Decentralized Machine Learning with Centralized Performance Guarantees via Gibbs Algorithms

Yaiza Bermudez*, Samir M. Perlaza*†§, and Iñaki Esnaola‡§
Emails: name.lastname@inria.fr and esnaola@sheffield.ac.uk
*Centre Inria d'Université Côte d'Azur, INRIA, Sophia Antipolis, France.
†Laboratoire GAATI, Université de la Polynésie française, Fa'a'ā, French Polynesia.
‡School of Electrical and Electronic Engineering, University of Sheffield, Sheffield, United Kingdom.
§ECE Dept. Princeton University, Princeton, 08544 NJ, USA.

***Abstract*—In this paper, it is shown, for the first time, that centralized performance is achievable in decentralized learning without sharing the local datasets. Specifically, when clients adopt an empirical risk minimization with relative-entropy regularization (ERM-RER) learning framework and a forward-backward communication between clients is established, it suffices to share the locally obtained Gibbs measures to achieve the same performance as that of a centralized ERM-RER with access to all the datasets. The core idea is that the Gibbs measure produced by client k is used, as reference measure, by client $k+1$. This effectively establishes a principled way to encode prior information through a reference measure. In particular, achieving centralized performance in the decentralized setting requires a specific scaling of the regularization factors with the local sample sizes. Overall, this result opens the door to novel decentralized learning paradigms that shift the collaboration strategy from sharing data to sharing the local inductive bias via the reference measures over the set of models.**

I. Introduction

Decentralized learning studies how a collection of clients can collaboratively tune a learning algorithm by communicating only over a network, without explicitly exchanging raw datasets. This setting extends early work on distributed and asynchronous optimization, where coordination is achieved through local computations and intermittent message passing [1], [2]. It becomes particularly relevant when a central coordinator is unavailable or undesirable, or when data transfers are impractical due to bandwidth, latency, ownership, privacy, or regulatory constraints [3]. A standard benchmark for collaborative learning is the centralized regime in which all local datasets are pooled and a single training procedure is run on the aggregated data. While conceptually simple, the pooled-data benchmark is often unachievable in decentralized environments due to communication constraints and/or restricted disclosure of local datasets [3], [4]. This benchmark is revisited through the lens of *Gibbs algorithms*, i.e., data-dependent Gibbs probability measures on the model space. Such Gibbs measures naturally arise as solutions to empirical risk minimizations with relative-entropy regularization (ERM-RER) [5]–[8]. This viewpoint also connects to exponential-weights predictors and PAC-Bayesian posteriors, which reason directly in terms of distributions on hypotheses [9]–[14]. Beyond their variational interpretation, Gibbs measures also capture the long-run distribution of stochastic gradient methods under suitable regimes [15]–[18]. From this standpoint, a complementary line of work studies Gibbs measures as solutions to ERM-RER problems and their extensions [6]–[8], [19]. Other studies focus on change-of-measure techniques to quantify the variation of an expectation when the underlying probability measure changes [5], [20]. These developments provide tools to interpret and manipulate Gibbs measures as first-class objects in learning systems, and to reason about how information is transported through probability measures rather than through datasets.

This paper shows that centralized performance guarantees can be achieved in a decentralized system through a strategic design of (i) the *reference measures* and (ii) the *regularization factors* that define the clients' Gibbs algorithm. More precisely, a peer-to-peer communication protocol is introduced, in which each client transmits its Gibbs probability measure to its successor, which adopts it as reference measure. This mechanism embeds information from datasets into the learning process without explicitly transmitting such datasets. A closed-form expression is obtained for the resulting decentralized Gibbs probability measures, together with conditions under which it coincides with the Gibbs measure induced by the centralized pooled-data benchmark.

The paper is organized as follows. Section II introduces the notation and formalizes the decentralized learning setting. Section III defines Gibbs conditional probability measures and their interpretation within the context of ERM-RER. Section IV presents the communication protocol and the main results establishing centralized-performance guarantees. Section V sketches the proof of the main result. Section VI concludes and discusses practical challenges, including the impact of distortions when probability measures are communicated under finite-rate constraints.

This work is supported in part by the European Commission through the H2020-MSCA-RISE-2019 project 872172; the French National Agency for Research (ANR) through the Project ANR-21-CE25-0013 and the project ANR-22-PEFT-0010 of the France 2030 program PEPR Réseaux du Futur; and in part by the Agence de l'innovation de défense (AID) through the project UK-FR 2024352.

II. Supervised Machine Learning

Consider a decentralized learning system in which K clients collaboratively tune their local learning algorithms by communicating with each other. For all $k \in \{1, 2, \ldots, K\}$, let $\mathcal{M}_k$, $\mathcal{X}_k$ and $\mathcal{Y}_k$, with $\mathcal{M}_k \subseteq \mathbb{R}^{d_k}$ and $d_k \in \mathbb{N}$, be sets of *models*, *patterns*, and *labels*, respectively, at client k. The training data available for client k consists of n_k data points $(x_{k,1}, y_{k,1})$, $(x_{k,2}, y_{k,2})$, ..., (x_{k,n_k}, y_{k,n_k}), which are elements of the set $\mathcal{Z}_k \triangleq \mathcal{X}_k \times \mathcal{Y}_k$. Such data points form the local training dataset, denoted by $z_k \in \mathcal{Z}_k^{n_k}$, which can be explicitly written as

$$z_k \triangleq \left((x_{k,1}, y_{k,1}), (x_{k,2}, y_{k,2}), \ldots, (x_{k,n_k}, y_{k,n_k})\right). \quad (1)$$

The dataset obtained by the aggregation of all local datasets, denoted by z_0, satisfies

$$z_0 \triangleq (z_1, z_2, \ldots, z_K) \in \mathcal{Z}_1^{n_1} \times \mathcal{Z}_2^{n_2} \times \ldots \times \mathcal{Z}_k^{n_K} \quad (2)$$

$$= \left((x_{0,1}, y_{0,1}), (x_{0,2}, y_{0,2}), \ldots, (x_{0,n_0}, y_{0,n_0})\right). \quad (3)$$

Hence, the total number of data points, denoted by $n_0 \in \mathbb{N}$, satisfies $n_0 \triangleq \sum_{k=1}^{K} n_k$. Given a model $\boldsymbol{\theta} \in \mathcal{M}_k$ for client k, the loss induced by such a model with respect to a data point $(x, y) \in \mathcal{Z}_k$ is $\ell_k(x, y, \boldsymbol{\theta})$, where the function

$$\ell_k : \mathcal{Z}_k \times \mathcal{M}_k \to [0, +\infty), \quad (4)$$

is referred to as the *loss function* of client k. Such a loss function is assumed to be Borel measurable. The *empirical risk* induced by such a model $\boldsymbol{\theta} \in \mathcal{M}_k$, with respect to the dataset z_k in (1), is determined by the function

$$\mathsf{L}_k : \begin{cases} \mathcal{Z}_k^{n_k} \times \mathcal{M}_k \longrightarrow [0, +\infty) \\ (z_k, \boldsymbol{\theta}) \longmapsto \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k\left(x_{k,i}, y_{k,i}, \boldsymbol{\theta}\right), \end{cases} \quad (5)$$

where the function ℓ_k is defined in (4).

The set of all probability measures on the measurable space $\left(\mathcal{M}_k, \mathscr{F}_{\mathcal{M}_k}\right)$ is denoted by $\triangle\left(\mathcal{M}_k, \mathscr{F}_{\mathcal{M}_k}\right)$, or simply $\triangle(\mathcal{M}_k)$. The set of all probability measures on $\mathcal{M}_k$ conditioned on an element of $\mathcal{Z}_k^{n_k}$ is denoted by $\triangle\left(\mathcal{M}_k \mid \mathcal{Z}_k^{n_k}\right)$. Moreover, the set of probability measures in $\triangle(\mathcal{M}_k)$ that are absolutely continuous with respect to Q_k is denoted by $\triangle_{Q_k}(\mathcal{M}_k)$. Using this notation, a supervised machine learning algorithm is represented by a conditional probability measure, as defined hereunder.

Definition 1 (Algorithm)**.** *For all $k \in \{1, 2, \ldots, K\}$, a conditional probability measure $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k} \in \triangle\left(\mathcal{M}_k \mid \mathcal{Z}_k^{n_k}\right)$ is said to represent a supervised machine learning algorithm.*

Let $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k} \in \triangle\left(\mathcal{M}_k \mid \mathcal{Z}_k^{n_k}\right)$ be an algorithm. Hence, the instance of such an algorithm trained upon the dataset z_k in (1) is denoted by $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}$, which is simply a probability measure in $\triangle(\mathcal{M}_k)$.

III. Gibbs Algorithms

The learning framework of client k, with $k \in \{1, 2, \ldots, K\}$, is defined by an ERM-RER problem. To formalize this optimization problem, consider a dataset $z_k \in \mathcal{Z}_k^{n_k}$ and the functional R_{k,z_k} defined as follows,

$$\mathsf{R}_{k,z_k} : \begin{cases} \triangle(\mathcal{M}_k) \longrightarrow [0, +\infty) \\ P \longmapsto \int \mathsf{L}_k(z_k, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}), \end{cases} \quad (6)$$

where the function L_k is defined in (5). The corresponding ERM-RER problem is

$$\min_{P \in \triangle_{Q_k}(\mathcal{M}_k)} \mathsf{R}_{k,z_k}(P) + \lambda_k D(P \parallel Q_k), \quad (7)$$

where $Q_k \in \triangle(\mathcal{M}_k)$ is a σ-finite measure; $\lambda_k \in (0, +\infty)$ is the regularization factor; and $D(\cdot \parallel \cdot)$ represents the relative entropy, [20, Definition 3]. As shown later in Lemma 1, the solution to (7), whenever it exists, admits a closed-form expression. This expression is a probability measure parametrized by the empirical risk function L_k; the σ-finite measure $Q_k \in \triangle(\mathcal{M}_k)$; and the dataset $z_k \in \mathcal{Z}_k^{n_k}$. Such measures are referred to as Gibbs probability measures. In order to define them, consider the following function:

$$\mathsf{K}_{k,Q_k,z_k} : \begin{cases} \mathbb{R} \longrightarrow \mathbb{R} \\ t \longmapsto \log \int \exp\left(t \, \mathsf{L}_k(z_k, \boldsymbol{\theta}_k)\right) \mathrm{d}Q_k(\boldsymbol{\theta}_k), \end{cases} \quad (8)$$

where the function L_k is defined in (5). Under the assumption that the reference measure Q_k is a probability measure, the function K_{k,Q_k,z_k} in (8) is the cumulant generating function of the random variable $\mathsf{L}_k(z_k, \boldsymbol{\theta}_k)$, for some fixed dataset $z_k \in \mathcal{Z}_k^{n_k}$, when the model $\boldsymbol{\theta}_k$ is sampled from Q_k. Using this notation, the definition of the Gibbs conditional probability measure is presented hereunder.

Definition 2. *Given the function L_k in (5); a σ-finite measure Q_k; and a $\lambda_k \in (0, +\infty)$, with $k \in \{1, 2, \ldots, K\}$, the probability measure $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k}^{(Q_k,\lambda_k)} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$ is said to be an $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs conditional probability measure if*

$$\forall z_k \in \mathcal{Z}^{n_k}, \ \mathsf{K}_{k,Q_k,z_k}\left(\frac{-1}{\lambda_k}\right) < +\infty; \quad (9)$$

and for all $(z_k, \boldsymbol{\theta}_k) \in \mathcal{Z}^{n_k} \times \operatorname{supp} Q_k$,

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}^{(Q_k,\lambda_k)}}{\mathrm{d}Q_k}(\boldsymbol{\theta}_k) = \exp\left(\frac{-1}{\lambda_k}\mathsf{L}_k(z_k, \boldsymbol{\theta}_k) - \mathsf{K}_{k,Q_k,z_k}\left(\frac{-1}{\lambda_k}\right)\right), \quad (10)$$

where the function K_{k,Q_k,z_k} is defined in (8).

Note that, while $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k}^{(Q_k,\lambda_k)}$ in (10) is referred to as a Gibbs conditional probability measure, the measure $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}^{(Q_k,\lambda_k)}$, obtained by conditioning upon a given dataset $z_k \in \mathcal{Z}_k^{n_k}$, is referred to as a Gibbs probability measure. The following lemma formalizes the connection stated above between Gibbs measures and the ERM-RER problem in (7).

Lemma 1. *Assume that the optimization problem in (7) admits a solution. Then, the $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs probability measure $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}^{(Q_k,\lambda_k)}$ in (10) is the unique solution.*

Proof: The proof follows from [20, Lemma 1]. ∎

This result has also been reported for other f-divergences in [7], [8]. Interestingly, the probability measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (10) is the long-run distribution of a stochastic gradient descent algorithm [18]. In statistical learning, such a distribution is often referred to as the *Gibbs algorithm* [21].

Another optimization problem that is closely related to the probability measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (10) is the following:

$$\min_{P\in\triangle_{Q_k}(\mathcal{M}_k)} \quad \mathsf{R}_{k,\boldsymbol{z}_k}(P) \tag{11a}$$

$$\text{s.t.} \quad D(P \parallel Q_k) \leqslant \gamma_k, \tag{11b}$$

for some $\gamma_k > 0$. The following lemma establishes the connection.

Lemma 2. *Assume that the optimization problem in* (11) *admits a solution and that λ_k is such that*

$$D\left(P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q_k\right) = \gamma_k. \tag{12}$$

Then, the $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs probability measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (10) *is the unique solution to* (11).

Proof: The proof follows from [20, Lemma 4]. ∎

Lemma 2 implies that the probability measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (10) minimizes the training empirical risk over all probability measures in the following neighborhood of Q_k,

$$\left\{P \in \triangle_{Q_k}(\mathcal{M}_k) : D(P \parallel Q_k) \leqslant \gamma_k\right\}. \tag{13}$$

This observation is important for presenting the main results.

IV. Main result

The main result of this work (Theorem 3) is presented in Subsection IV-B. In order to present such a result, the peer-to-peer communication protocol used by the clients is introduced. The section ends by stating the necessary conditions under which centralized performance is obtained.

A. Communication Protocol

Figure 1 depicts the forward–backward peer-to-peer communication protocol used in this work. In the forward direction (blue arrows), for all $k \in \{1, 2, \ldots, K-1\}$, client k transmits to client $k+1$ the $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs probability measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (10), obtained as the solution to the ERM-RER problem in (7). Client $k+1$ adopts this transmitted measure as its reference measure Q_{k+1} in (7). This choice induces a nested structure of the reference measure, as $Q_{k+1} = P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$, with $k \in \{1, 2, \ldots, K-1\}$. This is formalized later in Theorem 3. The backward direction (red arrows) disseminates the final Gibbs probability measure. More specifically, once client K has computed its $(\mathsf{L}_K, Q_K, \lambda_K)$-Gibbs probability measure $P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}$ in (10), this measure is transmitted back along the chain, from client K to client $K-1$, then from client $K-1$ to client $K-2$, and so on until client 1. This backward transmission provides all clients with access to the same final Gibbs algorithm. The following subsection describes such a final algorithm.

B. Decentralized Algorithms

The main result of this work is presented by the following theorem.

Theorem 3. *For all $k \in \{1, 2, \ldots, K\}$, consider an $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs conditional probability measure, denoted by $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$, where the reference measure Q_k satisfies*

$$Q_k = \begin{cases} Q_1 & \text{if } k = 1 \\ P^{(Q_{k-1},\lambda_{k-1})}_{\boldsymbol{\Theta}_{k-1}|\boldsymbol{Z}_{k-1}=\boldsymbol{z}_{k-1}} & \text{if } k \geqslant 2, \end{cases} \tag{14}$$

for some given Q_1. Then, for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\exp\left(\sum_{j=1}^{k} \frac{-1}{\lambda_j}\mathsf{L}_j(\boldsymbol{z}_j, \boldsymbol{\theta})\right)}{\int \exp\left(\sum_{i=1}^{k} \frac{-1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i, \boldsymbol{\nu})\right) \mathrm{d}Q_1(\boldsymbol{\nu})}. \tag{15}$$

Proof: The proof is presented in [22]. ∎

The choice of reference measures Q_1, Q_2, …, Q_K in (14) induces a *nested* structure. Under this structure, the training performed by client k uses only its local dataset, while the influence of the previous clients' datasets is carried out through the reference measure Q_k. The relevance of this nested structure, in which client k shares its Gibbs probability measure (algorithm) with its successor, client $k+1$, is made clear by Lemma 1. More specifically, the probability measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (15) is the unique solution to (7) and, simultaneously, the unique minimizer of the following optimization problem:

$$\min_{P\in\triangle_{Q_1}(\mathcal{M}_k)} \int \left(\sum_{j=1}^{k} \frac{1}{\lambda_j}\mathsf{L}_j(\boldsymbol{z}_j, \boldsymbol{\theta})\right) \mathrm{d}P(\boldsymbol{\theta}) + D(P \parallel Q_1). \tag{16}$$

The following corollary of Theorem 3 formalizes this observation.

Corollary 4. *Under the assumption that the measures Q_1, Q_2, …, Q_K satisfy* (14)*, the solutions to the optimization problems in* (7) *and* (16) *are unique and coincide with the $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs probability measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in* (15).

Given $k > 1$, the optimization problem in (7) depends exclusively on the local training dataset $\boldsymbol{z}_k$. While the reference Q_k in (14) depends on the training datasets of the $k-1$ previous clients, such datasets do not need to be explicitly known for solving (7). In particular, solving (7) requires access to the probability measure $Q_k \in \triangle(\mathcal{M}_k)$, but not access to the training datasets of all previous clients. This is because for fixed training datasets, Q_k is simply a probability measure on the model space. In contrast, solving the optimization problem in (16) requires knowing the training datasets of client j, for all $j \in \{1, 2, \ldots, k\}$. In this case, the reference measure, Q_1, does not depend on any training dataset. The fact that problems (7) and (16) share the same solution unveils an important observation: providing client k with a reference measure Q_k of the form in (14) reproduces the effect of having access to the training datasets of the $k-1$ previous clients. The following subsection unveils, under specific conditions, the centralized-type guarantees of the nested structure.

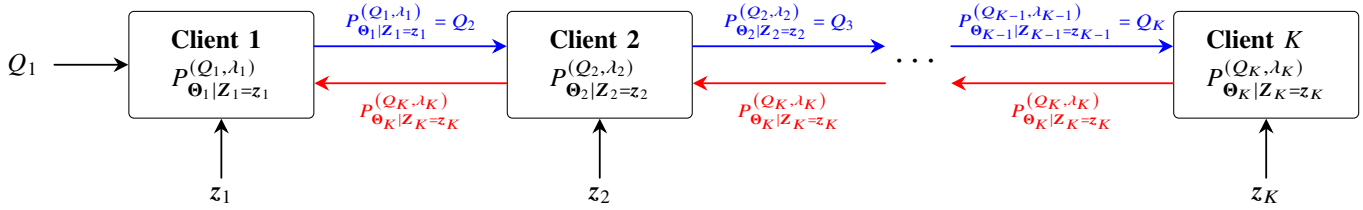

Figure 1. Nested Structure: the Gibbs measure produced by client k becomes the reference measure Q_{k+1} used by client $k+1$.

C. Centralized Performance Guarantees

An important observation is that a strategic choice of λ_1, λ_2, ..., λ_K in (15) can lead to achieving the same Gibbs probability distribution as in a setting in which the training datasets of all clients are available to all clients. This describes a decentralized system whose distributed nature does not prevent it from achieving the same Gibbs algorithm that would have been obtained if all the training datasets were available to all clients. The following theorem formalizes this observation.

Theorem 5. *Assume that the loss functions in* (4) *satisfy* $\ell_1 = \ell_2 = \ldots = \ell_K = \ell$ *and for all* $k \in \{1, \ldots, K\}$,

$$\lambda_k = \frac{n_0 \lambda_0}{n_k}, \tag{17}$$

for some $\lambda_0 > 0$ *and some loss function* ℓ. *Consider some measures* $Q_1, Q_2, \cdots, Q_K$ *satisfying* (14). *Then, for all* $\boldsymbol{\theta} \in \operatorname{supp} Q_1$, *it follows that,*

$$\frac{\mathrm{d}P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}}{\mathrm{d}P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}}(\boldsymbol{\theta}) = 1, \tag{18}$$

where the probability measure $P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}$ *is defined in* (15); *the measure* $P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\exp\left(\frac{-1}{n_0\lambda_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i}, y_{0,i}, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\frac{-1}{n_0\lambda_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i}, y_{0,i}, \boldsymbol{\nu}\right)\right)\mathrm{d}Q_1(\boldsymbol{\nu})}; \tag{19}$$

and $(x_{0,i}, y_{0,i})$ *are data points of the aggregated dataset* $\boldsymbol{z}_0$ *in* (2).

Proof: The proof is presented in [22]. ∎

The relevance of Theorem 5 is highlighted by the following observations. Under the assumptions of Theorem 5, in particular that the loss functions satisfy $\ell_1 = \ell_2 = \cdots = \ell_K = \ell$, the equality in (17) together with [21, Lemma 4] allows rewriting the optimization problem in (16) as

$$\min_{P \in \triangle_{Q_1}(\mathcal{M}_K)} \int \frac{1}{n_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i}, y_{0,i}, \boldsymbol{\theta}\right)\mathrm{d}P(\boldsymbol{\theta}) + \lambda_0 D(P \parallel Q_1), \tag{20}$$

which requires access to the training datasets of all clients. Interestingly, from Lemma 1, it follows that the probability measure $P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (19) is the solution to (20). More importantly, from Lemma 2, if λ_0 is chosen such that

$$D\left(P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, Q_1\right) = \gamma_0, \tag{21}$$

for some $\gamma_0 > 0$, the probability measure $P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (19) is also the solution to the following optimization problem:

$$\min_{P \in \triangle_{Q_1}(\mathcal{M}_K)} \int \frac{1}{n_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i}, y_{0,i}, \boldsymbol{\theta}\right)\mathrm{d}P(\boldsymbol{\theta}) \tag{22a}$$

$$\text{s.t.} \quad D(P \parallel Q_1) \leqslant \gamma_0. \tag{22b}$$

From (18), it follows that the measures $P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}$ and $P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ are identical, which implies that the nested structure, induced by the choice of reference measures in (14) and the regularization factors in (17), allows achieving in a decentralized system, the same learning algorithm that would have been obtained in a centralized system in which all training datasets are available to all clients.

Under the forward–backward communication protocol in Section IV-A, after $K-1$ forward messages (blue arrows in Figure 1) client K obtains the probability measure that minimizes the empirical risk with respect to all training datasets within the neighborhood of Q_1. The backward dissemination (red arrows in Figure 1) provides each client with the same Gibbs algorithm, minimizing within a neighborhood of the form in (13) around Q_1 the empirical risk with respect to the aggregated dataset.

V. Proof of Main Result

This section first introduces the preliminaries needed for the proof of Theorem 3, and then outlines the main steps. A more detailed proof appears in [22].

A. Preliminaries

Lemma 6. *For all* $k \in \{1, 2, \ldots, K\}$, *consider an* $(\mathsf{L}_k, Q_k, \lambda_k)$*-Gibbs conditional probability measure, denoted by* $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$, *where the reference measure* Q_k *satisfies* (14). *Then, for all* $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\exp(C_k), \tag{23}$$

where the measure $P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k}$ *is an* $(\mathsf{L}_k, Q_1, \lambda_k)$*-Gibbs conditional probability measure and* $C_k \in \mathbb{R}$ *satisfies*

$$C_k \triangleq \log\left(\frac{\exp\left(\mathsf{K}_{k,Q_1,\boldsymbol{z}_k}\left(\frac{-1}{\lambda_k}\right)\right)\int \exp\left(-\sum_{i=1}^{k-1}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\nu})\right)\mathrm{d}Q_1(\boldsymbol{\nu})}{\int \exp\left(-\sum_{j=1}^{k}\frac{1}{\lambda_j}\mathsf{L}_j(\boldsymbol{z}_j,\boldsymbol{\nu}')\right)\mathrm{d}Q_1(\boldsymbol{\nu}')}\right), \tag{24}$$

where the functional $\mathsf{K}_{k,Q_1,\boldsymbol{z}_k}$ *is defined in* (8).

Proof: The proof is presented in [22]. ∎

The relevance of this lemma lies in the fact that the Radon–Nikodym derivative of an $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs conditional probability measure with respect to its reference measure Q_k can be re-expressed relative to the fixed reference measure Q_1, up to a factor, C_k in (24), that does not depend on the model. Moreover, this factor admits an information-theoretic characterization in terms of Kullback–Leibler divergences; see [22, Lemma 9]. Lemma 6 constitutes a main step in the proof of Theorem 3 and is explained in the following subsection.

B. Sketched proof of Theorem 3

Using [23, Theorem 4] (chain rule), the Radon–Nikodym derivative in the left-hand side of (23) can be written as follows

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\,\frac{\mathrm{d}Q_1}{\mathrm{d}Q_k}(\boldsymbol{\theta}), \tag{25}$$

Then, from Lemma 6 and [23, Theorem 5], the equality in (25) yields

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\mathrm{d}Q_k}{\mathrm{d}Q_1}(\boldsymbol{\theta})\,\frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\exp(C_k). \tag{26}$$

Moreover, from [6, Lemma 3], the Radon–Nikodym derivatives in (25) and in (26) are well defined.

The next step consists in expressing $\dfrac{\mathrm{d}Q_k}{\mathrm{d}Q_1}$ using the nested definition of the reference measures in (14). This nested construction implies the absolute continuity assumptions required to apply a chain rule for Radon–Nikodym derivatives. The resulting product decomposition isolates successive changes of reference measure and therefore allows each factor to be handled separately using Lemma 6. Then, from [23, Theorem 4], it follows that

$$\frac{\mathrm{d}Q_k}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \prod_{j=2}^{k}\frac{\mathrm{d}Q_j}{\mathrm{d}Q_{j-1}}(\boldsymbol{\theta}) \tag{27}$$

$$= \frac{\mathrm{d}Q_2}{\mathrm{d}Q_1}(\boldsymbol{\theta})\prod_{j=3}^{k}\frac{\mathrm{d}Q_j}{\mathrm{d}Q_{j-1}}(\boldsymbol{\theta}) \tag{28}$$

$$= \frac{\mathrm{d}P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}_1|\boldsymbol{Z}_1=\boldsymbol{z}_1}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\prod_{j=2}^{k-1}\frac{\mathrm{d}P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}_j|\boldsymbol{Z}_j=\boldsymbol{z}_j}}{\mathrm{d}Q_j}(\boldsymbol{\theta}) \tag{29}$$

$$= \frac{\exp\left(-\sum_{i=1}^{k-1}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\theta})\right)}{\int\exp\left(-\sum_{i=1}^{k-1}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\nu})\right)\mathrm{d}Q_1(\boldsymbol{\nu})}, \tag{30}$$

where (29) follows from (14); and (30) follows from Definition 2 and Lemma 6. Substituting (30) in (26) yields

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\exp\left(-\sum_{i=1}^{k-1}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\theta})\right)}{\int\exp\left(-\sum_{i=1}^{k-1}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\nu})\right)\mathrm{d}Q_1(\boldsymbol{\nu})}\,\frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\exp(C_k) \tag{31}$$

$$= \frac{\exp\left(-\sum_{i=1}^{k}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\theta})\right)}{\int\exp\left(-\sum_{i=1}^{k-1}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\nu})\right)\mathrm{d}Q_1(\boldsymbol{\nu})}\,\frac{\int\exp\left(-\sum_{i=1}^{k-1}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\nu})\right)\mathrm{d}Q_1(\boldsymbol{\nu})}{\int\exp\left(-\sum_{j=1}^{k}\frac{1}{\lambda_j}\mathsf{L}_j(\boldsymbol{z}_j,\boldsymbol{\nu}')\right)\mathrm{d}Q_1(\boldsymbol{\nu}')} \tag{32}$$

$$= \frac{\exp\left(-\sum_{i=1}^{k}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\theta})\right)}{\int\exp\left(-\sum_{i=1}^{k}\frac{1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i,\boldsymbol{\nu}')\right)\mathrm{d}Q_1(\boldsymbol{\nu}')}, \tag{33}$$

where (32) follows from (24); and (33) yields (15). This completes the proof.

VI. Conclusions and Final Remarks

This work establishes that, in a decentralized machine learning scenario, an appropriate choice of reference measures $(Q_1, Q_2, \ldots, Q_K)$ and regularization factors $(\lambda_1, \lambda_2, \ldots, \lambda_K)$ allows guaranteeing the same performance as a centralized system in which all training datasets are aggregated and jointly available. The construction of such regularization factors is rather simple. The regularization factor of client k shall be the product of a strictly positive real (common to all clients) and the ratio of the sizes of the training dataset of client k and the aggregated training dataset. The reference measure of client k, $k > 1$, is the Gibbs measure (Gibbs algorithm) from which client $k-1$ samples its models. The first client uses a given reference Q_1. Under such a choice, client K obtains a Gibbs probability measure that solves the ERM-RER problem, with respect to the aggregated dataset, within a neighborhood of Q_1. Via backward dissemination, all clients obtain the same probability measure (algorithm). This choice of reference measures induces a nested structure whose construction requires transmitting $K-1$ probability measures with common support [6, Lemma 3]. Several practical challenges arise from this requirement. A main limitation is finite-rate communication, possibly under delay constraints, which implies that the probability measure transmitted by client k may be received by client $k+1$ with distortion. Characterizing the impact of such distortions on the nested construction remains an open problem and is not addressed here. A further limitation is the potentially large support of the involved Gibbs probability measures, which can make the communication requirement comparable to transmitting the training datasets. Nonetheless, the transmission of a probability measure is more privacy-preserving than the actual transmission of training datasets.

REFERENCES

[1] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Transactions on Automatic Control*, vol. 31, no. 9, pp. 803–812, Sep. 1986.

[2] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, 1st ed. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[3] P. Kairouz, B. H. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. d'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Ozgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu, and S. Zhao, *Advances and Open Problems in Federated Learning*. Now Publishers, 2021, vol. 14, no. 1–2.

[4] C. Dwork, "Differential privacy," in *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, vol. 4052, Venice, Italy, Jul. 2006, pp. 1–12.

[5] I. Csiszár, "I-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, Feb. 1975.

[6] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, "Empirical risk minimization with relative entropy regularization," *IEEE Transactions on Information Theory*, vol. 70, no. 7, pp. 5122 – 5161, Jul. 2024.

[7] F. Daunas, I. Esnaola, S. M. Perlaza, and H. V. Poor, "Equivalence of empirical risk minimization to regularization on the family of f-divergences," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, Jul. 2024, pp. 759–764.

[8] ——, "Asymmetry of the relative entropy in the regularization of empirical risk minimization," *IEEE Transactions on Information Theory*, vol. 71, no. 8, pp. 6198–6226, Aug. 2025.

[9] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*, 1st ed. New York, NY, USA: Cambridge University Press, 2006.

[10] D. A. McAllester, "Some PAC-Bayesian theorems," *Machine Learning*, vol. 37, no. 3, pp. 355–363, Dec. 1999.

[11] M. Seeger, "PAC-Bayesian generalisation error bounds for Gaussian process classification," *Journal of Machine Learning Research*, vol. 3, pp. 233–269, Oct. 2002.

[12] J. Langford and J. Shawe-Taylor, "PAC-Bayes and margins," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 15, Vancouver, Canada, Dec. 2002, pp. 439–446.

[13] O. Catoni, *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, 1st ed. Beachwood, OH, USA: Institute of Mathematical Statistics Lecture Notes - Monograph Series, 2007, vol. 56.

[14] P. Alquier, "User-friendly introduction to PAC-Bayes bounds," *Foundations and Trends in Machine Learning*, vol. 17, no. 2, pp. 174–303, 2024.

[15] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient Langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, Bellevue, Washington, USA, Jun. 2011, pp. 681–688.

[16] S. Mandt, M. D. Hoffman, and D. M. Blei, "Stochastic gradient descent as approximate Bayesian inference," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4873 – 4907, Jan. 2017.

[17] M. Raginsky, A. Rakhlin, and M. Telgarsky, "Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis," in *Proceedings of the Conference on Learning Theory (COLT)*, vol. 65, Amsterdam, Netherlands, Jul. 2017, pp. 1674–1703.

[18] W. Azizian, F. Lutzeler, J. Malick, and P. Mertikopoulos, "What is the long-run distribution of stochastic gradient descent? A large deviations analysis," in *Proceedings of the International Conference on Machine Learning (ICML)*, Vienna, Austria, Jul. 2024, pp. 2168 – 2229.

[19] Y. Bermudez, S. M. Perlaza, and I. Esnaola, "Machine unlearning for Gibbs supervised learning algorithms," in *Proceedings of the International Symposium on Information Theory (ISIT)*, Guangzhou, China, Jun. 2026.

[20] S. M. Perlaza and G. Bisson, "Variations on the expectation due to changes in the probability measure," *Entropy*, vol. 27, no. 8:865, pp. 1–20, Aug. 2025.

[21] S. M. Perlaza, I. Esnaola, G. Bisson, and H. V. Poor, "On the validation of Gibbs algorithms: Training datasets, test datasets and their aggregation," in *Proceedings of the International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023, pp. 328–333.

[22] Y. Bermudez, S. M. Perlaza, and I. Esnaola, "Decentralized machine learning with centralized performance guarantees via Gibbs algorithms," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9608, Jan. 2026.

[23] Y. Bermudez, G. Bisson, I. Esnaola, and S. M. Perlaza, "Proofs for folklore theorems on the Radon-Nikodym derivative," INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis, France, Tech. Rep. RR-9591, Jul. 2025.