theta}^K = \frac{1}{K} \sum_{k=1}^K \theta_k$ and $\lambda^*$ is the state of $\lambda^k$ when $\theta_k$ reaches the optimum $\theta_* = (\theta_1^*, \theta_2^*, \ldots, \theta_N^*)$ of (17). Under $(\epsilon, \delta)$-LDP, the utility loss is $O(\frac{\sqrt{Nd}\|\theta_0 - \theta_*\|_{G_x}}{\epsilon - \delta})$. 

12
Here, $B$ still stands for the $l_{\infty}$ 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 5), and then combine both to complete the proof.

Analysis of Algorithm 2 In comparison to ADMM, DySGD only captures consensus optimization. When $\|\nabla f_i\|$ is bounded, the following theorem shows the privacy-utility tradeoff of Algorithm 2. Here, we assume $\theta^* = (\theta^*, \theta^*, \ldots, \theta^*)$ is the optimum solution to (17).

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

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

When we select the step size $\eta_k = e^{\sqrt{c}}$ for some constant $c$, then $\frac{1}{N} \sum_{i=1}^{N} f_i(\frac{\sum_{k=0}^{K-1} (y_{K}^i + y_{K}^i)}{2K}) - F(\theta^*)$ is upper bounded by

$$
\frac{c(\|y_0^i - \theta^*\|^2 + \|y_0^i - \theta^*\|^2)}{4 \sqrt{K}} + \frac{(\log K + 2) \sqrt{K} + 1 (G^2 + V^2)}{2Kc}.
$$

In $\epsilon$-LDP, Algorithm 2 has utility loss $\tilde{O}(\sqrt{N} d^{3/2} / \epsilon)$. With $(\epsilon, \delta)$ relaxation, this bound can be sharpened to $\tilde{O}(\sqrt{N} d B / \epsilon)$. $\tilde{O}$ 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 E.

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, DGP19]. 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 $\gamma$ 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 $\alpha$-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 [FMT18].

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, HHG19, HTP17] either assume strong convexity or central aggregators.


A Simulation Results

A.1 Theoretical privacy amplification

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

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 (without $\gamma$).

Tests are run with different parameter settings. 100 independent runs of each algorithm for comparison are performed in Algorithm 1 and 2, respectively. One of the key observations is, with the same setup, the hybrid randomization 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

$\gamma$ and $\omega$ that $\beta$ is fixed to 0.001. With $\gamma$ ranging from 0.1 to 1 and $\omega$ from 2 to 10, clearly larger $\gamma$ and $\omega$, corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification.

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

$\gamma$ and $\omega$ that $\beta$ is fixed to 0.001. With $\gamma$ ranging from 0.1 to 1 and $\omega$ from 2 to 10, clearly larger $\gamma$ and $\omega$, corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification.

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  

$\gamma$ and $\omega$ that $\beta$ is fixed to 0.001. With $\gamma$ ranging from 0.1 to 1 and $\omega$ from 2 to 10, clearly larger $\gamma$ and $\omega$, corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification. 

$\gamma$ and $\omega$ that $\beta$ is fixed to 0.001. With $\gamma$ ranging from 0.1 to 1 and $\omega$ from 2 to 10, clearly larger $\gamma$ and $\omega$, corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification.

$\gamma$ and $\omega$ that $\beta$ is fixed to 0.001. With $\gamma$ ranging from 0.1 to 1 and $\omega$ from 2 to 10, clearly larger $\gamma$ and $\omega$, corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification.

$\gamma$ and $\omega$ that $\beta$ is fixed to 0.001. With $\gamma$ ranging from 0.1 to 1 and $\omega$ from 2 to 10, clearly larger $\gamma$ and $\omega$, corresponding to a longer interval length and noise of smaller variance, lead to better privacy amplification.
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 $\epsilon$ 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 $\theta_{[1:N]}$ (or $y_{[1:2]}$ in Algorithm 2) is larger. This is consistent with Theorem 4 where a larger interval length $\omega$ renders smaller $\gamma$.

We further consider synthetic datasets where each data point $(z_i, y_i)$ is i.i.d. generated by the model $y_i = \text{sign}[\frac{1}{1+e^{\langle \theta^*, z_i \rangle + e_i}} - \frac{1}{2}]$, where $z_i \in \mathbb{R}^2$ 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 $\mathcal{N}(0, 0.5^2)$; 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 $(\mu, \sigma)$ is $P(e_i = x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$. 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 Algorithm 2 each agent holds 200 samples.
B Proof of Theorem 3

Proof. W.l.o.g., we assume $\theta_1^{k-1} = 0$ and $\theta_2^{k-1} = \omega$. We use $x$ to denote $\theta_i^k$ in the following and the upper bound of $\epsilon$ can be reformulated as follows. For $x \in \mathbb{R}$, we consider

$$\max_{|t| \leq B} \left| \log \frac{\int_{t}^{\omega} e^{\beta x - c} dc}{\int_{0}^{t+\omega} e^{\beta x - c} dc} \right|$$

for some positive numbers $\omega$, $\beta$ and $B$, which correspond to the length of the interval, Laplace noise factor and sensitivity, respectively.

For a fixed $t$, $|t| \leq B$, if $x \not\in [0, \omega] \cup [t, \omega + t]$, then

$$\left| \log \frac{\int_{t}^{\omega} e^{\beta x - c} dc}{\int_{0}^{t+\omega} e^{\beta x - c} dc} \right| = \left| \log \frac{\int_{0}^{\omega} e^{\beta x - c} dc}{e^{\beta t} \int_{0}^{\omega} e^{\beta x - c} dc} \right| = |\beta t| \leq \beta B.$$  \hspace{1cm} (26)

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

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

it is equivalent to showing

$$\begin{cases}
2e^{\beta |t|} - e^{-\beta x + \beta |t|} - e^{-\beta (\omega - x) + \beta |t|} \geq 2 - e^{-\beta (x - t)} - e^{-\beta (\omega + t - x)}, \\
2 - e^{-\beta x} - e^{-\beta (\omega - x)} \leq 2e^{\beta |t|} - e^{-\beta (x - t) + \beta |t|} - e^{-\beta (\omega + t - x) + \beta |t|}.
\end{cases}$$  \hspace{1cm} (27)

\hspace{1cm} (28)
Due to the symmetry, we merely prove the case when $t \geq 0$, where (28) can be rewritten as,

$$\begin{align*}
2e^{\beta t} - e^{-\beta(x-t)} - e^{-\beta(\omega-x-t)} & \geq 2 - e^{-\beta(x-t)} - e^{-\beta(\omega+t-x)}, \\
2 - e^{-\beta x} - e^{-\beta(\omega-x)} & \leq 2e^{\beta t} - e^{-\beta(x-2t)} - e^{-\beta(\omega-x)}.
\end{align*}$$

(29)

Clearly, for the first inequality, it suffices to show

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

(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

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

(31)

which holds since $x > t$.

At last, we consider $x \not\in [t, t+\omega]$. Again, due to the symmetry, we can assume $t \geq 0$ and $x \leq t$. Then, it is equivalent to showing:

$$\begin{align*}
2e^{\beta t} - e^{-\beta(x-t)} - e^{-\beta(\omega-x)+\beta t} & \geq e^{-\beta(t-x)} - e^{-\beta(\omega+t-x)}, \\
2e^{-\beta x} - e^{-\beta(\omega-x)} & \leq e^{\beta x} - e^{-\beta(\omega-x)}.
\end{align*}$$

(32)

As for the first inequality, assume that $g(t) = 2e^{\beta t} - e^{-\beta(x-t)} - e^{-\beta(\omega-x)+\beta t} + e^{-\beta(\omega+t-x)}$. 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,

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

(33)

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

$$2e^{\beta t} - e^{-\beta(x-t)} + e^{-\beta(t-x)} - e^{-\beta(\omega-x)+\beta t} - e^{-\beta(\omega+t-x)} \geq 0.$$  (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, $e^{\beta t} \geq e^{-\beta(\omega-x)+\beta t} = e^{\beta x}$ 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 $\max_{|t| \leq B} |t/\beta| = \beta B$. 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 $\beta B$, which corresponds to the pure Laplace mechanism case where we fix all parameters to be constants.

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

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

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

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

which is upper bounded by $\frac{2}{1-e^{-\beta x}}$. On the other hand, if $x \not\in [B, \omega + B]$ but within $[0, \omega]$, the ratio becomes

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

which is upper bounded by $\frac{2e^{\beta(\omega-x)}}{1-e^{-\beta\omega}}$. 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 \(x = \theta^k_i\) is a Laplace distribution whose mean is uniformly distributed in \([0, \omega]\), conditional on all prior intermediate outputs. As a corollary of Theorem 3, we still assume \(\sup_{D} E_D \sup_{D'} c_D (\theta^k_i)\), it suffices to calculate

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\omega} \Theta(x) \frac{\beta}{2\omega} e^{-\beta|x-y|} dy dx,
\]

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

\[
\log \int_{-\infty}^{\infty} \int_{-\infty}^{\omega} \frac{\beta}{2\omega} e^{-\beta|x-y|} \max_{t=\pm B} \left\{ \int_0^\omega e^{-\beta|x-z|} dz, \int_0^{\omega+t} e^{-\beta|x-z|} dz \right\} dy dx.
\]

Now, we take a closer look into \(\max_{t=\pm B} \left\{ \int_0^\omega e^{-\beta|x-z|} dz, \int_0^{\omega+t} e^{-\beta|x-z|} dz \right\} = e^{\beta B}\).

Since we assume \(\omega > B\), it is not hard to observe that

\[
x \in [0, \frac{\omega-B}{2}], \max_{t=\pm B} \left( \int_0^\omega e^{-\beta|x-z|} dz, \int_0^{\omega+t} e^{-\beta|x-z|} dz \right) = \int_0^{\omega+t} e^{-\beta|x-z|} dz \quad \Rightarrow \quad t = B;
\]

\[
x \in [\frac{\omega-B}{2}, \frac{\omega}{2}], \max_{t=\pm B} \left( \int_0^\omega e^{-\beta|x-z|} dz, \int_0^{\omega+t} e^{-\beta|x-z|} dz \right) = \int_0^{\omega+t} e^{-\beta|x-z|} dz \quad \Rightarrow \quad t = B;
\]

\[
x \in [\frac{\omega}{2}, \frac{\omega+B}{2}], \max_{t=\pm B} \left( \int_0^\omega e^{-\beta|x-z|} dz, \int_0^{\omega+t} e^{-\beta|x-z|} dz \right) = \int_0^{\omega+t} e^{-\beta|x-z|} dz \quad \Rightarrow \quad t = B;
\]

\[
x \in [\frac{\omega+B}{2}, \omega], \max_{t=\pm B} \left( \int_0^\omega e^{-\beta|x-z|} dz, \int_0^{\omega+t} e^{-\beta|x-z|} dz \right) = \int_0^{\omega+t} e^{-\beta|x-z|} dz \quad \Rightarrow \quad t = B.
\]

Thus, fortunately, we can avoid the complicated integral at least in \(x \in [0, \frac{\omega-B}{2}], \) or \(x \in [\frac{\omega-B}{2} + B, \omega]\) where it is simplified to \(O(\int_0^{\omega+t} e^{-\beta|x-y|} dy)\). Now, we can split \(\mathbb{R}\) into three parts, \((-\infty, 0) \cup (\omega, \infty), [0, \frac{\omega-B}{2}] \cup [\frac{\omega-B}{2} + B, \omega]\) and \((\frac{\omega-B}{2}, \frac{\omega+B}{2})\). To avoid the tedious term when \(x \in (\frac{\omega-B}{2}, \frac{\omega+B}{2})\), here, we simply substitute the global upper bound to derive a closed-form expression but one may obtain the expression of \(\gamma\) exactly with the same reasoning. Note the symmetry on \(t = \pm B\), (36) is upper bounded by

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

**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

\[
\begin{cases}
\hat{\theta}^{k+1}_i := \arg \min_{\theta_i} f_i(\theta_i) - \lambda^{kT} \left( A_i \theta_i + \sum_{j \neq i} A_j \theta_j^k - c \right) \\
\quad + \frac{1}{2} \left\| A_i \theta_i + \sum_{j \neq i} A_j \theta_j^k - c \right\|_{\rho_i^{k+1}}^2 + \frac{1}{2} \left\| \theta_i - \theta_i^k \right\|^2_{\rho_i^{k+1}} + \Delta_i^{k+1},
\end{cases}
\]

(38)
and the Lagrangian multiplier is updated accordingly as \( \hat{\lambda}^{k+1} := \lambda^k - \gamma^{k+1}_i \rho^{k+1}_i (\sum_{j=1}^N A_j \hat{\theta}^{k+1}_j - c) \); and \( \lambda^{k+1} := \lambda^k - \gamma^{k+1}_i \rho^{k+1}_i (\sum_{j=1}^N A_j \hat{\theta}^{k+1}_j - c) \). \( \gamma^{k+1}_i \rho^{k+1}_i \) is a global constant set up at the beginning. Let \( u^{k+1} = [\theta^{k+1}_{[1:N]}, \lambda^{k+1}] \) and \( u^* = [\theta^*_{[1:N]}, \lambda^*] \), where \( \theta^*_{[1:N]} \) stand for the optimum to (17) and \( \lambda^* \) is an arbitrary point in \( \mathbb{R}^d \). In Algorithm 1, we further apply first-order approximation in (39) as

\[
\theta^{k+1}_i := H_i^{-1} [A_i^T (\lambda^k - \gamma^{k+1}_i \rho^{k+1}_i (\sum_{j \neq i} A_j \hat{\theta}^{k+1}_j - c)) + I_i^{k+1} \theta^k_i - \nabla f_i(\theta^k_i)] + \Delta^{k+1}_i.
\]

(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),

\[
2E[F(\theta^{k+1}) - F(\theta^*) - \lambda^T A(\theta^{k+1} - \theta^*)] \\
\leq -E[h^{k+1}] + \sum_{i=1}^N (M + H_i + \frac{1}{\zeta})E[\|\Delta^{k+1}_i\|^2] + E[\|u^{k+1} - u^*\|_G^2 -\|u^{k+1} - u^*\|_{\hat{G}}^2],
\]

(40)

where \( G = \text{diag}(H_1, ..., H_N, 1/\zeta) \) is a diagonal matrix. Here, \( H_i = H_i \cdot I = I_i^{k+1} + A_i^T \rho_i^{k+1} A_i \) is positive definite, \( h^{k+1} \) is some remainder term which is non-negative if for any \( i, j \in [1 : N] \)

\[
\left\{ \begin{array}{l}
\frac{H_i}{N} (\rho_i - \frac{1}{2}) \geq \sigma^{max, i}_i \rho^2 \\
\frac{1}{(N-1)^2} (\rho_i - \frac{1}{2}) (\rho_j - \frac{1}{2}) \geq (\rho_i - \rho_j)^2
\end{array} \right.
\]

(41)

where \( \rho_i \cdot I \leq \rho_i^{k+1} \leq \rho_i \cdot I \) for some constants \( 0 < \rho_i < \hat{\rho}_i \) and \( \sigma^{max, i} \) is the largest singular value of \( A_i \).

**Proof.** Since \( f_i \) is convex,

\[
\langle \nabla f_i(\hat{\theta}^{k+1}_i), \tilde{\theta}^{k+1}_i - \theta^*_i \rangle \geq f_i(\tilde{\theta}^{k+1}_i) - f_i(\theta^*_i).
\]

(42)

Due to the optimality condition satisfied, we have

\[
\nabla f_i(\tilde{\theta}^{k+1}_i) = A_i^T (\lambda^k - \rho^{k+1}_i (\sum_{j \neq i} A_j \hat{\theta}^{k+1}_j - c)) + I_i^{k+1} (\theta^k_i - \tilde{\theta}^{k+1}_i).
\]

(43)

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

\[
\langle A_i^T (\lambda^k - \rho^{k+1}_i (\sum_{j \neq i} A_j \hat{\theta}^{k+1}_j - c)) + I_i^{k+1} (\theta^k_i - \tilde{\theta}^{k+1}_i), \tilde{\theta}^{k+1}_i - \theta^*_i \rangle + \langle \tilde{\theta}^{k+1}_i - \theta^*_i, (\tilde{\theta}^{k+1}_i - \theta^*_i)^T I_i^{k+1} (\theta^k_i - \tilde{\theta}^{k+1}_i) \rangle \\
\geq f_i(\tilde{\theta}^{k+1}_i) - f_i(\theta^*_i) - \lambda^T A_i (\tilde{\theta}^{k+1}_i - \theta^*_i).
\]

(44)

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

\[
\langle \lambda^k - \lambda, A_i \tilde{\theta}^{k+1}_i - \theta^*_i \rangle + \langle \tilde{\theta}^{k+1}_i - \theta^*_i, (\tilde{\theta}^{k+1}_i - \theta^*_i)^T I_i^{k+1} (\theta^k_i - \tilde{\theta}^{k+1}_i) \rangle \\
\geq f_i(\tilde{\theta}^{k+1}_i) - f_i(\theta^*_i) - \lambda^T A_i (\tilde{\theta}^{k+1}_i - \theta^*_i).
\]

(45)

Summing up the above formulas for \( i = 1, 2, ..., N \), we have

\[
\langle \lambda^k - \lambda, \tilde{\theta}^{k+1}_i - \theta^*_i \rangle + \sum_{i=1}^N \langle \tilde{\theta}^{k+1}_i - \theta^*_i, (\tilde{\theta}^{k+1}_i - \theta^*_i)^T I_i^{k+1} (\theta^k_i - \tilde{\theta}^{k+1}_i) \rangle \\
\geq f_i(\tilde{\theta}^{k+1}_i) - f_i(\theta^*_i) - \lambda^T A_i (\tilde{\theta}^{k+1}_i - \theta^*_i).
\]

(46)
Let the matrix \( G = \text{diag}\{H_1, \ldots, H_N, 1/\zeta\} \), where \( H_k = H \cdot I \). With the identity: \( \|u^k - u^*\|^2_G + \|\tilde{u}^{k+1} - u^*\|^2_G = 2(\tilde{u}^{k+1} - u^*)^T G(u^k - \tilde{u}^{k+1}) + \|u^k - \tilde{u}^{k+1}\|^2_G \), we have

\[
\|u^k - u^*\|^2_G - \|\tilde{u}^{k+1} - u^*\|^2_G - \|u^k - \tilde{u}^{k+1}\|^2_G + \frac{2}{\zeta} \|\lambda^k - \tilde{\lambda}^{k+1}\|^2 - 2\langle A\theta^k - c, \sum_{i=1}^N \rho_i^{k+1} A_i(\tilde{\theta}_i^{k+1} - \theta_i^*) \rangle \geq 2[F(\theta^{k+1}) - F(\theta^*)] - \lambda^T A(\tilde{\theta}^{k+1} - \theta^*)].
\] (47)

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

\[
h^{k+1} = \sum_{i=1}^N H_i \|\theta_i^k - \tilde{\theta}_i^{k+1}\|^2 - \frac{1}{\zeta} \|\lambda^k - \tilde{\lambda}^{k+1}\|^2 + 2\sum_{i=1}^N A_i(\theta_i^k - \tilde{\theta}_i^{k+1} + \tilde{\theta}_i^*) \sum_{i=1}^N \rho_i^{k+1} A_i(\tilde{\theta}_i^{k+1} - \theta_i^*)
\]

\[
= \sum_{i=1}^N H_i \|\theta_i^k - \tilde{\theta}_i^{k+1}\|^2 - \zeta A(\tilde{\theta}^{k+1} - \theta^*)^2 + 2\sum_{i=1}^N A_i(\theta_i^k - \tilde{\theta}_i^{k+1}), \sum_{i=1}^N \rho_i^{k+1} A_i(\tilde{\theta}_i^{k+1} - \theta_i^*)
\]

\[
+ 2\rho_0 A(\tilde{\theta}^{k+1} - \theta^*)^2 + 2\sum_{i=1}^N A_i(\tilde{\theta}_i^{k+1} - \theta_i^*) \sum_{i=1}^N (\rho_i^{k+1} = \rho_0 I) A_i(\tilde{\theta}_i^{k+1} - \theta_i^*).
\] (48)

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

\[
\begin{aligned}
\frac{H_j}{N^2} (\rho_j - \frac{\zeta}{2}) &\geq \sigma_{\max,j}, \\
\frac{1}{(N-1)^2} (\rho_j - \frac{\zeta}{2}) (\rho_j - \frac{\zeta}{2}) &\geq (\tilde{\rho}_j - \rho_j)^2.
\end{aligned}
\] (49)

In the following, we generalize the above with respect to \( \theta_i^{k+1} = \tilde{\theta}_i^{k+1} + \Delta_i^{k+1} \). It is noted that

\[
f_i(\theta + \Delta_i^{k+1}) \leq f_i(\theta) + \langle \Delta_i^{k+1}, \nabla f_i(\theta) \rangle + \frac{M}{2} \|\Delta_i^{k+1}\|^2.
\]

In addition, \( \mathbb{E}[\|\tilde{u}^{k+1} - u^{k+1}\|^2_G] \) w.r.t. the randomness of \( \Delta_i^{k+1} \) is \( \|\tilde{u}^{k+1} - u^*\|^2_G + \mathbb{E}[\|u^{k+1} - u^*\|^2_G] \), 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 \( h^{k+1} \geq 0 \), let \( \theta^K = \frac{1}{K} \sum_{k=1}^K \theta_k \), then

\[
\mathbb{E}[F(\tilde{\theta}^K)] - F(\theta^*) \leq \frac{\|\theta_0 - \theta^*_G\| + 4\zeta \|\lambda^*\|^2}{2K} + \sum_{i=1}^N A_i(M + H_i + \frac{1}{\zeta}) \mathbb{E}[\|\Delta_i^{k+1}\|^2] + M \|\Delta_i^{k+1}\|^2,
\] (50)

where \( \lambda^* \) is the state of \( \lambda^k \) when \( \theta_k \) reaches the optimum. Under \((\epsilon, \delta)\)-LDP, the utility loss is \( O(\sqrt{N\frac{d\|G\|}{\epsilon}d\|G\|}) \).

Here, \( G_x = \text{diag}\{H_1, \ldots, H_N\} \).

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

\[
\mathbb{E}[K(\hat{F}(\tilde{\theta}^K) - F(\theta^*)) - \lambda^T A(\sum_{k=0}^K (\theta_{k+1} - \theta_k))] \leq \sum_{k=0}^{K-1} \mathbb{E}[F(\theta_{k+1}) - F(\theta_k)] + \lambda^T A(\theta_{k+1})
\]

\[
\leq \frac{1}{2} \sum_{k=0}^{K-1} \sum_{i=1}^N (M + H_i + \frac{1}{\zeta}) \mathbb{E}[\|\Delta_i^{k+1}\|^2] + \mathbb{E}[\|u^0 - u^*\|^2_G].
\] (51)

\( ^5 \) K-fold composition of \((\epsilon, \delta)\)-(L)DP mechanisms can produce a \((\delta, \tilde{K} + \delta)\)-DP, where \( \tilde{\epsilon} = K \epsilon + \epsilon \sqrt{-2K \log \delta} \).
Since $\lambda^0 = 0$ and $\lambda$ in $u^*$ is an arbitrary point in $\mathbb{R}^d$, by letting $\lambda = 0$, we have

$$
\mathbb{E}[F(\tilde{\theta}^K) - F(\theta^*)] \leq \frac{\|\theta_0 - \theta^*\|^2_{\mathcal{S}_G}}{2K} + \frac{\sum_{k=1}^K \sum_{i=1}^N (\sqrt{M} + H_i + \frac{1}{\xi}) \mathbb{E}[\|\Delta^k_i\|^2]}{K}.
$$

(52)

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

$$
F(\theta^*) \leq F(\tilde{\theta}^K) - \nabla F(\theta^*)^T (\tilde{\theta}^K - \theta^*).
$$

(53)

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

$$
-\lambda^*^T A \sum_{k=0}^K (\theta_{k+1} - \theta^*) \leq F(\tilde{\theta}^K) - F(\theta^*) - 2\lambda^*^T A \sum_{k=0}^K (\theta_{k+1} - \theta^*) \leq \mathbb{E}[\|\theta_0 - \theta^*\|^2_{\mathcal{S}_G}] + \frac{\sum_{k=1}^K \sum_{i=1}^N (M + H_i + \frac{1}{\xi}) \mathbb{E}[\|\Delta^k_i\|^2]}{K}.
$$

(54)

Following [BST14], $\mathbb{E}[\|\Delta^k_i\|^2] = O\left(\frac{d^2 K^2 \beta^2}{\epsilon^2}\right)$ under $(\epsilon, \delta)$-LDP guarantee. Putting the upper and lower bounds of $F(\tilde{\theta}^K) - F(\theta^*)$ together, let $K = O\left(\frac{1}{\epsilon Nd^2 \beta}\right)$, the claim follows.

Finally, we can bridge the above utility analysis and the case where we further apply first-order approximation in (48) as

$$
\theta_i^{k+1} := H_i^{-1} \left[ A_i^T (\lambda^k - \rho_i^{k+1} (\sum_{j \neq i} A_j \theta_j^k - c)) + F_i^{k+1} \theta_i^k - \nabla f_i(\theta_i^k) \right] + \Delta_i^{k+1}.
$$

(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 $i, j \in [1 : N]$

$$
\left\{ \begin{array}{l}
\frac{H_i - M}{N} (\rho_i - \frac{\xi}{2}) \geq \sigma_i^2_{\text{max,j}} \rho_i^2 \\
\frac{1}{(N-1)^2} (\rho_i - \frac{\xi}{2}) (\rho_j - \frac{\xi}{2}) \geq (\rho_i - \rho_j)^2.
\end{array} \right.
$$

(56)

**Proof.** With the smooth assumptions on $\nabla f_i$, i.e., for any $\theta$ and $y$,

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

we have the following fact: for any $z$

$$
|f_i(\theta) - f_i(y)| \leq \frac{M}{2} \|\theta - \Delta_i\|^2.
$$

Therefore, by replacing $\theta_i^{k+1}$ with $\theta_i^k$ in (43), all the deductions in Theorem[1] stay the same except that the term $\sum_{i=1}^N H_i \|\theta_i^{k+1} - \theta_i^k\|^2$ in the expression of $h^{k+1}$ in (48) becomes $\sum_{i=1}^N (H_i - M) \|\theta_i^{k+1} - \theta_i^k\|^2$. 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 $F(\theta) = \frac{1}{N} \sum_{i=1}^N f_i(\theta_i)$ and accordingly the updating rule of Algorithm[2] becomes,

$$
\theta_i := w_i y_i^k + (I_d - w_i) y_2^k - \eta_{k+1} \nabla f_i\left(\frac{y_i^k + y_2^k}{2}\right) + \Delta_i.
$$

(57)
For each $k \in [0 : K - 1]$,

\[
\|y_{k}^{i+1} - \theta^*\|^2 = \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} (w_i y_i^k + (I_d - w_i)(y_2^k - \theta^* - \zeta_{k+1} \nabla f_i(\frac{y_1^k + y_2^k}{2}))\|^2
\]

\[
= \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} (w_i (y_i^k - \theta^*) + (I_d - w_i)(y_2^k - \theta^*))\|^2
\]

\[
- 2\zeta_{k+1} \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} (w_i y_i^k + (I_d - w_i)y_2^k) - \theta^*, \sum_{i \in S_{2k+1}} \nabla f_i(\frac{y_1^k + y_2^k}{2} + \zeta_{k+1} \Delta_i)
\]

\[
+ \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} (\zeta_{k+1} \nabla f_i(\frac{y_1^k + y_2^k}{2} + \Delta_i))\|^2.
\]

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

\[
\sum_{i=1}^{N} (\omega_i r_i)^2 \leq \sum_{i=1}^{N} \omega_i r_i^2.
\]

It is noted that in (58), the sum of weights of $(y_1^k - \theta^*)$ and $(y_2^k - \theta^*)$ is always the identity. With (59), by taking expectation on both sides of (58), we have

\[
E[\|y_{k+1}^i - \theta^*\|^2] \leq E[\|y_1^k - \theta^*\|^2 + \|y_2^k - \theta^*\|^2 + (\eta_{k+1} G^2 + E\|\Delta^k_i\|^2)]
\]

\[
- 2\eta_{k+1} (\frac{y_1^k + y_2^k}{2}, \nabla F(\frac{y_1^k + y_2^k}{2})),
\]

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

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

for arbitrary $\theta$. On the other hand, $E[\frac{1}{|S_i|} \sum_{i \in S_i} \omega_i] = \frac{1}{2} I_d$, where the selection of $w_i$ is independent to the agent grouping scheme.

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

\[
E[\|y_{k+1}^i - \theta^*\|^2] \leq E[\|y_1^k - \theta^*\|^2 + \|y_2^k - \theta^*\|^2 + (\eta_{k+1} G^2 + E\|\Delta^k_i\|^2)] - 2\eta_{k+1} (\frac{y_1^k + y_2^k}{2}, \nabla F(\frac{y_1^k + y_2^k}{2}))
\]

where $\Delta_i^k = \frac{1}{|S_{2k+1}|} \sum_{i \in S_{2k+1}} \|\Delta_i\|$. Applying the fact that $(\frac{y_1^k + y_2^k}{2} - \theta^*, \nabla F(\frac{y_1^k + y_2^k}{2})) \geq F(\frac{y_1^k + y_2^k}{2}) - F(\theta^*)$ and taking the average on both sides of (60) and (61), we can bound $F(\frac{y_1^k + y_2^k}{2}) - F(\theta^*)$ as,

\[
F(\frac{y_1^k + y_2^k}{2}) - F(\theta^*) \leq \frac{\eta_{k+1} (\|y_1^k - \theta^*\|^2 + \|y_2^k - \theta^*\|^2 - \|y_1^{k+1} - \theta^*\|^2 - \|y_2^{k+1} - \theta^*\|^2)}{4}
\]

\[
+ \frac{\eta_{k+1} G^2 + E\|\Delta^k_i\|^2 + E\|\Delta^k_i\|^2}{4\eta_{k+1}}.
\]

Before we can derive a global convergence analysis, we need to give an upper bound on $\|y_{k+1}^i - \theta^*\|$ and $\|y_{k+1}^i - \theta^*\|$ with the initial divergence $\|y_0^i - \theta^*\|, \|y_1^i - \theta^*\|$ and the noise $E\|\Delta^k_i\|^2$ and $E\|\Delta^k_i\|^2$. It is noted that, with rearrangement on (62) and the fact $F(\theta) - F(\theta^*) \geq 0$,

\[
E[\|y_{1}^{k+1} - \theta^*\|^2] \leq E[\|y_1^k - \theta^*\|^2 + \|y_2^k - \theta^*\|^2 + \eta_{k+1} G^2 + E\|\Delta^k_i\|^2].
\]
When we select \( \eta_k = \frac{1}{\sqrt{c^2 K}} \), \( \sum_{k=1}^{K} \eta_k^2 = \sum_{k=1}^{N} \frac{1}{c^2 K} \leq \frac{\log K + 1}{c^2} \) since \( k \leq K \). The above renders an upper bound on 
\[
E\left[ \frac{\|y_{k+1}^i - \theta^*\|^2 + \|y_{k+1}^j - \theta^*\|^2}{2} \right] 
\]
that
\[
E\left[ \frac{\|y_1^i - \theta^*\|^2 + \|y_1^j - \theta^*\|^2}{2} \right] \leq E\left[ \frac{\|y_1^i - \theta^*\|^2 + \|y_1^j - \theta^*\|^2}{2} + \frac{\log K + 1}{c^2} (G^2 + V^2) \right],
\]
with the assumption \( \max_{k,j} E[(\Delta_j^k/\eta_k)^2] \leq V^2 \). Thus, (62) can be further formulated as
\[
\sum_{k=0}^{K-1} \frac{E[F(\frac{\sum_{i=0}^{K} y_k^i}{2}) - F(\theta^*)]}{K} 
\]
\[
\leq \sum_{k=1}^{K} (\eta_{k+1}^j - \eta_{k}^j) (\|y_1^i - \theta^*\|^2 + \|y_1^j - \theta^*\|^2) + \sum_{k=0}^{K-1} \eta_{k+1}^j (\|y_1^i - \theta^*\|^2 + \|y_1^j - \theta^*\|^2) + 2 \sum_{k=0}^{K-1} \eta_{k+1}^j (G^2 + V^2) \]
\[
= O \left( \sqrt{\frac{K}{c^2}} (\|y_1^i - \theta^*\|^2 + \|y_1^j - \theta^*\|^2) + c^{-1} (\log K + 2) \sqrt{K + 1} (G^2 + V^2) \right),
\]
and \( \sum_{i=1}^{N} f_i(\sum_{k=0}^{K} \theta_i^k / NK) - f_i(\theta^*) \leq \sum_{k=0}^{K} \sum_{i=1}^{N} \frac{E[f_i(\theta^k)] - f_i(\theta^*)}{K} \). 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 \( \epsilon \)-LDP setting, since the sensitivity is bounded by \( B \) in \( L_{\infty} \), on each dimension we may add a noise following \( \text{Lap}(0, \frac{d \epsilon}{dN B}) \) to produce a total \( \epsilon \) loss from \( d \) dimensions. Under the relaxed \( (\epsilon, \delta) \)-DP setting, with the strong composition theorem [KOV17], the variance \( E[(\Delta_k^i/\eta_k)^2] \) can be reduced to \( \tilde{O} \left( \frac{K}{N} d (\frac{2R}{\epsilon})^2 \right) \). Substituting those into (64), we complete the proof of the Theorem that
\[
\begin{align*}
\text{pure } \epsilon - \text{LDP : } & \tilde{O} \left( \frac{\sqrt{\|y_1^i - \theta^*\|^2 + \|y_1^j - \theta^*\|^2} (G + \sqrt{\frac{K}{c^2}} dN \mathcal{B})}{\sqrt{K}} \right) = \tilde{O} \left( \frac{\sqrt{\mathcal{B}}}{\sqrt{N} \epsilon} \right) \\
\text{relaxed } (\epsilon, \delta) - \text{LDP : } & \tilde{O} \left( \frac{d \mathcal{B}}{\sqrt{N} \epsilon} \right).
\end{align*}
\]

F Non-asymptotic Convergence Rate Analysis

Theorems 4 and 5 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
\[
\theta_{k+1} = W_{k+1} \theta_k - \xi_{k+1} \nabla F(\mathbb{E}[W_{k+1}] \theta_k) + \Delta_{k+1}
\]
(66)
where, recalling \( w_{ij} \) in (16), \( W_{k+1}(i, j) = w_{ij} \cdot I \) is a random stochastic matrix determined by the weights selected by each agent. Here, \( I \) 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 \( \xi_k = O \left( \frac{1}{\sqrt{\epsilon}} \right) \), when \( f_{[1:N]}(\cdot) \) are L-Lipschitz, and \( \mathbb{E}[W_k] \) is doubly stochastic,
\[
|F\left( \frac{\sum_{k=0}^{K-1} \theta_k}{K} \right) - F(\theta^*)| \leq \frac{\mathcal{R}}{\sqrt{K}} + L \sum_{i=1}^{N} \mathbb{E}[\mathcal{T}_i^K],
\]
(67)
where \( W_0 = I, \mathcal{R} \) is a term invariant to the randomness of \( W_k \), specified in the proof, and
\[
\mathcal{T}_i^K = \left\| \sum_{k=0}^{K-1} \sum_{i=1}^{N} \mathbb{E}[W_{k+1}(i, :)^T \theta_k] - \sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}(i, :)^T \theta_k] \right\| + \left\| \sum_{k=0}^{K-1} (\theta_k - \mathbb{E}[W_{k+1}(i, :)^T \theta_k]) \right\|.
\]
Proof. Applying the convexity of $F(\cdot)$ that $(\langle \nabla F(\mathbb{E}[W_{k+1}]\theta_k), \mathbb{E}[W_{k+1}]\theta_k - \theta_* \rangle \geq F(\mathbb{E}[W_{k+1}]\theta_k) - F(\theta_*)$, we can derive the following:

$$
\mathbb{E}[F(\mathbb{E}[W_{k+1}]\theta_k)] - F(\theta_*)
\leq \frac{1}{2}(\xi_{k+1}^{-1})^2 \left( \mathbb{E}[\|W_{k+1}\theta_k - \theta_*\|^2 + \|\xi_{k+1}^T \nabla F(\mathbb{E}[W_{k+1}]\theta_k) - \Delta_{k+1}\|^2 - \|\theta_{k+1} - \theta_*\|] \right)
\leq \frac{1}{2}(\xi_{k+1}^{-1})^2 \left( \mathbb{E}[\|\theta_k - \theta_*\|^2 - \|\theta_{k+1} - \theta_*\|^2 + \|\xi_{k+1}^T \nabla F(\mathbb{E}[W_{k+1}]\theta_k)\|^2] + \mathbb{E}[\|\Delta_{k+1}\|^2])
$$

Following the proof of Theorem 7, we may sum up and average out both sides of (68) for $k = 0, 1, ..., K - 1$, which turns to be,

$$
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k] - F(\theta_*) \geq F\left(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k}{K}\right) - F(\theta_*).
$$

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

$$
\sum_{i=1}^{N} f_i\left(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k^i}{KN}\right) - F(\theta_*)
\leq F\left(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k}{K}\right) - F(\theta_*) + L \sum_{i=1}^{N} \left\| \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k^i}{KN} - \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k^i}{KN} \right\|,
$$

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

$$
\left| F\left(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k}{K}\right) - F(\theta_*) \right| \leq \frac{\mathcal{R}}{\sqrt{K}} + L \sum_{i=1}^{N} \left\| \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k^i}{KN} - \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k^i}{KN} \right\|,
$$

where $\frac{\mathcal{R}}{\sqrt{K}}$ 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,

$$
\left| F\left(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k}{K}\right) - F\left(\frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k}{K}\right) \right| \leq L \sum_{i=1}^{N} \left\| \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k^i}{KN} - \frac{\sum_{k=0}^{K-1} \mathbb{E}[W_{k+1}]\theta_k^i}{KN} \right\|.
$$

Putting things together, the claim follows.

Theorem 7 indicates that, to study the utility loss, it suffices to consider the rate of $\theta_k$ towards the consensus, i.e., $\theta_k^i = \theta_k^j$ for $i, j \in [1 : N]$, where the deviation amongst $\theta_k$ controls $T^k_i$ 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 $\theta_k$ approaches the unique consensus optima $\theta_*$, the excess loss is then proportional to the divergence among $\theta_k$ in expectation. For quantification, we introduce the following metric $\phi(\theta)$, which denotes the largest deviation between any two elements of $\theta$ in $l_2$ norm. For example, $\phi(\theta) = \max_{i,j} \|\theta^i - \theta^j\|$. With the above understanding, we move our focus to $\phi(\sum_{k=0}^{K-1} \theta_k^i/K)$. To proceed, we rewrite $\sum_{k=0}^{K-1} \theta_k^i/K$ in the following form,

$$
\sum_{k=0}^{K-1} \theta_k^i/K = \frac{1}{K} \left( \sum_{k=0}^{K-1} \prod_{j=1}^{k} W_{j} \theta_0 + \sum_{j=1}^{K} \prod_{i=j+1}^{k} W_{i} R_{j} \right)/K.
$$

where for simplicity we rewrite $\theta_{k+1} = W_{k+1}\theta_k + R_{k+1}$ for some remainder term $R_{k+1}$ and $\prod_{j=k}^{k} W_{j} = I$ 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 $\sum_{k=0}^{K-1} \theta_k / K$, compared to the fixed $W_k = W$ case. From (72), we consider the extreme scenario where the communication graph is fully connected and $W(i, j) = \frac{1}{N} \cdot I$. The justification of this choice is as follows. Such fixed weight matrix $W$ with identical rows has the property that for any $\theta$, elements in $W \theta$ are identical, i.e., $\phi(W \theta) = 0$. Let $W_k = W$ in (72), then all the terms that are a multiple of $W$ reach consensus and thus $\phi(\sum_{k=0}^{K-1} \theta_k / K) = \phi((\theta_0 + \sum_{k=1}^{K-1} R_k) / K)$. In contrast, for randomized $W_k$, those terms of multiples of $W_k$, such as $\prod_{k} W_k \theta$ 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 $W_k$: $W_k(i, j) = r_i^{k+1} \times \frac{\Delta_i}{N} \cdot I$ if $j \leq \frac{N}{2}$, otherwise $W(i, j) = (1 - r_i^{k+1}) \times \frac{\Delta_i}{N} \cdot I$. $r_i^k$ are i.i.d. random variables in $(0, 1)$. Clearly, $E[W_{k+1}] = W$.

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

$$E[\phi(\{(W_1 W_2)(i, j)\})] \leq \frac{1}{2} E[\phi(W_2(1, j))].$$

**Proof.** For simplicity, we omit the $I$ in the elements of $W_k$. It is noted that $W(i_1, j) = \sum_{l=1}^{N} W(i_1, l) W(l, j)$ and $W(i_2, j) = \sum_{l=1}^{N} W(i_2, l) W(l, j)$. Without loss of generality, we assume $W_2(1, j) = \min_i \{ W_2(l, j) \}$. Thus,

$$\left| \tilde{W}(i_1, j) - \tilde{W}(i_2, j) \right| = \left| \sum_{l=2}^{N} (W_1(i_1, l) - W_1(i_2, l)) W_2(l, j) \right|$$

$$= \left| (1 - \sum_{l=1}^{N} W_1(i_1, l) - 1 + \sum_{l=1}^{N} W_1(i_2, l)) W_2(1, j) + \sum_{l=2}^{N} (W_1(i_1, l) - W_1(i_2, l)) W_2(l, j) \right|$$

$$= \sum_{l=2}^{N} (W_1(i_1, l) - W_1(i_2, l))(W_2(l, j) - W_2(1, j))].$$

It is noted that $W_2(l, j) - W_2(1, j) \geq 0$ for $l \geq 2$ which is no bigger than $\phi(W_2(1, j))$. 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),

$$\mathbb{E}[\left| \tilde{W}(i_1, j) - \tilde{W}(i_2, j) \right|] \leq \frac{N}{2} \mathbb{E}[\max_i W_1(i_1, l) - \min_i W_1(i_2, l)] \mathbb{E}[\phi(W_2(1, j))]

= (\frac{3}{4} - \frac{1}{4}) \mathbb{E}[\phi(W_2(1, j))] = \frac{1}{2} \mathbb{E}[\phi(W_2(1, j))].$$

**Lemma 4** indicates that even with randomness in $W_k$ where $\phi(W_k \theta) = 0$ 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 $\theta$, $\phi(\prod_k W_k \theta) = O(\max_j \phi((\prod_k W_k)(j, :)))$. Then, the gap can be bounded theoretically as follows.

**Theorem 8** Under $(\epsilon, \delta)$-LDP and the setup claimed above, for $W_k$ fixed to $W$, $\phi(\sum_{k=0}^{K-1} \theta_k / K) = O(\frac{1}{\sqrt{K}} + \frac{1}{K} + \frac{dB_{\infty}}{\epsilon})$; as for randomized $W_k$ generated as Lemma 4 $\phi(\sum_{k=0}^{K-1} \theta_k / K) = O(\frac{1}{\sqrt{K}} + \frac{1}{K} + (1 + \frac{1}{N}) \frac{dB_{\infty}}{\epsilon})$.

**Proof.** First, we rewrite the expression of $R_k$. Recalling (66), $R_k = (-\xi_k \nabla f_1(W \theta_{k-1}) + \Delta_k^1, -\xi_k \nabla f_2(W \theta_{k-1}) + \Delta_k^2, \ldots, -\xi_k \nabla f_N(W \theta_{k-1}) + \Delta_k^N)$. It is noted that, though $\nabla F(\theta^*) = \sum_{i=1}^{N} \nabla f_i(\theta^*) = 0$, $\nabla f_i(\theta^*)$ does not necessarily equal to 0. On the other hand, when we select $\xi_k = O(1/\sqrt{K})$, the sensitivity of $\xi_k \nabla f_i(\theta^*)$ is $O(1/\sqrt{K})$, where accordingly the noise $|\Delta_k^i|$ is also scaled in $O(\frac{1}{\sqrt{K}} \cdot \frac{dR_{\infty}}{\epsilon})$ for $(\epsilon, \delta)$ privacy guarantee when we ignore other terms. Thus, for fixed $W_k = W$,

$$\phi(\sum_{k=0}^{K-1} \theta_k / K) = \phi(\theta_0 + \sum_{k=1}^{K-1} R_k) = O(\frac{1}{\sqrt{K}} + \frac{1}{K} + \frac{dB_{\infty}}{\epsilon}).$$

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

28
In comparison, for randomized $W_k$, applying Lemma 4 on $\mathbb{E}[\phi(W_{k+1} \nabla F(\mathbb{E}[W_k]|\theta))]$, we have
\[
\mathbb{E}[\phi(W_{k+1} \nabla F(\mathbb{E}[W_k]|\theta))] = \mathbb{E}[\phi(W_k \nabla F(W \theta))] \leq \frac{1}{2} \phi(\nabla F(W \theta)).
\]

Moreover, since $\Delta^k_i$ are i.i.d. noise, due to the construction of $W_k$ in the setup, $\mathbb{E}[\phi(W_{k+1} \Delta^k)] = O(\frac{1}{\sqrt{N}} \mathbb{E}[||\Delta^k||])$, where $\Delta^k = (\Delta^k_1, \Delta^k_2, ..., \Delta^k_N)$. 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, $\phi(\frac{\sum_{k=0}^{K-1} \theta_k}{K}) = O(\frac{1}{\sqrt{K}} + \frac{2}{K} + (1 + \frac{1}{\sqrt{N}}) \cdot \frac{dB_{\infty}}{\epsilon})$, which asymptotically enjoys $O(\frac{dB_{\infty}}{\epsilon})$ as $N$ and $K$ increase.

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