Provable Mutual Benefits from Federated Learning in Privacy-Sensitive Domains

Nikita Tsoy
INSAIT, Sofia University
Sofia, Bulgaria

Anna Mihalkova
ETH Zurich
Zurich, Switzerland

Teodora Todorova
High School of Mathematics
Burgas, Bulgaria

Nikola Konstantinov
INSAIT, Sofia University
Sofia, Bulgaria

Abstract

Cross-silo federated learning (FL) allows data owners to train accurate machine learning models by benefiting from each others private datasets. Unfortunately, the model accuracy benefits of collaboration are often undermined by privacy defenses. Therefore, to incentivize client participation in privacy-sensitive domains, a FL protocol should strike a delicate balance between privacy guarantees and end-model accuracy. In this paper, we study the question of when and how a server could design a FL protocol provably beneficial for all participants. First, we provide necessary and sufficient conditions for the existence of mutually beneficial protocols in the context of mean estimation and convex stochastic optimization. We also derive protocols that maximize the total clients' utility, given symmetric privacy preferences. Finally, we design protocols maximizing end-model accuracy and demonstrate their benefits in synthetic experiments.

1 INTRODUCTION

The fundamental reliance of modern machine learning algorithms on diverse, high-quality data has recently led to a surge of interest in collaborative learning techniques. Among these, federated learning (FL) stands out because it enables collaborative training in a fully distributed manner by exchanging gradient updates via a central server (McMahan et al., 2017; Konečný et al., 2016; Kairouz et al., 2021). However, despite the distributed nature of these protocols, recent work has shown that the communicated gradients are often enough to reconstruct sensitive information about the participants (Geiping et al., 2020; Hu et al., 2022; Zhu et al., 2019).

These vulnerabilities emphasize the importance of privacy protection techniques in FL, e.g., in terms of differential privacy (DP) (Dwork and Roth, 2014). Such defenses usually add a certain level of noise to the communicated messages, with higher noise levels leading to stronger privacy guarantees. However, this noise often harms the end-model performance, potentially undermining all benefits of collaboration (Bassily et al., 2014; Abadi et al., 2016).

Such a delicate balance between model accuracy and data privacy brings into question the viability of FL, especially for entities managing sensitive data. While such entities can improve their services using FL, they also seek to provide strong privacy guarantees to their clients. Therefore, they may be unwilling to participate in a FL protocol that does not match their preferences in terms of the accuracy-privacy trade-off.

Contributions We address the question of whether and how FL can be made *provably mutually beneficial to all its participants*. To this end, we consider a formal framework that models the clients' privacy and accuracy preferences and the server's objectives. We provide necessary and sufficient conditions for the existence of mutually beneficial FL protocols in the context of two classic learning tasks: mean estimation and strongly convex stochastic optimization. Our results cover both DP and a privacy notion based on a local data reconstruction loss.

We also study the question of how the server can maximize two natural objectives, the total client utility or the end-model accuracy, using mutually beneficial protocols. For the first objective, in the case of symmetric client utilities, we provide rich theoretical descriptions of the optimal noise levels and the corresponding utility benefits from collaboration. For the second objective, we provide simulations on synthetic data, which

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

demonstrate the benefits of tailoring the FL protocol to the client incentives.

2 RELATED WORK

Accuracy-Privacy Trade-off in FL Privacy is a major concern in FL (Kairouz et al., 2021) and differential privacy mechanisms are popular tools for addressing this problem (Abadi et al., 2016; Ponomareva et al., 2023). However, often privacy protection decreases the resulting accuracy of the model. Thus, numerous works study the accuracy-privacy trade-off. In particular, Bassily et al. (2019) and Feldman et al. (2020) study the optimal rates for stochastic optimization given a fixed level of privacy protection, and Asi et al. (2022) and Cai et al. (2021) derive the min-max optimal rates for private mean estimation.

Our work diverges from these studies in three significant ways. First, our work incorporates the participation constraints of the clients (Section 4). Second, we provide a theoretically grounded methodology to pick up points from the accuracy-privacy frontier (Sections 3, 5, and 6). Third, our approach accounts for differences in privacy preferences and allows the server to personalize privacy protections as long as all clients benefit in utility (Section 3).

Privacy-Related Incentives in FL Several recent works consider privacy-related incentives in federated learning. In particular, Shen et al. (2023) consider how to move closer to the accuracy-privacy Pareto front under statistical client heterogeneity. However, they do not model heterogeneity in the client privacy preferences and do not provide tools to find the right level of privacy protection. Huang et al. (2022) study ways of incentivizing clients to contribute data to an FL protocol. However, they do not study how to construct an optimal protocol for the accuracy-privacy trade-off, focusing on incentivization schemes given a fixed learning protocol. Similarly to our work, Ghosh and Roth (2011) and Anjarlekar et al. (2023) also study how to construct a privacy protocol beneficial for the server given participation constraints. However, they only consider monetary payments as a way to compensate the clients, while we consider compensations based on end-model accuracy, which seems more fitting for the case of collaborative learning.

Other Incentives in Federated Learning Other incentives, apart from privacy, have also been considered in the context of FL. In particular, recent work studies incentives to free-ride and corresponding ways to design fair payment allocation schemes for the participants (Lyu et al., 2020; Donahue and Kleinberg, 2021; Blum et al., 2021; Karimireddy et al., 2022). Another lines of work studied incentives stemming from downstream competition (Tsoy and Konstantinov, 2023; Dorner et al., 2023), as well as ML games for delegated data collection Saig et al. (2023); Ananthakrishnan et al. (2023) and fine-tuning foundation models Laufer et al. (2023).

Also related is the field of data valuation (Wang et al., 2020; Pei, 2022; Rozemberczki et al., 2022). A key conceptual difference to our work is that such studies seek to quantify the usefulness of data relative to the other players, e.g., by evaluating the amount of informativeness or redundancy of a data source given the others' data. In contrast, we study the accuracy improvements with all clients' data and how it is affected by privacy defenses.

3 GENERAL FRAMEWORK

Our work analyzes a fundamental accuracy-privacy trade-off in federated learning from a multi-objective perspective. We consider a standard federated learning setup, where N potential participants with local datasets $(D_i)_{i=1}^N$ may train a model together. While collaboration may improve end-model accuracy, it also increases privacy risk. Therefore, the clients will join the protocol if and only if they expect a substantial increase in model's accuracy from collaboration. To study this trade-off, we propose a general framework that consists of four parts: the federated learning protocol, the clients' evaluations of the protocol, the clients' utility functions, and the server's objective.

3.1 Federated Learning Protocol

FL protocol p describes all details of the server-client interactions, including the distributed learning algorithm and privacy protection mechanism. It details all training parameters, such as optimization step count, batch size, and learning rate, and determines (possibly client-specific) privacy protection parameters, such as privacy-preserving noise levels, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)$. Our work focuses on adjusting privacy-preserving noise levels, $\boldsymbol{\alpha}$, since they are the most important in the accuracy-privacy trade-off. Given the noise levels, we optimize the remaining parameters for the client utilities. We denote the resulting protocol as $p_{\boldsymbol{\alpha}}$.

3.2 Clients' Evaluations

We assume that the clients could reason about the proposed protocol by estimating the end model error and training procedure privacy risk. Formally, we introduce two sets of functions: $(\texttt{err}_i)_{i=1}^N$ and $(\texttt{leak}_i)_{i=1}^N$.

$\texttt{err}_i \colon P \to [0,\infty)$ describes client i's evaluation of the end-model error given protocol p. Since final error is unknown before training, the clients may use theoretical guarantees to reason about it, such as *expected* values or *high probability upper bounds*. For example, in the case of collaborative mean estimation, the clients might use the expected mean squared error as an $\texttt{err}$ function.

$\texttt{leak}_i \colon P \to [0,\infty)$ describes client i's evaluation of the privacy violation given protocol p. Since privacy leaks are hard to predict, the client will again reason via theoretical guarantees. For example, the clients might quantify privacy protection using the popular notion of *differential privacy* (Dwork and Roth, 2014).

Definition 3.1. A randomized algorithm $A \colon \mathcal{D} \to \mathcal{S}$ is (ε,δ)-differentially private if, for all neighbouring datasets D and D' (i.e. datasets differing in one point),

$$\forall S \subseteq \mathcal{S}\ \Pr(A(D) \in S) \le \mathrm{e}^{\varepsilon} \Pr(A(D') \in S) + \delta.$$

DP requires that the outputs of an algorithm are almost indistinguishable, regardless of whether a certain individual data point included or not into the dataset. Intuitively, ϵ measures how many bits of information the algorithm's outputs reveal about the input, while δ represents the probability of such guarantees failing to hold. Therefore, it is a common practice to seek $\delta \ll 1/|D|$ (Ponomareva et al., 2023). In our work, we consider algorithms that ensure $\delta = \mathrm{O}(1/|D|^2)$.

3.3 Clients' Utilities and Rationality

We further assume that the clients use the evaluations from the previous section to reason about the benefits of a protocol. Formally, client i gets utility $u_i(\texttt{err}, \texttt{leak})$ from a protocol with estimated accuracy loss $\texttt{err}$ and privacy loss $\texttt{leak}$. To derive quantitative results about the accuracy-privacy trade-off, we introduce a specific form of this utility function. We also present results about general utility functions in Section 4.4.

Parametric Utility Form In our quantitative analysis, we assume the following form of utility:

$$u_i(\texttt{err}, \texttt{leak}) = u_i^{p,\lambda_i}(\texttt{err}, \texttt{leak}) := -\texttt{leak}^p - \lambda_i \texttt{err}^p,$$

which corresponds to a negative weighted l_p-norm of the vector $(\texttt{err}, \texttt{leak})$. Here, parameter λ_i measures the importance of accuracy for participant i and is responsible for the correct rescaling of $\texttt{err}$ and $\texttt{leak}$ units. Notice that this utility is monotonically decreasing and continuous in both $\texttt{err}$ and $\texttt{leak}$, which reflects the preferences of the clients towards better accuracy and privacy guarantees.

To motivate this parametric form, we use the classic result from the multi-objective optimization literature. Since the clients benefit from the protocols with low accuracy error and privacy leakage, we can treat the minimization of these objectives as a multi-objective optimization problem. To describe its Pareto front, one may analyze the linear combinations of the objectives, $F^{\mathrm{lin}} := \sum_{i=1}^N w_i^{\texttt{leak}} \texttt{leak}_i^p + w_i^{\texttt{err}} \texttt{err}_i^p$, since the minimizers of F^{lin} lie on the Pareto front (Emmerich and Deutz, 2018). However, since

$$F^{\mathrm{lin}} = \sum_{i=1}^N w_i^{\texttt{leak}} (\texttt{leak}_i^p + \lambda_i \texttt{err}_i^p) =$$

$$\sum_{i=1}^N -w_i^{\texttt{leak}} u_i^{p,\lambda_i}(\texttt{err}_i, \texttt{leak}_i), \text{ where } \lambda_i := \frac{w_i^{\texttt{err}}}{w_i^{\texttt{leak}}},$$

the problem of finding a Pareto frontier protocol will naturally lead to the problem of optimizing the utility functions of our parametric family.

In our work, we will focus on the family u_i^{2,λ_i}, as one the most natural choices among l_p-norms. To evaluate the robustness of our results to the choice of utility, we also discuss a different utility function in Appendix C.

Participation Constraints Based on the privacy, accuracy and utility analysis of the protocol, the clients can reason about the benefitis of joining it, compared to the an outside options for training (such as training a model alone). This creates contraints on what protocols are feasible in the first place. In general, additional contraints may reflect not only the individual desires of clients, but also possible legal or fairness requirements imposed by other entities (e.g. governments or the server itself).

In this work, we focus on constraints coming from the rationality of the clients. We therefore require protocols to be *mutually beneficial* as a natural notion of feasibility.

Definition 3.2. A protocol p_α is called mutually beneficial if for any player i, $u_i(\texttt{err}_i(p), \texttt{leak}_i(p)) \ge u_i^0$, where u_i^0 is the utility from local training.

Mutually beneficial protocol ensures that all clients will benefit from collaborative learning compared to individual learning. Therefore, it is rational for them to participate. We usually consider $u_i^0 = u^i(\texttt{err}_i^0, 0)$, where $\texttt{err}_i^0$ is the accuracy loss of local model. We study feasibility of this constraint in Section 4.

3.4 Server's Objectives

Since there might be several protocols that will satisfy participation constraints, the server may seek additional desirable properties. In our paper, we will study two possible objectives of the server. The first one is *maximization of the total utility* of all participants (Section 5). We study this objective for the case of symmetric client utilities and provide a precise theoretical description the optimal protocols. The second objective is the *maximization of the server model accuracy* (Section 6). We study this objective empirically and show how personalization of privacy protection can improve the accuracy of the global model.

4 FEASIBILITY OF COLLABORATION

We now focus on the existence of mutually-beneficial protocols. We quantitatively analyze two canonical learning tasks: mean estimation (Section 4.1) and strongly convex stochastic optimization (Section 4.2). First, we use DP as a privacy notion in both setups. To evaluate the robustness of our findings to the choice of privacy notion, we study also mean estimation with a privacy notion based on reconstruction loss in Section 4.3. Finally, we generalize our findings in the limit case as the number of clients grows to infinity to general utility functions in Section 4.4.

4.1 Feasibility of DP Mean Estimation

Setup We begin the quantitative exploration of our framework with the problem of DP mean estimation.

Formally, we assume that the clients are interested in parameter μ and have noisy observations of this parameter. Entity i's noisy observations of this parameter, $(x_i^j)_{j=1}^n$, satisfy

$$\mathbf{E}(x_i^j) = \mu,\ \mathrm{Var}(x_i^j) = \sigma^2,\ \mathrm{supp}\, x_i^j \subseteq \mu + [-B/2, B/2].$$

Let $\bar{x}_i \coloneqq \sum_{j=1}^n x_i^j / n$ be a local average of all samples. The server will choose among the protocols $p_{\boldsymbol{\alpha}}$ in which entity i reveals the following noisy message to other participants $m_i = \bar{x}_i + \epsilon_i$, where $\epsilon_i \sim \mathrm{N}(0, \alpha_i^2)$.

Accuracy Loss The entities' accuracy loss will be the root of the mean squared error

$$\mathtt{err}_i = \sqrt{\mathbf{E}_{\{D_i\}_{i=1}^N, \{\epsilon_i\}_{i=1}^N}((\hat{\mu}_i - \mu)^2)},$$

where $\hat{\mu}_i$ is the best unbiased linear predictor of μ that can be constructed from $D_i \cup \{m_1, \dots, m_N\}$.

Theorem 4.1 (Proof in Appendix B.1)**.** *$\hat{\mu}_i$ has the following property*

$$\mathbf{E}((\hat{\mu}_i - \mu)^2) = \frac{1}{\gamma_i + \rho},$$

where $\rho \coloneqq \frac{n}{\sigma^2}$, $\beta_i \coloneqq \frac{1}{\frac{1}{\rho} + \alpha_i^2}$, $\gamma_i \coloneqq \sum_{k \neq i} \beta_k$.

As we can see, the accuracy for client i will depend on the "informativeness" of the messages of other clients, $\beta_k, k \neq i$. "Informativeness" β_k is decreasing in noise α_k and is bounded by the "informativeness" of the noiseless message, ρ.

Privacy Loss We will use the standard result for DP of the Gaussian mechanism (Dwork and Roth, 2014, see Theorem A.1 in Appendix A), which gives the following bound for DP budget:

$$\varepsilon_i \leq \frac{\kappa}{\alpha_i},\ \text{where } \kappa = \frac{\sqrt{2\ln(1.25n^2)}B}{n}.$$

Here, κ describes the strength of privacy concerns. This strength is small when the observations of each client are similar (small B) or when their message is close to the true mean of the distribution (big n).

The entities will use this upper bound to reason about their privacy loss, $\mathtt{leak}_i = {}^{\kappa}/_{\alpha_i}$. We assume that the clients will use this upper bound even if $\varepsilon_i \geq 1$ and Theorem A.1 does not hold.

Participation Incentives These choices result in utility function

$$u_i = -\mathtt{leak}_i^2 - \lambda_i \mathtt{err}_i^2 = -\frac{\kappa^2}{\alpha_i^2} - \frac{\lambda_i}{\gamma_i + \rho}.$$

To find the utility of training alone, we consider limit $\forall k\ \alpha_k \to \infty$ when clients use only their local data. This limit gives $u_i^0 = -{}^{\lambda_i}/_{\rho}$. Therefore, the protocol is mutually beneficial when

$$\forall i\ u_i \geq u_i^0 \iff -\frac{\kappa^2 \rho \beta_i}{\rho - \beta_i} + \frac{\lambda_i \gamma_i}{\rho(\gamma_i + \rho)} \geq 0. \tag{1}$$

Now, we provide a necessary and sufficient condition for the existence of parameters β_k (and corresponding α_k) that yield a mutually beneficial protocol.

Theorem 4.2 (Proof in Appendix B.2)**.** *Equation* (1) *has solutions if and only if*

$$\sum_{i=1}^N \zeta_i > 1,\ \textit{where } \zeta_i \coloneqq \frac{\lambda_i}{\lambda_i + \kappa^2 \rho^2}.$$

Moreover, if the system has solutions, it has a series of solutions in this form,

$$\beta_i = \zeta_i b, b \in \left[0, \left(1 - \frac{1}{\sum_{i=1}^N \zeta_i}\right)\rho\right].$$

To interpret these results, we make two key observations. First, Equation (1) implies $\rho \alpha_i^2 \geq \frac{\kappa^2 \rho^2}{\lambda_i}$. Thus, ratio ${}^{\rho^2 \kappa^2}/_{\lambda_i}$ describes the minimally acceptable level of

DP noise for client i. This level of noise "minimally" trades off the privacy concerns (described by κ) for accuracy gains (bounded by $^1/_\rho$). The existence condition in Theorem 4.2 requires that preference towards accuracy λ_i is sufficiently large so that this minimal noise level $^{\rho^2\kappa^2}/_{\lambda_i}$ is small enough to allow accuracy benefits for everyone. Second, if all other parameters are fixed and λ_i are bounded from below, the sum in Theorem 4.2 diverges: collaboration becomes mutually beneficial. This property shows that the accuracy benefits will outweigh the potential privacy concerns as the number of players grows.

4.2 Feasibility of DP Stochastic Optimization

Setup Now, we analyze DP stochastic optimization.

Formally, we assume that the clients are interested in parameters $\boldsymbol{w}$ that minimize strongly convex objective function f on the closed and convex set W, such that $f^* := \min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}) = \min_{\boldsymbol{w}\in W} f(\boldsymbol{w})$ and $\operatorname{diam}(W) \le D$. Each entity i has access to gradient oracle $\boldsymbol{g}_i$, which allows it to use its data points to get the unbiased estimate of the gradient of the objective

$$\begin{aligned}
&\forall \boldsymbol{w}\ \nabla f(\boldsymbol{w}) = \mathbf{E}_{x_i^j}(\boldsymbol{g}_i(\boldsymbol{w}, x_i^j)),\\
&\mathbf{E}(\|\nabla f(\boldsymbol{w}) - \boldsymbol{g}_i(\boldsymbol{w}, x_i^j)\|^2) \le \sigma^2,\\
&\forall \boldsymbol{w}\ \operatorname{supp} \boldsymbol{g}_i(\boldsymbol{w}, \cdot) \subseteq \nabla f(\boldsymbol{w}) + \mathrm{B}(B/2),
\end{aligned}$$

where $\mathrm{B}(R) := \{\boldsymbol{y} \mid \|\boldsymbol{y}\| \le R\}$. We assume that the objective function is L-smooth on $\mathbb{R}^d$ and μ-strongly convex on W (see Appendix A for the definitions of smoothness and strong convexity).

The server will communicate with the clients for m rounds. At each round, the server gets the batch mean of local gradient estimates of the clients. The clients use $b = \lfloor ^n/_m \rfloor$ points per batch selected randomly without replacement. The server then optimally aggregates the gradient estimates and updates the global parameters (Algorithm 1). In this algorithm, we do not consider the additional randomization over the choice of clients because we assume that the server will be able to determine who participated in each round of computation. Moreover, we will pass through the data only once because even one pass through the data guarantees the optimal asymptotic behavior of the training objective.

Accuracy Loss For fixed noise levels $\boldsymbol{\alpha}$, one can show the following upper bound on the distance to optimal parameter $\boldsymbol{w}^*$. Here, aggregation weights $(a_i^*)_{i=1}^N$ and step sizes $(\eta_*^t)_{t=1}^m$ are chosen to minimize the bound.

Theorem 4.3 (Proof in Appendix B.4). *Optimization*

Algorithm 1 Collaborative Learning Protocol

Input: protocol $p_{\boldsymbol{\alpha}}$
Server randomly initializes $\boldsymbol{w}^0 \in W$
for $i = 1$ **to** N **do**
  Client i randomly chooses $\pi_i\colon [m] \to [m]$
end for
for $t = 1$ **to** m **do**
  for $i = 1$ **to** N **do**
    Client i samples $\boldsymbol{\xi}_i^t \sim \mathrm{N}(\mathbf{0}, \alpha_i^2 \mathbf{I})$
    Client i calculates

$$\boldsymbol{m}_i^t = \boldsymbol{\xi}_i^t + \frac{1}{b}\sum_{j=1}^{b} \boldsymbol{g}_i\Big(\boldsymbol{w}^{t-1}, x_i^{(\pi_i(t)-1)b+j}\Big)$$

    Client i sends $\boldsymbol{m}_i^t$ to the server
  end for
  Server computes $\boldsymbol{g}^t = \sum_{i=1}^N a_i \boldsymbol{m}_i^t$
  Server updates $\boldsymbol{w}^t = \mathit{\Pi}_W(\boldsymbol{w}^{t-1} - \eta^t \boldsymbol{g}^t)$
end for

outcome $\Delta\boldsymbol{w}^m := \mathbf{E}(\|\boldsymbol{w}^m - \boldsymbol{w}^*\|^2)$ *satisfies*

$$\Delta\boldsymbol{w}^m \le \begin{cases} \frac{1}{\left(1+\frac{\chi}{2-\chi}(m-T)\right)L\mu\mathit{\Gamma}}, & m \ge T,\\ \left(1-\frac{\chi}{2}\right)^m \frac{y_*^0}{L\mu\mathit{\Gamma}}, & m < T, \end{cases} =: \Delta\boldsymbol{w}^{m,ub},$$

where $T := \left\lceil -\frac{\ln(y_*^0)}{\ln\left(1-\frac{\chi}{2}\right)} \right\rceil$, $\chi := \frac{\mu}{L}$, $y_*^0 := \frac{L\mu n N D^2}{\sigma^2}$, $\mathit{\Gamma} := \sum_{i=1}^N \beta_i$, $\beta_i = \frac{1}{1/\rho + d\alpha_i^2}$, $\rho := {}^b/_{\sigma^2}$.

As we can see, we get a similar dependence to Theorem 4.1. However, in this case, the optimization outcome for client i will also deteriorate in their noise contribution α_i. Similarly to the previous case, we assume $\texttt{err}_i = \sqrt{\Delta\boldsymbol{w}^{m,\mathrm{ub}}}$.

Privacy Loss We use the result of Feldman et al. (2022), applied to the standard Gaussian mechanism (Dwork and Roth, 2014).

Theorem 4.4 (Proof in Appendix B.5). *If* $m \ge 2$ *and the server uses Algorithm* 1 *and chooses* $\varepsilon_{0,i} \in \left(0, \min\left(1, \ln\left(\frac{n}{16\ln(2/\delta)}\right)\right)\right)$, $\delta_0 = {}^1/_{mn^2}$, *and* $\delta = {}^1/_{n^2}$, *the client will have the local* $\left(\varepsilon_i, \frac{1+(e^{\varepsilon_i}+1)(1+e^{-\varepsilon_{0,i}/2})}{n^2}\right)$*-DP guarantee, where* ε_i *has the following property:*

$$\varepsilon_i \le \frac{16\sqrt{2e\ln(1.25mn^2)\ln(4n^2)}B}{\sqrt{m}b\alpha_i} =: \frac{\kappa}{\alpha_i}.$$

Here, κ behaves similarly to the mean estimation case but has an additional multiplier, $^1/_{\sqrt{m}}$, which comes from the shuffling of the data. Similarly to the previous section, we assume the clients use this bound to reason about privacy leak $\texttt{leak}_i = {}^\kappa/_{\alpha_i}$. (Here, we will again assume that the clients will ignore the restrictions on $\varepsilon_{0,i}$ and m.)

Participation Incentives To choose the optimal batch size b for training, we consider the behavior of $\texttt{err}_i$ and $\texttt{leak}_i$ when $m \to \infty$, $b \to \infty$, and $\alpha_i \approx 0$. In this limit, we get

$$\texttt{err}_i^2 \propto \frac{1}{m\Gamma} \propto \frac{1}{nN},\ \texttt{leak}_i^2 \propto \frac{\ln(n)^2}{b^2 m}$$

We see that the increase in batch size does not greatly affect the accuracy of estimates but strongly decreases the privacy leak. Thus, our analysis considers $m = T$, resulting in utility function

$$u_i = -\frac{\kappa^2}{\alpha_i^2} - \frac{\lambda_i}{L\mu\Gamma},$$

where $\kappa := \frac{16\sqrt{2e\ln(1.25Tn^2)\ln(4n^2)T}B}{n}$. We also ignore that T is an integer and assume that $T = \max\left(-\frac{\ln(y_*^0)}{\ln\left(1-\frac{\chi}{2}\right)}, 1\right)$, $y_*^0 = \frac{L\mu nND^2}{\sigma^2}$, $b = \frac{n}{T}$.

To model the utility of local training, u_i^0, we consider limit $\alpha_i \to 0$ and $\forall k \neq i\ \alpha_k \to \infty$ for the accuracy error and $\texttt{leak}_i \to 0$ for privacy leakage. This limit gives $u_i^0 = -\lambda_i/L\mu\rho$. Therefore, a protocol is mutually beneficial when

$$\forall i\ u_i \geq u_i^0 \iff \psi_i\left(1 - \frac{\rho}{\Gamma}\right) \geq \frac{\beta_i}{\rho - \beta_i}, \tag{2}$$

where $\psi_i := \frac{\lambda_i}{L\mu d\kappa^2\rho^2}$. The corresponding necessary and sufficient condition for the existence of a mutually beneficial protocol is the following.

Theorem 4.5 (Proof in Appendix B.6)**.** *Equation* (2) *has a solution if and only if the following inequality has a solution*

$$\sum_{i=1}^{N} \frac{\psi_i x}{(\psi_i + 1)x + 1} \geq x + 1,\ x \geq 0.$$

Corollary 4.6 (Proof in Appendix B.7)**.** *Equation* (2) *has a solution if* $\sum_{i=1}^N \frac{\psi_i}{\psi_i+2} \geq 2$ *and only if* $\sum_{i=1}^N \frac{\psi_i}{\sqrt{\psi_i+1}} \geq 4$.

Similarly to Theorem 4.2, $1/\psi_i$ describes the minimal level of noise for our problem. To make collaboration beneficial for everyone, we need these minimal noise levels to be small.

4.3 Feasibility of Bayesian Mean Estimation

Setup To evaluate the robustness of our findings to the choice of privacy notion, we also consider mean estimation with a different privacy loss.

Formally, we assume that the clients are interested in a parameter μ sampled from a prior distribution $\mathrm{N}(0, \frac{1}{\tau})$. Entity i's noisy observations of this parameter $(x_i^j)_{j=1}^n$ are sampled from $\mathrm{N}(\mu, \sigma^2)$. Let $\bar{x}_i := \frac{1}{n}\sum_{j=1}^n x_i^j$ be a local average of all local samples. The server will use protocols $p_{\boldsymbol{\alpha}}$, where entity i reveals noisy message $m_i := \bar{x}_i + \epsilon_i$, where $\epsilon_i \sim \mathrm{N}(0, \alpha_i^2)$, to others.

Accuracy Loss Similarly to the previous sections, the entities' accuracy loss will be the root of the mean squared error

$$\texttt{err}_i = \sqrt{\mathbf{E}_{\{D_i\}_{i=1}^N, \{\epsilon_i\}_{i=1}^N, \mu}((\hat{\mu}_i - \mu)^2)},$$

where $\hat{\mu}_i$ is the best predictor of μ that can be constructed from $D_i \cup \{m_1, \ldots, m_N\}$.

Theorem 4.7 (Proof in Appendix B.10)**.** $\hat{\mu}_i$ *satisfies*

$$\mathbf{E}((\hat{\mu}_i - \mu)^2) = \frac{1}{\gamma_i + \rho + \tau},$$

where $\rho := \frac{n}{\sigma^2}$, $\beta_i := \frac{1}{\frac{1}{\rho} + \alpha_i^2} \in [0, \rho]$, $\gamma_i := \sum_{k \neq i} \beta_k$.

Similarly to Theorem 4.1, the quality of the estimate for client i depends on the "informativeness" of the messages of the others.

Privacy Loss The entities' privacy loss will be the root of the negative mean squared error

$$\texttt{leak}_i = \sqrt{S - \mathbf{E}_{\{D_i\}_{i=1}^N, \{\epsilon_i\}_{i=1}^N, \mu}\left(\frac{1}{n}\sum_{j=1}^n (\hat{x}_i^j - x_i^j)^2\right)},$$

where $S := \mathbf{E}_{\{D_i\}_{i=1}^N, \{\epsilon_i\}_{i=1}^N, \mu}\left(\frac{1}{n}\sum_{j=1}^n (x_i^j)^2\right) = \sigma^2 + \frac{1}{\tau}$ is the data reconstruction error if the server uses a zero estimator for local data and $\hat{x}_i^j$ is the best predictor of x_i^j that can be constructed from $\{m_1, \ldots, m_N\}$.

Intuitively, this notion of privacy violation measures how well the private data points can be reconstructed from the revealed messages relative to a simple Bayesian estimate. We have the following formula for the amount of leakage given optimal reconstruction.

Theorem 4.8 (Proof in Appendix B.9)**.** $\{\hat{x}_i^j\}_{j=1}^n$ *have the following property*

$$\mathbf{E}\left(\sum_{j=1}^n (\hat{x}_i^j - x_i^j)^2\right) = \sigma^2(n-1) + n\left(\frac{1}{\alpha_i^2} + \frac{\rho(\gamma_i + \tau)}{\gamma_i + \rho + \tau}\right)^{-1}$$

As we can see, the privacy leak depends on two terms. Term $\frac{1}{\alpha_i^2}$ determines how close the client's message is to the true mean. Term $\frac{\rho(\gamma_i+\tau)}{\gamma_i+\rho+\tau}$ describes how well the messages of others estimate the true mean.

Participation Incentives The utility function will have the form

$$u_i = \Big(\frac{1}{\alpha_i^2} + \frac{\rho(\gamma_i+\tau)}{\gamma_i+\rho+\tau}\Big)^{-1} - \frac{1}{\rho} - \frac{1}{\tau} - \frac{\lambda_i}{\gamma_i+\rho+\tau}.$$

To compare this utility with the utility of training alone, we consider limit $\forall k\ \beta_k \to 0$ (i.e., $\alpha_k \to \infty$). This limit implies $\gamma_i \to 0$ and corresponds to utility $u_i^0 = -\frac{\lambda_i}{\rho+\tau}$. The clients agree to participate if the following constraint holds

$$\forall i\ \frac{(\rho-\beta_i)^2}{(\varGamma+\tau)\rho^2} - \frac{\beta_i}{\rho^2} - \frac{\lambda_i}{\gamma_i+\rho+\tau} \geq \frac{1}{\tau} - \frac{\lambda_i}{\rho+\tau}. \quad (3)$$

We will consider the behavior of this inequality in limit $\forall k\, \beta_k \to 0$. This limit corresponds to the small perturbation of the status quo case of non-collaboration or, alternatively, to the situation where participants have a minimal privacy loss.

Theorem 4.9 (Proof in Appendix B.11)**.** *A small beneficial deviation for everyone exists if and only if*

$$\forall i\ \xi_i \geq 0 \ \text{and} \ \sum_{i=1}^N \xi_i > 1,$$

where $\xi_i \coloneqq \left(1 + \frac{(\rho+\tau)^2}{\tau^2\rho^2\left(\frac{\lambda_i}{(\rho+\tau)^2} - \frac{1}{\tau^2}\right)}\right)^{-1}$.

Moreover, the proof of the theorem suggests that one possible beneficial deviation will have the following form $\beta_i = \xi_i b$, where b is sufficiently small.

To interpret the results, notice that, in the first order approximation, the following should hold

$$\xi_i\varGamma \geq \beta_i \iff \alpha_i^2 \geq \frac{1}{\xi_i\varGamma} - \frac{1}{\rho},\ \varGamma \coloneqq \sum_{i=1}^N \beta_i.$$

Since we consider the limit $\varGamma \to 0$, the first term will dominate the second one, and the value $1/\xi_i$ would again correspond to the "minimal level" of noise the entity should add compared to others. We want these "minimal levels" of noise to be small to make participation optimal for everyone.

4.4 Limit Results for General Utility

In all cases we analyzed, we observed that mutually beneficial protocols exist when the number of players $N \to \infty$. We now prove this for a general class of utility functions and accuracy and privacy losses. Consider any fixed protocol p_α^{sym} with the same noise parameter α for each player. We assume that $\texttt{err}_i$ and $\texttt{leak}_i$ as functions of N and α are the same across all participants. We also make the following minimal assumptions.

Assumption 4.10.
1. u_i is monotonically decreasing and continuous in both arguments and $u_i(0,0) > u_i(\texttt{err}, \texttt{leak}), \forall(\texttt{err}, \texttt{leak}) \neq (0,0)$.
2. $\texttt{err}_i(p_\alpha^{\text{sym}})$ decreases strictly and monotonically to 0 as $N \to \infty$ and $\texttt{leak}_i(p_\alpha^{\text{sym}})$ decreases strictly and monotonically to 0 as $\alpha \to \infty$.
3. $\texttt{err}_i^0 \geq \texttt{err}^0$ for all i for some $\texttt{err}^0 > 0$. Thus, no player can learn an arbitrary accurate model alone.

These natural assumptions describe the preferences of the clients towards improved accuracy and privacy and the fact that accuracy gains improve with the number of players. They also reflect the natural impact of noise on accuracy and privacy. We have the following result.

Theorem 4.11 (Proof in Appendix B.14)**.** *Assume that the utilities of all players belong to a finite set of possible utility functions U. Assume that every $u \in U$ and all functions $\texttt{err}_i$ and $\texttt{leak}_i$ satisfy Assumption 4.10. With all other parameters fixed, there exist values $N_1 \in \mathbb{N}$ and $\alpha \in (0,\infty)$, such that whenever $N \geq N_1$ players are available with utilities from U, setting $\alpha_i = \alpha$ for all players $i \in [N]$ ensures that the protocol p_α^{sym} is mutually beneficial.*

5 OPTIMAL PROTOCOLS FOR UTILITY

In this section, we consider how the server could optimize the utilities of all participants. To do this, we assume that the objective of the server is the linear combination of the clients utilities

$$\max_{p_{\boldsymbol{\alpha}}} F(p_{\boldsymbol{\alpha}}) \coloneqq \sum_{i=1}^N \nu_i u_i(p_{\boldsymbol{\alpha}}) \text{ s.t. } u_i(p_{\boldsymbol{\alpha}}) \geq u_i^0.$$

For simplicity, we consider the symmetric case $\forall i\ \lambda_i = \lambda, \alpha_i = \alpha$, which will imply the same utilities for all participants $\forall i\ u_i(p_{(\alpha,\dots,\alpha)^\intercal}) =: u^{\text{sym}}(\alpha)$. Setting $u^{0,\text{sym}} = \min_i u_i^0$ we arrive at the following objective

$$\max_{\alpha} F^{\text{sym}}(\alpha) \coloneqq u^{\text{sym}}(\alpha) \text{ s.t. } u^{\text{sym}}(\alpha) \geq u^{0,\text{sym}}.$$

DP Mean Estimation The next Theorem describes the choice of β (and therefore α) parameters that is optimal for the utility maximization problem.

Theorem 5.1 (Proof in Appendix B.3)**.** *If $(N-1)\lambda \leq \kappa^2\rho^2$, the collaboration becomes unprofitable, $\beta^* = 0$.*

If $(N-1)\lambda > \kappa^2\rho^2$, the collaboration becomes profitable, $(\alpha^*)^2 = \frac{N\kappa}{\sqrt{(N-1)\lambda} - \kappa\rho}$.

The optimal level of noise has the following properties

$$u_i(\alpha^*) - u_i^0 = \frac{(\sqrt{(N-1)\lambda} - \kappa\rho)^2}{N\rho}, \ \frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}\lambda} \leq 0,$$
$$\operatorname{sign} \frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}N} = \operatorname{sign}((N-2)\sqrt{\lambda} - 2\sqrt{N-1}\kappa\rho).$$

As we can see, collaboration becomes profitable if the number of clients N is large, the accuracy incentives λ are strong, the privacy concerns κ are small, or the variance of local estimates ρ is large. Naturally, the optimal noise level decrease when the accuracy incentives λ are large or the privacy concerns κ are small. At the same time, when the number of participants N is large, the optimal noise level increases, since sufficient accuracy gains are still present due to the large number of updates. We also discuss the dependence of α^* on other parameters the supplementary material.

Symmetric DP Stochastic Optimization We have

Theorem 5.2 (Proof in Appendix B.8)**.** *If* $\sqrt{(N-1)\lambda} < \sqrt{\frac{4NL\mu d}{N-1}}\kappa\rho$, *collaboration is unprofitable. If* $\sqrt{(N-1)\lambda} \geq \sqrt{\frac{4NL\mu d}{N-1}}\kappa\rho$, *collaboration is profitable with* $(\alpha^*)^2 = \frac{\sqrt{L\mu N}\kappa}{\sqrt{\lambda d}}$.

The optimal level of noise satisfies

$$\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}\lambda} \leq 0, \ \frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}\sigma^2} \leq 0, \ \frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}N} \geq 0,$$
$$u_i(\alpha^*) - u_i^0 = \frac{\sqrt{(N-1)\lambda}}{L\mu\rho N}\left(\sqrt{(N-1)\lambda} - \sqrt{\frac{4NL\mu d}{N-1}}\kappa\rho\right).$$

As we can see, the results are qualitatively similar to the results of Theorem 5.1.

Symmetric Bayesian Mean Estimation First, we give some necessary and sufficient conditions for the existence of non-trivial solution and some general insights about the optimal level of noise.

Theorem 5.3 (Proof in Appendix B.12)**.** *If* $\lambda < \frac{(N\rho+\tau)^2}{(N-1)^2\rho^2}$, *it is not profitable to collaborate,* $\beta^* = 0$.

If $\lambda > \frac{(N\rho^2+2\rho\tau+\tau^2)(\rho+\tau)^2}{(N-1)\rho^2\tau^2}$, *it is always profitable to collaborative,* $\beta^* > 0$. *If also* $\lambda < \frac{(N\rho+\tau)^2}{(N-1)\rho^2}$, β^* *will be the only solution of equation* $\frac{\mathrm{d}u_i}{\mathrm{d}\beta} = 0$ *on the interval* $[0, \rho]$ *and will have the following properties*

$$\frac{\mathrm{d}\beta^*}{\mathrm{d}\lambda} \geq 0, \ \frac{\mathrm{d}\beta^*}{\mathrm{d}\rho} \geq 0.$$

If $\lambda > \frac{(N\rho^2+2\rho\tau+\tau^2)(\rho+\tau)^2}{(N-1)\rho^2\tau^2}$ *and* $\lambda \geq \frac{(N\rho+\tau)^2}{(N-1)\rho^2}$, *collaboration will be profitable and moreover nobody will add noise to their messages,* $\beta^* = \rho$.

As we can see, when λ is small, collaboration becomes unprofitable for everybody. On the other hand, when λ is big, people would want to collaborate and moreover will not want to trade-off accuracy gain for privacy. As expected, an increase in accuracy concerns incentivize people to send less noise to the server. Interestingly, an increase in informativeness of local estimate (increase in ρ) also makes their messages more informative.

Now, we will look at the limit behavior of the solution when $N \to \infty$, $\rho \to \infty$, and $\rho \to 0$.

Theorem 5.4 (Proof in Appendix B.13)**.** *When* $\lambda > \frac{\rho+\tau}{\tau}$, *in the limit* $N \to \infty$, *it is profitable to collaborate,* $\beta^* = \sqrt{\frac{\lambda-1}{N}}\rho + \mathrm{o}\left(\frac{1}{\sqrt{N}}\right)$,

$$u_i(\beta^*) - u_i^0 = \frac{\lambda}{\rho+\tau} - \frac{1}{\tau} + \mathrm{o}(1).$$

When $\lambda < \frac{\rho+\tau}{\tau}$, *in the limit* $N \to \infty$, *it is unprofitable to collaborate,* $\beta^* = 0$. *In the limit* $\rho \to \infty$, *it is unprofitable to collaborate,* $\beta^* = 0$. *In the limit* $\rho \to 0$, *collaboration is also unprofitable,* $\beta^* = 0$.

Similarly to Theorem 5.1, when the number of participants is big enough, people will want to join the collaborative learning procedure and shrink their contributions with the number of participants. The second property shows that when people have very good local estimates they do not want to participate in collaborative learning: the potential gain in accuracy is small, while the privacy concerns become very big. The last property shows that, when potential gain from learning is very small, $\rho \to 0$, privacy concerns start to dominate learning concerns.

6 OPTIMAL PROTOCOLS FOR ACCURACY

Finally, we empirically find the optimal protocols for the server's end-model accuracy. The server's accuracy in all cases depends only on the value of $\mathit{\Gamma} = \sum \beta_i$ (essentially the inverse of the amount of noise) that the clients use. Thus, we will be interested in maximizing $\mathit{\Gamma}$ given the participation constraints.

We focus on the mean estimation case and compare two families of solutions. The first family are the symmetric solutions $\beta_i = b, b \in [0, \rho]$, which are standard in practice. The second family are the solutions from our existence proofs, Theorems 4.2 and 4.9. In all cases, we seek a protocol that maximizes $\mathit{\Gamma} = \sum_{i=1}^N \beta_i$,

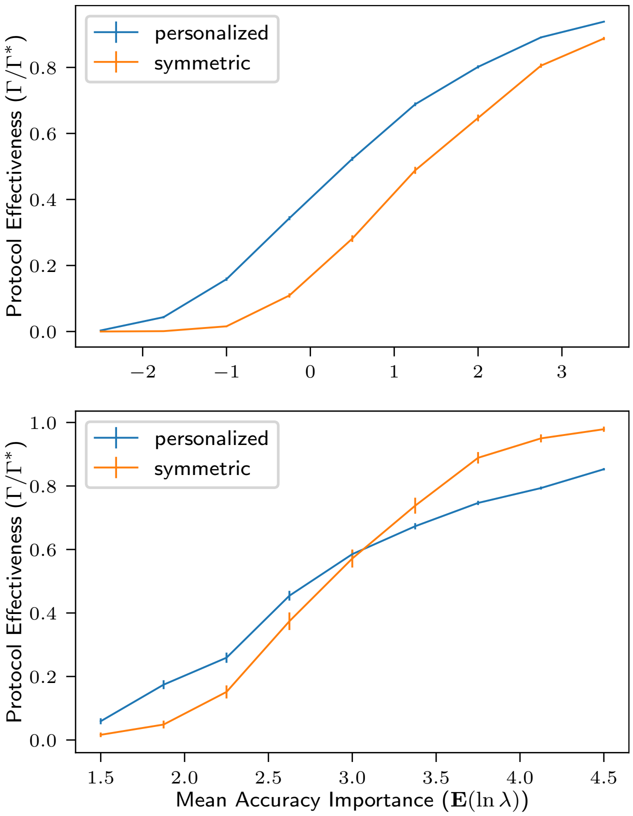

Figure 1: Personalized vs symmetric protocols. Error bars depict the standard deviation of the average effectiveness.

by optimizing the parameter β via a grid search. We focus on a setup with $N = 5$, $n = 100$, and $\sigma = 10$. In each experiment, we sample the preferences for accuracy from the log-normal distribution $\lambda_i \sim \exp(\omega)$, where $\omega \sim \mathrm{N}(m, 1)$. We repeat the experiment 1000 for each value of m and look at the average ratio of the resulting $\mathit{\Gamma}$ over the optimal value of $\mathit{\Gamma}^* = N\rho$.

DP Mean Estimation We compare the symmetric solutions to the solutions found in Theorem 4.2, $\beta_i = \zeta_i b, b \in [0, \rho/\max(\zeta_i)]$. We set $B = 20$. We depict our results on Figure 1 (upper half). As we can see, the personalized version of the protocol is always more beneficial than the symmetric version, especially for small λ.

Bayesian Mean Estimation We compare the symmetric solutions to the solutions inspired by Theorem 4.9, $\beta_i = \max(\xi_i, 0)b, b \in [0, \rho/\max(\xi_i)]$. We set $\tau = 1$. We depict our results on Figure 1 (lower half). The personalized version of the protocol is more efficient than the symmetric version of the protocol for small λ. However, for large λ the trend is reversed, as the first-order approximation used for Theorem 4.9 is less accurate for large β.

7 DISCUSSION

This work studied the accuracy-privacy trade-off in federated learning as a multi-objective utility maximization problem. First, we presented a general framework that formalizes our multi-objective problem and our notion of mutually beneficial protocols. Then, we derived necessary and sufficient conditions for the existence of mutually beneficial protocols in the context of mean estimation and strongly convex optimization. In the case of symmetric privacy preferences, we found mutually beneficial protocols that are optimal for total client utility. Finally, we studied optimal mutually beneficial protocols for the end-model accuracy via a synthetic data experiment. Our findings demonstrate that analyzing federated learning through the lens of utility-based analysis can guide the development of better FL protocols.

Future work We see further practical exploration of our framework as a promising direction for future work.

A key practical challenge is to apply our analysis to other privacy notions (e.g., Renyi DP Mironov, 2017) and non-convex problems. In particular, for large-scale applications, the known theoretical upper bounds for accuracy are often loose or inapplicable, complicating the `err`-function choice. A practical solution might involve running small-scale tests to get an empirical scaling law for error. We deem the exploration of such practical approaches to our multi-objective problem a promising direction for future work.

Another practical challenge is to study the variations of the classic Fed-SGD algorithm. In particular, it would be interesting to study weight communication (McMahan et al., 2017; Li et al., 2020, 2021; Kamp et al., 2023) and co-training methods (Bistritz et al., 2020; Abourayya et al., 2024), which have been developed to decrease privacy risks in FL. Even in such setups, privacy remains a concern (Geng et al., 2022; Dimitrov et al., 2022; Zhu et al., 2023), highlighting the fundamental nature of the privacy-accuracy tradeoff and the need for the instruments for its analysis.

Finally, our results assume that the agents know their privacy preferences in terms of λ parameter (e.g., due to innate preferences or personality traits). However, in cross-silo settings, clients' privacy preferences might instead come from profit maximization incentives (e.g., little privacy protection may result in the loss of clients and, therefore, revenue). We see the development of such micro-foundation models for privacy preferences as an important direction for future work.

Acknowledgments

This research has received funding by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure). Anna Mihalkova and Teodora Todorova conducted most of their work on this paper while on an internship at INSAIT, Sofia University. The authors would like to thank Florian Dorner and Dimitar I. Dimitrov for their helpful feedback and discussions in relation to this work.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318.

Amr Abourayya, Jens Kleesiek, Kanishka Rao, Erman Ayday, Bharat Rao, Geoff Webb, and Michael Kamp. Protecting Sensitive Data through Federated Co-Training, 2024.

Nivasini Ananthakrishnan, Stephen Bates, Michael I. Jordan, and Nika Haghtalab. Delegating Data Collection in Decentralized Machine Learning, 2023.

Ameya Anjarlekar, Rasoul Etesami, and R. Srikant. Striking a Balance: An Optimal Mechanism Design for Heterogenous Differentially Private Data Acquisition for Logistic Regression, 2023.

Hilal Asi, Vitaly Feldman, and Kunal Talwar. Optimal Algorithms for Mean Estimation under Local Differential Privacy. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1046–1056. PMLR, 17–23 Jul 2022.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473, 2014. doi: 10.1109/FOCS.2014.56.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private Stochastic Convex Optimization with Optimal Rates. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ilai Bistritz, Ariana Mann, and Nicholas Bambos. Distributed Distillation for On-Device Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22593–22604. Curran Associates, Inc., 2020.

Avrim Blum, Nika Haghtalab, Richard Lanas Phillips, and Han Shao. One for One, or All for All: Equilibria and Optimality of Collaboration in Federated Learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1005–1014. PMLR, 18–24 Jul 2021.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021. doi: 10.1214/21-AOS2058.

K. L. Chung. On a Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 25(3):463 – 483, 1954. doi: 10.1214/aoms/1177728716.

Dimitar Iliev Dimitrov, Mislav Balunovic, Nikola Konstantinov, and Martin Vechev. Data Leakage in Federated Averaging. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Kate Donahue and Jon Kleinberg. Model-sharing Games: Analyzing Federated Learning Under Voluntary Participation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5303–5311, May 2021. doi: 10.1609/aaai.v35i6.16669.

Florian E. Dorner, Nikola Konstantinov, Georgi Pashaliev, and Martin Vechev. Incentivizing Honesty among Competitors in Collaborative Learning and Optimization. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 7659–7696. Curran Associates, Inc., 2023.

Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9 (3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042.

Michael TM Emmerich and André H Deutz. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural computing*, 17:585–609, 2018.

Vitaly Feldman, Tomer Koren, and Kunal Talwar. Private stochastic convex optimization: optimal rates in linear time. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pages 439–449, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384335.

Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding Among the Clones: A Simple and Nearly Optimal Analysis of Privacy Amplification by Shuffling. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–964, 2022. doi: 10.1109/FOCS52979.2021.00096.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting Gradients - How easy is it to break privacy in federated learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16937–16947. Curran Associates, Inc., 2020.

Jiahui Geng, Yongli Mou, Feifei Li, Qing Li, Oya Beyan, Stefan Decker, and Chunming Rong. Towards General Deep Leakage in Federated Learning, 2022.

Arpita Ghosh and Aaron Roth. Selling privacy at auction. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 199–208, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450302616. doi: 10.1145/1993574.1993605.

Yuzheng Hu, Tianle Cai, Jinyong Shan, Shange Tang, Chaochao Cai, Ethan Song, Bo Li, and Dawn Song. Is Vertical Logistic Regression Privacy-Preserving? A Comprehensive Privacy Analysis and Beyond, 2022.

Chao Huang, Shuqi Ke, Charles Kamhoua, Prasant Mohapatra, and Xin Liu. Incentivizing Data Contribution in Cross-Silo Federated Learning, 2022.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. ISSN 1935-8237. doi: 10.1561/2200000083.

Michael Kamp, Jonas Fischer, and Jilles Vreeken. Federated Learning from Small Datasets. In *The Eleventh International Conference on Learning Representations*, 2023.

Sai Praneeth Karimireddy, Wenshuo Guo, and Michael I. Jordan. Mechanisms that Incentivize Data Sharing in Federated Learning, 2022.

Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated Optimization: Distributed Machine Learning for On-Device Intelligence, 2016.

Benjamin Laufer, Jon Kleinberg, and Hoda Heidari. Fine-Tuning Games: Bargaining and Adaptation for General-Purpose Models, 2023.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *International Conference on Learning Representations*, 2021.

Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. *Collaborative Fairness in Federated Learning*, pages 189–204. Springer International Publishing, Cham, 2020. ISBN 978-3-030-63076-8. doi: 10.1007/978-3-030-63076-8_14.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

Ilya Mironov. Rényi Differential Privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017. doi: 10.1109/CSF.2017.11.

Jian Pei. A Survey on Data Pricing: From Economics to Data Science. *IEEE Transactions on Knowledge and Data Engineering*, 34(10):4586–4608, 2022. doi: 10.1109/TKDE.2020.3045927.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to DP-fy ML: A Practical Guide to Machine Learning with Differential Privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023.

Benedek Rozemberczki, Lauren Watson, Péter Bayer, Hao-Tsung Yang, Olivér Kiss, Sebastian Nilsson, and Rik Sarkar. The Shapley Value in Machine Learning. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5572–5579.

International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/778. Survey Track.

Eden Saig, Inbal Talgam-Cohen, and Nir Rosenfeld. Delegated Classification. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 13200–13236. Curran Associates, Inc., 2023.

Zebang Shen, Jiayuan Ye, Anmin Kang, Hamed Hassani, and Reza Shokri. Share Your Representation Only: Guaranteed Improvement of the Privacy-Utility Tradeoff in Federated Learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Nikita Tsoy and Nikola Konstantinov. Strategic Data Sharing between Competitors. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 16483–16514. Curran Associates, Inc., 2023.

Tianhao Wang, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song. *A Principled Approach to Data Valuation for Federated Learning*, pages 153–167. Springer International Publishing, Cham, 2020. ISBN 978-3-030-63076-8. doi: 10.1007/978-3-030-63076-8_11.

Junyi Zhu, Ruicong Yao, and Matthew B Blaschko. Surrogate model extension (SME): A fast and accurate weight update attack on federated learning. *International Conference on Machine Learning*, 2023.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep Leakage from Gradients. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

The supplementary material is structured as follows.

- Appendix A contains several existing results on differential privacy, as well as definitions from convex optimization that are used in our analysis.
- Appendix B contains the proofs of all results in the main text.
- Appendix C presents further analysis with another choice of utility function.

A BACKGROUND

Differential privacy We first present a classic result about the DP guarantees of the Gaussian mechanism.

Theorem A.1 (Theorem A.1, Dwork and Roth 2014)**.** *Let $\varepsilon \in (0,1)$ be arbitrary. For $c^2 > 2\ln\left(\frac{1.25}{\delta}\right)$, the Gaussian Mechanism for function f with parameter $\sigma \geq \frac{c\Delta_2 f}{\epsilon}$, where $\Delta_2 f$ is the ℓ_2 sensitivity of f, is (ε, δ)-differentially private.*

In the case of mean estimation, f is the average over all samples, $\Delta_2 f = \frac{B}{n}$, $\delta = \frac{1}{n^2}$, and $\sigma = \alpha_i$. In the case of stochastic optimization, f is the average over all gradients, $\Delta_2 f = \frac{B}{n}$, and $\sigma = \alpha_i$.

Additionally, we state a result from Feldman et al. 2022, which provides DP guarantees for chaining several local-DP algorithms.

Theorem A.2 (Theorem 3.8, Feldman et al. 2022)**.** *For a domain $\mathcal{D}$, let $R^t \colon f \times \mathcal{D} \to S^t$ for $t \in [m]$ (where S^t is the range space of R^t) be a sequence of algorithms such that $R^t(z_{1:t-1}, \cdot)$ is a $(\varepsilon_0, \delta_0)$-DP local randomizer for all values of auxiliary inputs $z_{1:t-1} \in S^1 \times \dots \times S^{t-1}$. Let $A \colon \mathcal{D}^m \to S^1 \times \dots \times S^m$ be the algorithm that given a dataset $b_{1:m} \in B$, samples a uniformly random permutation π, then sequentially computes $z^t = R^t(z_{1:t-1}, b_{\pi(t)})$ for $t \in [m]$ and outputs $z_{1:m}$. Then for any $\delta \in [0,1]$ such that $\varepsilon_0 \leq \ln\left(\frac{n}{16\ln\left(\frac{2}{\delta}\right)}\right)$, A is $(\varepsilon, \delta + (e^{\varepsilon}+1)(1+e^{-\frac{\varepsilon_0}{2}})m\delta_0)$-DP, where ε has the following property*

$$\varepsilon \leq \ln\left(1 + \frac{e^{\varepsilon_0} - 1}{e^{\varepsilon_0} + 1}\left(\frac{8\sqrt{e^{\varepsilon_0}\ln(4/\delta)}}{\sqrt{m}} + \frac{8e^{\varepsilon_0}}{m}\right)\right).$$

In our case, the dataset B corresponds to the batches that the client will use during training. Two batches will be neighboring if they differ in one element and this elements are neighboors in the usual sense. At the same time, R^t corresponds to the computation of gradient using the client's batch. (Here, the contribution of other clients are treated as an additional randomization in the algorithm R^t.)

Assumptions for the SGD objective We assume that f is μ-strongly convex on W:

$$\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in W \quad f(\boldsymbol{w}_1) \geq f(\boldsymbol{w}_2) + \nabla f(\boldsymbol{w}_2)^\mathsf{T}(\boldsymbol{w}_1 - \boldsymbol{w}_2) + \frac{\mu}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2$$

as well as L-smooth on $\mathbb{R}^d$:

$$\forall \boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^d \quad f(\boldsymbol{w}_1) \leq f(\boldsymbol{w}_2) + \nabla f(\boldsymbol{w}_2)^\mathsf{T}(\boldsymbol{w}_1 - \boldsymbol{w}_2) + \frac{L}{2}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2.$$

B PROOFS

B.1 Proof of Theorem 4.1

We want to solve the following optimization problem

$$\min_{\{a_i\}_{i=1}^N} \mathbf{E}\left(\left(\sum_{k\neq i} a_k m_k + a_i \bar{x}_i - \mu\right)^2\right) \text{ s.t. } \sum_{i=1}^N a_i = 1.$$

Since $\{m_k\}_{k\neq i} \cup \{\bar{x}_i\}$ are independent, we get

$$\mathbf{E}\left(\left(\sum_{k\neq i} a_k m_k + a_i \bar{x}_i - \mu\right)^2\right) = \sum_{k\neq i} a_k^2\left(\alpha_k^2 + \frac{1}{\rho}\right) + \frac{a_i^2}{\rho},$$

where $\rho \coloneqq \frac{n}{\sigma^2}$. Therefore, we get the following first optimality order conditions

$$2a_k\left(\alpha_k^2 + \frac{1}{\rho}\right) = \nu \,\forall k \neq i,$$
$$\frac{2a_i}{\rho} = \nu,$$

where ν is a Lagrange multiplier. Solving these equations, we get

$$a_k = \frac{\beta_k}{\gamma_i + \rho}\,\forall k \neq i,$$
$$a_i = \frac{\rho}{\gamma_i + \rho},$$

where $\beta_k \coloneqq \frac{1}{\alpha_k^2 + \frac{1}{\rho}}$ and $\gamma_i \coloneqq \sum_{k\neq i}\beta_k$. These weights give the desired formula for the optimal estimator.

B.2 Proof of Theorem 4.2

Denote $\mathit{\Gamma} \coloneqq \sum_{i=1}^N \beta_i = \gamma_i + \beta_i$. Then we get the following expression for the system Equation (1)

$$\forall i\; \frac{\lambda_i(\mathit{\Gamma} - \beta_i)}{\rho(\mathit{\Gamma} - \beta_i + \rho)} \geq \frac{\kappa^2\rho\beta_i}{\rho - \beta_i} \iff \zeta_i\mathit{\Gamma}\rho - (\mathit{\Gamma} + \rho)\beta_i + \beta_i^2 \geq 0, \text{ where } \zeta_i = \frac{\lambda_i}{\lambda_i + \kappa^2\rho^2}.$$

Thus,

$$\left(\sum_{i=1}^N \zeta_i\right)\mathit{\Gamma}\rho - \mathit{\Gamma}(\mathit{\Gamma} + \rho) + \sum_{i=1}^N \beta_i^2 \geq 0 \implies \left(\sum_{i=1}^N \zeta_i\right)\mathit{\Gamma}\rho - \mathit{\Gamma}(\mathit{\Gamma} + \rho) + \mathit{\Gamma}^2 > 0,$$

which gives you the following necessary condition

$$\sum_{i=1}^N \zeta_i > 1.$$

This condition is also sufficient. To prove this property, we consider the following parametric family $\beta_i = \zeta_i b$. It gives the following version of Equation (1)

$$\zeta_i\left(\sum_{k=1}^N \zeta_k\right)b\rho - \zeta_i b\left(\rho + \left(\sum_{k=1}^N \zeta_k\right)b\right) + \zeta_i^2 b^2 \geq 0 \iff \zeta_i\left(\left(\sum_{k=1}^N \zeta_k\right) - 1\right)b\rho - \zeta_i\left(\sum_{k\neq i}\zeta_k\right)b^2 \geq 0$$

As we can see, this equation holds for $b \in \left[0, \left(1 - \frac{1}{\sum_{i=1}^N \zeta_i}\right)\rho\right]$.

B.3 Proof of Theorem 5.1

According to Theorem 4.2, we have solutions if and only if

$$(N-1)\lambda > \kappa^2\rho^2,$$

which proves the first statement in the theorem.

In symmetric case, utility gain has the following form

$$h_\lambda(\beta) \coloneqq -\frac{\kappa^2\rho\beta}{\rho - \beta} + \frac{(N-1)\lambda\beta}{\rho((N-1)\beta + \rho)} = \frac{\lambda}{\rho} + \kappa^2\rho - \frac{\lambda}{(N-1)\beta + \rho} - \frac{\kappa^2\rho^2}{\rho - \beta}.$$

To find optimal amount of noise β^*, we should solve the first order condition for the extremum of h_λ

$$0 = h'_\lambda(\beta^*) = \frac{\lambda(N-1)}{((N-1)\beta^* + \rho)^2} - \frac{\kappa^2\rho^2}{(\rho - \beta^*)^2},$$

which gives

$$\beta^* = \frac{(\sqrt{(N-1)\lambda} - \kappa\rho)\rho}{\sqrt{(N-1)\lambda} + (N-1)\kappa\rho}.$$

Using definition of β^*, we get the following expression for $(\alpha^*)^2$

$$(\alpha^*)^2 = \frac{1}{\beta^*} - \frac{1}{\rho} = \frac{N}{\frac{\sqrt{(N-1)\lambda}}{\kappa} - \rho} = \frac{N}{\frac{\sqrt{(N-1)\lambda}n}{\sqrt{2\ln(1.25n^2)}B} - \frac{n}{\sigma^2}}.$$

It is easy to see that α^* is decreasing in λ and σ. By direct calculations,

$$\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}N} = \frac{1}{\frac{\sqrt{(N-1)\lambda}}{\kappa} - \rho} - \frac{N\sqrt{\lambda}}{2\sqrt{N-1}\kappa\left(\frac{\sqrt{(N-1)\lambda}}{\kappa} - \rho\right)^2} = \frac{(N-2)\sqrt{\lambda} - 2\sqrt{N-1}\kappa\rho}{2\sqrt{N-1}\kappa\left(\frac{\sqrt{(N-1)\lambda}}{\kappa} - \rho\right)^2}$$

$$\implies \operatorname{sign}\left(\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}N}\right) = \operatorname{sign}\Big((N-2)\sqrt{\lambda} - 2\sqrt{N-1}\kappa\rho\Big).$$

To show find the sign of $\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}n}$, we will analyze the derivative of inverse

$$N\frac{\mathrm{d}\frac{1}{(\alpha^*)^2}}{\mathrm{d}n} = \frac{\sqrt{(N-1)\lambda}}{\sqrt{2\ln(1.25n^2)}B} - \frac{1}{\sigma^2} - \frac{\sqrt{(N-1)\lambda}}{\ln(1.25n^2)\sqrt{2\ln(1.25n^2)}B} = \frac{1}{n}\left(\frac{\sqrt{(N-1)\lambda}}{\kappa} - \rho - \frac{\sqrt{(N-1)\lambda}}{\ln(1.25n^2)\kappa}\right)$$

$$\implies \operatorname{sign}\left(\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}n}\right) = \operatorname{sign}\left(\kappa\rho + \frac{\sqrt{(N-1)\lambda}}{\ln(1.25n^2)} - \sqrt{(N-1)\lambda}\right).$$

Finally, we will calculate the utility gain form using the optimal amount of noise. First, notice that

$$\rho - \beta^* = \frac{N\kappa\rho^2}{\sqrt{(N-1)\lambda} + \kappa\rho},$$
$$(N-1)\beta^* + \rho = \frac{N\sqrt{(N-1)\lambda}\rho}{\sqrt{(N-1)\lambda} + \kappa\rho}.$$

Thus,

$$\frac{\beta^*}{\rho - \beta^*} = \frac{\sqrt{(N-1)\lambda} - \kappa\rho}{N\kappa\rho},$$
$$\frac{\beta^*}{(N-1)\beta^* + \rho} = \frac{\sqrt{(N-1)\lambda} - \kappa\rho}{N\sqrt{(N-1)\lambda}}.$$

And we finally get

$$h_\lambda(\beta^*) = \frac{(\sqrt{(N-1)\lambda} - \kappa\rho)^2}{N\rho}.$$

B.4 Proof of Theorem 4.3

First, we will find the weights a_i that will give the best unbiased estimate of the gradient.

Lemma B.1. *Optimal weights a_i have the following properties*

$$a_i = \frac{\beta_i}{\Gamma},\ \operatorname{Var}\boldsymbol{g}^{t+1} = \frac{1}{\Gamma},$$

where $\rho = \frac{b}{\sigma^2}$, $\beta_i \coloneqq \frac{1}{\frac{1}{\rho} + d\alpha_i^2}$, and $\Gamma = \sum_{i=1}^N \beta_i$.

Proof. Similarly to the differentially private estimation case, we want to solve the following optimization problem

$$\min_{\{a_i\}_{i=1}^N} \mathbf{E}\left(\left\|\sum_{k=1}^N a_k \boldsymbol{m}_k^t - \nabla f(\boldsymbol{w}^t)\right\|^2\right) \text{ s.t. } \sum_{i=1}^N a_i = 1.$$

Since messages are independent, we get

$$\mathbf{E}\left(\left\|\sum_{k=1}^N a_k \boldsymbol{m}_k^t - \nabla f(\boldsymbol{w}^t)\right\|^2\right) = \sum_{k=1}^N a_k^2\left(d\alpha_k^2 + \frac{1}{\rho}\right),$$

where $\rho \coloneqq \frac{b}{\sigma^2}$. Therefore, we get the following first optimality order conditions

$$2a_k\left(d\alpha_k^2 + \frac{1}{\rho}\right) = \nu \;\forall k,$$

where ν is a Lagrange multiplier. Solving these equations, we get

$$a_k = \frac{\beta_k}{\Gamma} \;\forall k,$$

where $\beta_k \coloneqq \frac{1}{d\alpha_k^2 + \frac{1}{\rho}}$ and $\Gamma \coloneqq \sum_{k=1}^N \beta_k$. These weights give the desired formula for the optimal estimator. $\square$

Using this estimate for the variance of gradients, we get the following upper bound on the objective.

Lemma B.2. *If $\eta^{t+1} \leq \frac{2}{L}$, the iterates will satisfy the following inequality*

$$\Delta \boldsymbol{w}^{t+1} \leq (1 - 2\eta^{t+1}\mu + L\mu(\eta^{t+1})^2)\Delta \boldsymbol{w}^t + \frac{(\eta^{t+1})^2}{\Gamma},$$

where $\Delta \boldsymbol{w}^\tau \coloneqq \mathbf{E}(\|\boldsymbol{w}^\tau - \boldsymbol{w}^*\|^2)$ and $\boldsymbol{w}^* \coloneqq \arg\min_{\boldsymbol{w}\in W} f(\boldsymbol{w})$.

Before proving it, we derive the following standard lemma.

Lemma B.3. *Let $f^* = \min_{\mathbb{R}^d} f(\boldsymbol{w})$. If f is L-smooth,*

$$\|\nabla f(\boldsymbol{w})\|^2 \leq 2L(f(\boldsymbol{w}) - f^*).$$

Proof. Denote $\boldsymbol{g} = \nabla f(\boldsymbol{w})$. The smoothness gives

$$f\left(\boldsymbol{w} - \frac{1}{L}\boldsymbol{g}\right) \leq f(\boldsymbol{w}) - \frac{1}{2L}\|\boldsymbol{g}\|^2.$$

Therefore,

$$f^* - f(\boldsymbol{w}) \leq f\left(\boldsymbol{w} - \frac{1}{L}\boldsymbol{g}\right) - f(\boldsymbol{w}) \leq -\frac{1}{2L}\|\boldsymbol{g}\|^2.$$

$\square$

Now, we will proof Lemma B.2.

Proof. We have

$$\begin{aligned}\Delta \boldsymbol{w}^{t+1} &= \mathbf{E}(\|\Pi_W(\boldsymbol{w}^t - \eta^{t+1}\boldsymbol{g}^{t+1}) - \boldsymbol{w}^*\|^2) \leq \mathbf{E}(\|\boldsymbol{w}^t - \eta^{t+1}\boldsymbol{g}^{t+1} - \boldsymbol{w}^*\|^2)\\
&= \Delta \boldsymbol{w}^t - 2\eta^{t+1}\,\mathbf{E}((\boldsymbol{g}^{t+1})^\mathsf{T}(\boldsymbol{w}^t - \boldsymbol{w}^*)) + (\eta^{t+1})^2\,\mathbf{E}(\|\boldsymbol{g}^{t+1}\|^2)\\
&= \Delta \boldsymbol{w}^t - 2\eta^{t+1}\,\mathbf{E}(\nabla f(\boldsymbol{w}^t)^\mathsf{T}(\boldsymbol{w}^t - \boldsymbol{w}^*)) + (\eta^{t+1})^2\left(\mathbf{E}(\|\nabla f(\boldsymbol{w}^t)\|^2) + \frac{1}{\Gamma}\right)\\
&\leq \Delta \boldsymbol{w}^t - 2\eta^{t+1}\left(\mathbf{E}(f(\boldsymbol{w}^t) - f^*) + \frac{\mu}{2}\Delta \boldsymbol{w}^t\right) + (\eta^{t+1})^2\left(2L\,\mathbf{E}(f(\boldsymbol{w}^t) - f^*) + \frac{1}{\Gamma}\right)\\
&\leq (1 - 2\eta^{t+1}\mu + \mu L(\eta^{t+1})^2)\Delta \boldsymbol{w}^t + \frac{(\eta^{t+1})^2}{\Gamma},\end{aligned}$$

where we used strong convexity and Lemma B.3 in the penultimate inequality and strong convexity and optimality in the last one. $\square$

To study the sequence $\Delta \boldsymbol{w}^t$, we consider the following sequence of upper bounds

$$y^0 := \frac{L\mu NnD^2}{\sigma^2},$$
$$y^{t+1} := (1 - 2\eta^{t+1}\mu + L\mu(\eta^{t+1})^2)y^t + L\mu(\eta^{t+1})^2.$$

Lemma B.4. *For any choice of the step-sizes, we have $y^t \geq L\mu\Gamma\Delta\boldsymbol{w}^t$. η_*^{t+1} that minimizes the sequence y^t has the following properties*

$$\eta_*^{t+1} = \frac{y^t}{L(y^t+1)} < \frac{2}{L},\ (1 - 2\eta_*^{t+1}\mu + L\mu(\eta_*^{t+1})^2)y^t + L\mu(\eta_*^{t+1})^2 = \left(1 - \frac{\chi y^t}{y^t+1}\right)y^t,$$

where $\chi = \frac{\mu}{L}$.

Proof. First order conditions for the problem $\min_{\eta^{t+1}} y^{t+1}$ gives

$$-\mu y^t + L\mu(1 + y^t)\eta^{t+1} = 0,$$

which gives optimal η_*^{t+1}. Substituting it into a formula for y^{t+1} gives the desired result. □

Now, we are ready to proof Theorem 4.3.

Proof. Notice that

$$\Delta\boldsymbol{w}^0 \leq D^2 \implies L\mu\Gamma\Delta\boldsymbol{w}^0 \leq L\mu\Gamma D^2 \leq \frac{L\mu nND^2}{\sigma^2} = y_*^0.$$

By induction,

$$\begin{aligned} L\mu\Gamma\Delta\boldsymbol{w}^{t+1} &\leq (1 - 2\eta_*^{t+1}\mu + \mu L(\eta_*^{t+1})^2)L\mu\Gamma\Delta\boldsymbol{w}^t + L\mu(\eta_*^{t+1})^2 \\ &\leq (1 - 2\eta^{t+1}\mu + \mu L(\eta^{t+1})^2)y_*^t + L\mu(\eta_*^{t+1})^2 = y_*^{t+1}. \end{aligned}$$

Which proves that $L\mu\Gamma\Delta\boldsymbol{w}^{t+1} \leq y_*^{t+1}$.

We know that

$$y_*^{t+1} \leq \left(1 - \frac{\chi y_*^t}{y_*^t + 1}\right)y_*^t = h(y_*^t).$$

Notice that $h(y)$ is decreasing in y.

Also notice that if $y_*^t \geq 1$, $y_*^{t+1} \leq \left(1 - \frac{\chi}{2}\right)y_*^t$. Then for $t < T := \max\left(\left\lceil -\frac{\ln(y_*^0)}{\ln\left(1-\frac{\chi}{2}\right)}\right\rceil, 1\right)$, we have

$$y_*^t \leq \left(1 - \frac{\chi}{2}\right)^t y_*^0.$$

For $t \geq T$, we have $y_*^t \leq 1$.

Consider the sequence $z^t := \frac{1}{1+\frac{\chi}{2-\chi}t}$. We want to show that $z^t \geq y_*^{t+T}$. We already know that $z^0 = 1 \geq y_*^T$. Now notice that

$$\begin{aligned} h(z^t) \leq z^{t+1} &\iff \left(1 - \frac{\chi}{2 + \frac{\chi}{2-\chi}t}\right)\frac{1}{1 + \frac{\chi}{2-\chi}t} \leq \frac{1}{1 + \frac{\chi}{2-\chi}(t+1)} \iff \\ &\left(2 + \frac{\chi}{2-\chi}t - \chi\right)\left(1 + \frac{\chi}{2-\chi}(t+1)\right) \leq \left(2 + \frac{\chi}{2-\chi}t\right)\left(1 + \frac{\chi}{2-\chi}t\right) \iff \frac{2\chi^2(1-\chi)}{(2-\chi)^2}t \geq 0, \end{aligned}$$

which shows that $z^{t+1} \geq h(z^t)$. By induction, we have

$$y_*^{t+1+T} \leq h(y_*^{t+T}) \leq h(z^t) \leq z^{t+1}.$$

Thus,

$$\Delta w^t \leq \frac{y_*^t}{L\mu\Gamma} \leq \begin{cases} \frac{1}{\left(1+\frac{\chi}{2-\chi}(t-T)\right)L\mu\Gamma}, & t \geq T, \\ \left(1 - \frac{\chi}{2}\right)^m \frac{y_*^0}{L\mu\Gamma}, & t < T. \end{cases}$$

□

B.5 Proof of Theorem 4.4

First, notice that

$$\frac{\mathrm{e}^x - 1}{\mathrm{e}^x + 1} \le x.$$

To prove it, we consider the following function

$$h(x) = x(\mathrm{e}^x + 1) - \mathrm{e}^x + 1.$$

Notice that $h(0) = 0$ and

$$h'(x) = \mathrm{e}^x + 1 + x\mathrm{e}^x - \mathrm{e}^x = 1 + x\mathrm{e}^x > 0.$$

Therefore $h(x) \ge 0$, which proves the desired property.

Also notice that

$$\frac{\mathrm{e}^{\varepsilon_{0,i}}}{m} \le \frac{\mathrm{e}}{2} \le \ln(4).$$

Now we would use Theorem A.2 to describe the global privacy budget, ε_i, and Theorem A.1 describe the privacy budget of each iteration, $\varepsilon_{0,i}$.

Therefore,

$$\begin{aligned}\varepsilon_i \le \ln\left(1 + \frac{\mathrm{e}^{\varepsilon_{0,i}} - 1}{\mathrm{e}^{\varepsilon_{0,i}} + 1}\left(\frac{8\sqrt{\mathrm{e}^{\varepsilon_{0,i}} \ln(\frac{4}{\delta})}}{\sqrt{m}} + \frac{8\mathrm{e}^{\varepsilon_{0,i}}}{m}\right)\right) &\le \frac{\mathrm{e}^{\varepsilon_{0,i}} - 1}{\mathrm{e}^{\varepsilon_{0,i}} + 1}\left(\frac{8\sqrt{\mathrm{e}^{\varepsilon_{0,i}} \ln(\frac{4}{\delta})}}{\sqrt{m}} + \frac{8\mathrm{e}^{\varepsilon_{0,i}}}{m}\right) \\ &\le \varepsilon_{0,i}\frac{16\sqrt{\mathrm{e}\ln(\frac{4}{\delta})}}{\sqrt{m}} \le \frac{\sqrt{2\ln(1.25mn^2)}B}{b\alpha_i}\frac{16\sqrt{\mathrm{e}\ln(4n^2)}}{\sqrt{m}}.\end{aligned}$$

B.6 Proof of Theorem 4.5

Notice that Equation (2) is equivalent to the system

$$\forall i \ \frac{\rho\psi_i\left(1 - \frac{\rho}{\Gamma}\right)}{1 + \psi_i\left(1 - \frac{\rho}{\Gamma}\right)} \ge \beta_i.$$

Denote $x \coloneqq \frac{\Gamma}{\rho} - 1$. We get the system

$$\forall i \ \frac{\rho\psi_i x}{(\psi_i + 1)x + 1} \ge \beta_i.$$

Summing all inequalities, we get the desired necessary condition

$$\sum_{i=1}^{N} \frac{\psi_i x}{(\psi_i + 1)x + 1} \ge \frac{\Gamma}{\rho} = x + 1.$$

To prove that this condition is also sufficient, we consider

$$\beta_i = \frac{\rho\psi_i x^*}{(\psi_i + 1)x^* + 1} \le \rho,$$

where x^* solves the necessary condition. Then we get

$$\Gamma = \sum_{i=1}^{N} \beta_i = \sum_{i=1}^{N} \frac{\psi_i \rho x^*}{(\psi_i + 1)x^* + 1} \ge \rho(x^* + 1).$$

Therefore, $\frac{\rho}{\Gamma} \le \frac{1}{x^*+1}$ and

$$\frac{\rho\psi_i\left(1 - \frac{\rho}{\Gamma}\right)}{1 + \psi_i\left(1 - \frac{\rho}{\Gamma}\right)} \ge \frac{\rho\psi_i x^*}{x^* + 1 + \psi_i x^*} = \beta_i,$$

which shows that the condition is also sufficient.

B.7 Proof of Corollary 4.6

To prove the sufficient condition, we will require that $x = 1$ is a solution. Then we should have

$$\sum_{i=1}^{N} \frac{\psi_i}{\psi_i + 2} \geq 2.$$

To prove the necessary condition, notice that

$$\sum_{i=1}^{N} \frac{\psi_i x}{(\psi_i + 1)x + 1} \leq \sum_{i=1}^{N} \frac{\psi_i x}{2\sqrt{(\psi_i + 1)x}}.$$

Therefore, it is necessary for the following system to have solutions

$$\sum_{i=1}^{N} \frac{\psi_i}{\sqrt{\psi_i + 1}} \geq 2\left(\sqrt{x} + \frac{1}{\sqrt{x}}\right).$$

However, the system above have solutions if and only if

$$\sum_{i=1}^{N} \frac{\psi_i}{\sqrt{\psi_i + 1}} \geq 4.$$

B.8 Proof of Theorem 5.2

According to theorem Theorem 4.5, Equation (2) has a solution if and only if inequality $\sum_{i=1}^{N} \frac{\psi x}{(\psi+1)x+1} \geq x + 1$, where $\psi = \frac{\lambda}{L\mu d\kappa^2\rho^2}$, has a solution. In symmetric case, this inequality is equivalent to

$$N\psi x \geq (x + 1)((\psi + 1)x + 1) \iff 0 \geq (\psi + 1)x^2 - ((N - 1)\psi - 2)x + 1.$$

The determinant of the left-hand part is equal to

$$((N - 1)\psi - 2)^2 - 4(\psi + 1) = ((N - 1)^2\psi - 4N)\psi.$$

Therefore, we should have

$$\sqrt{(N - 1)\psi} \geq \sqrt{\frac{4N}{N - 1}} \iff \sqrt{(N - 1)\lambda} \geq \sqrt{\frac{4NL\mu d}{N - 1}}\kappa\rho$$

for the system to have a solution.

To find the optimal level of noise, we notice that utility gain is equal to the following function in the symmetric case

$$h_\lambda(\beta) := u_i - u_i^0 = \kappa^2\rho d\left(\psi\left(1 - \frac{\rho}{N\beta}\right) - \frac{\beta}{\rho - \beta}\right).$$

First-order conditions for the optimal level of noise give

$$0 = h'_\lambda(\beta^*) = \kappa^2\rho d\left(\frac{\psi\rho}{N(\beta^*)^2} - \frac{\rho}{(\rho - \beta^*)^2}\right) \implies \beta^* = \frac{\sqrt{\psi}\rho}{\sqrt{\psi} + \sqrt{N}} \implies$$

$$(\alpha^*)^2 = \frac{\sqrt{N}}{\sqrt{\psi}\rho d} = \frac{\sqrt{NL\mu}\kappa}{\sqrt{\lambda}d} = \frac{16\sqrt{2eNL\mu\ln(1.25Tn^2)\ln(4n^2)}TB}{\sqrt{\lambda}dn},$$

where

$$T = \max\left(-\frac{\ln(y_*^0)}{\ln\left(1 - \frac{\chi}{2}\right)}, 1\right),\ y_*^0 = \frac{L\mu nND^2}{\sigma^2}.$$

It is easy to see that

$$\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}\lambda} \leq 0.$$

Moreover, since T is increasing in N and decreasing in σ^2, we get

$$\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}N} \geq 0, \ \frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}\sigma^2} \leq 0.$$

Also, we have

$$\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}n} = -(\alpha^*)^2\left(\frac{1}{n} - \frac{1}{2\ln(1.25Tn^2)}\left(\frac{2}{n} + \frac{\mathrm{d}T}{\mathrm{d}n}\right) - \frac{1}{\ln(4n^2)n} - \frac{1}{2T}\frac{\mathrm{d}T}{\mathrm{d}n}\right).$$

Since

$$\frac{\mathrm{d}T}{\mathrm{d}n} = -\frac{\left[\ln(y_*^0) \geq -\ln\left(1-\frac{\chi}{2}\right)\right]}{n\ln\left(1-\frac{\chi}{2}\right)},$$

we have

$$\begin{aligned}\frac{\mathrm{d}(\alpha^*)^2}{\mathrm{d}n} = -\frac{(\alpha^*)^2}{n}\left(1 - \frac{1}{\ln(1.25Tn^2)} + \frac{\left[\ln(y_*^0) \geq -\ln\left(1-\frac{\chi}{2}\right)\right]}{2\ln(1.25n^2T)\ln\left(1-\frac{\chi}{2}\right)} - \frac{1}{\ln(4n^2)} + \frac{\left[\ln(y_*^0) \geq -\ln\left(1-\frac{\chi}{2}\right)\right]}{2T\ln\left(1-\frac{\chi}{2}\right)}\right) \\ = -\frac{(\alpha^*)^2}{n}\left(1 - \mathrm{O}\left(\frac{1}{\ln(n)}\right)\right) \leq 0 \text{ for big enough } n.\end{aligned}$$

Finally, we have

$$h_\lambda(\beta^*) = \kappa^2\rho d\left(\psi\frac{N-1}{N} - \frac{2\sqrt{\psi}}{\sqrt{N}}\right) = \frac{\sqrt{(N-1)\lambda}}{L\mu\rho N}\left(\sqrt{(N-1)\lambda} - \sqrt{\frac{4NL\mu d}{N-1}}\kappa\rho\right).$$

B.9 Proof of Theorem 4.8

First, we want to show that it is optimal to use the same predictor for different local points.

Let $\boldsymbol{x}_i \coloneqq (x_i^1, \ldots, x_i^n)^\mathsf{T}$ and $\hat{\boldsymbol{x}}_i \coloneqq (\hat{x}_i^1, \ldots, \hat{x}_i^n)^\mathsf{T}$. Notice that

$$\boldsymbol{x}_i - \boldsymbol{\iota}\bar{x}_i = \mathbf{I}\boldsymbol{x}_i - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\mathsf{T}}{n}\boldsymbol{x}_i = \left(\mathbf{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\mathsf{T}}{n}\right)\boldsymbol{x}_i.$$

Therefore, $\boldsymbol{x}_i$ can be decomposed into the following sum

$$\boldsymbol{x}_i = \boldsymbol{\iota}\bar{x}_i + \sum_{j=1}^{n-1}\boldsymbol{v}^j\delta_i^j,$$

where $\{\boldsymbol{v}^j\}$ are orthonormal eigenvectors of the matrix $\mathbf{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}^\mathsf{T}}{n}$ that correspond to unit eigenvalue, and δ_i^j are independent normal variables $\mathrm{N}(0, \sigma^2)$.

Let $\hat{\delta}_i^j = (\boldsymbol{v}^j)^\mathsf{T}\hat{\boldsymbol{x}}_i$ and $\hat{\bar{x}}_i = \frac{1}{n}\sum_{j=1}^n \hat{x}_i^j$. We have the following decomposition

$$\sum_{j=1}^n(\hat{x}_i^j - x_i^j)^2 = (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i)^\mathsf{T}(\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i) = \sum_{j=1}^{n-1}(\hat{\delta}_i^j - \delta_i^j)^2 + n(\hat{\bar{x}}_i - \bar{x}_i)^2.$$

Notice that $\hat{\boldsymbol{x}}_i$ can not depend on $\{\delta_i^j\}_{j=1}^{n-1}$. Thus,

$$\mathbf{E}((\hat{\delta}_i^j - \delta_i^j)^2) = \mathbf{E}((\hat{\delta}_i^j)^2 + (\delta_i^j)^2) = \mathbf{E}((\hat{\delta}_i^j)^2) + \sigma^2.$$

Therefore, the optimal predictor should have $\hat{\delta}_i^j = 0$. It allow us to simplify privacy loss

$$\sum_{j=1}^n(\hat{x}_i^j - x_i^j)^2 = (n-1)\sigma^2 + n(\hat{\bar{x}}_i - \bar{x}_i)^2.$$

Corresponding best predictor should have a form $\hat{\boldsymbol{x}}_i = \boldsymbol{\iota}\hat{\bar{x}}_i$.

Now, we want to derive Bayes estimator for $\bar{x}_i$ that only uses variables $\{m_1, \dots, m_N\}$. By integrating over irrelevant noise, we get

$$\forall k \neq i \; m_k \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n} + \alpha_k^2\right), m_i \sim \mathrm{N}(\bar{x}_i, \alpha_i^2), \bar{x}_i \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \mu \sim \mathrm{N}\left(0, \frac{1}{\tau}\right).$$

Denote $\beta_k \coloneqq \frac{1}{\frac{\sigma^2}{n} + \alpha_k^2}$ and $\rho \coloneqq \frac{n}{\sigma^2}$. We get the following density function

$$\begin{aligned} p &\propto \exp\left(-\frac{(m_i - \bar{x}_i)^2}{2\alpha_i^2} - \sum_{k \neq i} \frac{\beta_k (m_k - \mu)^2}{2} - \frac{\rho(\bar{x}_i - \mu)^2}{2} - \frac{\tau\mu^2}{2}\right) \\ &\propto \exp\left(-\frac{(m_i - \bar{x}_i)^2}{2\alpha_i^2} - \frac{\sum_{k \neq i} \beta_k m_k^2}{2} - \frac{\rho \bar{x}_i^2}{2} + \left(\sum_{k \neq i} \beta_k m_k + \rho \bar{x}_i\right)\mu - \frac{(\sum_{k \neq i} \beta_k + \rho + \tau)\mu^2}{2}\right). \end{aligned}$$

As we can see, the only statistics we use from $\{m_k\}_{k \neq i}$ is $s_i \coloneqq \sum_{k \neq i} \beta_k m_k \sim \mathrm{N}(\gamma_i \mu, \gamma_i)$, where $\gamma_i \coloneqq \sum_{k \neq i} \beta_k$. Using this variable, we get

$$p \propto \exp\left(-\frac{(m_i - \bar{x}_i)^2}{2\alpha_i^2} - \frac{s_i^2}{2\gamma_i} - \frac{\rho \bar{x}_i^2}{2} + (s_i + \rho \bar{x}_i)\mu - \frac{(\gamma_i + \rho + \tau)\mu^2}{2}\right).$$

After we integrate over μ, we get

$$\begin{aligned} p &\propto \exp\left(-\frac{(m_i - \bar{x}_i)^2}{2\alpha_i^2} - \frac{s_i^2}{2\gamma_i} - \frac{\rho \bar{x}_i^2}{2} + \frac{(s_i + \rho \bar{x}_i)^2}{2(\gamma_i + \rho + \tau)}\right) \\ &\propto \exp\left(-\frac{m_i^2}{2\alpha_i^2} - \frac{s_i^2}{2\gamma_i} + \left(\frac{m_i}{\alpha_i^2} + \frac{\rho s_i}{\gamma_i + \rho + \tau}\right)\bar{x}_i - \left(\frac{1}{\alpha_i^2} + \frac{\rho(\gamma_i + \tau)}{\gamma_i + \rho + \tau}\right)\frac{\bar{x}_i^2}{2}\right). \end{aligned}$$

So, $\bar{x}_i$ is distributed as

$$\mathrm{N}\left(\left(\frac{1}{\alpha_i^2} + \frac{\rho(\gamma_i + \tau)}{\gamma_i + \rho + \tau}\right)^{-1}\left(\frac{m_i}{\alpha_i^2} + \frac{\rho s_i}{\gamma_i + \rho + \tau}\right), \left(\frac{1}{\alpha_i^2} + \frac{\rho(\gamma_i + \tau)}{\gamma_i + \rho + \tau}\right)^{-1}\right)$$

and the mean of this distribution is the optimal estimator for $\bar{x}_i$.

B.10 Proof of Theorem 4.7

Similarly to the previous section we have the following distributions of $\bar{x}_i$, m_i, and μ

$$\forall k \neq i \; m_k \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n} + \alpha_k^2\right), \bar{x}_i \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right), \mu \sim \mathrm{N}\left(0, \frac{1}{\tau}\right).$$

It implies the following density

$$p \propto \exp\left(-\sum_{k \neq i} \frac{\beta_k (m_k - \mu)^2}{2} - \frac{\rho(\bar{x}_i - \mu)^2}{2} - \frac{\tau\mu^2}{2}\right)$$

Again, the sufficient statistics of $\{m_k\}_{k \neq i}$ is s_i and we get the following density

$$\begin{aligned} p &\propto \exp\left(-\frac{(s_i - \gamma_i \mu)^2}{2\gamma_i} - \frac{\rho(\bar{x}_i - \mu)^2}{2} - \frac{\tau\mu^2}{2}\right) \\ &\propto \exp\left(-\frac{s_i^2}{2\gamma_i} - \frac{\rho \bar{x}_i^2}{2} + (s_i + \rho \bar{x}_i)\mu - \frac{(\gamma_i + \rho + \tau)\mu^2}{2}\right). \end{aligned}$$

Thus, μ is distributed as $\mathrm{N}\left(\frac{s_i + \rho \bar{x}_i}{\gamma_i + \rho + \tau}, \frac{1}{\gamma_i + \rho + \tau}\right)$ and the optimal estimator is the mean of this distribution.

B.11 Proof of Theorem 4.9

The first order Taylor approximation gives the following inequality

$$\left(\frac{1}{\tau}+\frac{1}{\rho}\right)\left(1-\frac{\beta_i}{\rho}+\frac{\gamma_i}{\rho+\tau}-\frac{\beta_i+\gamma_i}{\tau}\right)-\frac{\lambda_i}{\rho+\tau}\left(1-\frac{\gamma_i}{\rho+\tau}\right)\geq\left(\frac{1}{\tau}+\frac{1}{\rho}\right)-\frac{\lambda_i}{\rho+\tau}$$
$$\iff \left(\frac{\lambda_i}{(\rho+\tau)^2}-\frac{1}{\tau^2}\right)\gamma_i-\frac{(\rho+\tau)^2}{\rho^2\tau^2}\beta_i\geq 0 \iff \left(\frac{\lambda_i}{(\rho+\tau)^2}-\frac{1}{\tau^2}\right)\sum_{i=1}^N\beta_k-\left(\frac{\lambda_i}{(\rho+\tau)^2}+\frac{(2\rho+\tau)}{\rho^2\tau}\right)\beta_i\geq 0.$$

Notice that the inequality have a solution only if

$$\frac{\lambda_i}{(\rho+\tau)^2}-\frac{1}{\tau^2}\geq 0.$$

Now, we assume that all clients care about utility strongly enough, so that

$$\forall i\ \frac{\lambda_i}{(\rho+\tau)^2}-\frac{1}{\tau^2}\geq 0,$$

and derive necessary conditions for the feasibility of the participation of all clients. Notice that

$$\left(\frac{\lambda_i}{(\rho+\tau)^2}-\frac{1}{\tau^2}\right)\sum_{i=1}^N\beta_k-\left(\frac{\lambda_i}{(\rho+\tau)^2}+\frac{(2\rho+\tau)}{\rho^2\tau}\right)\beta_i\geq 0 \iff \xi_i\sum_{i=1}^N\beta_k-\beta_i\geq 0,$$

where

$$\xi_i := \frac{\frac{\lambda_i}{(\rho+\tau)^2}-\frac{1}{\tau^2}}{\frac{\lambda_i}{(\rho+\tau)^2}+\frac{(2\rho+\tau)}{\rho^2\tau}}=\frac{1}{1+\frac{(\rho+\tau)^2}{\tau^2\rho^2\left(\frac{\lambda_i}{(\rho+\tau)^2}-\frac{1}{\tau^2}\right)}}.$$

By summing inequalities for each client, we get

$$\left(\sum_{j=1}^N\xi_j\right)\left(\sum_{k=1}^N\beta_k\right)\geq\sum_{k=1}^N\beta_k.$$

So, the necessary condition for the existence of solution is

$$\sum_{j=1}^N\xi_j\geq 1.$$

This condition is also sufficient: when it holds $\beta_i=\xi_i b, b>0$ is a solution.

B.12 Proof of Theorem 5.3

In this case, we have $\gamma_i=(N-1)\beta$. The inequality will look like

$$\frac{(\rho-\beta)((N-1)\beta+\rho+\tau)}{(N\beta+\tau)\rho^2}-\frac{\lambda}{(N-1)\beta+\rho+\tau}\geq\frac{\rho+\tau}{\rho\tau}-\frac{\lambda}{\rho+\tau}$$
$$\iff -\frac{N-1}{N\rho^2}\beta-\frac{(\rho+\frac{\tau}{N})^2}{\tau\rho^2}+\frac{\lambda}{\rho+\tau}+\frac{(\rho+\frac{\tau}{N})^2}{N\rho^2(\beta+\frac{\tau}{N})}-\frac{\lambda}{(N-1)(\beta+\frac{\rho+\tau}{N-1})}\geq 0.$$

Denote left hand part of the inequality as $h_\lambda(\beta)$. Notice that $h_\lambda(0)=0$. Thus, $\max_\beta h_\lambda(\beta)\geq 0$. Now, we will consider the optimal β^* for the utility of all participants. To do it, we calculate the derivative of h_λ

$$h'_\lambda(\beta)=-\frac{N-1}{N\rho^2}-\frac{(\rho+\frac{\tau}{N})^2}{N\rho^2(\beta+\frac{\tau}{N})^2}+\frac{\lambda}{(N-1)(\beta+\frac{\rho+\tau}{N-1})^2}.$$

Notice that h_λ and h'_λ are increasing in λ.

To study this function, we will take the second derivative

$$h''_\lambda(\beta) = 2\frac{(\rho+\frac{\tau}{N})^2}{N\rho^2(\beta+\frac{\tau}{N})^3} - 2\frac{\lambda}{(N-1)(\beta+\frac{\rho+\tau}{N-1})^3}$$
$$= \frac{2}{(\beta+\frac{\rho+\tau}{N-1})^3}\left(\frac{(\rho+\frac{\tau}{N})^2}{N\rho^2}\left(1+\frac{N\rho+\tau}{N(N-1)(\beta+\frac{\tau}{N})}\right)^3 - \frac{\lambda}{N-1}\right).$$

As we can see, h''_λ can change sign only once from positive to negative on the interval $[0,\rho]$. Thus, h'_λ can change sign only twice: from negative to positive on the interval where $h''_\lambda > 0$ and from positive to negative on the interval where $h''_\lambda < 0$.

1. If h''_λ is always positive $\lambda \le \frac{(N\rho+\tau)^2}{(N-1)^2\rho^2}$, the minimum of h'_λ is $h'_\lambda(0)$ and the maximum is $h'_\lambda(\rho)$.
2. If h''_λ changes sign $\frac{(N\rho+\tau)^2(\rho+\tau)^3}{(N-1)^2\rho^2\tau^3} > \lambda > \frac{(N\rho+\tau)^2}{(N-1)^2\rho^2}$, then one of the points $h'_\lambda(0)$ or $h'_\lambda(\rho)$ is the minimum of h'_λ.
3. If h''_λ is always negative $\lambda \ge \frac{(N\rho+\tau)^2(\rho+\tau)^3}{(N-1)^2\rho^2\tau^3}$, the minimum of h'_λ is $h'_\lambda(\rho)$ and the maximum is $h'_\lambda(0)$.

Notice that in the first case

$$h'_\lambda(\rho) = -\frac{1}{\rho^2} + \frac{\lambda}{(N-1)\left(\frac{N\rho+\tau}{N-1}\right)^2} < -\frac{N-2}{(N-1)\rho^2} < 0.$$

Therefore, $\beta^* = 0$. Since we are not interested in this trivial solutions, we will consider only $\lambda > \frac{(N\rho+\tau)^2}{(N-1)^2\rho^2}$.

The previous analysis of the second derivative immediately gives the following lemma

Lemma B.5. *We can describe the behavior of β^* using the following five cases*

1. $h'_\lambda(0) > 0$ *and* $h'_\lambda(\rho) > 0$*. Then* h'_λ *is always positive and* $\beta^* = \rho$*.*
2. $h'_\lambda(0) \le 0$ *and* $h'_\lambda(\rho) > 0$*. Then* h'_λ *changes sign only once (from negative to positive) and* $\beta^* \in \{0, \rho\}$*.*
3. $h'_\lambda(0) > 0$ *and* $h'_\lambda(\rho) \le 0$*. Then* h'_λ *changes sign only once (from positive to negative) and* $\beta^* \in (h'_\lambda)^{-1}(0) \cap (h''_\lambda)^{-1}((-\infty, 0])$*.*
4. $h'_\lambda(0) \le 0$*,* $h'_\lambda(\rho) \le 0$*, and* $h'_\lambda(\beta) = 0$ *has roots. Then* h'_λ *changes sign twice and* $\beta^* \in \{0\} \cup ((h'_\lambda)^{-1}(0) \cap (h''_\lambda)^{-1}((-\infty, 0]))$*.*
5. $h'_\lambda(0) \le 0$*,* $h'_\lambda(\rho) \le 0$*, and* $h'_\lambda(\beta) = 0$ *does not have roots. Then* h'_λ *is negative and* $\beta^* = 0$*.*

Notice that when h''_λ is always negative, case 4 can not be realized. To distinguish between cases 4 and 5 when h''_λ changes sign, one could find the root of equation $h''_\lambda(\beta_{\text{test}}) = 0$ and look at the sign of $h'_\lambda(\beta_{\text{test}})$. If $h'_\lambda(\beta_{\text{test}}) < 0$, then case 5 is realized. Otherwise, case 4 is realized.

We can describe cases 1–5 in the following manner. In case 1, when λ is very big, the participants care a lot about accuracy and are ready to fully compromise their privacy to get a good machine learning model. In case 5, when λ is very small, the participants care a lot about privacy and are not ready to collaborate at all due to this constraint. Cases 2–4 are intermediate between these two extremes: people care about accuracy of the model and are ready to collaborate, but want to have some privacy protection.

The conditions in the theorem ensures that case 3 will be realized. To show this, we will analyze the restrictions for this case

$$\begin{cases} h'_\lambda(0) &= -\frac{N\rho^2+2\rho\tau+\tau^2}{\rho^2\tau^2} + \frac{\lambda(N-1)}{(\rho+\tau)^2} \ge 0, \\ h'_\lambda(\rho) &= -\frac{1}{\rho^2} + \frac{\lambda(N-1)}{(N\rho+\tau)^2} < 0, \end{cases} \iff \frac{(N\rho^2+2\rho\tau+\tau^2)(\rho+\tau)^2}{(N-1)\rho^2\tau^2} \le \lambda < \frac{(N\rho+\tau)^2}{(N-1)\rho^2}.$$

To make this case self-consistent, we should require

$$N\rho^3 + (2N+2)\rho^2\tau \le (N^2 - N - 5)\rho\tau^2 + (2N-4)\tau^3,$$

which holds when the number of participants is big enough.

In this case $h''_\lambda(\beta^*) \le 0$, $h'_\lambda(\beta^*) = 0$, and $h_\lambda(\beta^*) > 0$. Using implicit function theorem, we get

$$0 = \frac{\mathrm{d}h'_\lambda}{\mathrm{d}\xi} = \frac{\partial h'_\lambda}{\partial \xi} + h''_\lambda(\beta^*)\frac{\mathrm{d}\beta^*}{\mathrm{d}\xi} \implies \frac{\mathrm{d}\beta^*}{\mathrm{d}\xi} = -(h''_\lambda(\beta^*))^{-1}\frac{\partial h'_\lambda}{\partial \xi},$$

where ξ is any parameter. Thus, the sign of $\frac{\mathrm{d}\beta^*}{\mathrm{d}\xi}$ is the same as the sign of $\frac{\partial h'_\lambda}{\partial \xi}$. It allows us to get the following properties

$$\frac{\mathrm{d}\beta^*}{\mathrm{d}\lambda} \ge 0,\ \frac{\mathrm{d}\beta^*}{\mathrm{d}\rho} \ge 0.$$

To prove these properties, we need to determine the signs of $\frac{\partial h'_\lambda}{\partial \lambda}$ and $\frac{\partial h'_\lambda}{\partial \rho}$. For the first derivative, it is straightforward

$$\frac{\partial h'_\lambda}{\partial \lambda} = \frac{1}{(N-1)(\beta^* + \frac{\rho+\tau}{N-1})^2} > 0.$$

For the second, we will use that $h'_\lambda(\beta^*) = 0$ and $h''_\lambda(\beta^*) \le 0$

$$\begin{aligned}
\frac{\partial h'_\lambda}{\partial \rho} &= \frac{N-1}{N\rho^3} + \frac{2(\rho+\frac{\tau}{N})\frac{\tau}{N}}{N\rho^3(\beta+\frac{\tau}{N})^2} - \frac{2\lambda}{(N-1)^2(\beta+\frac{\rho+\tau}{N-1})^3} \\
&= \frac{\lambda}{(N-1)\rho(\beta+\frac{\rho+\tau}{N-1})^2} - \frac{(\rho+\frac{\tau}{N})^2}{N\rho^3(\beta+\frac{\tau}{N})^2} + \frac{2(\rho+\frac{\tau}{N})\frac{\tau}{N}}{N\rho^3(\beta+\frac{\tau}{N})^2} - \frac{2\lambda}{(N-1)^2(\beta+\frac{\rho+\tau}{N-1})^3} \\
&= \frac{\lambda(\beta+\frac{\tau-\rho}{N-1})}{(N-1)^2\rho(\beta+\frac{\rho+\tau}{N-1})^3} - \frac{\rho^2-\frac{\tau^2}{N^2}}{N\rho^3(\beta+\frac{\tau}{N})^2} \ge \frac{(\rho+\frac{\tau}{N})^2(\beta+\frac{\tau-\rho}{N-1})}{N\rho^3(\beta+\frac{\tau}{N})^3} - \frac{\rho^2-\frac{\tau^2}{N^2}}{N\rho^3(\beta+\frac{\tau}{N})^2} \\
&= \frac{(\rho+\frac{\tau}{N})^2(\beta+\frac{\tau-\rho}{N-1}) - (\rho^2-\frac{\tau^2}{N^2})(\beta+\frac{\tau}{N})}{N\rho^3(\beta+\frac{\tau}{N})^3} \ge 0,
\end{aligned}$$

where we have used that $(N-1)\beta + \tau - \rho \ge 0$, which follows from the lemma below.

Lemma B.6. *Assume that $h'_\lambda(\beta) = 0$ and $h''_\lambda(\beta) < 0$. Then $(N-1)\beta + \tau \ge \sqrt[3]{N-1}\rho$.*

To proof this, we notice the following

$$\begin{aligned}
&\frac{(\rho+\frac{\tau}{N})^2}{N\rho^2(\beta+\frac{\tau}{N})^3} \le \frac{\lambda}{(N-1)(\beta+\frac{\rho+\tau}{N-1})^3} \iff \frac{(\rho+\frac{\tau}{N})^2}{N\rho^2(\beta+\frac{\tau}{N})^3} \le \frac{1}{(\beta+\frac{\rho+\tau}{N-1})}\left(\frac{N-1}{N\rho^2} + \frac{(\rho+\frac{\tau}{N})^2}{N\rho^2(\beta+\frac{\tau}{N})^2}\right) \\
&\iff \left(\rho+\frac{\tau}{N}\right)^2\left(\beta+\frac{\rho+\tau}{N-1}\right) \le (N-1)\left(\beta+\frac{\tau}{N}\right)^3 + \left(\rho+\frac{\tau}{N}\right)^2\left(\beta+\frac{\tau}{N}\right) \\
&\iff \left(\rho+\frac{\tau}{N}\right)^3 \le (N-1)^2\left(\beta+\frac{\tau}{N}\right)^3 \iff \sqrt[3]{N-1}\left(\rho+\frac{\tau}{N}\right) \le (N-1)\beta+\tau-\frac{\tau}{N} \implies \rho \le (N-1)\beta+\tau.
\end{aligned}$$

B.13 Proof of Theorem 5.4

B.13.1 $N \to \infty$

In this limit, $h'_\lambda(\rho) \to -\frac{1}{\rho^2} < 0$. According to Lemma B.5, $\beta^* \in \{0\} \cup ((h'_\lambda)^{-1}(0) \cap (h''_\lambda)^{-1}((-\infty, 0]))$.

We start with the possible solution $h'_\lambda(\beta) = 0$ and $h''_\lambda(\beta) < 0$. Lemma B.6 implies that $(N-1)\beta + \tau \ge \sqrt[3]{N-1}\rho$. Thus, $\beta = \Omega\left(\frac{1}{\sqrt[3]{N^2}}\right)$. It allows to write the leading terms in equation $h'_\lambda(\beta) = 0$ in the following manner

$$h'_\lambda(\beta) = -\frac{1}{\rho^2} - \frac{1}{N\beta^2} + \frac{\lambda}{N\beta^2} + \mathrm{O}\left(\frac{1}{\sqrt[3]{N}}\right).$$

It gives $\beta = \sqrt{\frac{\lambda-1}{N}}\rho + \mathrm{o}\left(\frac{1}{\sqrt{N}}\right)$. It gives the following expression for the final gain in utility

$$h_\lambda(\beta) = \frac{(\rho-\beta)((N-1)\beta+\rho+\tau)}{\rho^2(N\beta+\tau)} - \frac{\lambda}{(N-1)\beta+\rho+\tau} - \frac{1}{\rho} - \frac{1}{\tau} + \frac{\lambda}{\rho+\tau} = \frac{\lambda}{\rho+\tau} - \frac{1}{\tau} + \mathrm{O}\left(\frac{1}{\sqrt{N}}\right).$$

Therefore, in the case $\lambda < \frac{\rho+\tau}{\tau}$, it is optimal to not collaborate, $\beta^* = 0$. And, in the case, $\lambda > \frac{\rho+\tau}{\tau}$, it is optimal to collaborate, $\beta^* = \sqrt{\frac{\lambda-1}{N}}\rho + \mathrm{o}\left(\frac{1}{\sqrt{N}}\right)$.

In the case $\lambda > \frac{\rho+\tau}{\tau}$, in the first approximation, β^* is increasing in λ and ρ and decreasing in N. Moreover,

$$(\alpha^*)^2 = \frac{1}{\beta^*} - \frac{1}{\rho} \approx \frac{\sqrt{\frac{N}{\lambda-1}} - 1}{\rho}$$

is decreasing in ρ.

B.13.2 $\rho \to \infty$

According to Lemma B.5, $\beta^* \in \{0, \rho\} \cup ((h'_\lambda)^{-1}(0) \cap (h''_\lambda)^{-1}((-\infty, 0]))$.

First, we consider the possible solution $h'_\lambda(\beta) = 0$ and $h''_\lambda(\beta) < 0$. Lemma B.6 ensures that $\beta \geq \frac{\rho-\tau}{N-1}$. In this case,

$$\begin{aligned} h_\lambda(\beta) &= \frac{(\rho-\beta)((N-1)\beta+\rho+\tau)}{(N\beta+\tau)\rho^2} - \frac{\lambda}{(N-1)\beta+\rho+\tau} - \frac{1}{\rho} - \frac{1}{\tau} + \frac{\lambda}{\rho+\tau} \\ &= \frac{(\rho-\beta)((N-1)\beta+\rho)}{N\beta\rho^2} - \frac{\lambda}{(N-1)\beta+\rho} - \frac{1}{\rho} - \frac{1}{\tau} + \frac{\lambda}{\rho} + \mathrm{o}\left(\frac{1}{\rho}\right) = -\frac{1}{\tau} + \mathrm{O}\left(\frac{1}{\rho}\right) < 0. \end{aligned}$$

Similarly, $h_\lambda(\rho) = -\frac{1}{\tau} + \mathrm{O}\left(\frac{1}{\rho}\right) < 0$. Therefore, $\beta^* = 0$.

B.13.3 $\rho \to 0$

According to Theorem 5.3, the following condition on λ should hold for the existence of solutions

$$\lambda \geq \frac{(N\rho+\tau)^2}{(N-1)^2\rho^2}.$$

However, the right-hand part goes to infinity when $\rho \to 0$. Thus, this condition is not fulfilled for the small ρ.

B.14 Proof of Theorem 4.11

Finally, we prove Theorem 4.11 which concerns the existence of mutually-beneficial protocols in the limit case when the number of participants $N \to \infty$.

Proof Let $U = \{\tilde{u}_1, \ldots, \tilde{u}_{|U|}\}$. For any utility function $\tilde{u}^j$, since $\mathtt{err}^0 > 0$, we have that $\tilde{u}^j\left(\mathtt{err}^0, 0\right) < u^i(0,0)$. Since the function $\tilde{u}^j(0,\epsilon) \in U$ is continuous in ϵ, there exists a value $\epsilon_j > 0$, such that $\tilde{u}^j\left(0, \epsilon_j\right) \in \left(\tilde{u}^j\left(\mathtt{err}^0, 0\right), \tilde{u}^j\left(0, 0\right)\right]$. Let $\tilde{\epsilon} := \min_j \epsilon_j$, so that $\tilde{u}^j\left(0, \tilde{\epsilon}\right) \in \left(\tilde{u}^j\left(\mathtt{err}^0, 0\right), \tilde{u}^j\left(0, 0\right)\right]$ for all $j \in [|U|]$.

Since $\mathtt{leak}$ is decreasing in α, there exists a value α_*, such that $\mathtt{leak} \leq \tilde{\epsilon}$ for all participants. Therefore, for any $i \in [N], j \in [|U|]$, $\tilde{u}^j\left(\mathtt{err}_i^0, 0\right) < \tilde{u}^i\left(0, \tilde{\epsilon}\right) \leq \tilde{u}^i\left(0, \mathtt{leak}\right)$. Now for any $i \in [N], j \in [|U|]$ we have that $\tilde{u}^j(a, \mathtt{leak})$ is a continuous function in a, so there exists a value $a_{i,j}$, such that $\tilde{u}^j(a_{i,j}, \mathtt{leak}) > \tilde{u}^j\left(\mathtt{err}_i^0, 0\right)$. Setting $\tilde{a} = \min_{i,j} a_{i,j}$, we have that $\tilde{u}^j(\tilde{a}, \mathtt{leak}) > \tilde{u}^j\left(\mathtt{err}_i^0, 0\right)$ for all $i \in [N], j \in [|U|]$.

Now setting N large enough so that $\mathtt{err} \leq \tilde{a}$ for all $i \in [N]$ ensures that $\tilde{u}^j(\mathtt{err}, \mathtt{leak}) \geq \tilde{u}^j(\tilde{a}, \mathtt{leak}) > \tilde{u}^j\left(\mathtt{err}_i^0, 0\right)$ for all $j \in [|U|]$ and any participant. Therefore, for any set of utility functions from U for the clients, we have that the protocol is mutually beneficial.

C ANALYSIS OF DIFFERENT UTILITY FUNCTIONS

Here we present additional analysis of our problem, in the case when the utility of the participants is given by:

$$u_i(\texttt{err}, \texttt{leak}) = -\texttt{leak}_i - \lambda_i \texttt{err}_i^2.$$

A rationnelle for this utility function is that the ϵ in DP privacy, can be thought about as the number of bits that the algorithm's outputs reveil about its input. Therefore, the loss in utility from a large ϵ may be linear in ϵ. At the same time, in our framework, $\texttt{err}$ corresponds to the root of the mean squared error of our estimate/end iterate. Therefore, $\texttt{err}^2$ corresponds to MSE more commonly used to quantify the loss of a statistical/ML procedure.

We present an analysis for mean estimation and SGD with DP used as a privacy notion.

C.1 Mean estimation with DP

We consider the mean estimation task with DP. We have N participants who sample n data points each from a global distribution $\mathcal{D}$ on $\mathbb{R}$, with mean μ, variance $\mathbb{E}_{X\sim\mathcal{D}}\left((X-\mu)^2\right) = \sigma^2$ and bounded (almost surely) in $\mu+[-B/2, B/2]$. Each participant then communicates a message $m_i = \bar{x}_i + \epsilon_i$, where $\epsilon_i \sim \mathrm{N}(0, \alpha_i^2)$. For simplicity we do not consider Bayesian-optimal actions of the players based on all messages, but a simplified version similar to Dorner et al. (2023), in which the server communicates the mean of all messages $\bar{m} = \frac{1}{N}\sum m_i$ to all players. The players then accept this estimate, only correcting for their own added noice by taking $\bar{m}_i = \frac{1}{N}(N\bar{m} - m_i + \bar{x}_i)$.

We study the accuracy of the end estimates $\bar{m}_i$ and the privacy of these protocols. For the accuracy, note that Theorem 4.1 in Dorner et al. (2023) implies that:

$$\texttt{err}_i^2 = \mathbb{E}((\bar{m}_i - \mu)^2) = \frac{\sigma^2}{Nn} + \frac{1}{N^2}\sum_{j\neq i}\alpha_j^2,$$

compared to $\texttt{err}_i^2 = \frac{\sigma^2}{n}$ when training a model alone.

As shown in the main text, the privacy guarantee for the Gaussian mechanism implies that

$$\texttt{leak}_i = \frac{\sqrt{2\ln(1.25n^2)}B}{n\alpha_i}.$$

We compare this to $\texttt{leak}_i = 0$ when training locally.

Therefore, the condition for player i being incentivized to join are as follows:

$$\begin{aligned}
\forall i \quad u_i \geq u_i^0 &\iff -\texttt{leak}_i - \lambda_i\texttt{err}_i^2 \geq -\lambda_i\texttt{err}_0^2 \\
&\iff -\frac{\sqrt{2\ln(1.25n^2)}B}{n\alpha_i} - \lambda_i\left(\frac{\sigma^2}{Nn} + \frac{1}{N^2}\sum_{j\neq i}\alpha_j^2\right) \geq -\lambda_i\frac{\sigma^2}{n} \\
&\iff \lambda_i\left(\frac{\sigma^2}{n} - \frac{\sigma^2}{Nn} - \frac{1}{N^2}\sum_{j\neq i}\alpha_j^2\right) \geq \frac{\sqrt{2\ln(1.25n^2)}B}{n\alpha_i}.
\end{aligned}$$

Consider values of λ_i bounded in some compact domain $\in [c, C]$ for some small constant c and large constant C. As predicted by Theorem 4.11, as $N \to \infty$, a mutually beneficial assignment of $\boldsymbol{\alpha}$ exists. In contrast, if $n \to \infty$, the left-hand side of the inequality becomes negative, so that no mutually beneficial alocation of $\boldsymbol{\alpha}$ exists.

C.2 SGD with DP

We consider the same setup as in Section 4.2, with batch size $b = 1$ and equal weights $a_i = \frac{1}{N}$. However, we make the following assumption on the variance:

$$\mathrm{Var}_x(\boldsymbol{g}_i(\boldsymbol{w}, x)) \leq M + M_V\|\nabla f(\boldsymbol{w})\|^2 \; \forall w \in W. \tag{4}$$

Additionally, we assume that $f(w) - f^*$ is bounded over W.

C.2.1 Accuracy guarantees

Given the assumption above, we apply the following improved version of Lemma E.1 from Dorner et al. (2023).

Lemma C.1 (Adapted from Lemma E.1 in Dorner et al. 2023). *Consider $\eta \in \mathbb{N}$, such that $\frac{4}{\mu(\eta+1)} \leq \frac{1}{L(M_V/N+1)}$, the learning rate schedule $\eta^t = \frac{4}{\mu(\eta+t)}$. We get the following upper-bound for the optimization error:*

$$\mathbf{E}(f(\boldsymbol{w}^t) - f(\boldsymbol{w}^*)) \leq \frac{8L\left(\frac{M}{N} + \frac{d\sum_{i=1}^N (\alpha_t^i)^2}{N^2}\right)}{3\mu^2 t} + \mathrm{O}\left(\frac{1}{t^4} + \frac{1}{Nt}\right),$$

as long as $\Pr(\exists t \leq T : \mathit{\Pi}_W(\boldsymbol{w}^{t-1} - \eta^t \boldsymbol{g}^t) \neq \boldsymbol{w}^{t-1} - \eta^t \boldsymbol{g}^t) = \mathrm{O}\left(\frac{1}{Nt}\right)$.

Proof. The proof follows the arguments in Dorner et al. (2023), however the dependence of the bound on the α parameters is traced explicitly and we improve the $\mathrm{O}\left(\frac{1}{t^2}\right)$ dependence to $\mathrm{O}\left(\frac{1}{t^4}\right)$ by carefully adapting a classic result from Chung (1954).

First assume that there is no $t \leq T$, such that $\mathit{\Pi}_W(\boldsymbol{w}^{t-1} - \eta^t \boldsymbol{g}^t) \neq \boldsymbol{w}^{t-1} - \eta^t \boldsymbol{g}^t$. Since the stochastic gradients and the Gaussian noise are independent, we have:

$$\begin{aligned}
\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N m_i^t\right) &= \mathbf{E}\left\|\frac{1}{N}\sum_{i=1}^N (g_i^t - \nabla f(\boldsymbol{w}^{t-1}) + \alpha_i \xi_i^t)\right\|^2 \\
&\leq \frac{1}{N^2}\left(\sum_{i=1}^N M + M_V \|\nabla f(\boldsymbol{w}^{t-1})\|^2 + d(\alpha_i)^2\right) \\
&= \frac{M + M_V \|\nabla f(\boldsymbol{w}^{t-1})\|^2}{N} + \frac{d\sum_{i=1}^N (\alpha_i)^2}{N^2}
\end{aligned}$$

Applying equation 4.23 from Bottou et al. (2018) with $\beta \to \frac{4}{\mu}, \eta \to \eta, \mu \to 1, M_V \to M_V/N, M \to \frac{M}{N} + \frac{d\sum_{i=1}^N (\alpha_t^i)^2}{N^2}, M_G \to M_V/N + 1$, we get

$$\mathbf{E}(f(\boldsymbol{w}^{t+1}) - f(\boldsymbol{w}^*)) \leq \left(1 - \frac{4}{\eta + t}\right)\mathbf{E}\big(f(\boldsymbol{w}^t) - f(\boldsymbol{w}^*)\big) + \frac{8L\left(\frac{M}{N} + \frac{d\sum_{i=1}^N (\alpha_i)^2}{N^2}\right)}{\mu^2(\eta+t)^2}$$

for any $t \geq 1$. To complete the proof, we will use the following version of a classic result by Chung:

Lemma C.2 (Modified from Chung (1954)). *Let $\{b_n\}_{n\geq 1}$ be a sequence of real numbers, such that for some $n_0 \in \mathbb{N}$, it holds that for all $n \geq n_0$,*

$$b_{n+1} \leq \left(1 - \frac{c}{n}\right) b_n + \frac{c_1}{n^2},$$

for some integer $c > 1$ and some real-valued constant $c_1 > 0$. Then

$$b_n \leq \frac{c_1}{c-1}\frac{1}{n} + \mathrm{O}\left(\frac{1}{n^c}\right).$$

Proof. Note that

$$\begin{aligned}
\frac{c_1}{c-1}\left(\frac{1}{n+1} - \left(1 - \frac{c}{n}\right)\frac{1}{n}\right) &= \frac{c_1}{c-1}\frac{n^2 - n(n+1) + c(n+1)}{n^2(n+1)} \\
&= \frac{c_1}{c-1}\frac{cn - n + c}{n^2(n+1)} \\
&> \frac{c_1}{c-1}\frac{n(c-1) + (c-1)}{n^2(n+1)} \\
&\geq \frac{c_1}{n^2}
\end{aligned}$$

Therefore, for any $n \geq n_0$,

$$b_{n+1} \leq \left(1 - \frac{c}{n}\right) b_n + \frac{c_1}{c-1}\left(\frac{1}{n+1} - \left(1 - \frac{c}{n}\right)\frac{1}{n}\right),$$

so that

$$b_{n+1} - \frac{c_1}{c-1}\frac{1}{n+1} \leq \left(1 - \frac{c}{n}\right)\left(b_n - \frac{c_1}{c-1}\frac{1}{n}\right),$$

Denote by $b'_n = b_n - \frac{c_1}{c-1}\frac{1}{n}$, so that $b'_{n+1} \leq \left(1 - \frac{c}{n}\right) b'_n$ for all $n \geq n_0$. If for some $n_1 > c$, $b'_{n_1} \leq 0$, then the same holds for any $n \geq n_1$, i.e.,

$$b_n \leq \frac{c_1}{c-1}\frac{1}{n},$$

whenever $n \geq n_1$. Otherwise, it holds that for all $n \geq c+1$,

$$0 < b'_n \leq b'_{c+1} \prod_{m=c+1}^{n-1}\left(1 - \frac{c}{m}\right) = b'_{c+1}\frac{1}{\binom{n-1}{c}} = \mathrm{O}\left(\frac{1}{n^c}\right).$$

□

Let $x_{t+\eta} \coloneqq \mathbf{E}\left(f(\boldsymbol{w}^t) - f^*\right)$ for $t \geq 1$ and $x_k \coloneqq \mathbf{E}\left(f(\boldsymbol{w}^0) - f^*\right)$ for $k < \eta + 1$. Using the inequality above, we get that for all $k \geq \eta + 1$

$$x_{k+1} \leq \left(1 - \frac{4}{k}\right) x_k + \frac{8L\left(\frac{M}{N} + \frac{d\sum_{i=1}^N (\alpha_i)^2}{N^2}\right)}{\mu^2 k^2} \implies x_t \leq \frac{8L\left(\frac{M}{N} + \frac{d\sum_{i=1}^N (\alpha_i)^2}{N^2}\right)}{3\mu^2 t} + \mathrm{O}\left(\frac{1}{t^4}\right).$$

Finally, the event that for some t we have $\Pi_W(\boldsymbol{w}^{t-1} - \eta^t \boldsymbol{g}^t) \neq \boldsymbol{w}^{t-1} - \eta^t \boldsymbol{g}^t$ happens with probability $\mathrm{O}(\frac{1}{Nt})$ and the loss $f(\boldsymbol{w}) - f^*$ is bounded by a constant by assumption. Therefore, an additional term of $\mathrm{O}\left(\frac{1}{Nt}\right)$ appears, which completes the proof. □

We would assume that the clients will use the following accuracy loss that corresponds to our upper bound:

$$\texttt{err}_i^2 = \frac{8L\left(\frac{M}{N} + \frac{d\sum_{i=1}^N (\alpha_i)^2}{N^2}\right)}{3\mu^2 n} + \frac{C}{Nn},$$

where C comes from the the $\mathrm{O}\left(\frac{1}{Nt}\right)$ in the theorem. The $\mathrm{O}\left(\frac{1}{t^4}\right)$-term is ignored as we consider a cross-silo setup, where n is expected to be large.

C.2.2 Privacy guarantees

We use Theorem 4.4, leading to the following bound for the privacy loss:

$$\texttt{leak}_i = \frac{16\sqrt{2\mathrm{e}\ln(1.25n^3)\ln(4n^2)}B}{\sqrt{n}\alpha_i}.$$

C.3 Utility Analysis

We would get the following system of participation constraints,

$$u_i - u_i^0 = -\frac{16\sqrt{2\mathrm{e}\ln(1.25n^3)\ln(4n^2)}B}{\sqrt{n}\alpha_i} + \lambda_i\left(\frac{8LM + 3\mu^2 C}{3\mu^2 n} - \frac{8L\left(\frac{M}{N} + \frac{d\sum_{i=1}^N (\alpha_i)^2}{N^2}\right)}{3\mu^2 n} - \frac{C}{Nt}\right).$$

As we can see, the privacy loss is of order $\frac{1}{\sqrt{n}}$ and the accuracy benefit is bounded by $\mathrm{O}\left(\frac{1}{n}\right)$. This discrepancy suggests that we should choose the α to be of order $\sqrt{n}$, but it will eliminate the accuracy benefits if N is bounded. Therefore, we do not expect this system to have solutions in the limit $n \to \infty$, $N = \mathrm{O}(1)$. On the other hand, in the limit $N \to \infty$, we can see that the contribution of α_i disappears, which suggests that the system will have solutions.