The Privacy Power of Correlated Noise in Decentralized Learning

Youssef Allouah [1] **Anastasia Koloskova** [1] **Aymane El Firdoussi** [1] **Martin Jaggi** [1] **Rachid Guerraoui** [1]

Abstract

Decentralized learning is appealing as it enables the scalable usage of large amounts of distributed data and resources without resorting to any central entity, while promoting privacy since every user minimizes the direct exposure of their data. Yet, without additional precautions, curious users can still leverage models obtained from their peers to violate privacy. In this paper, we propose DECOR, a variant of decentralized SGD with differential privacy (DP) guarantees. In DECOR, users securely exchange randomness seeds in one communication round to generate pairwise-canceling correlated Gaussian noises, which are injected to protect local models at every communication round. We theoretically and empirically show that, for arbitrary connected graphs, DECOR matches the central DP optimal privacy-utility trade-off. We do so under SecLDP, our new relaxation of local DP, which protects all user communications against an external eavesdropper and curious users, assuming that every pair of connected users shares a secret, i.e., an information hidden to all others. The main theoretical challenge is to control the accumulation of non-canceling correlated noise due to network sparsity. We also propose a companion SecLDP privacy accountant for public use.

1. Introduction

In numerous machine learning scenarios, the training dataset is dispersed among diverse sources, including individual users or distinct organizations responsible for generating each data segment. The nature of such data often involves privacy concerns, especially in applications like healthcare (Sheller et al., 2020), which can divulge sensitive information about an individual's health. Privacy issues make it either impractical or undesirable to transfer the data beyond their original sources, promoting the emergence of *federated* and *decentralized* learning (McMahan et al., 2017; Lian et al., 2017), where the training occurs directly on the data-holding entities. Decentralized learning additionally removes the assumption of a central server, with only the model updates being transmitted directly between users. A classical decentralized learning algorithm is decentralized stochastic gradient descent (D-SGD) (Koloskova et al., 2020), where users alternate between performing local gradient updates and averaging local models via gossiping.

When dealing with privacy-sensitive data, it is crucial not only to confine the sensitive information locally with decentralization, but also to ensure that the algorithm avoids leaking any sensitive information through its communicated updates or the final model. These can be observed by an *external eavesdropper* or even an *honest-but-curious user*, who follows the algorithm but may attempt to violate the privacy of other users. The notion of *differential privacy* (DP) (Dwork et al., 2014) serves as a widely accepted theoretical framework for measuring formal privacy guarantees. This notion has been extensively studied in centralized settings (Bassily et al., 2014; Abadi et al., 2016), i.e., assuming a trusted data curator or server. Yet, much less attention has been given to adapting DP to decentralized learning.

Several threat models have been considered in decentralized learning, the strongest corresponding to the classical notion of *local differential privacy* (LDP) (Kasiviswanathan et al., 2011). Under LDP, users do not trust any other entity and obfuscate all their communications independently. In contrast, *central differential privacy* (CDP) only protects the final model, exactly as if the learning was conducted on a single machine. Importantly, there is a significant gap in performance between LDP and CDP algorithms. In a system of n users, the optimal privacy-utility trade-off under LDP can be n times worse than CDP (Duchi et al., 2018). Indeed, the CDP baseline is the variant of D-SGD adding noise to protect the average of user models only, which is much less noise than that needed under LDP, to protect every local model before averaging. Some prior works aimed at reconciling this performance gap by investigating other relaxations of LDP. For example, in federated learning with an untrusted server, the shuffle model (Cheu et al., 2019) and distributed DP (Kairouz et al., 2021a) restrict the view of the

[1]EPFL, Switzerland. Correspondence to: Youssef Allouah <youssef.allouah@epfl.ch>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

server using cryptographic primitives and match the CDP optimal privacy-utility trade-off. However, these approaches are server-based and thus cannot be used in decentralized learning. Network DP (Cyffers & Bellet, 2022) considers honest-but-curious users whose view is restricted to their neighboring communications. As we discuss in Section 1.1, the privacy-utility trade-offs under Network DP match CDP only for well-connected graphs (Cyffers et al., 2022).

Contributions. We propose DECOR, a new algorithm for decentralized learning with differential privacy. DECOR is a variant of D-SGD, which additionally injects two types of privacy noise to protect local models: (i) uncorrelated Gaussian noise to protect the local model *after* gossip averaging, and (ii) correlated Gaussian noise, as a sum of pairwise cancelling noise terms for each neighbor, to protect local models *before* gossip averaging. In the presence of a server, after one round of DECOR, averaging all local models would cancel out the correlated Gaussian noise terms, and leave the uncorrelated Gaussian noise protecting the average of models, as was previously studied by Sabater et al. (2022) (see Section 1.1). However, on a sparse graph, the correlated noise terms do not all cancel out in DECOR. To obtain our main result, we control the accumulation of correlated noise across iterations in our convergence analysis, and show that its effect vanishes across iterations of DECOR.

We consider an external eavesdropper and honest-but-curious non-colluding users and show that DECOR matches the optimal CDP privacy-utility trade-off under our new relaxation of LDP we call *secret-based local differential privacy* (SecLDP). Our relaxation protects against an external eavesdropper and curious users who can observe all communications, assuming that every pair of connected users shares a secret, i.e., an information a priori hidden to all others, similar to secure aggregation (Bonawitz et al., 2017). For example, we consider the secrets to be shared randomness seeds exchangeable in one round of encrypted communications. Following the choice of the set of secrets, our relaxation can capture several threat models, e.g., including collusion of several users; or recovering LDP when no communications are secret.

We also demonstrate the empirical superiority of DECOR over the LDP baseline on simulated and real-world data and multiple network topologies, and provide a practical SecLDP privacy accountant for DECOR.[1]

1.1. Related Work

Most works on DP optimization have focused on the centralized setting (Chaudhuri et al., 2011; Bassily et al., 2014; Abadi et al., 2016), where a trusted curator collects user data. Also, several recent works tackle privacy in federated learning, where an honest-but-curious server coordinates the users. These works either use cryptographic primitives to only reveal the sum of updates (Jayaraman et al., 2018; Kairouz et al., 2021a; Agarwal et al., 2021) or to anonymize user identities through shuffling (Erlingsson et al., 2019; Cheu et al., 2019). Although these techniques provably achieve the centralized optimal privacy-utility trade-off, they are incompatible with fully decentralized settings, where only peer-to-peer communications are allowed, or induce large computational and communication costs.

Private decentralized learning. In decentralized settings, several distributed optimization algorithms (Bellet et al., 2018; Cheng et al., 2019; Huang et al., 2019; Li & Chi, 2023) have been adapted, by adding noise to gradient updates, to ensure LDP. However, these approaches yield a poor privacy-utility trade-off, which is a fundamental drawback of LDP (Duchi et al., 2013). Cyffers & Bellet (2022) consider a weaker privacy model than LDP where the threat comes from curious users solely, who can observe information exchanged with their communication graph neighbors only. Under this weaker privacy threat, it is possible to match the centralized privacy-utility trade-off for well-connected graphs only (Cyffers et al., 2022). In general, SecLDP and Network DP are orthogonal, since the latter restricts the view of users to local communications only, while the former hides part of the global communications—secrets—to an adversary observing all other communications. Yet, when considering honest-but-curious users, SecLDP is arguably stronger than Network DP as users in SecLDP have a larger view, i.e., all communications besides secrets outside their neighborhood. Also, the privacy-utility trade-off achieved under SecLDP is matches CDP for arbitrary connected topologies, unlike Network DP.

Correlated noise. Our correlated noise technique has been studied in various forms within secure multi-party computation, where the goal is to privately compute a function without a trusted central entity. A first form, called secret sharing (Shamir, 1979), consists in adding uniformly random noise terms which cancel out only if enough users collude. The same idea has also been analyzed for decentralized averaging (Li et al., 2019). However, these works guarantee the perfect security of the inputs, not the privacy of the average. Indeed, a curious adversary observing the average can infer the presence of an input or reconstruct it (Melis et al., 2019). In this direction, Imtiaz et al. (2019) proposed adding correlated Gaussian noise to the inputs, along with a smaller uncorrelated Gaussian noise to protect the average only. The correlated Gaussian noise is generated by having users sample Gaussian noise locally, and using secure aggregation (Bonawitz et al., 2017) to get the average of the noise terms, which is subtracted by users. Thus, averaging privatized inputs cancels out correlated

[1] Our code is available at https://github.com/elfirdoussilab1/DECOR.

noises and only leaves the smaller uncorrelated noise to protect the average. However, the algorithm requires a central entity for secure aggregation, which is not possible in decentralized learning and can be costly in communication. Sabater et al. (2022) further adapted the correlated Gaussian noise technique to decentralized settings, without using secure aggregation, by having connected users exchange pairwise cancelling Gaussian noise. However, their work only studies decentralized averaging, and does not cover the more challenging decentralized learning scenario, where the non-cancelled correlated noise accumulates across training iterations. Finally, we remark that correlated noise has also been studied in centralized settings with a different meaning, e.g., correlation is across iterations (Kairouz et al., 2021b), which is orthogonal to our work where noise is correlated across the users, but is uncorrelated across the iterations.

2. Problem Statement

We consider a set of users $[n] \coloneqq \{1, \dots, n\}$ who want to collaboratively solve a common machine learning task in a decentralized fashion. Each user $i \in [n]$ holds a local dataset $\mathcal{D}_i$ containing $m \in \mathbb{N}$ elements $\{\xi_i^1, \dots, \xi_i^m\}$ from data space $\mathcal{X}$.[2] The goal is to minimize the following global loss function:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{x}), \tag{1}$$

where the local loss functions $\mathcal{L}_i \colon \mathbb{R}^d \to \mathbb{R}, i \in [n]$, are distributed among n users and are given in empirical form:

$$\mathcal{L}_i(\mathbf{x}) \coloneqq \frac{1}{|\mathcal{D}_i|} \sum_{\xi \in \mathcal{D}_i} \ell(\mathbf{x}, \xi), \quad \forall \mathbf{x} \in \mathbb{R}^d, \tag{2}$$

where $\ell(\mathbf{x}, \xi) \in \mathbb{R}$ is the loss of parameter $\mathbf{x}$ on sample ξ. We study the fully decentralized setting where users are the nodes of an undirected communication graph $\mathcal{G} = ([n], \mathcal{E})$. Two nodes $i, j \in [n]$ can communicate directly if they are neighbors in $\mathcal{G}$, i.e., $\{i, j\} \in \mathcal{E}$.

Secret-based local DP. We aim to protect the privacy of user data against an adversary who can *eavesdrop* on all communications, while every pair of connected users $\{i, j\} \in \mathcal{E}$ shares a sequence of *secrets* $\mathbf{S}_{ij}$, which represent observations of random variables commonly known to the nodes sharing the secrets only. In practice, these are locally generated via *shared randomness seeds* exchanged after one round of encrypted communications (Bonawitz et al., 2017), and conceptually one can consider the secrets to be the shared randomness seeds only. We denote by $\mathcal{S}_{\text{all}} \coloneqq \{\mathbf{S}_{ij} \colon \{i, j\} \in \mathcal{E}\}$ the set of all secrets. While local DP (LDP) (Kasiviswanathan et al., 2011) protects the privacy of all communications without assuming the existence of secrets, at the price of a poor privacy-utility trade-off (Duchi et al., 2013), we propose to relax LDP into *secret-based local differential privacy* (SecLDP) as defined below.

Definition 1 (SecLDP)**.** *Let $\varepsilon \geq 0$, $\delta \in [0, 1]$. Consider a randomized decentralized algorithm $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathcal{Y}$, which outputs the transcript of all communications. Algorithm $\mathcal{A}$ satisfies $(\varepsilon, \delta, \mathcal{S})$-SecLDP if it satisfies (ε, δ)-DP given that the set of secrets $\mathcal{S}$ is unknown to the adversary. That is, for every adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^{m \times n}$,*

$$\mathbb{P}\left[\mathcal{A}(\mathcal{D}) \mid \mathcal{S} \text{ is hidden}\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[\mathcal{A}(\mathcal{D}') \mid \mathcal{S} \text{ is hidden}\right] + \delta,$$

where the event "$\mathcal{S}$ is hidden" conditions on the non-secret observations $\mathcal{S}_{\text{all}} \setminus \mathcal{S}$. We say that $\mathcal{A}$ satisfies (ε, δ)-SecLDP if it satisfies $(\varepsilon, \delta, \mathcal{S})$-SecLDP and $\mathcal{S}$ is clear from the context.

Our privacy definition can encode several levels of knowledge of the adversary, and the corresponding threat models, through the choice of the secrets $\mathcal{S}$. Essentially, the larger the set of secrets, the weaker is the adversary. To see this, we denote by $\mathcal{S}_i \coloneqq \{\mathbf{S}_{jk} \colon \{j, k\} \in \mathcal{E} \text{ and } j, k \neq i\}$ the set of secrets hidden from user $i \in [n]$, and by $\mathcal{S}_{\mathcal{I}} \coloneqq \cap_{i \in \mathcal{I}} \mathcal{S}_i$ the set of secrets hidden from the group of users $\mathcal{I} \subseteq [n]$, so that $\mathcal{S}_{\mathcal{I}} \subseteq \mathcal{S}_i \subseteq \mathcal{S}_{\text{all}}$ for every $i \in \mathcal{I} \subseteq [n]$. We consider the following adversaries in increasing strength:

I. External eavesdropper: the only adversary is not a user and ignores all the secrets $\mathcal{S}_{\text{all}}$, but can eavesdrop on all communications between users. This threat is covered by $(\varepsilon, \delta, \mathcal{S}_{\text{all}})$-SecLDP.

II. Honest-but-curious users without collusion: every user faithfully follows the protocol, but may try to infer private information from other users by eavesdropping on all communications, while knowing the secrets it shares with other users only. This threat is covered by having $(\varepsilon, \delta, \mathcal{S}_i)$-SecLDP for every $i \in [n]$.

III. Honest-but-curious users with partial collusion: every group of users of size $q < n$ may collude by disclosing the secrets they have access to. This threat is covered, at collusion level q, by having $(\varepsilon, \delta, \mathcal{S}_{\mathcal{I}})$-SecLDP for every $\mathcal{I} \subseteq [n], |\mathcal{I}| = q$.

IV. Full collusion: all users may collude against any other user in the system, as if the adversary can observe all communications and no secrets are hidden from them. This threat is covered by $(\varepsilon, \delta, \varnothing)$-SecLDP, which corresponds to LDP.

The adversaries above are in increasing strength in the sense that defending against adversary II consequently defends

[2] All datasets have the same size for simplicity; our theory can be directly extended to cover local datasets with different sizes.

against adversary I, and the same logic holds for the other adversaries. In this work, we consider the secrets to be shared randomness seeds, which allow every pair of users to keep the same observation of a random variable and generate correlated noise. In practice, such secrets, i.e., randomness seeds, can be shared securely and efficiently, as is common in secure aggregation for federated learning (Bonawitz et al., 2017; Kairouz et al., 2021a). Moreover, for ease of exposition, we focus on the adversaries of type I and II above and defer the extension of our results to types III and IV to the appendix.

Comparison with other relaxations. Recall from Section 1 that a common relaxation of LDP is central differential privacy (CDP), where the adversary can only access the final training model. In fact, CDP is recovered from SecLDP by considering the larger set of secrets consisting of all user communications. From a privacy point of view, CDP is equivalent to DP in the trusted curator model—the privacy model in the centralized setting—and thus allows achieving the best privacy-utility trade-off. In contrast, the best achievable mean squared error under LDP is n times worse than under CDP (Duchi et al., 2018; Allouah et al., 2023), for strongly convex optimization problems. Indeed, the CDP baseline is D-SGD with additional Gaussian noise magnitude $\Theta(\frac{1}{n\varepsilon^2})$, while the LDP baseline D-SGD with Gaussian noise magnitude $\Theta(\frac{1}{\varepsilon^2})$. We refer to these approaches as the CDP and LDP baselines, respectively.

However, CDP does not protect against honest-but-curious users, who can be expected in real-world scenarios. This limitation motivated Network DP (Cyffers & Bellet, 2022), which guarantees the privacy of all communications against honest-but-curious users whose view is restricted to communications with their neighbors, with privacy-utility trade-offs sometimes matching those of CDP (Cyffers et al., 2022). In general, SecLDP and Network DP are orthogonal, since the latter restricts the communications known to users, while the former restricts part of these communications—secrets—to an adversary observing all other communications. In the case of the aforementioned adversary II, SecLDP is arguably stronger than Network DP because honest-but-curious users in SecLDP have a larger view, i.e., all communications besides secrets outside their neighborhood.

3. DECOR: Decentralized SGD with Correlated Noise

We now present our algorithm DECOR, summarized in Algorithm 1. Overall, DECOR is a variant of D-SGD injecting the privacy noise each local model. This privacy noise consists of two parts: (i) correlated noise to protect the local communications before gossip averaging, and (ii) uncorrelated noise to protect the gossip average.

Algorithm 1 DECOR: DECENTRALIZED SGD WITH CORRELATED NOISE

Input: for each user $i \in [n]$ initialize $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations T, clipping threshold C, noise parameters σ_{cor} and σ_{cdp}.

1: **for** t **in** $0 \dots T-1$, i **in** $1 \dots n$, **in parallel do**
2: Sample $\xi_i^{(t)}$, compute $\mathbf{g}_i^{(t)} := \text{Clip}(\nabla \ell(\mathbf{x}_i^{(t)}; C), \xi_i^{(t)})$, where $\text{Clip}(\mathbf{g}; C) := \min\left\{1, \frac{C}{\|\mathbf{g}\|}\right\} \cdot \mathbf{g}$
3: Sample for all $j \in \mathcal{N}_i$, $\mathbf{v}_{ij}^{(t)} = -\mathbf{v}_{ji}^{(t)} \sim \mathcal{N}(0, \sigma_{\text{cor}}^2 \mathbf{I}_d)$, and $\overline{\mathbf{v}}_i^{(t)} \sim \mathcal{N}(0, \sigma_{\text{cdp}}^2 \mathbf{I}_d)$
4: $\tilde{\mathbf{g}}_i^{(t)} := \mathbf{g}_i^{(t)} + \sum_{j \in \mathcal{N}_i} \mathbf{v}_{ij}^{(t)} + \overline{\mathbf{v}}_i^{(t)}$ ▷ privacy noise
5: $\mathbf{x}_i^{(t+\frac{1}{2})} = \mathbf{x}_i^{(t)} - \eta_t \tilde{\mathbf{g}}_i^{(t)}$ ▷ stochastic gradient updates
6: $\mathbf{x}_i^{(t+1)} := \sum_{j=1}^n \mathbf{W}_{ij} \mathbf{x}_j^{(t+\frac{1}{2})}$ ▷ gossip averaging
7: **end for**

DECOR is an iterative decentralized algorithm proceeding in T iterations, whereby at each iteration $t \in [T]$, each user $i \in \{1, \dots, n\}$ first computes and clips a stochastic gradient at the current local model $\mathbf{x}_i^{(t)}$ (line 2 of Algorithm 1):

$$\mathbf{g}_i^{(t)} := \text{Clip}(\nabla \ell(\mathbf{x}_i^{(t)}, \xi_i^{(t)}); C),$$

where $\xi_i^{(t)}$ is a data point sampled at random from user i's dataset $\mathcal{D}_i$, and clipping with threshold C corresponds to $\text{Clip}(\mathbf{g}; C) := \min\left\{1, \frac{C}{\|\mathbf{g}\|}\right\} \cdot \mathbf{g}$ for any vector $\mathbf{g} \in \mathbb{R}^d$. The clipping operation ensures that the sensitivity of the gradient, to a change in data, is bounded as required by DP. Then, on line 4 of Algorithm 1, each user obfuscates the clipped gradient by adding privacy noise:

$$\tilde{\mathbf{g}}_i^{(t)} := \mathbf{g}_i^{(t)} + \sum_{j \in \mathcal{N}_i} \mathbf{v}_{ij}^{(t)} + \overline{\mathbf{v}}_i^{(t)}, \tag{3}$$

where is $\overline{\mathbf{v}}_i^{(t)} \sim \mathcal{N}(0, \sigma_{\text{cdp}}^2 \mathbf{I}_d)$ is independent Gaussian noise, $\mathcal{N}_i$ is the set of neighbors of i on graph $\mathcal{G}$, and $\{\mathbf{v}_{ij}^{(t)}\}_{j \in \mathcal{N}_i}$ are pairwise-cancelling correlated Gaussian noise terms; they satisfy $\mathbf{v}_{ij}^{(t)} = -\mathbf{v}_{ji}^{(t)} \sim \mathcal{N}(0, \sigma_{\text{cor}}^2 \mathbf{I}_d)$. Then, on line 5, each user makes a local update with the obfuscated stochastic gradient to obtain:

$$\mathbf{x}_i^{(t+\frac{1}{2})} = \mathbf{x}_i^{(t)} - \eta_t \tilde{\mathbf{g}}_i^{(t)},$$

where η_t is the iteration's learning rate. Finally, on line 6, each user broadcasts the obtained local model to its neighbors on graph $\mathcal{G}$, and updates its local model by performing a weighted average of the neighbors' local models:

$$\mathbf{x}_i^{(t+1)} := \sum_{j=1}^n \mathbf{W}_{ij} \mathbf{x}_j^{(t+\frac{1}{2})}, \tag{4}$$

where the weights are zero for non-neighboring users and form the mixing matrix $\mathbf{W} = [\mathbf{W}_{ij}]_{i,j\in[n]} \in \mathbb{R}^{n\times n}$, which is symmetric and doubly stochastic (see Definition 2 below). The motivation for injecting correlated noise in (3) is that the gossip averaging in (4) will cancel out part or all correlated noise terms. For example, if $\mathcal{G}$ is the fully connected graph and $\mathbf{W} = \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the matrix of ones times $\frac{1}{n}$, then (3) cancels out all correlated noise terms. Still, the uncorrelated noise term $\bar{\mathbf{v}}_i^{(t)}$ remains to protect the privacy of the gossip-averaged local model $\mathbf{x}_i^{(t+1)}$.

4. Privacy Analysis

In this section, we formalize the privacy guarantees of DECOR and introduce its privacy accountant.

First, we recall the notion of algebraic connectivity $a(\mathcal{G})$. Formally, algebraic connectivity $a(\mathcal{G})$ is equal to the second-smallest eigenvalue of the Laplacian matrix of the graph. Intuitively, it quantifies how well a graph is connected (Fiedler, 1973). For example, denoting by n the number of vertices, the algebraic connectivity of the fully-connected graph is equal to n; for the star graph, it is equal to 1; and for the ring graph it is equal to $2(1-\cos\frac{2\pi}{n}) = \Theta(\frac{1}{n^2})$ (see (De Abreu, 2007) for a survey).

In this section, it is more convenient to work with SecRDP, the stronger variant of SecLDP based on Rényi DP (RDP) (Mironov, 2017). We defer its formal definition to the appendix, and simply note that SecRDP implies SecLDP in the same way that RDP implies DP (Mironov, 2017).

For brevity, in Theorem 1 we state the privacy guarantees of a single step of DECOR against adversaries I and II (defined in Section 2). We defer the extension of the privacy guarantees for adversaries III and IV to the appendix.

Theorem 1. *Let $\alpha > 1$. Each iteration of* DECOR *(Algorithm 1) satisfies $(\alpha, \alpha\varepsilon)$-SecRDP (Definition 4) against*

- *an external eavesdropper with*

$$\varepsilon \leq 2C^2\left(\frac{1}{n\sigma_{\text{cdp}}^2} + \frac{1-\frac{1}{n}}{\sigma_{\text{cdp}}^2 + a(\mathcal{G})\sigma_{\text{cor}}^2}\right),$$

- *honest-but-curious non-colluding users with*

$$\varepsilon \leq 2C^2\left(\frac{1}{(n-1)\sigma_{\text{cdp}}^2} + \frac{1-\frac{1}{n-1}}{\sigma_{\text{cdp}}^2 + a_1(\mathcal{G})\sigma_{\text{cor}}^2}\right),$$

where $a_1(\mathcal{G})$ is the minimum algebraic connectivity across subgraphs obtained by deleting a single vertex from $\mathcal{G}$. Moreover, ε can be computed numerically with Algorithm 2.

The result of Theorem 1 implies that in order to obtain SecRDP and thus SecLDP, i.e., in order to bound ε, it is sufficient for the noise magnitudes to scale as $\sigma_{\text{cdp}}^2 = \Omega(\frac{1}{n})$, and $\sigma_{\text{cor}}^2 = \Omega(\frac{1}{a(\mathcal{G})})$ or $\sigma_{\text{cor}}^2 = \Omega(\frac{1}{a_1(\mathcal{G})})$ depending on the adversary strength. Also, as we shall see in the next section, σ_{cdp}^2 drives the dominant convergence terms, so its dependence on n is crucial, while that of σ_{cor}^2 only influences higher-order convergence terms. If $\mathcal{G}$ is disconnected, then its algebraic connectivity is zero (De Abreu, 2007), and one should set $\sigma_{\text{cdp}}^2 = \Theta(1)$ and $\sigma_{\text{cor}} = 0$ in DECOR, which corresponds to the LDP baseline.

Algorithm 2 SINGLE-STEP SECRDP ACCOUNTANT

Input: clipping threshold C, noise variances $\sigma_{\text{cdp}}, \sigma_{\text{cor}}$.
if external eavesdropper **then**
  Get Laplacian matrix $\mathbf{L}$ of the full graph $\mathcal{G}$
  Compute $\mathbf{\Sigma} = \left(\sigma_{\text{cdp}}^2\mathbf{I}_n + \sigma_{\text{cor}}^2\mathbf{L}\right)^{-1}$
  return $2C^2\max_{i\in[n]}\mathbf{\Sigma}_{ii}$
end if
if honest-but-curious non-colluding users **then**
  for i **in** $1\ldots n$ **do**
    Get Laplacian matrix $\mathbf{L}$ of the subgraph of $\mathcal{G}$ obtained by deleting vertex i
    Compute $\mathbf{\Sigma} = \left(\sigma_{\text{cdp}}^2\mathbf{I}_{n-1} + \sigma_{\text{cor}}^2\mathbf{L}\right)^{-1}$
    $\varepsilon_i = 2C^2\max_{j\in[n-1]}\mathbf{\Sigma}_{jj}$
  end for
  return $\max_{i\in[n]}\varepsilon_i$
end if

Finally, we compare our privacy analysis of a single step of DECOR with its counterpart in the work of Sabater et al. (2022). Their Theorem 1 states a DP guarantee, while DECOR guarantees RDP, which is stronger (Mironov, 2017). Moreover, our result applies generically to *arbitrary* graph topologies, while theirs necessitates a graph-dependent analysis to derive the theoretical values of σ_{cdp} and σ_{cor}. For worst-case connected graphs, such as the ring graph, their analysis has a tighter dependence in terms of σ_{cor} after converting from RDP to DP, although the latter only marginally influences the privacy-utility trade-off.

Privacy accountant. The theoretical privacy bound from Theorem 1 may be too loose for practical use. Thus, we devise a privacy accounting method, described in Algorithm 2, which allows computing tight privacy bounds for a single step of DECOR. The accounting procedure is simple, and mainly involves computing the inverse of a "modified" graph Laplacian matrix, which can be conducted efficiently for large sparse graphs (Vishnoi, 2012). It is straightforward to account the privacy for the full DECOR procedure using the composition and DP conversion properties of RDP (Mironov, 2017) in addition to Algorithm 2.

5. Utility Analysis

In this section, we present our theoretical convergence and privacy-utility trade-off results. We first state our optimization assumptions below.

5.1. Assumptions

For all our theoretical results, we assume that the local loss functions are smooth.

Assumption 1 (L-smoothness)**.** *Each function $\mathcal{L}_i$ is differentiable and there exists a constant $L \geq 0$ such that for each $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d, i \in [n]$:*

$$\|\nabla \mathcal{L}_i(\mathbf{y}) - \nabla \mathcal{L}_i(\mathbf{x})\|_2 \leq L \|\mathbf{x} - \mathbf{y}\|_2 \ . \tag{5}$$

Additionally, some of our results require the Polyak-Łojasiewicz (PL) inequality (Karimi et al., 2016). This condition does not require convexity, and is implied by strong convexity for example.

Assumption 2 (μ-PL)**.** *Function $\mathcal{L}$ satisfies the μ-Polyak-Łojasiewicz (PL) inequality. That is, for all $\mathbf{x} \in \mathbb{R}^d$:*

$$2\mu(\mathcal{L}(\mathbf{x}) - \mathcal{L}_\star) \leq \|\nabla \mathcal{L}(\mathbf{x})\|_2^2 \,, \tag{6}$$

where $\mathcal{L}_\star \coloneqq \inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x})$ denotes the infimum of $\mathcal{L}$.

We now formulate our conditions on the stochastic gradient noise and local loss functions heterogeneity.

Assumption 3 (Bounded noise and heterogeneity)**.** *We assume that there exist P, $\zeta_\star$ such that for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\tfrac{1}{n} \textstyle\sum_{i=1}^n \|\nabla \mathcal{L}_i(\mathbf{x})\|_2^2 \leq \zeta_\star^2 + P \|\nabla \mathcal{L}(\mathbf{x})\|_2^2 \ , \tag{7}$$

Also, we assume that there exist M, $\sigma_\star$ such that for all $\mathbf{x}_1, \dots \mathbf{x}_n \in \mathbb{R}^d$,

$$\Psi(\mathbf{x}_1, \dots, \mathbf{x}_n) \leq \sigma_\star^2 + \tfrac{M}{n} \textstyle\sum_{i=1}^n \|\nabla \mathcal{L}(\mathbf{x}_i)\|_2^2 \,, \tag{8}$$

where we introduced $\Psi(\mathbf{x}_1, \dots, \mathbf{x}_n) \coloneqq \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\xi_i} \|\nabla \ell(\mathbf{x}_i, \xi_i) - \nabla \mathcal{L}_i(\mathbf{x}_i)\|_2^2$.

Our noise assumption recovers the uniformly bounded noise assumption when $M = 0$ and $n = 1$, which is common for the non-convex analysis of SGD (Bottou et al., 2018). Our gradient heterogeneity assumption is one of the weakest in the literature (Karimireddy et al., 2020). For the smooth convex (or PL) case, these assumptions hold with $\zeta_\star^2$ and $\sigma_\star^2$ being the gradient heterogeneity and noise, respectively, at the minimum only (Vaswani et al., 2019).

We additionally assume that gradients are bounded. This is a common assumption in private optimization to ignore the effect of clipping (Agarwal et al., 2018; Noble et al., 2022; Allouah et al., 2023), which is not the focus of our work.

Assumption 4 (Bounded Gradients)**.** *We assume that there exists $C \geq 0$ such that for each $i \in [n], \mathbf{x} \in \mathbb{R}^d, \xi \in \mathcal{D}_i$,*

$$\|\nabla \ell(\mathbf{x}, \xi)\| \leq C. \tag{9}$$

As is typical in decentralized optimization algorithms, we make use of a mixing matrix $\mathbf{W}$, as defined below.

Definition 2 (Mixing matrix)**.** *A matrix $\mathbf{W} \in [0,1]^{n \times n}$ is a mixing matrix if it is symmetric and stochastic ($\mathbf{W}\mathbf{1} = \mathbf{1}$).*

Finally, we assume that the mixing matrix $\mathbf{W}$ brings any set of vectors closer to their average with factor at least $1 - p$.

Assumption 5 (Consensus rate)**.** *We assume that there exists $p \in (0, 1]$ such that for every matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$,*

$$\left\|\mathbf{X}\mathbf{W} - \bar{\mathbf{X}}\right\|_F^2 \leq (1 - p) \left\|\mathbf{X} - \bar{\mathbf{X}}\right\|_F^2 , \tag{10}$$

where we define the average $\bar{\mathbf{X}} := \mathbf{X}\frac{\mathbf{1}\mathbf{1}^\top}{n}$.

This assumption holds with $1 - p$ being the second-largest eigenvalue value of $\mathbf{W}\mathbf{W}^\top$ (Boyd et al., 2006), e.g., $p = 1$ for the complete graph, $p = \Theta(\frac{1}{n^2})$ for the ring graph.

5.2. Convergence Analysis

We now present the convergence rate of DECOR, showing how the two the privacy noises affect its convergence speed. First, we introduce the following quantity

$$\mathrm{H}_{\mathcal{G}}(\mathbf{W}) \coloneqq \frac{\sum_{i,k=1}^n \|\mathbf{W}_i - \mathbf{W}_k\|^2 \mathbf{1}_{k \in \mathcal{N}_i}}{2 \sum_{i,k=1}^n \mathbf{1}_{k \in \mathcal{N}_i}}, \tag{11}$$

where $\mathbf{W}_i$ denotes the i-th column of $\mathbf{W}$ and "$k \in \mathcal{N}_i$" denotes $\{i, k\} \in \mathcal{E}$. This quantity naturally appears when analyzing the correlated noise reduction after one gossiping step. It measures the heterogeneity of mixing weights $\mathbf{W}$ across connected users of the graph $\mathcal{G}$. The smaller $\mathrm{H}_{\mathcal{G}}(\mathbf{W})$, the closer are mixing weights, and the less correlated noise remains after one gossiping step, e.g., $\mathrm{H}_{\mathcal{G}}(\mathbf{W}) = 0$ for the complete graph with uniform mixing weights $\mathbf{W} = \frac{\mathbf{1}\mathbf{1}^\top}{n}$. More generally, for any graph $\mathcal{G}$ with minimal degree $k_{\min} \geq 1$, we can show that $\mathrm{H}_{\mathcal{G}}(\mathbf{W})$ decreases with the minimal degree as $\mathrm{H}_{\mathcal{G}}(\mathbf{W}) \leq \frac{2}{k_{\min}}$, when using uniform mixing weights, i.e., if $\mathbf{W}_{ij} = \frac{\mathbf{1}_{j \in \mathcal{N}_i}}{\deg(i)+1}, \forall i, j \in [n]$, where $\deg(i) = |\mathcal{N}_i|$ is the degree of user i in the graph.

We now state our convergence result in Theorem 2 below.

Theorem 2. *Let Assumptions 1, 3, 4, 5 hold. Consider Algorithm 1. Denote $\bar{\mathbf{x}}^{(t)} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^{(t)}$, $\mathcal{L}_0 \coloneqq \mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star$, $\Xi_0 \coloneqq \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i^{(0)} - \bar{\mathbf{x}}^{(0)}\|_2^2$, and $c \coloneqq \max\{4\sqrt{3(1-p)(3P + pM)}, \frac{\mu}{L}, 2p, \frac{4pM}{n}\}$. For $T \geq 1$:*

1. If $\mathcal{L}$ is μ-PL (Assumption 2) and $\eta_t = \frac{16}{\mu(t+c\frac{L}{\mu p})}$, then

$$\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(T)}) - \mathcal{L}_\star \lesssim \frac{L(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}{\mu^2 nT} + \frac{c^2L^2\mathcal{L}_0}{\mu^2p^2T^2} + \frac{cL^3\Xi_0}{\mu^2p^2T^2} + \frac{L^2\log T}{\mu^3pT^2}\Big((1-p)(\frac{\zeta_\star^2}{p} + \sigma_\star^2) + \frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2}{n} + \Big\|\mathbf{W} - \tfrac{\mathbf{1}\mathbf{1}^\top}{n}\Big\|_F^2 d\sigma_{\text{cdp}}^2\Big).$$

2. If $\eta_t = \min\{\frac{p}{2cL}, 2\sqrt{\frac{\mathcal{L}_0 n}{LT(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}}\}$, then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\|_2^2 \lesssim \sqrt{\frac{L\mathcal{L}_0(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}{nT}} + \frac{cL\mathcal{L}_0}{pT} + \frac{L^2\Xi_0}{pT} + \frac{L\mathcal{L}_0 n}{pT(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}\Big((1-p)(\frac{\zeta_\star^2}{p} + \sigma_\star^2) + \frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2}{n} + \Big\|\mathbf{W} - \tfrac{\mathbf{1}\mathbf{1}^\top}{n}\Big\|_F^2 d\sigma_{\text{cdp}}^2\Big).$$

In the above, $\lesssim$ denotes inequality up to absolute constants.

In Theorem 2, the leading (slowest) terms of the convergence rates are in $\mathcal{O}(\frac{1}{T})$ and $\mathcal{O}(\frac{1}{\sqrt{T}})$ in the PL and non-convex cases, respectively. Thus, our analysis recovers the optimal asymptotic convergences rates in stochastic optimization (Agarwal et al., 2009; Arjevani et al., 2023). Moreover, it features a linear speedup in the number of users n, like vanilla decentralized SGD (Koloskova et al., 2020).

The main difference with decentralized SGD is in the non-dominant terms due to the injected correlated noise. In the PL case for example, the term that depends on the correlated noise scales as

$$\widetilde{\mathcal{O}}\Big(\frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2}{nT^2}\Big), \tag{12}$$

by ignoring privacy-independent constants and logarithmic terms. The above term quantifies a *slowdown* effect of correlated noise. Interestingly, it is non-dominant in T, and proportional to $\mathrm{H}_{\mathcal{G}}(\mathbf{W})$. For example, this term is zero for the complete graph with $\mathbf{W} = \frac{\mathbf{1}\mathbf{1}^\top}{n}$, because $\mathrm{H}_{\mathcal{G}}(\mathbf{W}) = 0$ in this case, which is expected as all correlated noise terms should be cancelled after one gossiping step.

5.3. Privacy-utility Trade-off

We now combine our privacy and convergence analyses to quantify the privacy-utility trade-off of DECOR. We recall that a graph is 2-connected if it remains connected after removing any vertex. We present the privacy-utility trade-off of DECOR in Corollary 3 below by focusing on the PL case. We defer the non-convex result to the appendix because lower bounds are unknown in this case.

Corollary 3. *Let Assumptions 1-5 hold. Let $\varepsilon > 0, \delta \in (0,1)$ be such that $\varepsilon \leq \log(1/\delta)$. Algorithm 1 satisfies (ε,δ)-SecLDP (Definition 1) with expected error*

$$\mathcal{O}\left(\frac{C^2 d\log(1/\delta)}{n^2\varepsilon^2}\right),$$

against the following adversaries:

- *an external eavesdropper: if $\mathcal{G}$ is connected, $\sigma_{\text{cdp}}^2 = \frac{32C^2T\log(1/\delta)}{n\varepsilon^2}$ and $\sigma_{\text{cor}}^2 = \frac{32C^2T\log(1/\delta)}{a(\mathcal{G})\varepsilon^2}$,*
- *honest-but-curious non-colluding users: if $\mathcal{G}$ is 2-connected, $\sigma_{\text{cdp}}^2 = \frac{32C^2T\log(1/\delta)}{(n-1)\varepsilon^2}$ and $\sigma_{\text{cor}}^2 = \frac{32C^2T\log(1/\delta)}{a_1(\mathcal{G})\varepsilon^2}$, where $a_1(\mathcal{G})$ is the minimum algebraic connectivity across subgraphs obtained by deleting a single vertex from $\mathcal{G}$.*

In the above, $\mathcal{O}$ omits absolute constants, vanishing terms in T, and privacy-independent multiplicative constants L, μ.

Tightness. The lower bound on the privacy-utility trade-off under user-level CDP is $\Omega\big(\frac{d}{n^2\varepsilon^2}\big)$ (Bassily et al., 2014).[3] Under LDP, the lower bound on the privacy-utility trade-off is $\Omega\big(\frac{d}{n\varepsilon^2}\big)$ (Duchi et al., 2018). Therefore, following the result of Corollary 3, DECOR matches the optimal CDP privacy-utility trade-off, under SecLDP against both an external eavesdropper and non-colluding curious users. We recall that this improves by factor n over the trade-off achieved by LDP algorithms (Bellet et al., 2018; Cheng et al., 2019; Huang et al., 2019; Li & Chi, 2023). Besides, for comparison, Cyffers et al. (2022) derive a privacy-utility trade-off in $\mathcal{O}\Big(\frac{k_{\max}}{\sqrt{p}n^2\varepsilon^2}\Big)$, where $k_{\max}$ is the maximum degree of the graph, for a relaxation of Network DP (Cyffers & Bellet, 2022). Their trade-off matches CDP for well-connected graphs such as expanders (Ying et al., 2021), but degrades with poorer connectivity, e.g., $\mathcal{O}\big(\frac{1}{n\varepsilon^2}\big)$ for the ring graph. In contrast, our trade-off matches CDP for arbitrary connected graphs, albeit the privacy definitions are orthogonal in general, as we discuss in Section 2. We also extend Corollary 3 to colluding curious users (adversary III in Section 2) in the appendix and match the optimal privacy-utility trade-off when there is a constant fraction of colluding users. Naturally, if the group of colluding users is too large, the threat model of SecLDP approaches that of LDP, and thus cannot match the privacy-utility trade-off of CDP in such cases.

Although the trade-off achieved by DECOR is tight asymptotically in the number of iterations T, we recall that the convergence speed suffers from a slowdown due to correlated noise following our discussion of the term (12) of

[3] We refer to the strongly convex lower bound of Bassily et al. (2014), which also applies to the (larger) PL class of functions.

Theorem 2. Moreover, our analysis shows a tight trade-off by making stronger assumptions on the connectivity of the graph for stronger adversaries. That is, for non-colluding curious users, Corollary 3 assumes the graph to be 2-connected, i.e., it remains connected after removing any vertex, while it only assumes the graph to be connected for the external eavesdropper. We believe this condition to be necessary: in the worst-case where a curious user i is the unique neighbor of another user j, then i can substract the correlated noise injected by user j, and leave the latter with the uncorrelated noise only, which is insufficient to protect its local model.

6. Empirical Evaluation

In this section, we empirically show that DECOR achieves a privacy-utility trade-off matching the CDP baseline, and surpassing the LDP baseline. Recall that LDP is the strongest threat model in decentralized learning, while CDP is the weakest, and thus they represent lower and upper bounds in terms of performance. We compare these algorithms on three strongly convex and non-convex tasks with synthetic and real-world datasets, across various user-level privacy budgets and network topologies. For simplicity, we focus on adversary I of SecLDP, i.e., an external eavesdropper observing all user communications. We provide a summary of our empirical evaluations in Figure 1.

Setup. We consider $n = 16$ users on three usual network topologies in increasing connectivity: ring, grid (2d torus), and fully-connected. We use the Metropolis-Hastings (Boyd et al., 2006) mixing matrix, i.e., $\mathbf{W}_{ij} = \frac{\mathbf{1}_{j \in \mathcal{N}_i}}{\deg(i)+1}, \forall i, j \in [n]$, where $\deg(i) = |\mathcal{N}_i|$ is the degree of user i in the graph. We tune all hyperparameters for each algorithm individually, and run each experiment with four seeds for reproducibility. We account for the privacy budget using our SecLDP privacy accountant (Algorithm 2). We defer the full experimental setup to the appendix. [4]

6.1. Strongly Convex Tasks

We study two strongly convex tasks: least-squares regression on synthetic data and regularized logistic regression on the *a9a* LIBSVM dataset (Chang & Lin, 2011). Both of these tasks are covered by our theory in the PL case.

Least-squares regression. In this task, the empirical loss function is given for every $i \in [n]$ as follows:

$$\mathcal{L}_i(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}_i\mathbf{x} - \mathbf{b}_i\|_2^2, \ \forall \mathbf{x} \in \mathbb{R}^d,$$

where $\mathbf{A}_i^2 := \frac{i^2}{n}\mathbf{I}_d$ and $\mathbf{b}_i \sim \mathcal{N}(0, \frac{1}{i^2}\mathbf{I}_d)$. As shown in Figure 1a, DECOR achieves an order of magnitude lower training loss than LDP across all privacy levels and topologies. Moreover, the performance of DECOR is comparable to that of CDP, especially towards the low privacy regime.

Logistic regression. In this task, the loss function at data point $(\mathbf{a}, b) \in \mathbb{R}^d \times \{0, 1\}$ is given as:

$$\ell(\mathbf{x}, (\mathbf{a}, b)) = \log(1 + b\exp(-\mathbf{x}^\top \mathbf{a})) + \lambda\|\mathbf{x}\|_2^2, \forall \mathbf{x} \in \mathbb{R}^d.$$

In Figure 1b, similar to the previous task, DECOR closely matches the performance of CDP across all considered topologies and privacy budgets, while being an order of magnitude better than LDP.

6.2. Non-convex Task

We consider the MNIST (LeCun & Cortes, 2010) task with a one-hidden-layer neural network. Since this task is more challenging for the LDP baseline, we consider example-level DP, i.e., privacy is ensured when changing any single data point, to which our theoretical results can be extended straightforwardly using privacy amplification by subsampling (Wang et al., 2019).

In Figure 1c, we once again observe that DECOR matches CDP, and surpasses LDP, on all considered topologies and privacy budgets. Indeed, the gap of CDP with LDP is almost 10 accuracy points for the lowest privacy budget, as suggested by the theory, while the gap between DECOR on the ring topology and the CDP baseline, or DECOR on the grid topology, is less than 1 accuracy point.

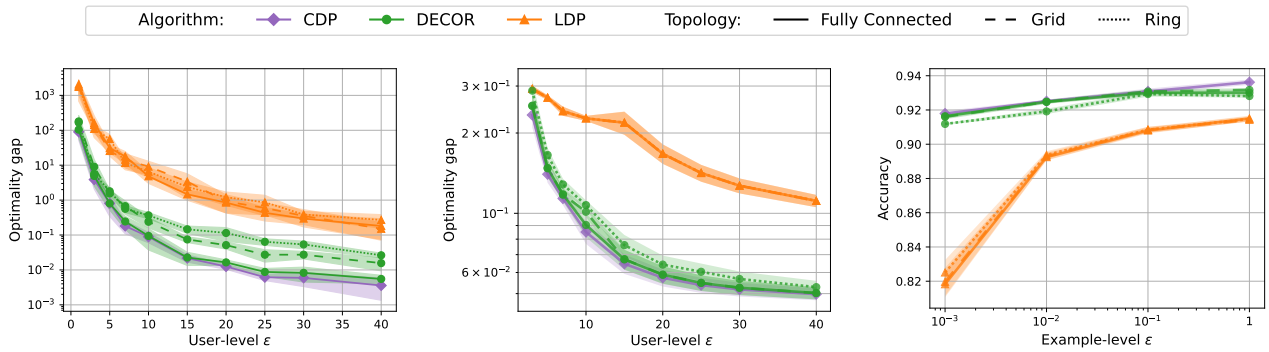
7. Conclusion

Decentralized learning is a promising paradigm for scalable data usage without compromising privacy. However, the risk of privacy breaches by curious users remains. Our solution DECOR addresses this challenge by securely exchanging randomness seeds and injecting pairwise correlated Gaussian noises to protect local models. Our theoretical and empirical results show that DECOR matches the optimal centralized privacy-utility trade-offs, while ensuring differential privacy within the strong threat model of SecLDP, our new relaxation of LDP. Our proposed SecLDP offers protection against external eavesdroppers and curious users, assuming shared secrets among connected pairs. We propose a privacy accountant with our algorithm in order to foster reproducibility and encourage further research and deployment of private decentralized learning algorithms.

Acknowledgments

This work was supported in part by SNSF grant 200021_200477. The authors are thankful to the anonymous reviewers for their constructive comments.

[4] Our code is available at https://github.com/elfirdoussilab1/DECOR.



(a) Least-squares regression on synthetic data (b) Logistic regression on a9a (c) Neural network training on MNIST

Figure 1. Privacy-utility trade-offs for DECOR and the CDP and LDP baselines on least-squares regression, logistic regression, and neural network training under $(\varepsilon, 10^{-5})$-SecLDP against an external eavesdropper observing all communications. DECOR closely matches the performance of CDP, and considerably surpasses LDP, across all considered tasks, privacy budgets, and topologies.

Impact Statement

Our work is mainly theoretical, we do not think any specific societal consequences must be highlighted here.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Agarwal, A., Wainwright, M. J., Bartlett, P., and Ravikumar, P. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22, 2009.

Agarwal, N., Suresh, A. T., Yu, F. X. X., Kumar, S., and McMahan, B. cpsgd: Communication-efficient and differentially-private distributed sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

Agarwal, N., Kairouz, P., and Liu, Z. The skellam mechanism for differentially private federated learning. *Advances in Neural Information Processing Systems*, 34: 5052–5064, 2021.

Allouah, Y., Guerraoui, R., Gupta, N., Pinot, R., and Stephan, J. On the privacy-robustness-utility trilemma in distributed learning. In *International Conference on Machine Learning*, 2023.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199 (1-2):165–214, 2023.

Arora, R., Bassily, R., González, T., Guzmán, C., Menart, M., and Ullah, E. Faster rates of convergence to stationary points in differentially private optimization. *arXiv preprint arXiv:2206.00846*, 2022.

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

Bellet, A., Guerraoui, R., Taziki, M., and Tommasi, M. Personalized and private peer-to-peer machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 473–481. PMLR, 2018.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1175–1191, 2017.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Chang, C.-C. and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Cheng, H.-P., Yu, P., Hu, H., Zawad, S., Yan, F., Li, S., Li, H., and Chen, Y. Towards decentralized deep learning with differential privacy. In *International Conference on Cloud Computing*, pp. 130–145. Springer, 2019.

Cheu, A., Smith, A., Ullman, J., Zeber, D., and Zhilyaev, M. Distributed differential privacy via shuffling. In *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*, pp. 375–403. Springer, 2019.

Cyffers, E. and Bellet, A. Privacy amplification by decentralization. In *International Conference on Artificial Intelligence and Statistics*, pp. 5334–5353. PMLR, 2022.

Cyffers, E., Even, M., Bellet, A., and Massoulié, L. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. *Advances in Neural Information Processing Systems*, 35:15889–15902, 2022.

De Abreu, N. M. M. Old and new results on algebraic connectivity of graphs. *Linear algebra and its applications*, 423(1):53–73, 2007.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE, 2013.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Erlingsson, Ú., Feldman, V., Mironov, I., Raghunathan, A., Talwar, K., and Thakurta, A. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2468–2479. SIAM, 2019.

Fiedler, M. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

Gil, M., Alajaji, F., and Linder, T. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

Huang, Z., Hu, R., Guo, Y., Chan-Tin, E., and Gong, Y. Dp-admm: Admm-based distributed learning with differential privacy. *IEEE Transactions on Information Forensics and Security*, 15:1002–1012, 2019.

Imtiaz, H., Mohammadi, J., and Sarwate, A. D. Distributed differentially private computation of functions with correlated noise. *arXiv preprint arXiv:1904.10059*, 2019.

Jayaraman, B., Wang, L., Evans, D., and Gu, Q. Distributed learning without distress: Privacy-preserving empirical risk minimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Kairouz, P., Liu, Z., and Steinke, T. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *International Conference on Machine Learning*, pp. 5201–5212. PMLR, 2021a.

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., and Xu, Z. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR, 2021b.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

LeCun, Y. and Cortes, C. MNIST handwritten digit database. 2010.

Li, B. and Chi, Y. Convergence and privacy of decentralized nonconvex optimization with gradient clipping and communication compression. *arXiv preprint arXiv:2305.09896*, 2023.

Li, Q., Cascudo, I., and Christensen, M. G. Privacy-preserving distributed average consensus based on additive secret sharing. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. IEEE, 2019.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in neural information processing systems*, 30, 2017.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 691–706. IEEE, 2019.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10110–10145. PMLR, 2022.

Sabater, C., Bellet, A., and Ramon, J. An accurate, scalable and verifiable protocol for federated differentially private averaging. *Machine Learning*, 111(11):4249–4293, 2022.

Shamir, A. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific reports*, 10(1):1–12, 2020.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019.

Vishnoi, N. K. Laplacian solvers and their algorithmic applications. *Theoretical Computer Science*, 8(1-2):1–141, 2012.

Wang, Y.-X., Balle, B., and Kasiviswanathan, S. P. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019.

Ying, B., Yuan, K., Chen, Y., Hu, H., Pan, P., and Yin, W. Exponential graph is provably efficient for decentralized deep training. *Advances in Neural Information Processing Systems*, 34:13975–13987, 2021.

APPENDIX

The appendix is organized as follows. The proofs and extensions of our privacy analysis, including Theorem 1, are in Appendix A. The proofs and extensions of our convergence analysis, including Theorem 2, are in Appendix B. The proofs and extensions of our privacy-utility trade-off, including Corollary 3, are in Appendix C. Our detailed experimental setup is in Appendix D.

A. Privacy Analysis

In this section, we prove our main privacy result stated in Theorem 1 and extend it to the general privacy adversaries discussed in Section 2. We first recall some useful facts around Rényi divergences and linear algebra.

Definition 3 (α-Rényi divergence)**.** *Let $\alpha > 0, \alpha \neq 1$. The α-Rényi divergence between two probability distributions P and Q is defined as*

$$\mathrm{D}_\alpha(P \parallel Q) \coloneqq \frac{1}{\alpha - 1} \log \mathbb{E}_{X \sim Q} \left(\frac{P(X)}{Q(X)} \right)^\alpha.$$

Lemma 4 ((Gil et al., 2013))**.** *Let $\alpha > 0, \alpha \neq 1, \mu_1, \mu_2 \in \mathbb{R}^n$, and $\mathbf{\Sigma} \in \mathbb{R}^{n \times n}$. Assume that $\mathbf{\Sigma}$ is positive definite. The α-Rényi divergence between the multivariate Gaussian distributions $\mathcal{N}(\mu_1, \mathbf{\Sigma})$ and $\mathcal{N}(\mu_2, \mathbf{\Sigma})$ is*

$$\mathrm{D}_\alpha(\mathcal{N}(\mu_1, \mathbf{\Sigma}) \parallel \mathcal{N}(\mu_2, \mathbf{\Sigma})) = \frac{\alpha}{2}(\mu_1 - \mu_2)^\top \mathbf{\Sigma}^{-1}(\mu_1 - \mu_2).$$

We recall the folklore result below, which is a consequence of the Courant-Fischer min-max theorem (De Abreu, 2007).

Lemma 5. *Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ be a real symmetric matrix and $\mathbf{u}_n \in \mathbb{R}^n$ be an eigenvector associated to the largest eigenvalue of $\mathbf{M}$. The second-largest eigenvalue of $\mathbf{M}$ is*

$$\lambda_{n-1}(\mathbf{M}) = \sup_{\substack{\mathbf{u} \neq \mathbf{0} \\ \langle \mathbf{u}, \mathbf{u}_n \rangle = 0}} \frac{\mathbf{u}^\top \mathbf{M} \mathbf{u}}{\|\mathbf{u}\|_2^2}.$$

We now define *secret-based local Rényi differential privacy* (SecRDP), a strong variant of SecLDP based on Rényi DP.

Definition 4 (SecRDP)**.** *Let $\varepsilon \geq 0$, $\delta \in [0, 1]$, $\alpha > 1$. Consider a randomized decentralized algorithm $\mathcal{A} : \mathcal{X}^{m \times n} \to \mathcal{Y}$, which outputs the transcript of all communications. Algorithm $\mathcal{A}$ is said to satisfy $(\alpha, \varepsilon, \mathcal{S})$-SecRDP if $\mathcal{A}$ satisfies (α, ε)-RDP given that $\mathcal{S}$ is unknown to the adversary. That is, for every adjacent datasets $\mathcal{D}, \mathcal{D}' \in \mathcal{X}^{m \times n}$, we have*

$$\mathrm{D}_\alpha(\mathcal{A}(\mathcal{D}) \mid \mathcal{S} \text{ is hidden} \parallel \mathcal{A}(\mathcal{D}') \mid \mathcal{S} \text{ is hidden}) \leq \varepsilon,$$

where the left-hand side is the Rényi divergence (Definition 4) between the probability distributions of $\mathcal{A}(\mathcal{D})$ and $\mathcal{A}(\mathcal{D}')$, conditional on the secrets $\mathcal{S}$ being hidden from the adversary. We simply say that $\mathcal{A}$ satisfies (α, ε)-SecRDP if it satisfies $(\alpha, \varepsilon, \mathcal{S})$-SecRDP for a certain $\mathcal{S}$.

Both SecLDP and SecRDP preserve the properties of DP and RDP, respectively, since these relaxations only condition the probability space of the considered distributions.

Proof of Theorem 1. For convenience, we restate Theorem 1 below, whose proof is a special case of the extended privacy result given next.

Theorem 1. *Let $\alpha > 1$. Each iteration of* DECOR *(Algorithm 1) satisfies $(\alpha, \alpha\varepsilon)$-SecRDP (Definition 4) against*

- *an external eavesdropper with*

$$\varepsilon \leq 2C^2 \left(\frac{1}{n\sigma_{\mathrm{cdp}}^2} + \frac{1 - \frac{1}{n}}{\sigma_{\mathrm{cdp}}^2 + a(\mathcal{G})\sigma_{\mathrm{cor}}^2} \right),$$

- *honest-but-curious non-colluding users with*

$$\varepsilon \leq 2C^2 \left(\frac{1}{(n-1)\sigma_{\mathrm{cdp}}^2} + \frac{1 - \frac{1}{n-1}}{\sigma_{\mathrm{cdp}}^2 + a_1(\mathcal{G})\sigma_{\mathrm{cor}}^2} \right),$$

Algorithm 3 GENERAL SECRDP ACCOUNTANT FOR DECOR

Input: clipping threshold C, noise variances $\sigma_{\text{cdp}}, \sigma_{\text{cor}}$, collusion level q.

for $\mathcal{I} \subseteq [n], |\mathcal{I}| = q$ **do**
  Get Laplacian matrix $\mathbf{L}$ of the subgraph of $\mathcal{G}$ after deleting vertices $\mathcal{I}$
  Compute $\mathbf{\Sigma} = \left(\sigma_{\text{cdp}}^2 \mathbf{I}_{n-q} + \sigma_{\text{cor}}^2 \mathbf{L}\right)^{-1}$
  $\varepsilon_{\mathcal{I}} = 2C^2 \max_{i \in [n-q]} \mathbf{\Sigma}_{ii}$
end for
return $\max_{\mathcal{I} \subseteq [n], |\mathcal{I}|=q} \varepsilon_{\mathcal{I}}$

where $a_1(\mathcal{G})$ is the minimum algebraic connectivity across subgraphs obtained by deleting a single vertex from $\mathcal{G}$. Moreover, ε can be computed numerically with Algorithm 2.

Proof. The result is a special case of Theorem 6, by taking $q = 0$ for the external eavesdropper and $q = 1$ for the honest-but-curious non-colluding users. □

Extended privacy analysis. We now state and prove a general privacy analysis of DECOR to all considered adversaries in Section 2, which includes collusion. We additionally provide a SecRDP accountant in Algorithm 3, which generalizes Algorithm 2 to the aforementioned adversaries.

Theorem 6. *Let $\alpha > 1$ and $q < n$. Each iteration of Algorithm 1 satisfies $(\alpha, \alpha\varepsilon)$-SecRDP (Definition 4) against honest-but-curious users colluding at level q with*

$$\varepsilon \leq 2C^2 \left(\frac{1}{(n-q)\sigma_{\text{cdp}}^2} + \frac{1 - \frac{1}{n-q}}{\sigma_{\text{cdp}}^2 + a_q(\mathcal{G})\sigma_{\text{cor}}^2} \right), \tag{13}$$

where $a_q(\mathcal{G})$ is the minimum algebraic connectivity across subgraphs obtained by deleting q vertices from $\mathcal{G}$. Moreover, ε can be computed numerically using Algorithm 3.

Proof. Let $\alpha > 1$, $q > 1$, and $\mathcal{I} \subseteq [n]$ be an arbitrary group of $|\mathcal{I}| = q$ users. Recall that we denote by $\mathcal{S}_{\mathcal{I}} := \{s_{jk} \colon \{j,k\} \in \mathcal{E}, j,k \notin \mathcal{I}\}$ the set of secrets hidden from all users in $\mathcal{I}$. We will prove that Algorithm 1 satisfies $(\alpha, \varepsilon, \mathcal{S}_{\mathcal{I}})$-SecRDP, which protects against honest-but-curious users colluding at level q, as discussed in Section 2. For ease of exposition, we consider the one-dimensional case $d = 1$. Extending the proof to the general case is straightforward.

Formally, at each iteration of Algorithm 1, users possess private inputs (gradients) in the form of vector $\mathbf{x} \in [-C, C]^n$, given that gradients are clipped at threshold C. Each user $i \in [n]$ shares the following privatized quantity:

$$\widetilde{\mathbf{x}}_i := \mathbf{x}_i + \sum_{j \in \mathcal{N}_i} \mathbf{v}_{ij} + \bar{\mathbf{v}}_i, \tag{14}$$

where $\mathbf{v}_{ij} = -\mathbf{v}_{ji} \sim \mathcal{N}(0, \sigma_{\text{cor}}^2)$ for all $j \in \mathcal{N}_i$, and $\bar{\mathbf{v}}_i \sim \mathcal{N}(0, \sigma_{\text{cdp}}^2)$. Note that each neighborhood $\mathcal{N}_i$ does not include i.

Denote by $\mathcal{H} := [n] \setminus \mathcal{I}$ the set of the $|\mathcal{H}| = n - q$ honest (non-colluding) users. Our goal is to show that the mechanism producing $\widetilde{\mathbf{X}}_{\mathcal{H}} := \left[\widetilde{\mathbf{x}}_i\right]_{i \in \mathcal{H}}$ satisfies SecRDP when a single entry of $\mathbf{X} := \left[\mathbf{x}_i\right]_{i \in \mathcal{H}}$ is arbitrarily changed; i.e., one user's input differs. To do so, we first rewrite (14) to discard the noise terms known to the colluding curious users who can simply substract them to get for every $i \in \mathcal{H}$:

$$\widetilde{\mathbf{x}}_i = \mathbf{x}_i + \sum_{j \in \mathcal{N}_i \cap \mathcal{H}} \mathbf{v}_{ij} + \bar{\mathbf{v}}_i, \tag{15}$$

Denote by $\mathcal{G}_{\mathcal{H}} := (\mathcal{H}, \mathcal{E}_{\mathcal{H}})$ the subgraph of $\mathcal{G}$ restricted to honest users. We now rewrite the above in matrix form as:

$$\widetilde{\mathbf{X}}_{\mathcal{H}} = \mathbf{X}_{\mathcal{H}} + \mathbf{K}\mathbf{N}_{\mathcal{E}} + \bar{\mathbf{N}}, \tag{16}$$

where $\mathbf{K} \in \mathbb{R}^{(n-q)\times|\mathcal{E}_{\mathcal{H}}|}$ is the oriented incidence matrix of the graph $\mathcal{G}_{\mathcal{H}}$ and $\mathbf{N}_{\mathcal{E}_{\mathcal{H}}} = [\mathbf{v}_{ij}]_{1\leq i<j\leq n-q} \in \mathbb{R}^{|\mathcal{E}_{\mathcal{H}}|}$ is the vector of pairwise noises. Now, consider two input vectors $\mathbf{X}_A, \mathbf{X}_B \in [-C, C]^{n-q}$ which differ maximally in an arbitrary coordinate $i \in [n-q]$ without loss of generality:

$$\mathbf{X}_A - \mathbf{X}_B = 2C\mathbf{e}_i \in \mathbb{R}^{n-q}, \tag{17}$$

where $\mathbf{e}_i$ is the vector of $\mathbb{R}^{n-q}$ where the only nonzero element is 1 in the i-th coordinate.

We will then show that the α-Rényi divergence between $\widetilde{\mathbf{X}}_A$ and $\widetilde{\mathbf{X}}_B$, which are respectively produced by input vectors $\mathbf{X}_A$ and $\mathbf{X}_B$, is bounded. To do so, by looking at Equation (16), we can see that $\widetilde{\mathbf{X}}_A, \widetilde{\mathbf{X}}_B$ follow a multivariate Gaussian distribution of means $\mathbf{X}_A, \mathbf{X}_B$ respectively and of variance

$$\mathbf{\Sigma} \coloneqq \mathbb{E}(\widetilde{\mathbf{X}}_A - \mathbf{X}_A)(\widetilde{\mathbf{X}}_A - \mathbf{X}_A)^\top = \mathbb{E}(\widetilde{\mathbf{X}}_B - \mathbf{X}_B)(\widetilde{\mathbf{X}}_B - \mathbf{X}_B)^\top = \sigma_{\text{cor}}^2\mathbf{L} + \sigma_{\text{cdp}}^2\mathbf{I}_{n-q} \in \mathbb{R}^{(n-q)\times(n-q)}, \tag{18}$$

where $\mathbf{L} = \mathbf{K}\mathbf{K}^\top \in \mathbb{R}^{(n-q)\times(n-q)}$ is the Laplacian matrix of the graph $\mathcal{G}_{\mathcal{H}}$ (De Abreu, 2007). Note that $\mathbf{\Sigma}$ is positive definite when $\sigma_{\text{cdp}}^2 > 0$ because $\mathbf{L}$ is positive semi-definite.

Therefore, following Lemma 4, the α-Rényi divergence between the distributions of $\widetilde{\mathbf{X}}_A$ and $\widetilde{\mathbf{X}}_B$ is

$$D_\alpha(\widetilde{\mathbf{X}}_A \parallel \widetilde{\mathbf{X}}_B) = \frac{\alpha}{2}(\mathbf{X}_A - \mathbf{X}_B)^\top\mathbf{\Sigma}^{-1}(\mathbf{X}_A - \mathbf{X}_B). \tag{19}$$

Now, recall that the spectrum of $\mathbf{L}$ is $0 = \lambda_1(\mathbf{L}) \leq \ldots \leq \lambda_{n-q}(\mathbf{L})$ because it is the Laplacian matrix of the graph $\mathcal{G}_{\mathcal{H}}$. Moreover, the eigenvector corresponding to the zero eigenvalue is $\mathbf{1} \in \mathbb{R}^{n-q}$ the vector of ones. Thus, since $\mathbf{\Sigma}$ is a real symmetric (positive definite) matrix, the spectrum of $\mathbf{\Sigma}^{-1}$ in ascending order is $\left(\frac{1}{\sigma_{\text{cdp}}^2+\sigma_{\text{cor}}^2\lambda_{n-q-i+1}(\mathbf{L})}\right)_{i\in[n-q]}$, and $\mathbf{1}$ the vector of ones is associated to its largest eigenvalue:

$$\mathbf{\Sigma}^{-1}\mathbf{1} = \frac{1}{\sigma_{\text{cdp}}^2}\mathbf{1}. \tag{20}$$

Define $\mathbf{x}_i \coloneqq \mathbf{e}_i - \frac{1}{n-q}\mathbf{1}$ and observe that $\langle \mathbf{x}_i, \mathbf{1}\rangle = 0$. Therefore, we can decompose the vector $\mathbf{X}_A - \mathbf{X}_B$ from Equation (17) as a sum of orthogonal vectors as follows:

$$\mathbf{X}_A - \mathbf{X}_B = 2C\mathbf{e}_i = 2C\left(\frac{1}{n-q}\mathbf{1} + \mathbf{x}_i\right).$$

Going back to (19), we can write

$$\begin{aligned} D_\alpha(\widetilde{\mathbf{X}}_A \parallel \widetilde{\mathbf{X}}_B) &= \frac{\alpha}{2}(\mathbf{X}_A - \mathbf{X}_B)^\top\mathbf{\Sigma}^{-1}(\mathbf{X}_A - \mathbf{X}_B) = \frac{4C^2\alpha}{2}(\frac{1}{n-q}\mathbf{1} + \mathbf{x}_i)^\top\mathbf{\Sigma}^{-1}(\frac{1}{n-q}\mathbf{1} + \mathbf{x}_i) \\ &= 2\alpha C^2\left(\frac{1}{(n-q)^2}\mathbf{1}^\top\mathbf{\Sigma}^{-1}\mathbf{1} + \mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i\right) = 2\alpha C^2\left(\frac{1}{(n-q)\sigma_{\text{cdp}}^2} + \mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i\right), \end{aligned} \tag{21}$$

where we have used that $\langle \mathbf{x}_i, \mathbf{1}\rangle = 0$, Equation (20) and $\langle \mathbf{1}, \mathbf{1}\rangle = n-q$ successively in the last two steps. Now, using Lemma 5 and the facts that $\langle \mathbf{x}_i, \mathbf{1}\rangle = 0$ and $\|\mathbf{x}_i\|_2^2 = 1 - \frac{1}{n-q}$, we have that

$$\mathbf{x}_i^\top\mathbf{\Sigma}^{-1}\mathbf{x}_i \leq \sup_{\substack{\mathbf{u}\neq\mathbf{0} \\ \langle\mathbf{u},\mathbf{1}\rangle=0}} \frac{\mathbf{u}^\top\mathbf{\Sigma}^{-1}\mathbf{u}}{\|\mathbf{u}\|_2^2} \cdot \|\mathbf{x}_i\|_2^2 \leq \lambda_{n-q-1}(\mathbf{\Sigma}^{-1})\,\|\mathbf{x}_i\|_2^2 = (1 - \frac{1}{n-q})\lambda_{n-q-1}(\mathbf{\Sigma}^{-1}) = \frac{1 - \frac{1}{n-q}}{\sigma_{\text{cdp}}^2 + \lambda_2(\mathbf{L})\sigma_{\text{cor}}^2}.$$

Plugging the bound above back in (21), we obtain

$$D_\alpha(\widetilde{\mathbf{X}}_A \parallel \widetilde{\mathbf{X}}_B) \leq 2\alpha C^2\left(\frac{1}{(n-q)\sigma_{\text{cdp}}^2} + \frac{1 - \frac{1}{n-q}}{\sigma_{\text{cdp}}^2 + \lambda_2(\mathbf{L})\sigma_{\text{cor}}^2}\right).$$

Recall that $\lambda_2(\mathbf{L})$ is the algebraic connectivity of the graph $\mathcal{G}_{\mathcal{H}}$, by definition. Moreover, since $\mathcal{I}$ and thus $\mathcal{H}$ are taken arbitrarily, in the worst case $\lambda_2(\mathbf{L})$ is $a_q(\mathcal{G})$ the minimum algebraic connectivity across subgraphs obtained by deleting q vertices from $\mathcal{G}$. This concludes the proof of (13) the main result.

Finally, it is easy to see from Equation (19) that the exact privacy bound ε can be computed numerically using Algorithm 2. Indeed, the maximal difference in inputs is $\mathbf{X}_A - \mathbf{X}_B = 2C\mathbf{e}_i$ for some $i \in [n-q]$ as in (17), so the maximal privacy bound, given that $\mathcal{I}$ is the set of colluding users, is

$$\varepsilon_{\mathcal{I}} = \max_{i\in[n-q]} \frac{1}{2}(2C\mathbf{e}_i)^\top \mathbf{\Sigma}^{-1}(2C\mathbf{e}_i) = 2C^2 \max_{i\in[n-q]} \mathbf{e}_i^\top \mathbf{\Sigma}^{-1}\mathbf{e}_i = 2C^2 \max_{i\in[n-q]} \mathbf{\Sigma}^{-1}_{ii},$$

where $\mathbf{\Sigma}^{-1}_{ii}$ is the i-th entry in the diagonal of the inverse of $\mathbf{\Sigma} = \sigma^2_{\text{cor}}\mathbf{L} + \sigma^2_{\text{cdp}}\mathbf{I}_{n-q}$. Thus, to get the maximal privacy loss across all possible colluding user groups of size q, we take $\varepsilon = \max_{\mathcal{I}\subseteq[n],|\mathcal{I}|=q} \varepsilon_{\mathcal{I}}$. Observing that the latter is exactly the output of Algorithm 3 concludes the proof. □

B. Convergence Analysis

In this section, we prove our convergence analysis stated in Theorem 2 and extend it to the general privacy adversaries discussed in Section 2. We introduce some useful notation in Section B.1, overview the main elements of the proof in Section B.2, prove the main theorem in Section B.3, and finally prove the intermediate lemmas in Section B.4.

B.1. Notation

We can rewrite the procedure of DECOR (Algorithm 1) using the following matrix notation, extending the definition used in Section 3:

$$\begin{aligned}
&\mathbf{X}^{(t)} := \left[\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_n^{(t)}\right] \in \mathbb{R}^{d\times n}, \qquad \bar{\mathbf{X}}^{(t)} := \left[\bar{\mathbf{x}}^{(t)}, \dots, \bar{\mathbf{x}}^{(t)}\right] \in \mathbb{R}^{d\times n},\\
&\partial\ell(\mathbf{X}^{(t)}, \xi^{(t)}) := \left[\nabla\ell(\mathbf{x}_1^{(t)}, \xi_1^{(t)}), \dots, \nabla\ell(\mathbf{x}_n^{(t)}, \xi_n^{(t)})\right] \in \mathbb{R}^{d\times n},\\
&\mathbf{N}^{(t)} := \left[\sum_{j\in\mathcal{N}_1} \mathbf{v}_{1j}^{(t)}, \dots, \sum_{j\in\mathcal{N}_n} \mathbf{v}_{nj}^{(t)}\right] \in \mathbb{R}^{d\times n}, \qquad \bar{\mathbf{N}}^{(t)} := \left[\bar{\mathbf{v}}_1^{(t)}, \dots, \bar{\mathbf{v}}_n^{(t)}\right] \in \mathbb{R}^{d\times n}.
\end{aligned} \tag{22}$$

We recall that under the bounded gradient assumption (Assumption 4), clipping leaves gradients unaffected, and thus we discard the clipping operator in this section.

Algorithm 4 DECOR IN MATRIX NOTATION

Input: for each user $i \in [n]$ initialize $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$, stepsizes $\{\eta_t\}_{t=0}^{T-1}$, number of iterations T, mixing matrix $\mathbf{W}$, noise parameters σ_{cor} and σ_{cdp}.

1: **for** t **in** $0 \dots T-1$ **do**
2: $\quad \mathbf{X}^{(t+\frac{1}{2})} = \mathbf{X}^{(t)} - \eta_t \left(\partial\ell(\mathbf{X}^{(t)}, \xi_i^{(t)}) + \mathbf{N}^{(t)} + \bar{\mathbf{N}}^{(t)}\right)$ ▷ stochastic gradient updates
3: $\quad \mathbf{X}^{(t+1)} = \mathbf{X}^{(t+\frac{1}{2})} \cdot \mathbf{W}$ ▷ gossip averaging
4: **end for**

B.2. Proof Overview

Our convergence analysis relies upon three elements: descent bound, pairwise noise reduction, and consensus distance recursion. We first state the corresponding lemmas, and defer their proofs to Section B.4.

The first proof element is the descent bound of Lemma 7. It quantifies the progress made after each DECOR step. In particular, compared to the error due to stochastic gradient variance $\sigma_\star^2$ as in vanilla SGD (Bottou et al., 2018), there are two additional quantities involved: and (uncorrelated) privacy noise variance σ^2_{cdp}, and the consensus distance Ξ_t defined for every $t \geq 1$ as

$$\Xi_t := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2 = \frac{1}{n}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2. \tag{23}$$

Lemma 7 (Descent bound)**.** *Under Assumptions 1, 3 and 5, the averages* $\bar{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ *of the iterates of Algorithm 1 with* $\eta_t \leq \frac{1}{2L}\min\{1, \frac{n}{2M}\}$ *satisfy*

$$\mathbb{E}\left[\mathcal{L}(\bar{\mathbf{x}}^{(t+1)}) - \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right] \leq -\frac{\eta_t}{4}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{3\eta_t L^2}{4}\Xi_t + \frac{L\eta_t^2}{2}\frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n}. \tag{24}$$

Interestingly, the descent bound does not involve the correlated noise variance σ_{cor}^2. This is thanks to the correlated noise terms cancelling out pairwise, so that the correlated noise disappears when analyzing the average model $\overline{\mathbf{x}}^{(t)}$.

Next, in order to bound the consensus distance Ξ_t, we first quantify in Lemma 8 the effect of correlated noise in a single step of DECOR on the consensus distance Ξ_t.

Lemma 8 (Correlated noise reduction)**.** *Consider Algorithm 1. For any undirected graph* $\mathcal{G} = (\{1,\ldots,n\},\mathcal{E})$ *and any matrix* $\mathbf{W} \in \mathbb{R}^{n\times n}$ *and at every iteration t, we have*

$$\mathbb{E}\left\|\mathbf{N}^{(t)}\mathbf{W}\right\|_F^2 = \mathrm{H}_{\mathcal{G}}(\mathbf{W})\cdot\mathbb{E}\left\|\mathbf{N}^{(t)}\right\|_F^2 = 2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2, \tag{25}$$

where we define $\mathrm{H}_{\mathcal{G}}(\mathbf{W}) := \frac{\sum_{i,k=1}^n \|\mathbf{W}_i - \mathbf{W}_k\|^2 \mathbf{1}_{k\in\mathcal{N}_i}}{2\sum_{i,k=1}^n \mathbf{1}_{k\in\mathcal{N}_i}}$*, and* $\mathbf{1}_{k\in\mathcal{N}_i}$ *denotes* $\{i,k\}\in\mathcal{E}$*, and* $|\mathcal{E}| = \frac{1}{2}\sum_{i,k=1}^n \mathbf{1}_{k\in\mathcal{N}_i}$ *is the number of edges on the graph* $\mathcal{G}$*. Moreover, if* $\mathbf{W}_{ij} = \frac{\mathbf{1}_{j\in\mathcal{N}_i}}{\deg(i)+1}, \forall i,j\in[n]$*, where* $\deg(i) = |\mathcal{N}_i|$ *is the degree of user i in the graph, we have* $\mathrm{H}_{\mathcal{G}}(\mathbf{W}) \leq \frac{2}{k_{\min}}$*, where* $k_{\min} \geq 1$ *is the minimal degree of graph* $\mathcal{G}$*.*

The analysis of the error due to correlated noise in Lemma 8 is exact, in the sense that it is an equality. Recall that $\mathrm{H}_{\mathcal{G}}(\mathbf{W})$ is graph- and mixing matrix-dependent. Broadly speaking, the average expected error per edge (due to correlated noise) is $d\sigma_{\text{cor}}^2$ (the variance of one correlated noise term), reduced by factor $\mathrm{H}_{\mathcal{G}}(\mathbf{W})$, which decreases with the connectivity with the graph. Using this lemma, we can now prove a powerful recursion on the consensus distance in Lemma 9 below.

Lemma 9 (Consensus distance recursion)**.** *Under Assumptions 1, 3, and 5, if in addition stepsizes satisfy* $\eta_t \leq \frac{p}{L\sqrt{6(1-p)(3+pM)}}$*, then*

$$\begin{aligned}\Xi_{t+1} \leq{}& (1-\frac{p}{2})\Xi_t + 2\eta_t^2(1-p)(\frac{3P}{p}+M)\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &+ \eta_t^2\left[6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2}{n} + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2\right],\end{aligned}$$

where $\Xi_t := \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ *is the consensus distance.*

The effects of the privacy noises are apparent in the lemma above, and correspond mainly to the quantity analyzed in Lemma 8, in addition to the effects of stochastic variance and heterogeneity, which are similar to vanilla D-SGD (Koloskova et al., 2020). It is indeed intuitive that the non-cancelled correlated noise should pull the local models away, and this worsens for poorly-connected graphs.

B.3. Main Proof

We now restate and prove Theorem 2 below, using the intermediate lemmas from the previous section.

Theorem 2. *Let Assumptions 1, 3, 4, 5 hold. Consider Algorithm 1. Denote* $\bar{\mathbf{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$*,* $\mathcal{L}_0 := \mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star$*,* $\Xi_0 := \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i^{(0)} - \bar{\mathbf{x}}^{(0)}\|_2^2$*, and* $c := \max\{4\sqrt{3(1-p)(3P+pM)}, \frac{\mu}{L}, 2p, \frac{4pM}{n}\}$*. For* $T \geq 1$*:*

1. If $\mathcal{L}$ *is* μ*-PL (Assumption 2) and* $\eta_t = \frac{16}{\mu(t + c\frac{L}{\mu p})}$*, then*

$$\begin{aligned}\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(T)}) - \mathcal{L}_\star \lesssim{}& \frac{L(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}{\mu^2 nT} + \frac{c^2L^2\mathcal{L}_0}{\mu^2p^2T^2} \\ &+ \frac{cL^3\Xi_0}{\mu^2p^2T^2} + \frac{L^2\log T}{\mu^3pT^2}\Big((1-p)(\frac{\zeta_\star^2}{p} + \sigma_\star^2) \\ &+ \frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2}{n} + \left\|\mathbf{W} - \tfrac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2\Big).\end{aligned}$$

2. If $\eta_t = \min\{\frac{p}{2cL}, 2\sqrt{\frac{\mathcal{L}_0 n}{LT(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}}\}$ *, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \lesssim \sqrt{\frac{L\mathcal{L}_0(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}{nT}} + \frac{cL\mathcal{L}_0}{pT} + \frac{L^2\Xi_0}{pT} + \frac{L\mathcal{L}_0 n}{pT(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}\Big((1-p)(\frac{\zeta_\star^2}{p} + \sigma_\star^2) + \frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2}{n} + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2\Big).$$

In the above, $\lesssim$ *denotes inequality up to absolute constants.*

B.3.1. PL CASE

Proof. Let assumptions 1-5 hold. Consider Algorithm 1 with the stepsize sequence defined for every $t \geq 0$ as:

$$\eta_t \coloneqq \frac{16}{\mu(t + c\frac{L}{\mu p})}, \tag{26}$$

where $c \coloneqq \max\{4\sqrt{3(1-p)(3P+pM)}, \frac{\mu}{L}, 2p, \frac{4pM}{n}\}$. Clearly, this sequence is decreasing and we have for every $t \geq 0$:

$$\eta_t \leq \eta_0 = \min\{\frac{p}{4L\sqrt{3(1-p)(3P+pM)}}, \frac{p}{\mu}, \frac{1}{2L}\min\{1, \frac{n}{2M}\}\}.$$

This ensures that the conditions of lemmas 7 and 9 are verified.

Consider the sequence defined for every $t \geq 0$ as:

$$V_t \coloneqq \mathbb{E}\left[\mathcal{L}(\bar{\mathbf{x}}^{(t)}) - \mathcal{L}_\star\right] + \frac{3L^2\eta_t}{p}\Xi_t, \tag{27}$$

where $\mathcal{L}_\star \coloneqq \inf_{\mathbf{x}\in\mathbb{R}^d}\mathcal{L}(\mathbf{x})$ denotes the infimum of $\mathcal{L}$. Clearly, since Ξ_t is also non-negative as a sum of squared distances, we have $V_t \geq 0$ for every $t \geq 0$. We also define the following auxiliary sequence for every $t \geq 0$:

$$W_t \coloneqq \frac{1}{\eta_t^2}V_t. \tag{28}$$

Fix $t \geq 0$. First, to analyze W_t, we write

$$W_{t+1} - W_t = \frac{1}{\eta_{t+1}^2}V_{t+1} - \frac{1}{\eta_t^2}V_t = \frac{1}{\eta_{t+1}^2}(V_{t+1} - \frac{\eta_{t+1}^2}{\eta_t^2}V_t).$$

Moreover, denoting $\hat{t} \coloneqq t + \tau$, we have $\frac{\eta_{t+1}^2}{\eta_t^2} = \frac{\hat{t}^2}{(\hat{t}+1)^2} = 1 - \frac{1+2\hat{t}}{(\hat{t}+1)^2}$. Thus, we have

$$W_{t+1} - W_t = \frac{1}{\eta_{t+1}^2}(V_{t+1} - (1 - \frac{1+2\hat{t}}{(\hat{t}+1)^2})V_t). \tag{29}$$

On the other hand, to analyze V_t, we use the fact that stepsizes are non-increasing and satisfy the conditions of lemmas 7 and 9:

$$\begin{aligned}
V_{t+1} - V_t &= \mathbb{E}\left[\mathcal{L}(\bar{\mathbf{x}}^{(t+1)}) - \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right] + \frac{3L^2}{p}(\eta_{t+1}\Xi_{t+1} - \eta_t\Xi_t) \\
&\leq \mathbb{E}\left[\mathcal{L}(\bar{\mathbf{x}}^{(t+1)}) - \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right] + \frac{3L^2\eta_t}{p}(\Xi_{t+1} - \Xi_t) \\
&\leq \left(-\frac{\eta_t}{4} + \frac{6L^2\eta_t^3}{p}(1-p)(\frac{3P}{p} + M)\right)\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + (\frac{3\eta_t L^2}{4} - \frac{3L^2\eta_t}{2})\Xi_t + \eta_t^2 A + \eta_t^3 B,
\end{aligned}$$

where we introduced $A \coloneqq \frac{L}{2}\frac{\sigma_\star^2+d\sigma_{\text{cdp}}^2}{n}$ and $B \coloneqq \frac{3L^2}{p}[6(1-p)\frac{\zeta_\star^2}{p}+(1-p)\sigma_\star^2+\frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2}{n}+\left\|\mathbf{W}-\frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2]$ for simplicity. Recall that we have $\eta_t \le \eta_0 \le \frac{p}{4L\sqrt{3(1-p)(3P+pM)}}$, so that $\frac{6L^2}{p}\eta_t^2(1-p)(\frac{3P}{p}+M) \le \frac{1}{8}$. Consequently, we have

$$V_{t+1}-V_t \le -\frac{\eta_t}{8}\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \frac{3L^2\eta_t}{4}\Xi_t + \eta_t^2 A + \eta_t^3 B. \tag{30}$$

Now, recall that $\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \ge 2\mu(\mathcal{L}(\bar{\mathbf{x}}^{(t)})-\mathcal{L}_\star)$ following Assumption 2, so that the bound above becomes

$$\begin{aligned}V_{t+1}-V_t &\le -\frac{\mu\eta_t}{4}(\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(t)})-\mathcal{L}_\star) - \frac{3L^2\eta_t}{4}\Xi_t + \eta_t^2 A + \eta_t^3 B\\ &= -\frac{\mu\eta_t}{4}\left(\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(t)})-\mathcal{L}_\star + \frac{3L^2}{\mu}\Xi_t\right) + \eta_t^2 A + \eta_t^3 B.\end{aligned}$$

Recall also that $\eta_t \le \eta_0 \le \frac{p}{\mu}$. Therefore, we have

$$V_{t+1}-V_t \le -\frac{\mu\eta_t}{4}\left(\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(t)})-\mathcal{L}_\star + \frac{3L^2\eta_t}{p}\Xi_t\right) + \eta_t^2 A + \eta_t^3 B = -\frac{\mu\eta_t}{4}V_t + \eta_t^2 A + \eta_t^3 B.$$

Plugging the above bound back in (29) and then substituting $\eta_t = \frac{16}{\mu\hat{t}}$, we get

$$\begin{aligned}W_{t+1}-W_t &= \frac{1}{\eta_{t+1}^2}(V_{t+1}-(1-\frac{1+2\hat{t}}{\hat{t}^2})V_t) \le \frac{1}{\eta_{t+1}^2}\left(-\frac{\mu\eta_t}{4}V_t + \eta_t^2 A + \eta_t^3 B + \frac{1+2\hat{t}}{\hat{t}^2}V_t\right)\\ &= -\mu^2(\hat{t}+1)^2\left(\frac{4}{\hat{t}}-\frac{1+2\hat{t}}{\hat{t}^2}\right)V_t + \frac{(\hat{t}+1)^2}{\hat{t}^2}A + \frac{16(\hat{t}+1)^2}{\mu\hat{t}^3}B.\end{aligned}$$

Observe that $\hat{t} = t + c\frac{L}{\mu p} \ge c\frac{L}{\mu p}$, so that $\frac{4}{\hat{t}} - \frac{1+2\hat{t}}{\hat{t}^2} = \frac{2}{\hat{t}} - \frac{1}{\hat{t}^2} \le \frac{1}{\hat{t}}$ and $\frac{(\hat{t}+1)^2}{\hat{t}^2} = 1 + \frac{1+2\hat{t}}{\hat{t}^2} \le 4$. Therefore, the bound above becomes

$$W_{t+1}-W_t \le -\mu^2\frac{(\hat{t}+1)^2}{\hat{t}}V_t + 4A + \frac{64}{\mu\hat{t}}B \le 4A + \frac{64}{\mu\hat{t}}B.$$

By summing over $t \in \{0,\dots,T-1\}$ and substituting A and B, we get

$$\begin{aligned}W_T - W_0 \le{}& 2LT\frac{\sigma_\star^2+d\sigma_{\text{cdp}}^2}{n}\\ &+\left(\sum_{t=0}^{T-1}\hat{t}\right)\frac{192L^2}{\mu p}\left(6(1-p)\frac{\zeta_\star^2}{p}+(1-p)\sigma_\star^2+\frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2}{n}+\left\|\mathbf{W}-\frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2\right).\end{aligned}$$

We now substitute W_T and W_0 to obtain $W_T - W_0 = \frac{1}{\eta_T^2}V_T - \frac{1}{\eta_0^2}V_0 = \frac{\mu^2}{256}((T+c\frac{L}{\mu p})^2 V_T - \left(\frac{cL}{\mu p}\right)^2 V_0)$. Also, as $c\frac{L}{\mu p} \ge \frac{2L}{\mu} \ge 2$, we have $\sum_{t=0}^{T-1}\frac{1}{\hat{t}} = \sum_{t=0}^{T-1}\frac{1}{t+c\frac{L}{\mu p}} \le \ln(T+1)$. Thus, after rearranging terms, the inequality above becomes

$$\begin{aligned}(T+c\tfrac{L}{\mu p})^2 V_T - \left(\tfrac{cL}{\mu p}\right)^2 V_0 \le{}& \frac{512LT}{\mu^2}\frac{\sigma_\star^2+d\sigma_{\text{cdp}}^2}{n}\\ &+\ln(T+1)\frac{49152L^2}{\mu^3 p}\left(6(1-p)\frac{\zeta_\star^2}{p}+(1-p)\sigma_\star^2+\frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\text{cor}}^2}{n}+\left\|\mathbf{W}-\frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2\right).\end{aligned}$$

Upon dividing both sides by $(T + c\frac{L}{\mu p})^2$, rearranging terms and recalling that $c\frac{L}{\mu p} \geq 1$, we obtain

$$
\begin{aligned}
V_T &\leq \frac{\left(\frac{cL}{\mu p}\right)^2}{(T + c\frac{L}{\mu p})^2} V_0 + \frac{512LT}{(T + c\frac{L}{\mu p})^2 \mu^2} \frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n} \\
&\quad + \frac{\ln(T+1)}{(T + c\frac{L}{\mu p})^2} \frac{49152L^2}{\mu^3 p} \left(6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\, d\sigma_{\text{cor}}^2}{n} + \left\| \mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right\|_F^2 d\sigma_{\text{cdp}}^2 \right) \\
&\leq \frac{c^2 L^2}{\mu^2 p^2 T^2} V_0 + \frac{512L}{\mu^2 T} \frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n} \\
&\quad + \frac{\ln(T+1)}{T^2} \frac{49152L^2}{\mu^3 p} \left(6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\, d\sigma_{\text{cor}}^2}{n} + \left\| \mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right\|_F^2 d\sigma_{\text{cdp}}^2 \right).
\end{aligned}
$$

Finally, we obtain the final result by substituting V_T, V_0 and η_T, η_0 and rearranging terms:

$$
\begin{aligned}
\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(T)}) - \mathcal{L}_\star + \frac{48L^2}{\mu p (T + c\frac{L}{\mu p})} \Xi_T &\leq \frac{512L}{\mu^2 T} \frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n} \\
&\quad + \frac{\ln(T+1)}{T^2} \frac{49152L^2}{\mu^3 p} \left((1-p)(6\frac{\zeta_\star^2}{p} + \sigma_\star^2) + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\, d\sigma_{\text{cor}}^2}{n} + \left\| \mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right\|_F^2 d\sigma_{\text{cdp}}^2 \right) \\
&\quad + \frac{c^2 L^2 (\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)}{\mu^2 p^2 T^2} + \frac{48cL^3}{\mu^2 p^2 T^2} \Xi_0.
\end{aligned}
$$

□

B.3.2. Non-convex case

Proof. Let assumptions 1, 3, 4, and 5 hold. Consider Algorithm 1 with the constant stepsize sequence defined for every $t \geq 0$ as:

$$
\eta_t = \eta := \min\left\{ \frac{p}{2cL}, 2\sqrt{\frac{(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)n}{LT(\sigma_\star^2 + d\sigma_{\text{cdp}}^2)}} \right\}, \tag{31}
$$

where $c := \max\{4\sqrt{3(1-p)(3P + pM)}, \frac{\mu}{L}, 2p, \frac{4pM}{n}\}$. This ensures that the conditions of lemmas 7 and 9 are verified.

Consider the sequence defined for every $t \geq 0$ as:

$$
V_t := \mathbb{E}\left[\mathcal{L}(\bar{\mathbf{x}}^{(t)}) - \mathcal{L}_\star \right] + \frac{3L^2\eta}{p} \Xi_t, \tag{32}
$$

where $\mathcal{L}_\star := \inf_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x})$ denotes the infimum of $\mathcal{L}$. Clearly, since Ξ_t is also non-negative as a sum of squared distances, we have $V_t \geq 0$ for every $t \geq 0$.

Denote $A := \frac{L}{2} \frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n}$ and $B := \frac{3L^2}{p} \left(6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2}{n} + \left\| \mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n} \right\|_F^2 d\sigma_{\text{cdp}}^2 \right)$. Following the same steps of the PL case until (30), we have

$$
V_{t+1} - V_t \leq -\frac{\eta}{8} \mathbb{E} \left\| \nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 - \frac{3L^2\eta}{4} \Xi_t + \eta^2 A + \eta^3 B \leq -\frac{\eta}{8} \mathbb{E} \left\| \nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 + \eta^2 A + \eta^3 B.
$$

By averaging over $t \in \{0, \ldots, T-1\}$, multiplying by $\frac{8}{\eta}$ and rearranging terms we obtain

$$
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)}) \right\|_2^2 \leq \frac{8(V_0 - V_T)}{\eta T} + 8\eta A + 8\eta^2 B.
$$

By recalling that $V_T \geq 0$ and substituting the values of V_0 and A, we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{8(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star+\frac{3L^2\eta}{p}\Xi_0)}{\eta T}+4\eta L\frac{\sigma_\star^2+d\sigma_{\text{cdp}}^2}{n}+8\eta^2 B$$
$$=\frac{8(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)}{\eta T}+4\eta L\frac{\sigma_\star^2+d\sigma_{\text{cdp}}^2}{n}+8\eta^2 B+\frac{24L^2}{pT}\Xi_0.$$

Now, recalling the value of η, and that $\frac{1}{\eta}=\max\left\{\frac{2cL}{p},\frac{1}{2}\sqrt{\frac{TL(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}{(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)n}}\right\}\leq\frac{2cL}{p}+\frac{1}{2}\sqrt{\frac{TL(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}{(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)n}}$. Therefore, the bound above becomes

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{16cL(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)}{pT}+4\sqrt{\frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}{nT}}$$
$$+8\sqrt{\frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}{nT}}+\frac{32(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)n}{LT(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}B+\frac{24L^2}{pT}\Xi_0.$$

By rearranging terms and substituting B, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \leq 12\sqrt{\frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}{nT}}+\frac{16cL(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)}{pT}+\frac{24L^2}{pT}\Xi_0$$
$$+\frac{96L(\mathcal{L}(\bar{\mathbf{x}}^{(0)})-\mathcal{L}_\star)n}{pT(\sigma_\star^2+d\sigma_{\text{cdp}}^2)}\left(6(1-p)\frac{\zeta_\star^2}{p}+(1-p)\sigma_\star^2+\frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\left|\mathcal{E}\right|d\sigma_{\text{cor}}^2}{n}+\left\|\mathbf{W}-\frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2\right).$$

The above concludes the proof. □

B.4. Proof of Lemmas

We now restate and prove the intermediate lemmas from the previous sections.

Lemma 7 (Descent bound)**.** *Under Assumptions 1, 3 and 5, the averages* $\bar{\mathbf{x}}^{(t)}:=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_i^{(t)}$ *of the iterates of Algorithm 1 with* $\eta_t\leq\frac{1}{2L}\min\left\{1,\frac{n}{2M}\right\}$ *satisfy*

$$\mathbb{E}\left[\mathcal{L}(\bar{\mathbf{x}}^{(t+1)})-\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right]\leq-\frac{\eta_t}{4}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2+\frac{3\eta_t L^2}{4}\Xi_t+\frac{L\eta_t^2}{2}\frac{\sigma_\star^2+d\sigma_{\text{cdp}}^2}{n}. \tag{24}$$

Proof. Let assumptions 1, 3, and 5 hold. Because mixing matrices preserve the average, as a direct consequence of Definition 2, we have

$$\bar{\mathbf{x}}^{(t+1)}=\bar{\mathbf{x}}^{(t)}-\frac{\eta_t}{n}\sum_{i=1}^n\tilde{\mathbf{g}}_i^{(t)}=\bar{\mathbf{x}}^{(t)}-\frac{\eta_t}{n}\sum_{i=1}^n\left(\nabla\ell(\mathbf{x}_i^{(t)},\xi_i^{(t)})+\sum_{j\in\mathcal{N}_i}\mathbf{v}_{i,j}^{(t)}+\overline{\mathbf{v}}_i^{(t)}\right)$$
$$=\bar{\mathbf{x}}^{(t)}-\frac{\eta_t}{n}\sum_{i=1}^n\nabla\ell(\mathbf{x}_i^{(t)},\xi_i^{(t)})-\frac{\eta_t}{n}\sum_{i=1}^n\sum_{j\in\mathcal{N}_i}\mathbf{v}_{i,j}^{(t)}-\frac{\eta_t}{n}\sum_{i=1}^n\overline{\mathbf{v}}_i^{(t)}.$$

Recall that for all $i\in[n],j\in\mathcal{N}_i$, we have $\mathbf{v}_{i,j}^{(t)}=-\mathbf{v}_{j,i}^{(t)}$, so that $\sum_{i=1}^n\sum_{j\in\mathcal{N}_i}\mathbf{v}_{i,j}^{(t)}=0$. Reporting this in the equation above yields:

$$\bar{\mathbf{x}}^{(t+1)}=\bar{\mathbf{x}}^{(t)}-\frac{\eta_t}{n}\sum_{i=1}^n\nabla\ell(\mathbf{x}_i^{(t)},\xi_i^{(t)})-\frac{\eta_t}{n}\sum_{i=1}^n\overline{\mathbf{v}}_i^{(t)}. \tag{33}$$

Also, since function $\mathcal{L}$ is L-smooth as the average of smooth functions (Assumption 1), by taking conditional expectation $\mathbb{E}_t$ on all randomness up to iteration t, we have (see (Bottou et al., 2018))

$$\mathbb{E}_t\,\mathcal{L}(\bar{\mathbf{x}}^{(t+1)})\leq\mathcal{L}(\bar{\mathbf{x}}^{(t)})+\underbrace{\mathbb{E}_t\left\langle\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)}),\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\rangle}_{=:A}+\frac{L}{2}\eta_t^2\underbrace{\mathbb{E}_t\left\|\bar{\mathbf{x}}^{(t+1)}-\bar{\mathbf{x}}^{(t)}\right\|_2^2}_{=:B}. \tag{34}$$

We start by bounding A, by using (33) and the smoothness of $\mathcal{L}_i$, as follows:

$$
\begin{aligned}
A &= -\eta_t \left\langle \nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)}), \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \nabla \ell(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) + \frac{1}{n}\sum_{i=1}^n \overline{\mathbf{v}}_i^{(t)}\right]\right\rangle = -\eta_t \left\langle \nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\rangle \\
&= \frac{\eta_t}{2}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)}) - \nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2\right] \\
&\le \frac{\eta_t}{2}\left[\frac{1}{n}\sum_{i=1}^n \left\|\nabla \mathcal{L}_i(\mathbf{x}_i^{(t)}) - \nabla \mathcal{L}_i(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2\right] \\
&\le -\frac{\eta_t}{2}\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \frac{\eta_t}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{\eta_t L^2}{2}\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2.
\end{aligned}
$$

For the last term B, using (33) and Assumption 3, we obtain

$$
\begin{aligned}
B &= \mathbb{E}_t \left\|\frac{1}{n}\sum_{i=1}^n \nabla \ell(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) + \frac{1}{n}\sum_{i=1}^n \overline{\mathbf{v}}_i^{(t)}\right\|_2^2 = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla \ell(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|_2^2 + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n \overline{\mathbf{v}}_i^{(t)}\right\|_2^2 \\
&= \mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^n \left(\nabla \ell(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right)\right\|_2^2 + \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{d\sigma_{\text{cdp}}^2}{n} \\
&\le \frac{\sigma_\star^2}{n} + \frac{M}{n^2}\sum_{i=1}^n \left\|\nabla \mathcal{L}(\mathbf{x}_i^{(t)})\right\|_2^2 + \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{d\sigma_{\text{cdp}}^2}{n} \\
&\le \frac{\sigma_\star^2}{n} + \frac{2M}{n^2}\sum_{i=1}^n \left\|\nabla \mathcal{L}(\mathbf{x}_i^{(t)}) - \nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{2M}{n}\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{d\sigma_{\text{cdp}}^2}{n} \\
&\le \frac{\sigma_\star^2}{n} + \frac{2ML^2}{n^2}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{2M}{n}\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{d\sigma_{\text{cdp}}^2}{n}.
\end{aligned}
$$

Combining the bounds on A and B in (34), we obtain

$$
\begin{aligned}
\mathbb{E}_t\, \mathcal{L}(\bar{\mathbf{x}}^{(t+1)}) &\le \mathcal{L}(\bar{\mathbf{x}}^{(t)}) - \frac{\eta_t}{2}\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \frac{\eta_t}{2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{\eta_t L^2}{2}\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 \\
&\quad + \frac{L}{2}\eta_t^2\left[\frac{\sigma_\star^2}{n} + \frac{2ML^2}{n^2}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{2M}{n}\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{d\sigma_{\text{cdp}}^2}{n}\right] \\
&\le \mathcal{L}(\bar{\mathbf{x}}^{(t)}) - \frac{\eta_t}{2}(1 - \frac{2ML}{n}\eta_t)\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 - \frac{\eta_t}{2}(1 - L\eta_t)\left\|\frac{1}{n}\sum_{i=1}^n \nabla \mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 \\
&\quad + \frac{\eta_t L^2}{2}(1 + \frac{2ML}{n}\eta_t)\frac{1}{n}\sum_{i=1}^n \left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{L\eta_t^2}{2}\frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n}.
\end{aligned}
$$

By using $\eta_t \le \frac{1}{2L}\min\{1, \frac{n}{2M}\}$ and taking total expectations, we conclude:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{L}(\bar{\mathbf{x}}^{(t+1)}) - \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right] &\le -\frac{\eta_t}{4}\,\mathbb{E}\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{3\eta_t L^2}{4}\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{L\eta_t^2}{2}\frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n} \\
&= -\frac{\eta_t}{4}\,\mathbb{E}\left\|\nabla \mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{3\eta_t L^2}{4}\Xi_t + \frac{L\eta_t^2}{2}\frac{\sigma_\star^2 + d\sigma_{\text{cdp}}^2}{n}.
\end{aligned}
$$

□

Lemma 8 (Correlated noise reduction)**.** *Consider Algorithm 1. For any undirected graph $\mathcal{G} = (\{1, \ldots, n\}, \mathcal{E})$ and any matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ and at every iteration t, we have*

$$\mathbb{E}\left\|\mathbf{N}^{(t)}\mathbf{W}\right\|_F^2 = \mathrm{H}_{\mathcal{G}}(\mathbf{W}) \cdot \mathbb{E}\left\|\mathbf{N}^{(t)}\right\|_F^2 = 2\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\mathrm{cor}}^2, \tag{25}$$

where we define $\mathrm{H}_{\mathcal{G}}(\mathbf{W}) := \frac{\sum_{i,k=1}^n \|\mathbf{W}_i - \mathbf{W}_k\|^2 \mathbf{1}_{k\in\mathcal{N}_i}}{2\sum_{i,k=1}^n \mathbf{1}_{k\in\mathcal{N}_i}}$, and $\mathbf{1}_{k\in\mathcal{N}_i}$ denotes $\{i,k\} \in \mathcal{E}$, and $|\mathcal{E}| = \frac{1}{2}\sum_{i,k=1}^n \mathbf{1}_{k\in\mathcal{N}_i}$ is the number of edges on the graph $\mathcal{G}$. Moreover, if $\mathbf{W}_{ij} = \frac{\mathbf{1}_{j\in\mathcal{N}_i}}{\deg(i)+1}, \forall i,j \in [n]$, where $\deg(i) = |\mathcal{N}_i|$ is the degree of user i in the graph, we have $\mathrm{H}_{\mathcal{G}}(\mathbf{W}) \leq \frac{2}{k_{\min}}$, where $k_{\min} \geq 1$ is the minimal degree of graph $\mathcal{G}$.

Proof. Let $\mathcal{G} = ([n], \mathcal{E})$ be an arbitrary undirected graph, and $\mathbf{W} \in \mathbb{R}^{n\times n}$ be an arbitrary matrix (not necessarily a mixing matrix nor dependent upon $\mathcal{G}$). First, we prove that for every $j \in [n]$, we have

$$\mathbf{N}^{(t)}\mathbf{W}_j = \frac{1}{2}\sum_{i,k=1}^n (\mathbf{W}_{ij} - \mathbf{W}_{kj})\mathbf{1}_{k\in\mathcal{N}_i}\mathbf{v}_{ik}^{(t)}, \tag{35}$$

where $\mathbf{W}_j \in \mathbb{R}^n$ denotes the j-th column of $\mathbf{W}$. Indeed, we have

$$\mathbf{N}^{(t)}\mathbf{W}_j = \sum_{i=1}^n \mathbf{W}_{ij}N_i^{(t)} = \sum_{i=1}^n\sum_{k\in\mathcal{N}_i}\mathbf{W}_{ij}\mathbf{v}_{ik}^{(t)} = \sum_{i,k=1}^n \mathbf{W}_{ij}\mathbf{1}_{k\in\mathcal{N}_i}\mathbf{v}_{ik}^{(t)} \tag{36}$$

$$= \sum_{i,k=1}^n \mathbf{W}_{ij}\mathbf{1}_{i\in\mathcal{N}_k}\mathbf{v}_{ik}^{(t)} = -\sum_{i,k=1}^n \mathbf{W}_{ij}\mathbf{1}_{i\in\mathcal{N}_k}\mathbf{v}_{ki}^{(t)} = -\sum_{i,k=1}^n \mathbf{W}_{kj}\mathbf{1}_{k\in\mathcal{N}_i}\mathbf{v}_{ik}^{(t)}, \tag{37}$$

where the last three equalities were successively obtained by using the facts that $\mathcal{G}$ is undirected so $\mathbf{1}_{i\in\mathcal{N}_k} = \mathbf{1}_{k\in\mathcal{N}_i}$, that $\mathbf{v}_{ik}^{(t)} = -\mathbf{v}_{ik}^{(t)}, \forall i,k \in [n]$, and exchanging symbols i,k in the double summation. Thus, averaging equalities (36) and (37) proves Equation (35).

Now, using Equation (35), we can write

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{N}^{(t)}\mathbf{W}\right\|_F^2 &= \sum_{j=1}^n \mathbb{E}\left\|\mathbf{N}^{(t)}\mathbf{W}_j\right\|^2 = \frac{1}{4}\sum_{j=1}^n \mathbb{E}\left\|\sum_{i,k=1}^n (\mathbf{W}_{ij} - \mathbf{W}_{kj})\mathbf{1}_{k\in\mathcal{N}_i}\mathbf{v}_{ik}^{(t)}\right\|^2 \\
&= \frac{1}{4}\sum_{j=1}^n \mathbb{E}\left\|\sum_{\substack{i,k=1\\ i<k}}^n \left[(\mathbf{W}_{ij} - \mathbf{W}_{kj})\mathbf{1}_{k\in\mathcal{N}_i}\mathbf{v}_{ik}^{(t)} + (\mathbf{W}_{kj} - \mathbf{W}_{ij})\mathbf{1}_{i\in\mathcal{N}_k}\mathbf{v}_{ki}^{(t)}\right]\right\|^2 \\
&= \frac{1}{4}\sum_{j=1}^n \mathbb{E}\left\|2\sum_{\substack{i,k=1\\ i<k}}^n (\mathbf{W}_{ij} - \mathbf{W}_{kj})\mathbf{1}_{k\in\mathcal{N}_i}\mathbf{v}_{ik}^{(t)}\right\|^2 = \sum_{j=1}^n\sum_{\substack{i,k=1\\ i<k}}^n (\mathbf{W}_{ij} - \mathbf{W}_{kj})^2\mathbf{1}_{k\in\mathcal{N}_i}\mathbb{E}\left\|\mathbf{v}_{ik}^{(t)}\right\|^2 \\
&= \sum_{j=1}^n\sum_{\substack{i,k=1\\ i<k}}^n (\mathbf{W}_{ij} - \mathbf{W}_{kj})^2\mathbf{1}_{k\in\mathcal{N}_i}d\sigma_{\mathrm{cor}}^2 = \frac{1}{2}\sum_{i,j,k=1}^n (\mathbf{W}_{ij} - \mathbf{W}_{kj})^2\mathbf{1}_{k\in\mathcal{N}_i}d\sigma_{\mathrm{cor}}^2 \\
&= \frac{1}{2}\sum_{i,k=1}^n \|\mathbf{W}_i - \mathbf{W}_k\|^2\,\mathbf{1}_{k\in\mathcal{N}_i}d\sigma_{\mathrm{cor}}^2,
\end{aligned}$$

where in the fourth equality we used that $\mathbf{v}_{ki}^{(t)} = -\mathbf{v}_{ik}^{(t)}$ and that $\mathbf{1}_{i\in\mathcal{N}_k} = \mathbf{1}_{k\in\mathcal{N}_i}$, on the fifth equality we used that $\mathbf{v}_{ik}^{(t)}$ are independent for $i < k$, and on the sixth equality that $\mathbb{E}\left\|\mathbf{v}_{il}^{(t)}\right\|^2 = d\sigma_{\mathrm{cor}}^2$.

Also, taking $\mathbf{W} = \mathbf{I}_n$ in the equation above (which holds for arbitrary $\mathbf{W}$), we have $\|\mathbf{W}_i - \mathbf{W}_k\|^2 = 2 \cdot \mathbf{1}_{k\neq i}$, and thus

$$\mathbb{E}\left\|\mathbf{N}^{(t)}\right\|_F^2 = \sum_{i,k=1}^n \mathbf{1}_{k\in\mathcal{N}_i}d\sigma_{\mathrm{cor}}^2 = 2\,|\mathcal{E}|\,d\sigma_{\mathrm{cor}}^2.$$

The last two equations directly lead to the main result of the lemma.

Now, denote by $k_{\min} \geq 1$ the minimal degree of $\mathcal{G}$ and assume that $\mathbf{W}_{ij} = \frac{\mathbf{1}_{j\in\mathcal{N}_i}}{\deg(i)+1}, \forall i,j \in [n]$, where $\deg(i) = |\mathcal{N}_i|$ is the degree of user i in the graph. Thus, we have $\|\mathbf{W}_i\|^2 = \frac{\deg(i)}{(\deg(i)+1)^2}$. Using Jensen's inequality, we have

$$\begin{aligned}\sum_{i,k=1}^{n} \|\mathbf{W}_i - \mathbf{W}_k\|^2 \mathbf{1}_{k\in\mathcal{N}_i} &\leq 2\sum_{i,k=1}^{n} \left(\|\mathbf{W}_i\| + \|\mathbf{W}_k\|^2\right) \mathbf{1}_{k\in\mathcal{N}_i} = 4\sum_{i,k=1}^{n} \|\mathbf{W}_i\|^2 \mathbf{1}_{k\in\mathcal{N}_i} \\ &= 4\sum_{i,k=1}^{n} \frac{\deg(i)}{(\deg(i)+1)^2}\mathbf{1}_{k\in\mathcal{N}_i} = 4\sum_{i=1}^{n} \frac{\deg(i)^2}{(\deg(i)+1)^2}.\end{aligned}$$

On the other hand, we have $2\sum_{i,k=1}^{n} \mathbf{1}_{k\in\mathcal{N}_i} = 2\sum_{i=1}^{n} \deg(i)$, so that

$$\mathrm{H}_{\mathcal{G}}(\mathbf{W}) = \frac{\sum_{i,k=1}^{n} \|\mathbf{W}_i - \mathbf{W}_k\|^2 \mathbf{1}_{k\in\mathcal{N}_i}}{2\sum_{i,k=1}^{n} \mathbf{1}_{k\in\mathcal{N}_i}} \leq 2\frac{\sum_{i=1}^{n} \frac{\deg(i)^2}{(\deg(i)+1)^2}}{\sum_{i=1}^{n} \deg(i)} \leq 2\max_{i\in[n]} \frac{\deg(i)}{(\deg(i)+1)^2} \leq \frac{2k_{\min}}{(k_{\min}+1)^2} \leq \frac{2}{k_{\min}}.$$

This concludes the second statement of the lemma. □

Lemma 9 (Consensus distance recursion)**.** *Under Assumptions 1, 3, and 5, if in addition stepsizes satisfy* $\eta_t \leq \frac{p}{L\sqrt{6(1-p)(3+pM)}}$*, then*

$$\begin{aligned}\Xi_{t+1} \leq (1-\frac{p}{2})\Xi_t &+ 2\eta_t^2(1-p)(\frac{3P}{p}+M)\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &+ \eta_t^2\left[6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\mathrm{cor}}^2}{n} + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\mathrm{cdp}}^2\right],\end{aligned}$$

where $\Xi_t := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|^2$ *is the consensus distance.*

Proof. Let assumptions 1, 3, and 5 hold. Also, assume that stepsizes verify $\eta_t \leq \frac{p}{96\sqrt{6}\tau L}$ for each iteration t. Denote

$$\partial\ell(\mathbf{X}^{(t)}) := \left[\nabla\mathcal{L}_1(\mathbf{x}_1^{(t)}), \ldots, \nabla\mathcal{L}_n(\mathbf{x}_n^{(t)})\right] \in \mathbb{R}^{d\times n}.$$

We first write

$$\begin{aligned}\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)} &= \mathbf{X}^{(t+\frac{1}{2})}\mathbf{W} - \mathbf{X}^{(t+\frac{1}{2})}\frac{\mathbf{1}\mathbf{1}^\top}{n} = \mathbf{X}^{(t+\frac{1}{2})}(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}) \\ &= \left[\mathbf{X}^{(t)} - \eta_t\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) + \mathbf{N}^{(t)} + \bar{\mathbf{N}}^{(t)}\right)\right](\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}) \\ &= \mathbf{X}^{(t)}(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}) - \eta_t\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) + \mathbf{N}^{(t)} + \bar{\mathbf{N}}^{(t)}\right)(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}) \\ &= (\mathbf{X}^{(t)} - \eta_t\partial\ell(\mathbf{X}^{(t)}))(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}) \\ &\quad - \eta_t\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)}) + \mathbf{N}^{(t)} + \bar{\mathbf{N}}^{(t)}\right)(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}).\end{aligned}$$

By independence, taking squared Frobenius norms and total expectations yields

$$\begin{aligned}n\Xi_{t+1} = \mathbb{E}\left\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)}\right\|_F^2 &= \mathbb{E}\left\|(\mathbf{X}^{(t)} - \eta_t\partial\ell(\mathbf{X}^{(t)}))(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 \\ &\quad + \eta_t^2\,\mathbb{E}\left\|\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)}) + \mathbf{N}^{(t)} + \bar{\mathbf{N}}^{(t)}\right)(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2. \qquad (38)\end{aligned}$$

The first term on the RHS of (38) can be bounded, by first using Assumption 5 and then Young's inequality, as follows:

$$\begin{aligned}
&\mathbb{E}\left\|(\mathbf{X}^{(t)} - \eta_t \partial\ell(\mathbf{X}^{(t)}))(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 \leq (1-p)\,\mathbb{E}\left\|\mathbf{X}^{(t)} - \eta_t\partial\ell(\mathbf{X}^{(t)}) - \bar{\mathbf{X}}^{(t)} + \eta_t\partial\ell(\mathbf{X}^{(t)})\frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 \\
&= (1-p)\,\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)} - \eta_t(\partial\ell(\mathbf{X}^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 \\
&\leq (1-p)(1+\frac{p}{3(1-p)})\,\mathbb{E}\left\|\mathbf{X}^{(t)} - \bar{\mathbf{X}}^{(t)}\right\|_F^2 + (1-p)(1+\frac{3(1-p)}{p})\eta_t^2\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 \\
&= (1-\frac{2p}{3})n\Xi_t + \frac{(1-p)(3-2p)}{p}\eta_t^2\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 \\
&\leq (1-\frac{2p}{3})n\Xi_t + \frac{3(1-p)}{p}\eta_t^2\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)})\right\|_F^2,
\end{aligned}$$

where the last inequality is due to $p \geq 0$ and also that for any $A \in \mathbb{R}^{d\times n}$, $B \in \mathbb{R}^{n\times n}$, we have $\|AB\|_F \leq \|A\|_F \|B\|_2$, along with the fact that $\left\|\mathbf{I}_n - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_2 = 1$.

The second term on the RHS of (38) can be bounded, using independence, as follows:

$$\begin{aligned}
&\mathbb{E}\left\|\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)}) + \mathbf{N}^{(t)} + \bar{\mathbf{N}}^{(t)}\right)(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 = \\
&\quad \mathbb{E}\left\|\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\right)(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 + \mathbb{E}\left\|\mathbf{N}^{(t)}(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 + \mathbb{E}\left\|\bar{\mathbf{N}}^{(t)}(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2.
\end{aligned}$$

We note that since $\bar{\mathbf{N}}^{(t)}$ is a matrix of $d \times n$ i.i.d. Gaussian variables of variance σ_{cdp}^2, we have $\mathbb{E}\left\|\bar{\mathbf{N}}^{(t)}(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 = \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 dn\sigma_{\text{cdp}}^2$. Moreover, using the fact that the sum of correlated noise terms is zero and then Lemma 8, we have $\mathbb{E}\left\|\mathbf{N}^{(t)}(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 = \mathbb{E}\left\|\mathbf{N}^{(t)}\mathbf{W}\right\|_F^2 = 2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2$. Plugging these last two results above, and then using Assumption 5, yields:

$$\begin{aligned}
&\mathbb{E}\left\|\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)}) + \mathbf{N}^{(t)} + \bar{\mathbf{N}}^{(t)}\right)(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 = \\
&\quad \mathbb{E}\left\|\left(\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\right)(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n})\right\|_F^2 + 2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2 + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 dn\sigma_{\text{cdp}}^2 \\
&\leq (1-p)\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)}) - \nabla\mathcal{L}(\mathbf{X}^{(t)},\xi^{(t)}) + \nabla\mathcal{L}(\mathbf{X}^{(t)})\right\|_F^2 + 2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2 + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 dn\sigma_{\text{cdp}}^2 \\
&\leq (1-p)\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\right\|_F^2 + 2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2 + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 dn\sigma_{\text{cdp}}^2.
\end{aligned}$$

Reporting the previous bounds back in (38) gives

$$\begin{aligned}
n\Xi_{t+1} \leq (1-\frac{2p}{3})n\Xi_t + \frac{3(1-p)}{p}\eta_t^2\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)})\right\|_F^2 + \eta_t^2\Big((1-p)\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\right\|_F^2 \\
+ 2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2 + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 dn\sigma_{\text{cdp}}^2\Big).
\end{aligned}$$

Rearranging and dividing by n yields

$$\begin{aligned}
\Xi_{t+1} \leq (1-\frac{2p}{3})\Xi_t + \frac{\eta_t^2}{n}\Big[\frac{3(1-p)}{p}\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)})\right\|_F^2 + (1-p)\,\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\right\|_F^2 \\
+ \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\text{cor}}^2}{n} + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\text{cdp}}^2\Big]. \qquad (39)
\end{aligned}$$

On the one hand, by using assumptions 1 and 3 and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)})\right\|_F^2 &= \sum_{i=1}^n \mathbb{E}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 \leq 2\sum_{i=1}^n \mathbb{E}\left\|\nabla\mathcal{L}_i(\mathbf{x}_i^{(t)}) - \nabla\mathcal{L}_i(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + 2\sum_{i=1}^n \mathbb{E}\left\|\nabla\mathcal{L}_i(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &\leq 2L^2\sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + 2n\zeta_\star^2 + 2nP\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|^2 \\ &= 2L^2 n\Xi_t + 2n\zeta_\star^2 + 2nP\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|^2.\end{aligned}$$

On the other hand, by using assumptions 3 and 1 and Jensen's inequality, we obtain that

$$\begin{aligned}\mathbb{E}\left\|\partial\ell(\mathbf{X}^{(t)},\xi^{(t)}) - \partial\ell(\mathbf{X}^{(t)})\right\|_F^2 &= \sum_{i=1}^n \mathbb{E}\left\|\nabla\ell(\mathbf{x}_i^{(t)},\xi^{(t)}) - \nabla\mathcal{L}_i(\mathbf{x}_i^{(t)})\right\|_2^2 \leq n\sigma_\star^2 + M\sum_{i=1}^n \mathbb{E}\left\|\nabla\mathcal{L}(\mathbf{x}_i^{(t)})\right\|_2^2 \\ &\leq n\sigma_\star^2 + 2M\sum_{i=1}^n \mathbb{E}\left\|\nabla\mathcal{L}(\mathbf{x}_i^{(t)}) - \nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 + 2Mn\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &\leq n\sigma_\star^2 + 2ML^2\sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\right\|_2^2 + 2Mn\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &= n\sigma_\star^2 + 2ML^2 n\Xi_t + 2Mn\,\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2.\end{aligned}$$

By reporting the two bounds above back into (39) and rearranging terms, and using $\eta_t \leq \frac{p}{L\sqrt{6(1-p)(3+pM)}}$ we obtain

$$\begin{aligned}\Xi_{t+1} &\leq \left[1 - \frac{2p}{3} + 2(1-p)L^2\eta_t^2(\frac{3}{p} + M)\right]\Xi_t + 2\eta_t^2(1-p)(\frac{3P}{p} + M)\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &\quad + \eta_t^2\left[6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\mathrm{cor}}^2}{n} + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\mathrm{cdp}}^2\right] \\ &\leq (1 - \frac{p}{2})\Xi_t + 2\eta_t^2(1-p)(\frac{3P}{p} + M)\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 \\ &\quad + \eta_t^2\left[6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{2\mathrm{H}_{\mathcal{G}}(\mathbf{W})|\mathcal{E}|d\sigma_{\mathrm{cor}}^2}{n} + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\mathrm{cdp}}^2\right].\end{aligned}$$

The above concludes the proof. □

C. Privacy-utility Trade-off

In this section, we prove our main privacy result stated in Corollary 3 and extend it to the general privacy adversaries discussed in Section 2. We first recall some useful facts around Rényi differential privacy (RDP) (Mironov, 2017).

Lemma 10 (RDP Composition, (Mironov, 2017))**.** *If a privacy mechanism $\mathcal{M}_1$ that takes the dataset as input is (α, ε_1)-RDP, and a privacy mechanism $\mathcal{M}_2$ that takes the dataset and the output of $\mathcal{M}_1$ as input is (α, ε_2)-RDP, then their composition $\mathcal{M}_2 \circ \mathcal{M}_1$ is $(\alpha, \varepsilon_1 + \varepsilon_2)$-RDP.*

Lemma 11 (RDP to DP conversion, (Mironov, 2017))**.** *If a privacy mechanism $\mathcal{M}$ is (α, ε)-RDP, then $\mathcal{M}$ is $(\varepsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-DP for all $\delta \in (0, 1)$.*

Proof of Corollary 3. For convenience, we restate Corollary 3 below, whose proof is a special case of the extended privacy-utility trade-off result given next.

Corollary 3. *Let Assumptions 1-5 hold. Let $\varepsilon > 0, \delta \in (0, 1)$ be such that $\varepsilon \leq \log(1/\delta)$. Algorithm 1 satisfies (ε, δ)-SecLDP (Definition 1) with expected error*

$$\mathcal{O}\left(\frac{C^2 d\log(1/\delta)}{n^2\varepsilon^2}\right),$$

against the following adversaries:

- *an external eavesdropper: if $\mathcal{G}$ is connected, $\sigma^2_{\text{cdp}} = \frac{32C^2T\log(1/\delta)}{n\varepsilon^2}$ and $\sigma^2_{\text{cor}} = \frac{32C^2T\log(1/\delta)}{a(\mathcal{G})\varepsilon^2}$,*
- *honest-but-curious non-colluding users: if $\mathcal{G}$ is 2-connected, $\sigma^2_{\text{cdp}} = \frac{32C^2T\log(1/\delta)}{(n-1)\varepsilon^2}$ and $\sigma^2_{\text{cor}} = \frac{32C^2T\log(1/\delta)}{a_1(\mathcal{G})\varepsilon^2}$, where $a_1(\mathcal{G})$ is the minimum algebraic connectivity across subgraphs obtained by deleting a single vertex from $\mathcal{G}$.*

In the above, $\mathcal{O}$ omits absolute constants, vanishing terms in T, and privacy-independent multiplicative constants L, μ.

Proof. This result is a special case of Corollary 12, by taking $q = 0$ for the external eavesdropper and $q = 1$ for the honest-but-curious non-colluding users in the PL case, and omitting vanishing terms in T. □

Extended privacy-utility trade-off. We now state and prove a general privacy-utility trade-off analysis of DECOR to all considered adversaries in Section 2, which includes collusion, as well as the non-convex case.

Corollary 12. *Let the assumptions of theorems 1 and 2 hold and assume that $\mathcal{G}$ is $(q+1)$-connected. Let $\varepsilon > 0, \delta \in (0,1)$ be such that $\varepsilon \leq \log(1/\delta)$. Consider Algorithm 1 with $\sigma^2_{\text{cdp}} = \frac{32C^2T\log(1/\delta)}{(n-q)\varepsilon^2}$ and $\sigma^2_{\text{cor}} = \frac{32C^2T\log(1/\delta)}{a_q(\mathcal{G})\varepsilon^2}$. Denote $\mathcal{L}_0 := \mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star$. Then, Algorithm 1 satisfies (ε, δ)-SecLDP and the following holds:*

1. *Assume that $\mathcal{L}$ is μ-PL:*

$$\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(T)}) - \mathcal{L}_\star = \widetilde{\mathcal{O}}\left(\frac{LC^2d\log(1/\delta)}{\mu^2 n(n-q)\varepsilon^2} + \frac{L}{\mu^2 nT}\left[\sigma_\star^2 + \frac{LC^2d\log(1/\delta)}{\mu p\varepsilon^2}\left(\frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|}{a(\mathcal{G}_{\mathcal{H}})} + \frac{n}{(n-q)}\left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2\right)\right]\right).$$

2. *In the general non-convex case:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 = \mathcal{O}\left(\frac{C\sqrt{d\log(1/\delta)}}{\sqrt{n(n-q)}\varepsilon} + \sqrt{\frac{L\mathcal{L}_0\sigma_\star^2}{nT}}\right).$$

Proof. Let the assumptions of theorems 1 and 2 hold. Let $\varepsilon > 0, \delta \in (0,1)$ be such that $\varepsilon \leq \log(1/\delta)$ and assume that $\mathcal{G}$ is $(q+1)$-connected. Consider Algorithm 1 with $\sigma^2_{\text{cdp}} = \frac{32C^2T\log(1/\delta)}{(n-q)\varepsilon^2}$ and $\sigma^2_{\text{cor}} = \frac{32C^2T\log(1/\delta)}{a_q(\mathcal{G})\varepsilon^2}$. The latter quantity is well-defined as $\mathcal{G}$ is $(q+1)$-connected and thus has positive algebraic connectivity after deleting any set of q vertices (De Abreu, 2007).

Privacy. We first show the privacy claim. Recall from Theorem 1 that each iteration of Algorithm 1 satisfies $(\alpha, \alpha\varepsilon_{\text{step}})$-SecRDP against collusion at level q for every $\alpha > 1$ where

$$\varepsilon_{\text{step}} \leq 2C^2\left(\frac{1}{(n-q)\sigma^2_{\text{cdp}}} + \frac{1}{a_q(\mathcal{G})\sigma^2_{\text{cor}}}\right). \tag{40}$$

Thus, following the composition property of RDP from Lemma 11, the full Algorithm 1 satisfies $(\alpha, T\alpha\varepsilon_{\text{step}})$-SecRDP for any $\alpha > 1$. From Lemma 11, we deduce that Algorithm 1 satisfies $(\varepsilon'(\alpha), \delta)$-SecLDP for any $\delta \in (0,1)$ and any $\alpha > 1$, where

$$\varepsilon'(\alpha) = T\alpha\varepsilon_{\text{step}} + \frac{\log(1/\delta)}{\alpha - 1} \leq 2\alpha C^2T\left(\frac{1}{(n-q)\sigma^2_{\text{cdp}}} + \frac{1}{a_q(\mathcal{G})\sigma^2_{\text{cor}}}\right) + \frac{\log(1/\delta)}{\alpha - 1}.$$

Optimizing the above bound over $\alpha > 1$ yields the solution $\alpha_\star = 1 + \frac{\sqrt{\log(1/\delta)}}{C\sqrt{2T\left(\frac{1}{(n-q)\sigma^2_{\text{cdp}}} + \frac{1}{a_q(\mathcal{G})\sigma^2_{\text{cor}}}\right)}}$ which gives the bound

$$\varepsilon_\star = \varepsilon'(\alpha_\star) \leq 2C^2T\left(\frac{1}{(n-q)\sigma^2_{\text{cdp}}} + \frac{1}{a_q(\mathcal{G})\sigma^2_{\text{cor}}}\right) + 2C\sqrt{2T\log(1/\delta)\left(\frac{1}{(n-q)\sigma^2_{\text{cdp}}} + \frac{1}{a_q(\mathcal{G})\sigma^2_{\text{cor}}}\right)}.$$

Now, recall that the choice of $\sigma_{\mathrm{cdp}}^2, \sigma_{\mathrm{cor}}^2$ implies that

$$\frac{1}{(n-q)\sigma_{\mathrm{cdp}}^2} + \frac{1}{a_q(\mathcal{G})\sigma_{\mathrm{cor}}^2} = \frac{\varepsilon^2}{16C^2T\log(1/\delta)}.$$

Therefore, using the assumption $\varepsilon \leq \log(1/\delta)$, Algorithm 1 satisfies $(\varepsilon_\star, \delta)$-DP where

$$\varepsilon_\star \leq \frac{\varepsilon^2}{8\log(1/\delta)} + \frac{\varepsilon}{\sqrt{2}} \leq \varepsilon.$$

This concludes the proof of the privacy claim.

Upper bound—PL case. Plugging the expressions of σ_{cdp}^2 and σ_{cor}^2 in the PL bound of Theorem 2 and rearranging terms yields

$$\begin{aligned}
\mathbb{E}\,\mathcal{L}(\bar{\mathbf{x}}^{(T)}) - \mathcal{L}_\star &= \mathcal{O}\Bigg(\frac{L}{\mu^2 T}\frac{\sigma_\star^2 + d\frac{C^2T\log(1/\delta)}{(n-q)\varepsilon^2}}{n} + \frac{c^2L^2(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)}{\mu^2p^2T^2} + \frac{cL^3}{\mu^2p^2T^2}\Xi_0 \\
&\quad + \frac{\ln T}{T^2}\frac{L^2}{\mu^3 p}\left[(1-p)(\frac{\zeta_\star^2}{p} + \sigma_\star^2) + \frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\frac{C^2T\log(1/\delta)}{a(\mathcal{G}_{\mathcal{H}})\varepsilon^2}}{n} + \left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\frac{C^2T\log(1/\delta)}{(n-q)\varepsilon^2}\right]\Bigg) \\
&= \widetilde{\mathcal{O}}\Bigg(\frac{LC^2d\log(1/\delta)}{\mu^2 n(n-q)\varepsilon^2} + \frac{L}{\mu^2 nT}\left[\sigma_\star^2 + \frac{LC^2d\log(1/\delta)}{\mu p\varepsilon^2}\left(\frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|}{a(\mathcal{G}_{\mathcal{H}})} + \frac{n}{n-q}\left\|\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2\right)\right] \\
&\qquad + \frac{L^2}{\mu^3p^2T^2}\left[(1-p)(\zeta_\star^2 + p\sigma_\star^2) + c^2\mu(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star) + c\mu L\Xi_0\right]\Bigg).
\end{aligned}$$

Upper bound—Non-convex case. Plugging the expressions of σ_{cdp}^2 and σ_{cor}^2 in the non-convex bound of Theorem 2 and rearranging terms yields

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\mathcal{L}(\bar{\mathbf{x}}^{(t)})\right\|_2^2 &= \mathcal{O}\Bigg(\sqrt{\frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)(\sigma_\star^2 + d\sigma_{\mathrm{cdp}}^2)}{nT}} + \frac{cL(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)}{pT} + \frac{L^2}{pT}\Xi_0 \\
&\quad + \frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)n}{pT(\sigma_\star^2 + d\sigma_{\mathrm{cdp}}^2)}\left(6(1-p)\frac{\zeta_\star^2}{p} + (1-p)\sigma_\star^2 + \frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\sigma_{\mathrm{cor}}^2}{n} + \left\|\mathbf{W} - \tfrac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\sigma_{\mathrm{cdp}}^2\right)\Bigg) \\
&= \mathcal{O}\Bigg(\sqrt{\frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)(\sigma_\star^2 + d\frac{C^2T\log(1/\delta)}{(n-q)\varepsilon^2})}{nT}} + \frac{cL(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)}{pT} + \frac{L^2}{pT}\Xi_0 \\
&\quad + \frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)n}{pT(\sigma_\star^2 + d\frac{C^2T\log(1/\delta)}{(n-q)\varepsilon^2})}\left((1-p)(\frac{\zeta_\star^2}{p} + \sigma_\star^2) + \frac{\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|\,d\frac{C^2T\log(1/\delta)}{a(\mathcal{G}_{\mathcal{H}})\varepsilon^2}}{n} + \left\|\mathbf{W} - \tfrac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2 d\frac{C^2T\log(1/\delta)}{(n-q)\varepsilon^2}\right)\Bigg) \\
&= \mathcal{O}\Bigg(\frac{C\sqrt{d\log(1/\delta)}}{\sqrt{n(n-q)}\varepsilon} + \sqrt{\frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)\sigma_\star^2}{nT}} + \frac{cL(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)}{pT} + \frac{L^2}{pT}\Xi_0 \\
&\quad + \frac{L(\mathcal{L}(\bar{\mathbf{x}}^{(0)}) - \mathcal{L}_\star)n}{pT}\left((1-p)(\frac{\zeta_\star^2}{p\sigma_\star^2} + 1) + \frac{(n-q)\mathrm{H}_{\mathcal{G}}(\mathbf{W})\,|\mathcal{E}|}{na_q(\mathcal{G})} + \left\|\mathbf{W} - \tfrac{\mathbf{1}\mathbf{1}^\top}{n}\right\|_F^2\right)\Bigg).
\end{aligned}$$

We conclude by ignoring higher-order terms in T: in $\frac{1}{T^2}$ for the PL case and $\frac{1}{T}$ for the non-convex case. □

In the PL case, observe that our privacy-utility trade-off matches CDP whenever there is at most a constant fraction of colluding user, i.e., the level of collusion is $q = \mathcal{O}(n)$. In the extreme scenario where almost all users are colluding, i.e., $n - q = \mathcal{O}(1)$, then the trade-off matches LDP only, which cannot be improved in general when $q = n - 1$ (Duchi et al., 2018). In the non-convex case, while it is not possible to discuss the tightness of our privacy-utility trade-off because lower bounds on the CDP trade-off are unknown, the error $\mathcal{O}\left(\frac{\sqrt{d}}{n\varepsilon}\right)$ matches the CDP baseline error without variance reduction (Arora et al., 2022).

D. Detailed Experimental Setup

In this section, we provide the full experimental setup of our empirical evaluation in Section 6.

Datasets. We conduct our evaluation on three datasets: synthetic data for least-squares regression, *a9a* LibSVM (Chang & Lin, 2011) and MNIST (LeCun & Cortes, 2010), that we distribute among $n = 16$ users, as explained in Section 6.

Privacy parameters. We consider user-level privacy for the first two tasks, and example-level privacy for the last task. For all our experiments, we set the privacy parameter δ to 10^{-5}, this ensures that $\delta \ll \frac{1}{nm} \leq \frac{1}{n}$.

D.1. Privacy Noise Parameters Search for DECOR

For a pre-specified SecLDP privacy budget ε, we would like to find a corresponding couple of privacy noises $(\sigma_{\text{cdp}}, \sigma_{\text{cor}})$ to be used in DECOR. However, Algorithm 2 does the reverse process, i.e., it computes the per-step SecRDP budget, denoted $\varepsilon_{\text{iter}}^{\text{RDP}}$ here, given the privacy noise couple $(\sigma_{\text{cdp}}, \sigma_{\text{cor}})$. Moreover, it is straightforward to obtain the desired per-step RDP budget $\varepsilon_{\text{iter}}^{\text{RDP}}$ given the full DP budget ε using composition and conversion properties of RDP (Mironov, 2017). Hence, we only need to search for $(\sigma_{\text{cdp}}, \sigma_{\text{cor}})$, given a pre-specified $\varepsilon_{\text{iter}}^{\text{RDP}}$. To do so, we fix σ_{cdp}, and we look for the other parameter σ_{cor}, using *binary search*, since the function $\varepsilon_{\text{iter}}^{\text{RDP}}(\sigma_{\text{cor}})$ is monotonous (non-increasing), as shown in Figure 2. Specifically, we use the following steps in our search:

1. Given the global (user-level) SecLDP privacy budget ε, we determine the per-step SecRDP privacy budget $\varepsilon_{\text{iter}}^{\text{RDP}}$ using the RDP composition and conversion properties

2. We know that the uncorrelated noise variance σ_{cdp} is bounded between the privacy noise variance used for the baseline CDP algorithm $\frac{C\sqrt{2}}{\sqrt{n\varepsilon_{\text{iter}}^{\text{RDP}}}}$, and the one used for the LDP baseline, that is $\frac{C\sqrt{2}}{\sqrt{\varepsilon_{\text{iter}}^{\text{RDP}}}}$. So we start by fixing σ_{cdp} in the interval $[\frac{C\sqrt{2}}{\sqrt{n\varepsilon_{\text{iter}}^{\text{RDP}}}}, \frac{C\sqrt{2}}{\sqrt{\varepsilon_{\text{iter}}^{\text{RDP}}}}]$

3. For every fixed σ_{cdp}, we search for the corresponding σ_{cor} in a sufficiently large interval ($[1, 10^3]$ in our experiments) using binary search on the outputs of our SecRDP accountant (Algorithm 2).

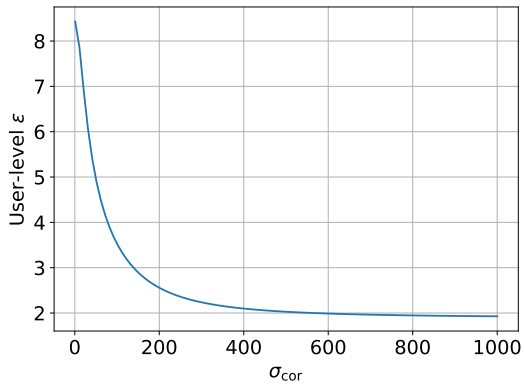


Figure 2. User-level SecLDP privacy budget ε, using Algorithm 2, as a function of σ_{cor} given a fixed σ_{cdp} in the center of the search interval, a total number of iterations $T = 1000$ and a clipping threshold $C = 1$.

Example-level privacy. The procedure to get the privacy noise parameters is slightly different for example-level privacy. Indeed, we use RDP privacy amplification by subsampling (Wang et al., 2019) after using Algorithm 2. However, the RDP privacy amplification by subsampling does not have a closed-form expression, so we cannot directly get the desired per-step SecRDP budget from the full DP budget ε. Therefore, we again fix σ_{cdp} in a grid, this time in $[\frac{C}{1000}, \frac{C}{20}]$, and we look for the other parameter σ_{cor} (this time in $[\frac{C}{2000}, \frac{C}{10}]$) using *binary search*, since the function $\varepsilon_{\text{iter}}^{\text{RDP}}(\sigma_{\text{cor}})$ is monotonous (non-increasing), as shown in Figure 3.

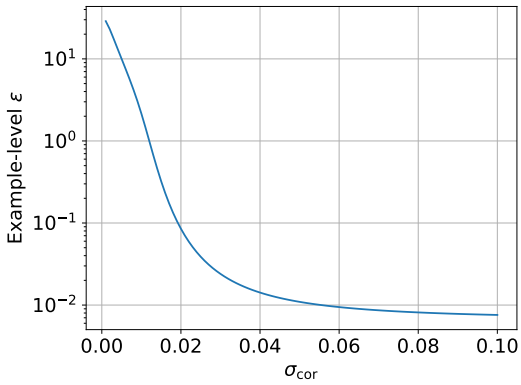


Figure 3. Example-level SecLDP privacy budget ε, using Algorithm 2 and RDP amplification by subsampling (Wang et al., 2019), as function of σ_{cor} given a fixed $\sigma_{\text{cdp}} = \frac{5C}{1000}$, a total number of iterations $T = 1000$, clipping threshold $C = 1$ and batch size 64.

D.2. Hyperparameter Tuning

For all considered tasks, we tune the hyperparameters of each algorithm individually, following the same steps, to obtain: the learning rate η, the clipping threshold C and the noise parameters σ_{cdp} and σ_{cor}. It is important to note that the couple of privacy noise parameters $(\sigma_{\text{cdp}}, \sigma_{\text{cor}})$ is not unique: we can find many couples that yield the same SecRDP budget, which is also visible in the theoretical bound from Theorem 1. However, in the CDP and LDP baselines (D-SGD with uncorrelated privacy noise), they are determined uniquely by the RDP guarantee for the Gaussian mechanism (Mironov, 2017).

For our tuning, we choose a grid of learning rates and clipping thresholds. First, we simply evaluate the CDP and LDP baselines with the desired topology on all the learning rate and clipping couples (η, C), and then we pick the best hyperparameter couple at the end. For DECOR, we do the same procedure for (η, C). However, there are many possible noise couples $(\sigma_{\text{cdp}}, \sigma_{\text{cor}})$ following the privacy noise search in the previous section, we choose three among them that yield the same privacy budget: the one with the lowest σ_{cdp} (first couple found by binary search), the largest σ_{cdp} (last couple) and the one in the middle. After evaluating these noises with every couple (η, C), we choose at the end the best quadruplet of hyperparameters.