entropy

MDPI

Article

An Information-Theoretic Analysis of the Cost of Decentralization for Learning and Inference under Privacy Constraints

Sharu Theresa Jose * and Osvaldo Simeone

Department of Engineering, King's College London, London WC2R 2LS, UK; osvaldo.simeone@kcl.ac.uk
* Correspondence: sharu.jose@kcl.ac.uk

Abstract: In vertical federated learning (FL), the features of a data sample are distributed across multiple agents. As such, inter-agent collaboration can be beneficial not only during the learning phase, as is the case for standard horizontal FL, but also during the inference phase. A fundamental theoretical question in this setting is how to quantify the cost, or performance loss, of decentralization for learning and/or inference. In this paper, we study general supervised learning problems with any number of agents, and provide a novel information-theoretic quantification of the cost of decentralization in the presence of privacy constraints on inter-agent communication within a Bayesian framework. The cost of decentralization for learning and/or inference is shown to be quantified in terms of conditional mutual information terms involving features and label variables.

Keywords: vertical federated learning; Bayesian learning; information-theoretic analysis

Citation: Jose, S.T.; Simeone, O. An Information-Theoretic Analysis of the Cost of Decentralization for Learning and Inference under Privacy Constraints. *Entropy* **2022**, *24*, 485. https://doi.org/10.3390/e24040485

Academic Editors: Sergio Cruces, Iván Durán-Díaz, Rubén Martín-Clemente and Andrzej Cichocki

Received: 28 February 2022
Accepted: 28 March 2022
Published: 30 March 2022

1. Introduction

Consider a digital bank interested in building a prediction model for credit scoring based on data features of given individuals, such as saving information and spending habits, that are distributed across other banks, fintech companies, and online retail shops (see Figure 1). Data labels indicating loan approval or rejection reside at a trusted third-party credit bureau, which keeps track of the approved loans [1]. This setting exemplifies vertical federated learning (FL), in which data features are scattered across different participating agents, with data barriers between them preventing a direct exchange of information.

Unlike conventional horizontal FL, in which agents have independent data points, in vertical FL settings, inter-agent collaboration can be beneficial not only during the learning phase but also during the inference phase [2,3]. It is therefore important to understand at a fundamental theoretical level whether decentralization, wherein agents use only local data for learning and/or inference, entails a significant performance loss as compared to collaborative learning and/or inference. This is the subject of this paper.

As a first attempt in this direction, Chen et al. [3] address this problem by studying a binary classification problem in which each class corresponds to a bivariate Gaussian distribution over two input features, which are vertically distributed between two agents. The authors identify four collaboration settings depending on whether collaboration is done during learning and/or inference phases as collaborative learning–collaborative inference (CL/CI), collaborative learning–decentralized inference (CL/DI), decentralized learning–collaborative inference (DL/CI), and decentralized learning–decentralized inference (DL/DI). By taking a frequentist approach, the authors compare the classification error rates achieved under these four settings.

In this work, inspired by [3], we develop a novel *information-theoretic* approach to quantify the cost of decentralization for *general* supervised learning problems with *any* number of agents and under *privacy* constraints. Specifically, we consider a supervised learning problem defined by an arbitrary joint distribution $P_{\mathbf{X},Y|W}$ involving the feature

vector $\mathbf{X}$ and label Y, with the feature vector vertically partitioned between any number of local agents. A trusted central server, also called a data scientist or aggregator [4], holds the labels, which it shares with the agents upon request (see Figure 1). The agents collaborate through the aggregator during learning and/or inference. To limit the information leakage from the shared feature to an adversarial eavesdropper, unlike [3], privacy constraints are imposed on the aggregation mapping. By adopting a Bayesian framework, we characterize the average predictive performance of the four settings—CL/CI, CL/DI, DL/CI, and DL/DI—under privacy constraints via information-theoretic metrics. Finally, we illustrate the relation between the four collaboration settings with/without privacy constraints on two numerical examples.

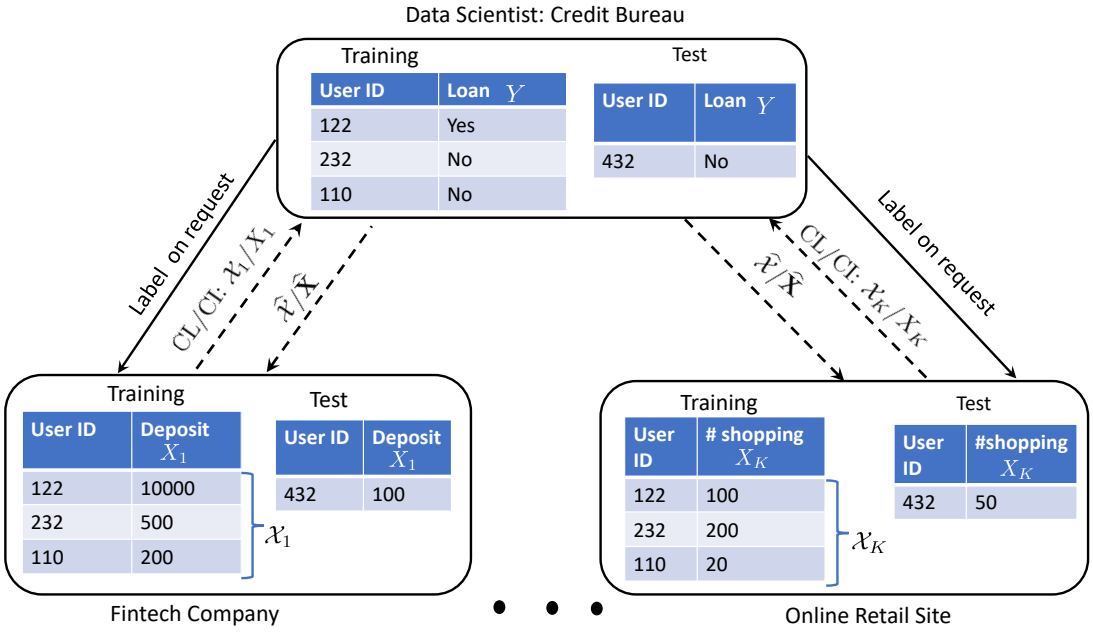

Figure 1. Illustration of the vertical federated learning (FL) setup under study for a prototypical credit scoring application.

In line with the recent works of [5,6], this work relates information-theoretic measures to learning centric performance metrics with the goal of providing theoretical insights. Specifically, we leverage information-theoretic tools to gain insights into the performance degradation resulting from decentralized learning and/or inference for general supervised learning problems. The main contribution is hence of theoretical nature, as it provides a connection between information-theoretic metrics and practically relevant measures of generalization in decentralized Bayesian learning and inference.

2. Problem Formulation

Setting: We study a vertical federated learning (FL) setting with K agents that can cooperate during the learning and/or inference phases of operation of the system. Our main goal is to quantify, using information-theoretic metrics, the benefits of cooperation for learning and/or inference. We focus on a supervised learning problem, in which each data point corresponds to a tuple $(\mathbf{X}, Y)$ encompassing the K-dimensional feature vector $\mathbf{X} = (X_1, \ldots, X_K)$ and the scalar output label Y. As illustrated in Figure 1, each kth feature X_k in vector $\mathbf{X}$ is observed only by the kth agent. A trusted central server, referred to as the aggregator, holds the output label Y, which it shares with the agents on request [4,7]. Features and labels can take values in arbitrary alphabets. The unknown data distribution is assumed to belong to a model class $\{P_{\mathbf{X},Y|W} : W \in \mathscr{W}\}$ of joint distributions that are identified by a model parameter vector W taking values in some space $\mathscr{W}$. Adopting a Bayesian approach, we endow the model parameter vector with a prior distribution P_W.

As illustrated in Figure 1, let $(\mathcal{X}, \mathcal{Y}) = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_N, Y_N)\}$ denote a training data set of N labelled samples, which, when conditioned on model parameter W, are assumed to be generated i.i.d. according to distribution $P_{\mathbf{X},Y|W}$. The $N \times K$ matrix $\mathcal{X}$ collects the K-dimensional feature vectors $\{\mathbf{X}_n\}_{n=1}^N$ by rows. We denote as $X_{n,k}$, the (n,k)th element of matrix $\mathcal{X}$, for $n = 1, \ldots, N$, and $k = 1, \ldots, K$; and as $\mathcal{X}_k = [X_{1,k}, \ldots, X_{N,k}]^T$ ($[\cdot]^T$ is the transpose operation), the kth column of the data matrix, which corresponds to the observations of agent k. The goal of the system is to use the training data set $(\mathcal{X}, \mathcal{Y})$ to infer the model parameter W, which enable the agents to predict the label of a new, previously unseen, test feature input $\mathbf{X}$. The joint distribution of model parameter W, training data $(\mathcal{X}, \mathcal{Y})$, and test data $(\mathbf{X}, Y)$ can be written as follows ([8], Chapter 3.3):

$$P_{W,\mathcal{X},\mathcal{Y},\mathbf{X},Y} = P_W \otimes_{i=1}^N \underbrace{(P_{\mathcal{X}_{i,1},\ldots,\mathcal{X}_{i,K},\mathcal{Y}_i|W})}_{\text{training}} \otimes \underbrace{P_{X_1,\ldots,X_K,Y|W}}_{\text{testing}}, \tag{1}$$

with $\otimes$ representing the product of distributions, and conditional distribution $P_{\mathcal{X}_{i,1},\ldots,\mathcal{X}_{i,K},\mathcal{Y}_i|W}$ being equal to $P_{X_1,\ldots,X_K,Y|W}$ for $i = 1, \ldots, N$.

Collaborative/decentralized learning/inference: In the learning phase, training data is used to infer the model parameter W, enabling the agents in the inference phase to make predictions about test label Y given the test feature vector $\mathbf{X}$ based on the model $P_{\mathbf{X},Y|W}$. Either or both learning and inference phases can be carried out collaboratively by the agents or in a decentralized fashion (i.e., separately by each agent). When collaborating for learning or inference, the K agents share their locally observed feature data via the aggregator. The operation of the aggregator is modelled as a stochastic aggregation mapping $P_{\widehat{\mathbf{X}}|X_1,\ldots,X_K} = P_{\widehat{\mathbf{X}}|\mathbf{X}}$ from the input K local features to an output shared feature $\widehat{\mathbf{X}}$, to be used by each of the K local agents. As detailed next, for learning, the mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ is applied independently to each data point. Furthermore, as we also detail later in this section, we impose privacy constraints on the aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ so that the shared feature $\widehat{\mathbf{X}}$ does not reveal too much information about the local agents' features.

We specifically distinguish the following four settings:

- *Collaborative learning–collaborative inference* (CL/CI): Agents collaborate during both learning and inference phases by sharing information about their respective features. Accordingly, during learning, each agent has access to the shared training data features $\widehat{\mathcal{X}} = (\widehat{\mathbf{X}}_1, \ldots, \widehat{\mathbf{X}}_N)$, where each nth component $\widehat{\mathbf{X}}_n \sim P_{\widehat{\mathbf{X}}|\mathbf{X}=\mathbf{X}_n}$ is generated independently by the aggregator in response to the observed feature vector $\mathbf{X}_n$, in addition to its own observed local feature data $\mathcal{X}_k$. Furthermore, during inference, agent k can use the shared test feature $\widehat{\mathbf{X}} \sim P_{\widehat{\mathbf{X}}|\mathbf{X}=\mathbf{X}}$, obtained by aggregating the test feature vector $\mathbf{X}$, in addition to its own observation X_k, in order to predict the test label Y.
- *Collaborative learning–decentralized inference* (CL/DI): Agents collaborate only during learning by sharing information about their respective features as explained above, while inference is decentralized. Accordingly, during inference, each kth agent uses the kth feature X_k of test feature vector $\mathbf{X}$ in order to predict the test label Y.
- *Decentralized learning–collaborative inference* (DL/CI): Agents collaborate for inference, while each kth agent is allowed to use only its observed training data $\mathcal{X}_k$, along with the labels $\mathcal{Y}$ shared by the aggregator, during learning.
- *Decentralized learning–decentralized inference* (DL/DI): Agents operate independently, with no cooperation in either learning or inference phases.

Privacy constraints: The aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ shares the output feature $\widehat{\mathbf{X}}$ with each of the K local agents during collaborative learning and/or inference. To account for privacy constraints concerning agents' data, we limit the amount of information that a "curious" eavesdropper may be able to obtain about the local features' data from observing $\widehat{\mathbf{X}}$. To this end, we impose the following privacy constraint on the aggregation mapping so that the shared feature $\widehat{\mathbf{X}}$ does not leak too much information about the local features X_k of all agents $k = 1, \ldots, K$.

The aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ is said to be ϵ- individually private if

$$I(\widehat{\mathbf{X}}; X_k|X^{(-k)}) \leq \epsilon, \quad \text{for all } \ k = 1, \ldots, K, \tag{2}$$

where $X^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K)$ and

$$I(\widehat{\mathbf{X}}; X_k|X^{(-k)}) = \mathbb{E}_{P_{\widehat{\mathbf{X}}\mathbf{X}}}\left[\log \frac{P_{\widehat{\mathbf{X}},X_k|X^{(-k)}}}{P_{\widehat{\mathbf{X}}|X^{(-k)}} P_{X_k|X^{(-k)}}}\right]$$

is the conditional mutual information under the joint distribution $P_{\widehat{\mathbf{X}},\mathbf{X}} = P_{\mathbf{X}} P_{\widehat{\mathbf{X}}|\mathbf{X}}$, with $P_{\mathbf{X}}$ being the marginal of $P_{\mathbf{X},Y,W}$. The constraint (2) measures privacy against a strong eavesdropper that knows all features except the kth feature X_k. Specifically, the conditional mutual information $I(\widehat{\mathbf{X}}; X_k|X^{(-k)})$ quantifies the additional information about X_k gained by the eavesdropper upon observing the shared feature $\widehat{\mathbf{X}}$. As such, the metric is also relevant as a privacy measure against "curious" agents.

We note that although the privacy constraint in (2) bears a resemblance to the MI-differential privacy (MI-DP) constraint introduced in [9], the condition (2) does not have the same operational meaning. In fact, the MI-DP constraint in [9,10] or the f-divergence-based DP constraint in [11] ensure differential privacy for individual i.i.d. data samples of a training data set, and they rely on a mechanism that applies to the entire data set during learning. In contrast, the constraint (2) accounts for the privacy of correlated local features via a per-sample masking mechanism, and it applies to both learning and inference phases.

Predictive loss under privacy constraints: In all the four settings described above, any agent k uses the available training data $(\widetilde{\mathcal{X}}_k, \mathcal{Y})$, with $\widetilde{\mathcal{X}}_k$ being equal to $\mathcal{X}_k$ for decentralized learning and to $(\mathcal{X}_k, \widehat{\mathcal{X}})$ for collaborative learning, in order to infer the model parameter W. The inferred model is then used to predict the label Y given the test feature input $\widetilde{\mathbf{X}}_k$, with $\widetilde{\mathbf{X}}_k$ being equal to X_k for decentralized inference and to $(X_k, \widehat{\mathbf{X}})$ for collaborative learning. We impose that the aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ must satisfy the privacy constraint in (2).

The joint operation of learning and inference at agent k can be accordingly described via a stochastic predictive distribution $Q_{Y|\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k}$ on the test label Y given the training data $(\widetilde{\mathcal{X}}_k, \mathcal{Y})$ and test feature input $\widetilde{\mathbf{X}}_k$. The predictive distribution can be thought of as the result of a two-step application of learning and inference, where a model parameter is first learned using the input training data $(\widetilde{\mathcal{X}}_k, \mathcal{Y})$ and is subsequently used to infer the label corresponding to the test feature input $\widetilde{\mathbf{X}}_k$. Note that this stochastic mapping can account for arbitrary choices of learning and inference algorithms. By optimizing over aggregation mapping as well as over learning and inference algorithms, we define the ϵ-private predictive loss as

$$\begin{aligned}\mathcal{R}(\epsilon) = \min_{\substack{P_{\widehat{\mathbf{X}}|\mathbf{X}} \\ \in \mathcal{P}(\widehat{\mathbf{X}}|\mathbf{X})}} \max_{k=1,\ldots,K} \min_{\substack{Q_{Y|\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k} \\ \in \mathcal{Q}(Y|\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k)}} & \mathbb{E}_{P_{Y,\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k}}\left[-\log Q_{Y|\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k}\right] \\ \text{s.t } \ I(\widehat{\mathbf{X}}; X_k|X^{(-k)}) \leq \epsilon \quad & \text{for all } \ k = 1, \ldots, K. \end{aligned} \tag{3}$$

In (3), the aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ is optimized over some specified family $\mathcal{P}(\widehat{\mathbf{X}}|\mathbf{X})$ of conditional distributions $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ in order to minimize the worst-case predictive loss across the agents under constraint (2). Furthermore, the inner optimization is over a class of predictive distributions $\mathcal{Q}(Y|\widetilde{\mathcal{X}}_k, \mathcal{Y}, \widetilde{\mathbf{X}}_k)$.

In the absence of privacy constraints (i.e., when $\epsilon = \infty$), assuming that the distribution family $\mathcal{P}(\widehat{\mathbf{X}}|\mathbf{X})$ is sufficiently large, the optimal aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ puts its entire mass on the output shared feature $\widehat{\mathbf{X}} = \mathbf{X}$. As such, under collaborative learning, each

agent k uses the entire feature data (i.e., $\widetilde{\mathcal{X}}_k = \mathcal{X}$) , and under collaborative inference, it uses the entire test feature vector $\widetilde{\mathbf{X}}_k = \mathbf{X}$. The predictive loss (3) in the absence of privacy constraints is evaluated as

$$\mathcal{R}(\infty) = \max_{k=1,\ldots,K} \min_{\substack{Q_{Y|\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k} \\ \in \mathcal{Q}(Y|\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k)}} \mathbb{E}_{P_{Y,\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k}} \left[-\log Q_{Y|\widetilde{\mathcal{X}}_k,\mathcal{Y},\widetilde{\mathbf{X}}_k} \right]. \tag{4}$$

The predictive loss (4) represents the worst-case minimum average cross-entropy loss across all agents, which can be obtained given the information about the training data set and the test input feature [5].

3. Preliminaries and Fully Collaborative Benchmark

In this section, we first provide a brief explanation of the main information-theoretic metrics used in this work. Then, we define and derive the average predictive loss for the benchmark case in which both learning and inference are collaborative.

Information-theoretic metrics: Let A and B denote two (discrete or continuous) random variables with joint distribution $P_{A,B}$, and with corresponding marginals P_A and P_B. The joint entropy of A and B, denoted $H(A,B)$, is defined as $H(A,B) = \mathbb{E}_{P_{A,B}}[-\log P_{A,B}]$, with $\mathbb{E}_P[\cdot]$ denoting the expectation with respect to distribution P. More generally, the conditional entropy of A given B is defined as $H(A|B) = \mathbb{E}_{P_{A,B}}[-\log P_{A|B}]$, where $P_{A|B} = P_{A,B}/P_B$ is the conditional distribution of A given B. By the chain rule, we have the relationship $H(A,B) = H(B) + H(A|B)$; we also have the property that conditioning does not increase entropy [12] (i.e., $H(A|B) \leq H(A)$). The mutual information $I(A;B)$ between the random variables is defined as $I(A;B) = \mathbb{E}_{P_{A,B}}\left[\log\left(\frac{P_{A,B}}{P_A P_B}\right)\right]$. Finally, for random variables A, B, and C with joint distribution $P_{A,B,C}$, the conditional mutual information $I(A;B|C)$ between A and B given C is defined as $I(A;B|C) = \mathbb{E}_{P_{A,B,C}}\left[\log\left(\frac{P_{A,B|C}}{P_{A|C}P_{B|C}}\right)\right]$.

Private collaborative learning–collaborative inference (CL/CI): As a benchmark, we now study the predictive loss (3) for the CL/CI setting. The ϵ-private predictive loss (3) of CL/CI is given as

$$\mathcal{R}^{\mathrm{CL/CI}}(\epsilon) = \min_{P_{\widehat{\mathbf{X}}|\mathbf{X}} \in \mathcal{F}(\widehat{\mathbf{X}}|\mathbf{X})} \max_{k=1,\ldots,K} \min_{\substack{Q_{Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k} \\ \in \mathcal{Q}(Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k)}} \mathbb{E}_{P_{Y,\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k}} \left[-\log Q_{Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k} \right] \tag{5}$$

where

$$\mathcal{F}(\widehat{\mathbf{X}}|\mathbf{X}) = \{P_{\widehat{\mathbf{X}}|\mathbf{X}} \in \mathcal{P}(\widehat{\mathbf{X}}|\mathbf{X}) : \text{constraint (2) holds}\} \tag{6}$$

is the feasible space of conditional distributions satisfying the privacy constraint (2). The following lemma presents an information-theoretic characterization of the loss $\mathcal{R}^{\mathrm{CL/CI}}(\epsilon)$.

Lemma 1. *Assume that the family $\mathcal{Q}(Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k)$ comprises the set of all predictive distributions $Q_{Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k}$. Then, the ϵ-private predictive loss (5) for the CL/CI setting evaluates as*

$$\mathcal{R}^{\mathrm{CL/CI}}(\epsilon) = \min_{\substack{P_{\widehat{\mathbf{X}}|\mathbf{X}} \\ \in \mathcal{F}(\widehat{\mathbf{X}}|\mathbf{X})}} \max_{k=1,\ldots,K} H(Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k). \tag{7}$$

In addition, if $\epsilon = \infty$, and $\mathcal{P}(\widehat{\mathbf{X}}|\mathbf{X})$ includes the space of all conditional distributions $P_{\widehat{\mathbf{X}}|\mathbf{X}}$, then the predictive loss (4) in the absence of privacy constraints for CL/CI is evaluated as

$$\mathcal{R}^{\mathrm{CL/CI}}(\infty) = H(Y|\mathbf{X},\mathcal{X},\mathcal{Y}). \tag{8}$$

Proof. For a fixed aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$, and an agent k, the predictive distribution that minimizes the inner cross entropy term in (5), $\mathbb{E}_{P_{Y,\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k}}\left[-\log Q_{Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k}\right]$, is the posterior distribution, $P_{Y|\widehat{\mathcal{X}},\mathcal{X}_k,\mathcal{Y},\widehat{\mathbf{X}},X_k}$ [12], resulting in the conditional entropy term in (7). When $\epsilon = \infty$ and $\mathcal{P}(\widehat{\mathbf{X}}|\mathbf{X})$ includes the space of all conditional distributions, we have $\widehat{\mathcal{X}} = \mathcal{X}$ and $\widehat{\mathbf{X}} = \mathbf{X}$, yielding (8). □

4. Cost of Decentralization Under Privacy Constraints

In this section, we use the benchmark predictive loss (7) observed under the ideal CL/CI setting to evaluate the cost of decentralization in the learning and/or inference phases under privacy constraints.

Lemma 2. *The ϵ-private predictive losses of decentralized learning and/or inference are given as*

$$\mathcal{R}^{\mathrm{CL/DI}}(\epsilon) = \min_{P_{\widehat{\mathbf{X}}|\mathbf{X}} \in \mathcal{F}(\widehat{\mathbf{X}}|\mathbf{X})} \max_{k=1,\ldots,K} H(Y|X_k, \mathcal{X}_k, \widehat{\mathcal{X}}, \mathcal{Y}) \tag{9}$$

$$\mathcal{R}^{\mathrm{DL/CI}}(\epsilon) = \min_{P_{\widehat{\mathbf{X}}|\mathbf{X}} \in \mathcal{F}(\widehat{\mathbf{X}}|\mathbf{X})} \max_{k=1,\ldots,K} H(Y|X_k, \widehat{\mathbf{X}}, \mathcal{X}_k, \mathcal{Y}) \tag{10}$$

$$\mathcal{R}^{\mathrm{DL/DI}}(\epsilon) = \max_{k=1,\ldots,K} H(Y|X_k, \mathcal{X}_k, \mathcal{Y}), \tag{11}$$

where set $\mathcal{F}(\widehat{\mathbf{X}}|\mathbf{X})$ is as defined in (6).

Proof. The result is a direct extension of Lemma 1 to CL/DI, DL/CI, and DL/DI. □

Note that the predictive loss (11) of the fully decentralized DL/DI setting does not depend on the privacy parameter ϵ, since decentralization does not entail any privacy loss. Therefore, in the absence of privacy constraints, we have $\mathcal{R}^{\mathrm{DL/DI}}(\infty) = \mathcal{R}^{\mathrm{DL/DI}}(\epsilon)$, while the predictive losses in (9)–(10) evaluate as

$$\mathcal{R}^{\mathrm{CL/DI}}(\infty) = \max_{k=1,\ldots,K} H(Y|X_k, \mathcal{X}, \mathcal{Y}), \tag{12}$$

$$\mathcal{R}^{\mathrm{DL/CI}}(\infty) = \max_{k=1,\ldots,K} H(Y|\mathbf{X}, \mathcal{X}_k, \mathcal{Y}), \tag{13}$$

under the assumption of sufficiently large $\mathcal{P}(\widehat{\mathbf{X}}|\mathbf{X})$. Furthermore, using the property that conditioning does not increase entropy [8] results in the following relation between the predictive losses of the four schemes—CL/CI, CL/DI, DL/CI and DL/DI—in the absence of privacy constraints:

$$\begin{aligned} \mathcal{R}^{\mathrm{CL/CI}}(\infty) &\leq \min\{\mathcal{R}^{\mathrm{CL/DI}}(\infty), \mathcal{R}^{\mathrm{DL/CI}}(\infty)\} \\ &\leq \max\{\mathcal{R}^{\mathrm{CL/DI}}(\infty), \mathcal{R}^{\mathrm{DL/CI}}(\infty)\} \\ &\leq \mathcal{R}^{\mathrm{DL/DI}}(\infty). \end{aligned} \tag{14}$$

The difference between the ϵ-private predictive risks of the decentralized and collaborative schemes captures the *cost of decentralization*. Specifically, given two schemes $a, b \in \{\mathrm{CL/CI}, \mathrm{CL/DI}, \mathrm{DL/CI}, \mathrm{DL/DI}\}$ such that $\mathcal{R}^a(\epsilon) \geq \mathcal{R}^b(\epsilon)$, we define the cost of a with respect to b as

$$\mathcal{C}^{a-b}(\epsilon) = \mathcal{R}^a(\epsilon) - \mathcal{R}^b(\epsilon). \tag{15}$$

In the absence of privacy constraints ($\epsilon = \infty$) and assuming symmetric agents so that the maximum in (4) is attained for any $k = 1, \ldots, K$, the cost of decentralization can be exactly characterized as in the following result.

Proposition 1. *The cost of decentralization* (15) *for* $\epsilon = \infty$ *and symmetric agents can be characterized for the kth learning agent as detailed in Table 1, where* $X^{(-k)} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_K)$ *and* $\mathcal{X}^{(-k)} = (\mathcal{X}_1, \ldots, \mathcal{X}_{k-1}, \mathcal{X}_{k+1}, \ldots, \mathcal{X}_K)$.

Proof. We illustrate the derivation of the cost of decentralization between CL/DI and CL/CI, as the proof can be similarly completed. In the absence of privacy constraints and assuming symmetric agents, we have from (8) and (12), $\mathcal{C}^{\text{CL/DI}-\text{CL/CI}}(\infty) = H(Y|X_k, \mathcal{X}, \mathcal{Y}) - H(Y|\mathbf{X}, \mathcal{X}, \mathcal{Y}) = I(Y; X^{(-k)}|X_k, \mathcal{X}, \mathcal{Y})$. □

The results in Table 1 have intuitive interpretations. For instance, the cost $\mathcal{C}^{\text{CL/DI}-\text{CL/CI}}(\infty) = I(Y; X^{(-k)}|X_k, \mathcal{X}, \mathcal{Y})$ corresponds to the additional information about label Y that can be obtained from observing the features $X^{(-k)}$ of other agents, given $\mathcal{X}, \mathcal{Y}$, and X_k. Examples will be provided in the next section in which the cost of decentralization is evaluated also in the presence of privacy constraints based on (7), (9)–(11).

Table 1. Cost of decentralization $\mathcal{C}^{a-b}(\infty)$ (a defines the column and b the row).

	CL/CI	**CL/DI**	**DL/CI**	**DL/DI**
CL/CI	–	$I(Y; X^{(-k)}\|X_k, \mathcal{X}, \mathcal{Y})$	$I(Y; \mathcal{X}^{(-k)}\|\mathbf{X}, \mathcal{X}_k, \mathcal{Y})$	$I(Y; X^{(-k)}, \mathcal{X}^{(-k)}\|X_k, \mathcal{X}_k, \mathcal{Y})$
CL/DI	–	–		$I(Y; \mathcal{X}^{(-k)}\|\mathcal{X}_k, X_k, \mathcal{Y})$
DL/CI	–		–	$I(Y; X^{(-k)}\|X_k, \mathcal{X}_k, \mathcal{Y})$
DL/DI	–	–	–	–

5. Examples and Remarks

In this section, we consider two simple numerical examples to illustrate the cost of decentralization for learning and/or inference with and without the privacy constraints that were quantified in Section 4 for general models. We note that evaluating the derived metrics for real-world examples would generally require the implementation of mutual information estimators, and is left for future work.

5.1. Two-Agent Non-Private Collaborative Learning (CL) and/or Inference (CI)

Consider two agents ($K = 2$) observing binary joint features $X_1, X_2 \in \{0, 1\}$, which have the joint distribution defined by the probability r of the two features X_1 and X_2 being equal, that is, $\Pr[X_1 = X_2|X_2 = x_2] = r/2$, with $\Pr[X_1 = 1] = \Pr[X_2 = 1] = 0.5$. Parameter r quantifies the statistical dependencies between features X_1 and X_2 through the MI $I(X_1; X_2) = \log 2 - H_b(r)$, where $H_b(r) = -r \log(r) - (1 - r) \log(1 - r)$ denotes the binary entropy with parameter r. Note that the MI takes the maximum value of $I(X_1; X_2) = 1$ when $r = 0$ or 1, and the minimum value of $I(X_1; X_2) = 0$ when $r = 0.5$. The output binary label $Y \in \{0, 1\}$ depends on the feature vector $\mathbf{X}$ through the model

$$P_{Y=1|X_1,X_2,W} = \begin{cases} W_1 & \text{if } X_1 \oplus X_2 = 0 \\ W_2 & \text{if } X_1 \oplus X_2 = 1 \end{cases}, \tag{16}$$

with model parameters $W = (W_1, W_2)$, where $\{W_1, W_2\} \in [0, 1]$. Accordingly, W_1 and W_2 are the probabilities of the event $Y = 1$ when X_1 and X_2 are equal or different, respectively. We assume that the model parameters are a priori independent and distributed according to beta distributions ([8], Section 2.4.2) as $P_{W_1,W_2} = \text{Beta}(W_1|\alpha_1, \beta_1)\text{Beta}(W_2|\alpha_2, \beta_2)$, where $\alpha_1, \beta_1, \alpha_2, \beta_2 > 0$ are fixed hyperparameters.

Figure 2 compares the predictive loss derived in Lemma 2 with no privacy constraints ($\epsilon = \infty$) under the four schemes—CL/CI, CL/DI, DL/CI and DL/DI—as a function of the mutual information $I(X_1; X_2)$ between the components of the bivariate feature vector. The number of data samples is $N = 3$, and other hyperparameters are set to $\alpha_1 = 2$, $\beta_1 = 1.5$, $\alpha_2 = 1.5$, and $\beta_2 = 2$. When the MI $I(X_1; X_2)$ is large, the predictive risks under

collaborative and decentralized schemes are similar, and the cost of decentralization is negligible. This is because a larger MI $I(X_1; X_2)$ implies that each local agent's feature X_k, for $k = 1, 2$, is highly informative about the local feature $X^{(-k)}$ of the other agent, and no significant additional information can be obtained via collaboration. This applies to both learning and inference phases. Conversely, when the MI is small, decentralization entails a significant cost. In this example, centralized inference is more effective than centralized learning due to the importance of having access to both X_1 and X_2 in order to infer Y by (16).

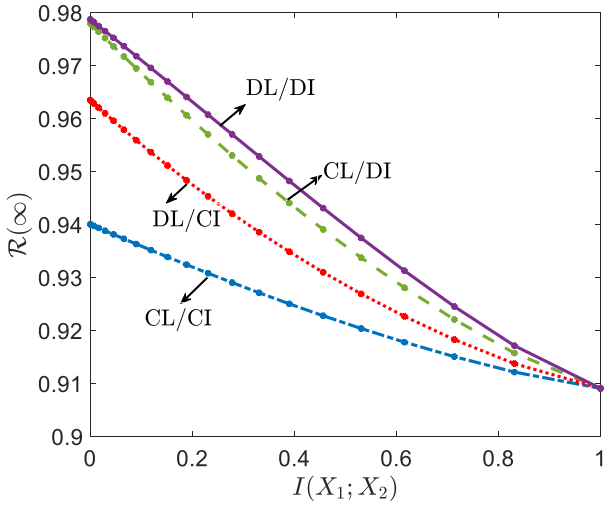

Figure 2. Predictive losses (7), (9)–(11) for the four schemes under no privacy constraints ($\epsilon = \infty$) as a function of the mutual information $I(X_1; X_2)$. ($\alpha_1 = 2$, $\beta_1 = 1.5$, $\alpha_2 = 1.5$, $\beta_2 = 2$, and $N = 3$).

5.2. Three-Agent Private CL and/or CI

We now extend the example in Section 5.1 by considering three agents ($K = 3$) and by imposing privacy constraints during collaboration in the learning and inference phases. The feature vector $\mathbf{X} = (X_1, X_2, X_3)$ consists of three binary features $X_k \in \{0, 1\}$ for $k = 1, 2, 3$, where X_1 and X_2 are distributed as in Section 5.1, and we have $\Pr[X_3|X_1 = x_1, X_2 = x_2] = \Pr[X_3|X_2 = x_2]$ with $\Pr[X_3 \neq X_2|X_2 = x_2] = 1 - r$. Generalizing the previous example, the output binary label $Y \in \{0, 1\}$ depends on the feature vector $\mathbf{X}$ through the model

$$P_{Y=1|\mathbf{X},W} = \begin{cases} W_1 & \text{if } X_1 \oplus X_2 \oplus X_3 = 0 \\ W_2 & \text{if } X_1 \oplus X_2 \oplus X_3 = 1 \end{cases}, \tag{17}$$

where model parameters have the same prior distribution. The aggregation mapping $P_{\widehat{\mathbf{X}}|\mathbf{X}}$ produces a binary random variable $\widehat{X} \in \{0, 1\}$ as $\widehat{X} = X_1 \oplus X_2 \oplus X_3 \oplus \xi$, with $\xi \sim \text{Bern}(s)$, where the noise variable $\xi \sim \text{Bern}(s)$ is chosen independently of the feature vector $\mathbf{X}$, and the parameter $s \in [0, 1]$ is selected so as to guarantee the privacy constraints in (2), which can be written as

$$\begin{aligned} \epsilon \geq \max\Big\{ & -H_b(s) + H_b(s(1-r) + r(1-s)), -H_b(s) + \\ & H_b(sr + (1-r)(1-s)), -H_b(s) + 2r(1-r)\log(2) \\ & + ((1-r)^2 + r^2) H_b\left(\frac{(1-r)^2 s + r^2(1-s)}{(1-r)^2 + r^2}\right)\Big\}. \end{aligned}$$

Figure 3 compares the predictive loss $\mathcal{R}(\epsilon)$ derived in Lemma 2 of the four schemes—CL/CI, CL/DI, DL/CI and DL/DI—as a function of the privacy parameter ϵ for fixed

$r = 0.5$. In the high-privacy regime, where ϵ is small, the shared feature $\widehat{X}$ is not informative about the local observed features, and collaborative learning/inference brings little benefit over the decentralized schemes. However, as ϵ increases, thereby weakening privacy requirements, the shared feature $\widehat{X}$ becomes more informative about the observed feature vector $\mathbf{X}$, and the cost of decentralization becomes increasingly significant, reaching its maximum value under no privacy (i.e., when $\epsilon = 1$).

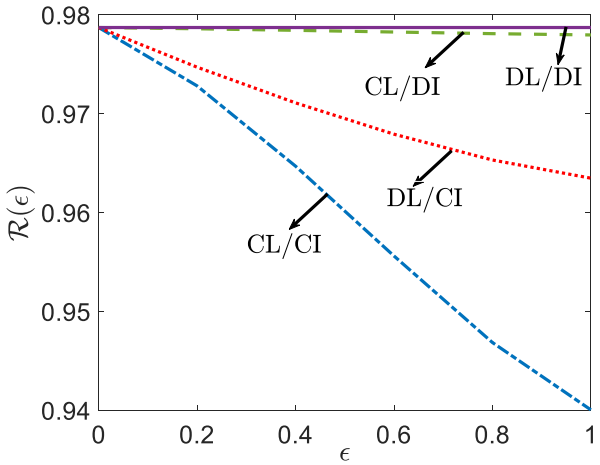

Figure 3. Predictive losses (7), (9)–(11) for the four schemes as a function of privacy measure ϵ. ($\alpha_1 = 2$, $\beta_1 = 1.5$, $\alpha_2 = 1.5$, $\beta_2 = 2$ and $N = 3$).

The examples studied in this section are simple enough to exactly evaluate the MI terms, but sufficiently rich to clearly demonstrate the cost of decentralization arising in the four collaboration settings of CL/CI, CL/DI, DL/CI, and DL/DI. They elucidate a simple vertical FL setting with features partitioned across agents and a discriminative model as given in (16).

6. Conclusions

This paper presents a novel information-theoretic characterization of the cost of decentralization during learning and/or inference in a vertical FL setting. Under privacy constraints on the aggregation mechanism that enables inter-agent communications, we show, by adopting a Bayesian framework, that the average predictive performance of the four schemes can be quantified in terms of conditional entropy terms. Furthermore, when no privacy constraints are imposed, the cost of decentralization for symmetric agents is shown to be exactly characterized by conditional mutual information terms.

The proposed information-theoretic framework is relevant for real-world vertical FL settings, such as credit scoring in banking [13], healthcare [14], and smart retailing. We leave the investigation of practical implications of the analysis via efficient MI estimators, such as the mutual information neural estimators (MINE) [15], to future research.

Author Contributions: Formal analysis, S.T.J. Supervision, O.S.; Writing—original draft, S.T.J. Writing—review & editing, S.T.J. and O.S. All authors have read and agreed to the published version of the manuscript.

Funding: The authors have received funding from the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 725731).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Luo, X.; Wu, Y.; Xiao, X.; Ooi, B.C. Feature inference attack on model predictions in vertical federated learning. In Proceedings of the 2021 IEEE 37th International Conference on Data Engineering (ICDE), Chania, Greece, 19–22 April 2021; pp. 181–192.
2. Verma, D.; Calo, S.; Witherspoon, S.; Bertino, E.; Jabal, A.A.; Swami, A.; Cirincione, G.; Julier, S.; White, G.; de Mel, G.; et al. Federated Learning for Coalition Operations. In Proceedings of the AAAI FSS-19: Artificial Intelligence in Government and Public Sector, Arlington, VA, USA, 7–8 November 2019.
3. Chen, Y.Z.J.; Towsley, D.; Verma, D. On Collaboration in Machine Learning. Available online: https://www.comsoc.org/publications/journals/ieee-tnse/cfp/collaborative-machine-learning-next-generation-intelligent (accessed on 27 March 2022).
4. Romanini, D.; Hall, A.J.; Papadopoulos, P.; Titcombe, T.; Ismail, A.; Cebere, T.; Sandmann, R.; Roehm, R.; Hoeh, M.A. Pyvertical: A vertical federated learning framework for multi-headed splitnn. *arXiv* **2021**, arXiv:2104.00489.
5. Xu, A.; Raginsky, M. Minimum Excess Risk in Bayesian Learning. *arXiv* **2020**, arXiv:2012.14868.
6. Hafez-Kolahi, H.; Moniri, B.; Kasaei, S.; Baghshah, M.S. Rate-Distortion Analysis of Minimum Excess Risk in Bayesian Learning. In Proceedings of the 38th International Conference on Machine Learning, Long Beach, CA, USA, 18–24 July 2021; Meila, M., Zhang, T., Eds.; PMLR: New York, NY, USA, 2021; Volume 139, pp. 3998–4007.
7. Cheng, K.; Fan, T.; Jin, Y.; Liu, Y.; Chen, T.; Papadopoulos, D.; Yang, Q. Secureboost: A lossless federated learning framework. *IEEE Intell. Syst.* **2021**, *36*, 87–98. [CrossRef]
8. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.
9. Cuff, P.; Yu, L. Differential privacy as a mutual information constraint. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, 24–28 October 2016; pp. 43–54.
10. Yagli, S.; Dytso, A.; Poor, H.V. Information-theoretic bounds on the generalization error and privacy leakage in federated learning. In Proceedings of the Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Atlanta, GA, USA, 26–29 May 2020; pp. 1–5.
11. Asoodeh, S.; Chen, W.N.; Calmon, F.P.; Özgür, A. Differentially private federated learning: An information-theoretic perspective. In Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Australia, 12–20 July 2021; pp. 344–349.
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
13. Zheng, F.; Li, K.; Tian, J.; Xiang, X. A vertical federated learning method for interpretable scorecard and its application in credit scoring. *arXiv* **2020**, arXiv:2009.06218.
14. Vepakomma, P.; Gupta, O.; Swedish, T.; Raskar, R. Split learning for health: Distributed deep learning without sharing raw patient data. *arXiv* **2018**, arXiv:1812.00564.
15. Belghazi, M.I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; Hjelm, D. Mutual information neural estimation. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 531–540.