Decentralized Machine Learning with Centralized Performance Guarantees via Gibbs Algorithms

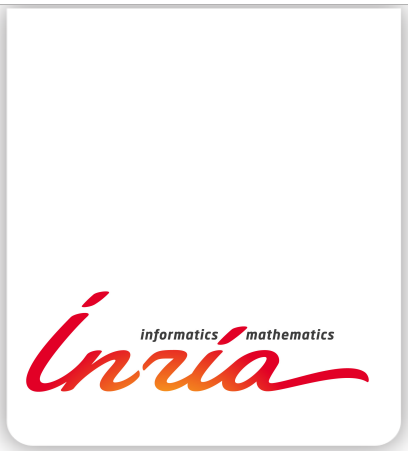
Yaiza Bermudez, Samir M. Perlaza, Iñaki Esnaola

Decentralized Machine Learning with Centralized Performance Guarantees via Gibbs Algorithms

Yaiza Bermudez, Samir M. Perlaza, and Iñaki Esnaola

RESEARCH REPORT
N° 9608
January 2026
Project-Team NEO

ISSN 0249-6399 ISRN INRIA/RR--9608--FR+ENG



Decentralized Machine Learning with Centralized Performance Guarantees via Gibbs Algorithms

Yaiza Bermudez, Samir M. Perlaza, and Iñaki Esnaola

Project-Team NEO

Research Report n° 9608 — January 2026 — 25 pages

Abstract: In this report, it is shown, for the first time, that centralized learning 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 a centralized ERM-RER with access to all the datasets. The core mechanism 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.

Key-words: Decentralized machine learning; Gibbs algorithms; Gibbs measures; Empirical risk minimization; Relative-entropy regularization; KL divergence; Reference measure; Centralized performance guarantees; Peer-to-peer communication.

Yaiza Bermudez and Samir M. Perlaza are with INRIA, Centre Inria d'Université Côte d'Azur, Sophia Antipolis 06902, France. Samir M. Perlaza is also with the GAATI Laboratory at the Université de la Polynésie Française, Faaa 98702, French Polynesia. Iñaki Esnaola is with the School of Electrical and Electronic Engineering at The University of Sheffield, Sheffield S1 3JD, UK. Samir M. Perlaza and Iñaki Esnaola also are with the Electrical and Computer Engineering Department at Princeton University, Princeton N.J. 08544, USA.
This research was supported in part by the European Commission through the H2020MSCA-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.

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Apprentissage décentralisé avec garanties de performance centralisée via des algorithmes de Gibbs

Résumé : Dans ce rapport, il est montré, pour la première fois, que des performances d'apprentissage centralisé sont atteignables en apprentissage décentralisé sans partage des jeux de données locaux. Plus précisément, lorsque les clients adoptent un cadre d'apprentissage de minimisation du risque empirique avec régularisation par entropie relative (ERM-RER) et qu'une communication aller-retour (forward-backward) est établie entre eux, il suffit de partager les mesures de Gibbs obtenues localement pour atteindre les mêmes performances qu'un ERM-RER centralisé ayant accès à l'ensemble des jeux de données. Le mécanisme clé est que la mesure de Gibbs produite par le client k est utilisée, comme mesure de référence, par le client $k+1$. Cela établit une manière principielle d'encoder une information a priori au moyen d'une mesure de référence. En particulier, l'obtention de performances centralisées dans le cadre décentralisé requiert une mise à l'échelle spécifique des facteurs de régularisation en fonction des tailles d'échantillon locales. Dans l'ensemble, ce résultat ouvre la voie à de nouveaux paradigmes d'apprentissage décentralisé qui déplacent la stratégie de collaboration : il ne s'agit plus de partager les données, mais de partager le biais inductif local via les mesures de référence sur l'ensemble des modèles.

Mots-clés : Apprentissage décentralisé; Algorithmes de Gibbs ; Mesures de Gibbs ; Minimisation du risque empirique ; Régularisation par entropie relative ; Divergence de Kullback–Leibler ; Mesure de référence ; Garanties de performance centralisée ; Communication pair-à-pair.

Contents

1 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., probability measures on the model space. Gibbs measures arise naturally as solutions of 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 its extensions [6–8]. Other studies focus on change-of-measure techniques to quantify the variation of an expectation when the underlying probability measure changes [5, 19]. 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.

1.1 Contributions

This report 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 that 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 report is organized as follows. Section 2 presents the supervised learning setting in a decentralized system and fixes the problem formulation. Section 3 defines Gibbs conditional probability measures and their interpretation within the context of ERM-RER. Section 4 presents the communication protocol and the main results establishing centralized-performance guarantees. Finally, Section 5

concludes and discusses practical considerations, including the impact of finite-rate communication distortions on the transmission of probability measures. Detailed proofs are provided in the appendices.

1.2 Notation and Preliminaries

The set of all probability measures on the measurable space $(\mathcal{X}, \mathscr{F}_{\mathcal{X}})$ is denoted by $\triangle(\mathcal{X}, \mathscr{F}_{\mathcal{X}})$, or simply $\triangle(\mathcal{X})$. Moreover, the set of probability measures in $\triangle(\mathcal{X})$ that are absolutely continuous with respect to Q is denoted by $\triangle_Q(\mathcal{X})$. Using this notation, a conditional probability measure is defined hereunder.

Definition 1 (Conditional Probability)**.** *A family $P_{Y|X} \triangleq (P_{Y|X=x})_{x\in\mathcal{X}}$ of elements of $\triangle(\mathcal{Y}, \mathscr{F}_{\mathcal{Y}})$ indexed by $\mathcal{X}$ is said to be a conditional probability measure if, for all sets $\mathcal{B} \in \mathscr{F}_{\mathcal{Y}}$, the map*

$$\begin{cases} \mathcal{X} \longrightarrow [0,1] \\ x \longmapsto P_{Y|X=x}(\mathcal{B}) \end{cases} \tag{1}$$

is Borel measurable. The set of such conditional probability measures is denoted by $\triangle(\mathcal{Y}|\mathcal{X})$.

Given two measures P and Q on the same measurable space, the notation $P \ll Q$ stands for "the measure P is absolutely continuous with respect to Q". Using this notation, the relative entropy or KL-divergence is defined hereunder.

Definition 2 (Relative Entropy)**.** *Given a probability measure P and a σ-finite measure Q, both on the same measurable space, with $P \ll Q$. The relative entropy of P with respect to Q is*

$$D(P \parallel Q) \triangleq \int \frac{\mathrm{d}P}{\mathrm{d}Q}(x) \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(x)\right) \mathrm{d}Q(x). \tag{2}$$

Lemma 1. *Consider three probability measures P,Q, and R on the same measurable space such that $R \ll P \ll Q$, it then follows that*

$$\int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}R(\boldsymbol{\theta}) = D(R \parallel Q) - D(R \parallel P). \tag{3}$$

Proof: Under the absolute continuity assumptions, *i.e.*, $R \ll P \ll Q$, from Definition 2 it follows that

$$\begin{aligned} \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}R(\boldsymbol{\theta}) &= \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}R}(\boldsymbol{\theta}) \frac{\mathrm{d}R}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}R(\boldsymbol{\theta}) && (4) \\ &= \int -\log\left(\frac{\mathrm{d}R}{\mathrm{d}P}(\boldsymbol{\theta})\right) + \log\left(\frac{\mathrm{d}R}{\mathrm{d}Q}(\boldsymbol{\theta})\right) \mathrm{d}R(\boldsymbol{\theta}) && (5) \\ &= D(R \parallel Q) - D(R \parallel P), && (6) \end{aligned}$$

where the equality in (4) follows from [20, Theorem 5] and the equality in (5) follows from [20, Theorem 4]. ∎

2 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 $\mathbf{z}_k \in \mathcal{Z}_k^{n_k}$, which can be explicitly written as

$$\mathbf{z}_k \triangleq \left((x_{k,1}, y_{k,1}), (x_{k,2}, y_{k,2}), \ldots, (x_{k,n_k}, y_{k,n_k})\right). \tag{7}$$

The dataset obtained by the aggregation of all local datasets, denoted by $\mathbf{z}_0$, satisfies

$$\begin{aligned} \mathbf{z}_0 &\triangleq (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K) \in \mathcal{Z}_1^{n_1} \times \mathcal{Z}_2^{n_2} \times \ldots \times \mathcal{Z}_k^{n_K} && (8)\\ &= \left((x_{0,1}, y_{0,1}), (x_{0,2}, y_{0,2}), \ldots, (x_{0,n_0}, y_{0,n_0})\right). && (9) \end{aligned}$$

Hence, the total number of data points, denoted by $n_0 \in \mathbb{N}$, satisfies

$$n_0 \triangleq \sum_{k=1}^{K} n_k. \tag{10}$$

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

$$\ell_k : \mathcal{Z}_k \times \mathcal{M}_k \to [0, +\infty), \tag{11}$$

is referred to as the *loss function* of client k. Such a loss function is measurable with respect to the measurable spaces $(\mathcal{Z}_k \times \mathcal{M}_k, \mathscr{F}_k)$ and $([0, +\infty), \mathscr{B}([0, +\infty)))$, where $\mathscr{F}_k$ is a given σ-field on $\mathcal{Z}_k \times \mathcal{M}_k$ and $\mathscr{B}([0, +\infty))$ is the Borel σ-field on $[0, +\infty)$. The *empirical risk* induced by such a model $\boldsymbol{\theta} \in \mathcal{M}_k$, with respect to the dataset $\mathbf{z}_k$ in (7), is determined by the function

$$\mathsf{L}_k : \begin{cases} \mathcal{Z}_k^{n_k} \times \mathcal{M}_k \longrightarrow [0, +\infty) \\ \quad (\mathbf{z}_k, \boldsymbol{\theta}) \longmapsto \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k\left(x_{k,i}, y_{k,i}, \boldsymbol{\theta}\right), \end{cases} \tag{12}$$

where the function ℓ_k is defined in (11). In the following, a supervised machine learning algorithm is represented by a conditional probability measure, as defined hereunder.

Definition 3 (Algorithm)**.** *For all $k \in \{1, 2, \ldots, K\}$, a conditional probability measure $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k} \in \triangle\left(\mathcal{M}_k \mid (\mathcal{X}_k \times \mathcal{Y}_k)^{n_k}\right)$ is said to represent a supervised machine learning algorithm.*

Let $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k} \in \triangle\left(\mathcal{M}_k \mid (\mathcal{X}_k \times \mathcal{Y}_k)^{n_k}\right)$ be an algorithm. Hence, the instance of such an algorithm trained upon the dataset $\mathbf{z}_k$ in (7) is denoted by $P_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$, which is simply a probability measure in $\triangle\left(\mathcal{M}_k\right)$.

A class of algorithms that are central in this work are known as Gibbs algorithms. These algorithms are introduced in the following section.

3 Gibbs Algorithms

Gibbs algorithms are represented by Gibbs conditional probability measures. Such conditional probability measures are parametrized by the empirical risk function L_k; a σ-finite measure $Q_k \in \triangle(\mathcal{M}_k)$; and a dataset $z_k \in \mathcal{Z}_k^{n_k}$, with $k \in \{1, 2, \ldots, K\}$. In order to define such Gibbs conditional measures, consider the following function:

$$\mathsf{K}_{k,Q_k,z_k} : \begin{cases} \mathbb{R} \longrightarrow \mathbb{R} \\ t \longmapsto \log\left(\int \exp\left(t\,\mathsf{L}_k\left(z_k, \boldsymbol{\theta}_k\right)\right) \mathrm{d}Q_k\left(\boldsymbol{\theta}_k\right)\right). \end{cases} \tag{13}$$

Under the assumption that the reference measure Q_k is a probability measure, the function K_{k,Q_k,z_k} in (13) is the cumulant generating function of the random variable $\mathsf{L}_k\left(z_k, \boldsymbol{\theta}_k\right)$, 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 Gibbs conditional probability measures is presented hereunder.

Definition 4. *Given the function L_k in (12); a σ-finite measure Q_k; and a $\lambda_k \in (0, +\infty)$, with $k \in \{1, 2, \ldots, K\}$, the probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k}^{(Q_k,\lambda_k)} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$ is said to be an $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs conditional probability measure if*

$$\forall z_k \in \mathcal{S}_k,\ \mathsf{K}_{k,Q_k,z_k}\left(\frac{-1}{\lambda_k}\right) < +\infty; \tag{14}$$

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

$$\frac{\mathrm{d}P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}^{(Q_k,\lambda_k)}}{\mathrm{d}Q_k}\left(\boldsymbol{\theta}_k\right) = \exp\left(\frac{-1}{\lambda_k}\mathsf{L}_k\left(z_k, \boldsymbol{\theta}_k\right) - \mathsf{K}_{k,Q_k,z_k}\left(\frac{-1}{\lambda_k}\right)\right), \tag{15}$$

where the function K_{k,Q_k,z_k} is defined in (13).

In the following, if the probability measure from which datasets are sampled by client k is $P_{\boldsymbol{Z}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\right)$, the set $\mathcal{S}_k$ in Definition 4 is chosen to be the support of $P_{\boldsymbol{Z}_k}$. Note that, while $P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k}^{(Q_k,\lambda_k)}$ in (15) is referred to as a Gibbs conditional probability measure, the measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}^{(Q_k,\lambda_k)}$, obtained by conditioning upon a given dataset $z_k \in \mathcal{Z}_k^{n_k}$, is referred to as a Gibbs probability measure. Consider the functional R_{k,z_k} defined as follows,

$$\mathsf{R}_{k,z_k} : \begin{cases} \triangle(\mathcal{M}_k) \longrightarrow [0, +\infty) \\ P \longmapsto \int \mathsf{L}_k\left(z_k, \boldsymbol{\theta}\right) \mathrm{d}P\left(\boldsymbol{\theta}\right), \end{cases} \tag{16}$$

where the function L_k is defined in (12). The Gibbs probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}^{(Q_k,\lambda_k)}$ in (15) can be related to the following optimization problem:

$$\min_{P \in \triangle_{Q_k}(\mathcal{M}_k)} \mathsf{R}_{k,z_k}(P) + \lambda_k D\left(P \parallel Q_k\right). \tag{17}$$

The following lemma formalizes this connection.

Lemma 2. *Assume that the optimization problem in* (17) *admits a solution. Then the probability measure* $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k}$ *in* (15) *is the unique solution.*

Proof: The proof follows from [19, Lemma 1]. ∎

This result has also been reported for other f-divergences in [7, 8]. Interestingly, the conditional probability measure $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k}$ in (15) 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 $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (15) is the following:

$$\min_{P\in\triangle_{Q_k}(\mathcal{M}_k)} \quad \mathsf{R}_{k,\mathbf{z}_k}(P) \tag{18a}$$

$$\text{s.t.} \quad D(P \parallel Q_k) \leqslant \gamma_k, \tag{18b}$$

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

Lemma 3. *Assume that the optimization problem in* (18) *admits a solution and that* λ_k *is such that*

$$D\left(P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_k\right) = \gamma_k. \tag{19}$$

Then, the probability measure $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k}$ *in* (15) *is the unique solution.*

Proof: The proof follows from [19, Lemma 4]. ∎

Lemma 3 implies that the probability measure $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (15) minimizes the training empirical risk over all probability measures in the following neighborhood of Q_k,

$$\left\{P \in \triangle_{Q_k}(\mathcal{M}_k) : D(P \parallel Q_k) \leqslant \gamma_k\right\}. \tag{20}$$

4 Main Result

The main result of this work (Theorem 4) is presented in Subsection 4.2. 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.

4.1 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 $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (15), obtained as the solution to the ERM-RER problem in (17). Client $k+1$ adopts this transmitted measure as its reference measure Q_{k+1} in (17). This choice induces a nested structure of the reference measure, as

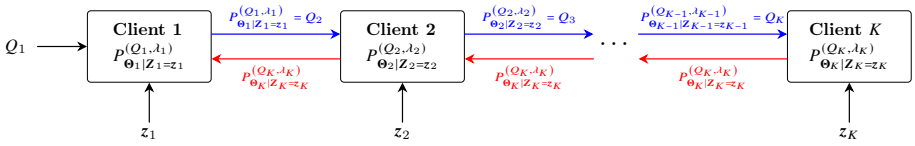


Figure 1: Nested Structure: the Gibbs measure produced by client k becomes the reference measure Q_{k+1} used by client $k+1$.

$Q_{k+1} = P^{(Q_k,\lambda_k)}_{\Theta_k|Z_k=z_k}$, with $k \in \{1,2,\ldots,K-1\}$. This is formalized later in Theorem 4. 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 $P^{(Q_K,\lambda_K)}_{\Theta_K|Z_K=z_K}$ in (15), 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.

4.2 Decentralized Algorithms

The main result of this work is presented in terms of the Gibbs algorithms obtained through the communication protocol described in Section 4.1. In particular, for all $k \in \{1,2,\ldots,K\}$, in such a protocol, the $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs probability measure

$$P^{(Q_k,\lambda_k)}_{\Theta_k|Z_k=z_k} \in \triangle(\mathcal{M}_k), \tag{21a}$$

with $\lambda_k > 0$, represents the Gibbs algorithm of client k trained upon its local dataset z_k defined in (7). The corresponding reference measure Q_k is specified recursively as

$$Q_k = \begin{cases} Q_1, & \text{if } k = 1, \\ P^{(Q_{k-1},\lambda_{k-1})}_{\Theta_{k-1}|Z_{k-1}=z_{k-1}}, & \text{if } k \geqslant 2, \end{cases} \tag{21b}$$

From Definition 4, for all $\theta \in \operatorname{supp} Q_1$, it follows that

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\Theta_k|Z_k=z_k}}{\mathrm{d}Q_k}(\theta) = \frac{\exp\left(-\frac{1}{\lambda_k}\mathsf{L}_k(z_k,\theta)\right)}{\int \exp\left(-\frac{1}{\lambda_k}\mathsf{L}_k(z_k,\nu)\right)\mathrm{d}Q_k(\nu)}. \tag{21c}$$

From Lemma 3, for all $k \in \{1,2,\ldots,K\}$, it follows that such an algorithm is optimal in the sense that the corresponding probability measure $P^{(Q_k,\lambda_k)}_{\Theta_k|Z_k=z_k}$ in (21) minimizes the expected training empirical risk with respect to z_k, over all probability measures in the following neighborhood of Q_k,

$$\mathcal{G}_k \triangleq \left\{P \in \triangle_{Q_1}(\mathcal{M}_k) : D(P \parallel Q_k) \leqslant D\left(P^{(Q_k,\lambda_k)}_{\Theta_k|Z_k=z_k} \,\middle\|\, Q_k\right)\right\}. \tag{22}$$

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

Theorem 4. *For all $k \in \{1, 2, \ldots, K\}$, consider 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 (21). Then, for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,*

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\exp\left(\sum_{j=1}^{k} \frac{-1}{\lambda_j} \mathsf{L}_j\left(\boldsymbol{z}_j, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{i=1}^{k} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})}. \tag{23}$$

Proof: The proof is presented in Appendix B. ∎

The choice of reference measures Q_1, Q_2, ..., Q_K in (21b) 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 $k+1$ neighbor, is made clear by [19, Lemma 1]. More specifically, the conditional measure $P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (23) is the unique solution to (17) and, simultaneously, the unique minimizer of the following optimization problem:

$$\min_{P \in \triangle_{Q_1}(\mathcal{M}_k)} \int \left(\sum_{j=1}^{k} \frac{1}{\lambda_j} \mathsf{L}_j\left(\boldsymbol{z}_j, \boldsymbol{\theta}\right)\right) \mathrm{d}P(\boldsymbol{\theta}) + D\left(P \parallel Q_1\right). \tag{24}$$

The following corollary of Theorem 4 formalizes this observation.

Corollary 5. *Under the assumption that the measures Q_1, Q_2, ..., Q_K satisfy* (21b)*, the solutions to the optimization problems in* (17) *and* (24) *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* (23)*.*

Given $k > 1$, the optimization problem in (17) depends exclusively on the local training dataset $\boldsymbol{z}_k$. While the reference Q_k in (21b) depends on the training datasets of the $k-1$ previous clients, the dependence is not direct. Solving (17), 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, the optimization problem in (24) depends on the training datasets of client j, for all $j \in \{1, 2, \ldots, k\}$, while the reference measure Q_1 does not depend on any training dataset. The fact that problems (17) and (24) share the same solution unveils an important observation: providing client k with a reference measure Q_k of the form in (21b) reproduces the effect of having access to the training datasets of the $k-1$ previous clients, without transmitting those datasets.

4.3 Centralized performance guarantees

An important observation is that a strategic choice of λ_1, λ_2, ..., λ_K in (23) can lead to achieve the same Gibbs probability distribution as in a setting in which the training datasets of all clients are available at all clients. This describes a decentralized system whose distributed nature does not limit achieving the same Gibbs algorithm that one would have obtained if all the training data were available at all clients. The following theorem formalizes this observation.

Theorem 6. *Assume that the loss functions in* (11) *satisfy* $\ell_1 = \ell_2 = \ldots = \ell_K = \ell$ *and for all* $k \in \{1, \ldots, K\}$,

$$\lambda_k = \frac{n_0}{n_k}\lambda_0, \tag{25}$$

for some $\lambda_0 > 0$ *and some loss function* ℓ. *Consider some measures* Q_1, Q_2, $\cdots$, Q_K *satisfying* (21b). *Then, for all* $\boldsymbol{\theta} \in \operatorname{supp} Q_1$, *it follows that,*

$$\frac{\mathrm{d}P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}}{\mathrm{d}P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}}(\boldsymbol{\theta}) = 1, \tag{26}$$

where the probability measure $P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}$ *is defined in* (23); *the measure* $P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ *satisfies for all* $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\exp\left(\frac{-1}{n_0\lambda_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i},y_{0,i},\boldsymbol{\theta}\right)\right)}{\int\exp\left(\frac{-1}{n_0\lambda_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i},y_{0,i},\boldsymbol{\nu}\right)\right)\mathrm{d}Q_1(\boldsymbol{\nu})}; \tag{27}$$

and $\left(x_{0,i}, y_{0,i}\right)$ *are data points of the aggregated dataset* $\boldsymbol{z}_0$ *in* (8).

Proof: The proof is presented in Appendix C. ∎

The relevance of Theorem 6 is highlighted by the following observations. Under the assumptions of Theorem 6, in particular that the loss functions satisfy $\ell_1 = \ell_2 = \cdots = \ell_K = \ell$, the equality in (25) together with [21, Lemma 4] allows rewriting the optimization problem in (24) as

$$\min_{P\in\triangle_{Q_1}(\mathcal{M}_K)}\int\frac{1}{n_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i},y_{0,i},\boldsymbol{\theta}\right)\mathrm{d}P(\boldsymbol{\theta}) + \lambda_0 D(P \parallel Q_1), \tag{28}$$

which requires access to the training datasets of all clients. Interestingly, from [19, Lemma 1], it follows that the probability measure $P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (27) is the solution to (28). More importantly, from [19, Lemma 4], if λ_0 is chosen such that

$$D\left(P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}\,\middle\|\,Q_1\right) = \gamma_0, \tag{29}$$

for some $\gamma_0 > 0$, the probability measure $P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}$ in (27) is also the solution to the following optimization problem:

$$\min_{P\in\triangle_{Q_1}(\mathcal{M}_K)} \quad \int\frac{1}{n_0}\sum_{i=1}^{n_0}\ell\left(x_{0,i},y_{0,i},\boldsymbol{\theta}\right)\mathrm{d}P\left(\boldsymbol{\theta}\right) \tag{30a}$$

$$\text{s.t.} \quad D\left(P \parallel Q_1\right) \leqslant \gamma_0. \tag{30b}$$

Hence, given that from (26), 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, the nested structure induced by the choice of reference measures in (21b) achieves the same Gibbs probability measure as a centralized system in which all training

datasets are available at client K. More specifically, under the forward–backward communication protocol in Section 4.1, after $K-1$ forward messages (blue arrows in Figure 1) client K obtains the Gibbs algorithm that minimizes the empirical risk with respect to all training datasets within the following neighborhood of Q_1,

$$\mathcal{G}_0 \triangleq \left\{ P \in \triangle_{Q_1}(\mathcal{M}_K) : D\left(P \parallel Q_1\right) \leqslant D\left(P^{(Q_1,\lambda_K)}_{\mathbf{\Theta}_K|\mathbf{Z}_0=\mathbf{z}_0} \,\middle\|\, Q_1\right) \right\}. \tag{31}$$

The backward dissemination (red arrows in Figure 1) then provides each client such a Gibbs algorithm.

Figure 2 provides a representation of the nested structure induced by the choice of reference measure in (21) under the assumption in (25): it depicts the successive displacement of the reference measure along the forward pass of the communication protocole, together with the corresponding sets $\mathcal{G}_k$ in (22), with $k \in \{1, 2, \ldots, K\}$, induced by the relative-entropy regularization. Starting from the reference measure Q_1, client 1 outputs the Gibbs probability measure $P^{(Q_1,\lambda_1)}_{\mathbf{\Theta}_1|\mathbf{Z}_1=\mathbf{z}_1}$ in (27), represented by the point Q_2, which belongs to the set $\mathcal{G}_1$. This locally obtained measure then becomes the reference measure for client 2, whose Gibbs probability measure $P^{(Q_2,\lambda_2)}_{\mathbf{\Theta}_2|\mathbf{Z}_2=\mathbf{z}_2}$ in (27) is represented by Q_3 and lies in $\mathcal{G}_2$. Iterating this construction yields the sequence $Q_1 \to Q_2 \to Q_3 \to Q_4 \to Q_5$, where each point Q_{k+1} represents $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (27) and each dashed region $\mathcal{G}_k$ describes the admissible neighborhood around Q_k enforced by (22).

In particular, the figure shows the sets $\mathcal{G}_0$ in (31) and $\mathcal{G}_k$ in (22) together with the probability measures $P^{(Q_k,\lambda_k)}_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (23), with $k \in \{1, 2, \ldots, K\}$, and $P^{(Q_1,\lambda_K)}_{\mathbf{\Theta}_K|\mathbf{Z}_0=\mathbf{z}_0}$ in (27). The endpoint Q_5 is annotated as $Q_5 = P^{(Q_1,\lambda_K)}_{\mathbf{\Theta}_K|\mathbf{Z}_0=\mathbf{z}_0}$, highlighting that the nested reference-measure construction produced by the communication protocol reaches the same Gibbs probability measure as the direct construction on the aggregated dataset within the set $\mathcal{G}_0$ in (31).

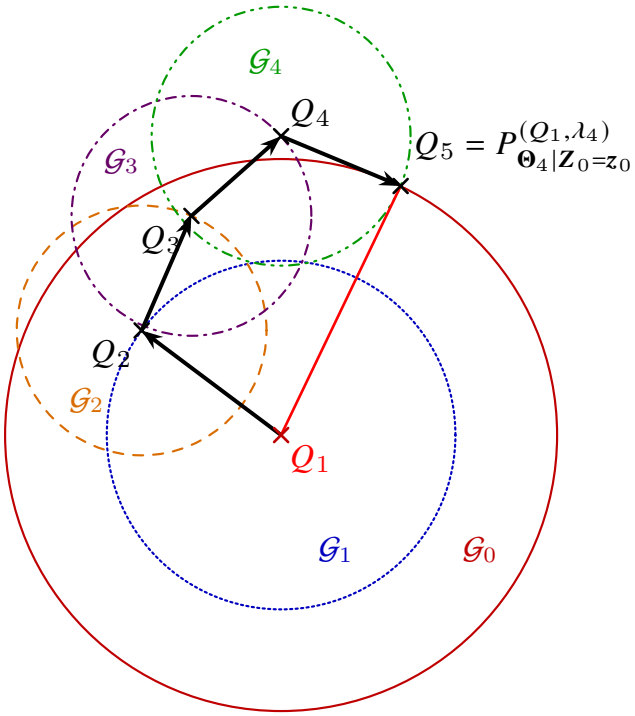


Figure 2: Geometric representation, under the assumptions in (25), of the sets $\mathcal{G}_0$ in (31) and $\mathcal{G}_k$ in (22), with $k \in \{1, 2, \ldots, 4\}$. The points Q_{k+1} represent the Gibbs probability measures $P^{(Q_k,\lambda_k)}_{\Theta_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (23), and the endpoint is $Q_5 = P^{(Q_1,\lambda_4)}_{\Theta_4|\mathbf{Z}_0=\mathbf{z}_0}$ as in (27).

5 Conclusion and Final Remarks

In this work, it has been shown that there exists a choice of reference measures $(Q_1, Q_2, \ldots, Q_K)$ and regularization factors $(\lambda_1, \lambda_2, \ldots, \lambda_K)$ in a decentralized machine learning scenario such that the same performance of a centralized system in which all training datasets are aggregated and made available is achievable. 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 data points in the training dataset of client k and the total number of data points across the training datasets of all clients. The reference measure of client k, $k > 1$, is the Gibbs measure (Gibbs algorithm) used by client $k-1$ to sample its models. The first client uses a given reference Q_1. Client K can back propagate its achieved Gibbs probability measure (Gibbs algorithm) across all clients in such a way that all clients achieve the same performance. That is, they sample their models from the same probability measure, which minimizes the empirical risk with respect to the training datasets of all clients within a neighborhood of Q_1.

This choice of reference measures leads to a nested structure whose construction requires the transmission of $K-1$ probability measures with common support [6, Lemma 3]. The practical construction of this nested structure faces several challenges. The most prominent is a consequence of data transmission within a finite number of bits and delay constraints. This implies that the Gibbs probability measure transmitted by client k might be received by client $k+1$ subject to distorsion. The impact of such a distorsion in the construction of the nested structure is an open problem and has not been taken into account in this work. Another challenge is related to the fact that the Gibbs measures to be transmitted might have supports that are significantly large and thus, the amount of data to be transmitted might be comparable to transmitting the training datasets. Nonetheless, the transmission of a probability measure is more privacy preserving than the actual transmission of training datasets. In a nutshell, several important theoretical discoveries have been reported in this work concerning decentralized learning.

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, Sept. 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, p. 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, July 2024, pp. 2168 – 2229.

[19] 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.

[20] 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, July 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, June 2023, pp. 328–333.

[22] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, Sept. 1946.

Appendices

A Preliminaries for the Proof of Theorem 4

The following lemma establishes a relationship between the $(\mathsf{L}_k, Q_k, \lambda_k)$-Gibbs probability measure in (21) and an $(\mathsf{L}_k, Q_1, \lambda_k)$-Gibbs probability measure $P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \in \triangle(\mathcal{M}_k)$.

Lemma 7. *For all $k \in \{1, 2, \ldots, K\}$, consider 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 (21). Then, for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,*

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \exp(C_k), \tag{32}$$

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

$$C_k = \log\left(\frac{\int \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i}\mathsf{L}_i(\boldsymbol{z}_i, \tilde{\boldsymbol{\nu}})\right) \mathrm{d}Q_1(\tilde{\boldsymbol{\nu}}) \left(\exp\left(\mathsf{K}_{k,Q_1,\boldsymbol{z}_k}\left(\frac{-1}{\lambda_k}\right)\right)\right)}{\int \exp\left(\sum_{j=1}^{k} \frac{-1}{\lambda_j}\mathsf{L}_j(\boldsymbol{z}_j, \boldsymbol{\nu})\right) \mathrm{d}Q_1(\boldsymbol{\nu})}\right), \tag{33}$$

where the functional $\mathsf{K}_{k,Q_1,\boldsymbol{z}_k}$ is defined as in (13).

Proof: The proof of (32) is done by induction on k.
Base case ($k = 2$): From Definition 4 and Q_2 in (21b), for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$, it follows that

$$\begin{aligned}
&\frac{\mathrm{d}P^{\left(P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}_1|\boldsymbol{Z}_1=\boldsymbol{z}_1},\lambda_2\right)}_{\boldsymbol{\Theta}_2|\boldsymbol{Z}_2=\boldsymbol{z}_2}}{\mathrm{d}P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}_1|\boldsymbol{Z}_1=\boldsymbol{z}_1}}(\boldsymbol{\theta}) \\
&= \exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2, \boldsymbol{\theta}) - \mathsf{K}_{2,P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}_1|\boldsymbol{Z}_1=\boldsymbol{z}_1},\boldsymbol{z}_2}\left(\frac{-1}{\lambda_2}\right)\right) && (34)\\
&= \frac{\exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2, \boldsymbol{\theta})\right)}{\int \exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2, \boldsymbol{\nu})\right) \frac{\mathrm{d}P^{(Q_1,\lambda_1)}_{\boldsymbol{\Theta}_1|\boldsymbol{Z}_1=\boldsymbol{z}_1}}{\mathrm{d}Q_1}(\boldsymbol{\nu}) \mathrm{d}Q_1(\boldsymbol{\nu})} && (35)\\
&= \frac{\exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2, \boldsymbol{\theta})\right)}{\int \exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2, \boldsymbol{\nu})\right) \left(\frac{\exp\left(\frac{-1}{\lambda_1}\mathsf{L}_1(\boldsymbol{z}_1, \boldsymbol{\nu})\right)}{\int \exp\left(\frac{-1}{\lambda_1}\mathsf{L}_1(\boldsymbol{z}_1, \boldsymbol{\theta}')\right) \mathrm{d}Q_1(\boldsymbol{\theta}')}\right) \mathrm{d}Q_1(\boldsymbol{\nu})} && (36)\\
&= \frac{\exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2, \boldsymbol{\theta})\right) \int \exp\left(\frac{-1}{\lambda_1}\mathsf{L}_1(\boldsymbol{z}_1, \boldsymbol{\theta}')\right) \mathrm{d}Q_1(\boldsymbol{\theta}')}{\int \exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2(\boldsymbol{z}_2, \boldsymbol{\nu})\right) \exp\left(\frac{-1}{\lambda_1}\mathsf{L}_1(\boldsymbol{z}_1, \boldsymbol{\nu})\right) \mathrm{d}Q_1(\boldsymbol{\nu})} && (37)
\end{aligned}$$

$$= \frac{\exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2\left(z_2,\boldsymbol{\theta}\right)\right)\int \exp\left(\frac{-1}{\lambda_1}\mathsf{L}_1\left(z_1,\boldsymbol{\theta}'\right)\right)\mathrm{d}Q_1(\boldsymbol{\theta}')}{\int \exp\left(\frac{-1}{\lambda_2}\mathsf{L}_2\left(z_2,\boldsymbol{\nu}\right)\right)\exp\left(\frac{-1}{\lambda_1}\mathsf{L}_1\left(z_1,\boldsymbol{\nu}\right)\right)\mathrm{d}Q_1(\boldsymbol{\nu})}\frac{\exp\left(\mathsf{K}_{2,Q_1,z_2}\left(\frac{-1}{\lambda_2}\right)\right)}{\exp\left(\mathsf{K}_{2,Q_1,z_2}\left(\frac{-1}{\lambda_2}\right)\right)} \tag{38}$$

$$= \frac{\mathrm{d}P^{(Q_1,\lambda_2)}_{\boldsymbol{\Theta}_2|\mathbf{Z}_2=z_2}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\frac{\exp\left(\mathsf{K}_{1,Q_1,z_1}\left(\frac{-1}{\lambda_1}\right)+\mathsf{K}_{2,Q_1,z_2}\left(\frac{-1}{\lambda_2}\right)\right)}{\int \exp\left(\frac{-1}{\lambda_1}\mathsf{L}_1\left(z_1,\boldsymbol{\nu}\right)+\frac{-1}{\lambda_2}\mathsf{L}_2\left(z_2,\boldsymbol{\nu}\right)\right)\mathrm{d}Q_1(\boldsymbol{\nu})}, \tag{39}$$

where (35) follows from [20, Theorem 2]; and (36) follows from Definition 4. The equality in (39) corresponds to (32) in the case of $k = 2$.

Induction hypothesis: Assume that (32) holds for $k \geqslant 2$.

Induction step ($k \to k+1$): It remains to prove (32) for $k+1$. Consider client $k+1$ and its reference measure Q_k in (21b). From Definition 4, it holds that for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_{k+1},\lambda_{k+1})}_{\boldsymbol{\Theta}_{k+1}|\mathbf{Z}=z_{k+1}}}{\mathrm{d}Q_{k+1}}(\boldsymbol{\theta}) = \frac{\exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(z_{k+1},\boldsymbol{\theta}\right)\right)}{\int \exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(z_{k+1},\boldsymbol{\nu}\right)\right)\mathrm{d}Q_{k+1}(\boldsymbol{\nu})} \tag{40}$$

$$= \frac{\exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(z_{k+1},\boldsymbol{\theta}\right)\right)}{\int \exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(z_{k+1},\boldsymbol{\nu}\right)\right)\frac{\mathrm{d}Q_{k+1}}{\mathrm{d}Q_k}(\boldsymbol{\nu})\mathrm{d}Q_k(\boldsymbol{\nu})} \tag{41}$$

$$= \frac{\exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}(z_{k+1},\boldsymbol{\theta})\right)}{\int \exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}(z_{k+1},\boldsymbol{\nu})\right)\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}}{\mathrm{d}Q_k}(\boldsymbol{\nu})\mathrm{d}Q_k(\boldsymbol{\nu})}, \tag{42}$$

where the equality in (41) follows from [20, Theorem 2]; and the equality in (42) follows from (21b). By the induction hypothesis, it holds that, for all $\boldsymbol{\nu} \in \operatorname{supp} Q_1$

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}}{\mathrm{d}Q_k}(\boldsymbol{\nu})$$

$$= \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\nu})\frac{\int \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(z_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})\left(\exp\left(\mathsf{K}_{k,Q_1,z_k}\left(\frac{-1}{\lambda_k}\right)\right)\right)}{\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(z_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})} \tag{43}$$

$$= \frac{\exp\left(\frac{-1}{\lambda_k}\mathsf{L}_k\left(z_k,\boldsymbol{\nu}\right)\right)\int \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(z_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})}{\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(z_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}. \tag{44}$$

Substituting (44) into the denominator of (42) yields for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_{k+1},\lambda_{k+1})}_{\boldsymbol{\Theta}_{k+1}|\mathbf{Z}=z_{k+1}}}{\mathrm{d}Q_{k+1}}(\boldsymbol{\theta})$$

$$= \frac{\exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(\boldsymbol{z}_{k+1},\boldsymbol{\theta}\right)\right)}{\int \exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(\boldsymbol{z}_{k+1},\boldsymbol{\nu}\right)\right)\exp\left(\frac{-1}{\lambda_k}\mathsf{L}_k\left(\boldsymbol{z}_k,\boldsymbol{\nu}\right)\right)\mathrm{d}Q_k(\boldsymbol{\nu})} \cdot \frac{\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}{\int \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})} \tag{45}$$

$$= \frac{\exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(\boldsymbol{z}_{k+1},\boldsymbol{\theta}\right)\right)}{\int \exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(\boldsymbol{z}_{k+1},\boldsymbol{\nu}\right)+\frac{-1}{\lambda_k}\mathsf{L}_k\left(\boldsymbol{z}_k,\boldsymbol{\nu}\right)\right)\frac{\mathrm{d}Q_k}{\mathrm{d}Q_{k-1}}(\boldsymbol{\nu})\cdots\frac{\mathrm{d}Q_3}{\mathrm{d}Q_2}(\boldsymbol{\nu})\frac{\mathrm{d}Q_2}{\mathrm{d}Q_1}(\boldsymbol{\nu})\,\mathrm{d}Q_1(\boldsymbol{\nu})} \cdot \frac{\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}{\int \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})}, \tag{46}$$

where (46) follows from [20, Theorem 4]. Moreover, for all $\boldsymbol{\nu} \in \operatorname{supp} Q_1$, it follows that,

$$\prod_{j=2}^{k}\left(\frac{\mathrm{d}Q_j}{\mathrm{d}Q_{j-1}}(\boldsymbol{\nu})\right) \tag{47}$$

$$= \prod_{j=2}^{k}\left(\frac{\mathrm{d}P^{(Q_{j-1},\lambda_{j-1})}_{\boldsymbol{\Theta}_{j-1}|\mathbf{Z}_{j-1}=\mathbf{z}_{j-1}}}{\mathrm{d}Q_{j-1}}(\boldsymbol{\nu})\right) \tag{48}$$

$$= \prod_{j=1}^{k-1}\left(\frac{\mathrm{d}P^{(Q_j,\lambda_j)}_{\boldsymbol{\Theta}_j|\mathbf{Z}_j=\mathbf{z}_j}}{\mathrm{d}Q_j}(\boldsymbol{\nu})\right) \tag{49}$$

$$= \prod_{j=1}^{k-1}\left(\frac{\exp\left(\frac{-1}{\lambda_j}\mathsf{L}_j\left(\boldsymbol{z}_j,\boldsymbol{\nu}\right)\right)\int \exp\left(\sum_{i=1}^{j-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})}{\int \exp\left(\sum_{i=1}^{j}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}\right) \tag{50}$$

$$= \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\boldsymbol{\nu}\right)\right)\prod_{j=1}^{k-1}\left(\frac{\int \exp\left(\sum_{i=1}^{j-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})}{\int \exp\left(\sum_{i=1}^{j}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}\right) \tag{51}$$

$$= \frac{\exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\boldsymbol{\nu}\right)\right)}{\int \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}, \tag{52}$$

where the equality in (49) follows from (21b); the equality in (49) follows by changing the indices in (47); and the equality in (50) follows from the induction hypothesis.

Substituting (52) into (46) yields for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_{k+1},\lambda_{k+1})}_{\boldsymbol{\Theta}_{k+1}|\mathbf{Z}=\mathbf{z}_{k+1}}}{\mathrm{d}Q_{k+1}}(\boldsymbol{\theta})$$

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

$$\cdot\left(\int \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})\right)\frac{\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}{\int \exp\left(\sum_{i=1}^{k-1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\nu}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\nu}})} \tag{53}$$

$$= \frac{\exp\left(\frac{-1}{\lambda_{k+1}}\mathsf{L}_{k+1}\left(\boldsymbol{z}_{k+1},\boldsymbol{\theta}\right)\right)\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})}{\int \exp\left(\sum_{i=1}^{k+1}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\boldsymbol{\nu}\right)\right)\mathrm{d}Q_1(\boldsymbol{\nu})} \tag{54}$$

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

$$\cdot\frac{\exp\left(\mathsf{K}_{k+1,Q_1,\boldsymbol{z}_{k+1}}\left(\frac{-1}{\lambda_{k+1}}\right)\right)}{\exp\left(\mathsf{K}_{k+1,Q_1,\boldsymbol{z}_{k+1}}\left(\frac{-1}{\lambda_{k+1}}\right)\right)} \tag{55}$$

$$= \frac{\mathrm{d}P^{(Q_1,\lambda_{k+1})}_{\boldsymbol{\Theta}_{k+1}|\boldsymbol{Z}_{k+1}=\boldsymbol{z}_{k+1}}(\boldsymbol{\theta})}{\mathrm{d}Q_1}$$

$$\frac{\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})\left(\exp\left(\mathsf{K}_{k+1,Q_1,\boldsymbol{z}_{k+1}}\left(\frac{-1}{\lambda_{k+1}}\right)\right)\right)}{\int \exp\left(\sum_{j=1}^{k+1}\frac{-1}{\lambda_j}\mathsf{L}_j\left(\boldsymbol{z}_j,\boldsymbol{\nu}\right)\right)\mathrm{d}Q_1(\boldsymbol{\nu})}, \tag{56}$$

where (54) follows by simplifying the ratio of products in the second factor in (53).

The second factor in (56) does not depend on $\boldsymbol{\theta}$ and is denoted by

$$C_{k+1} \triangleq \log\left(\frac{\int \exp\left(\sum_{i=1}^{k}\frac{-1}{\lambda_i}\mathsf{L}_i\left(\boldsymbol{z}_i,\tilde{\boldsymbol{\theta}}\right)\right)\mathrm{d}Q_1(\tilde{\boldsymbol{\theta}})\left(\exp\left(\mathsf{K}_{k+1,Q_1,\boldsymbol{z}_{k+1}}\left(\frac{-1}{\lambda_{k+1}}\right)\right)\right)}{\int \exp\left(\sum_{j=1}^{k+1}\frac{-1}{\lambda_j}\mathsf{L}_j\left(\boldsymbol{z}_j,\boldsymbol{\nu}\right)\right)\mathrm{d}Q_1(\boldsymbol{\nu})}\right). \tag{57}$$

Combining (56) and (57) yields for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_{k+1},\lambda_{k+1})}_{\boldsymbol{\Theta}_{k+1}|\boldsymbol{Z}=\boldsymbol{z}_{k+1}}}{\mathrm{d}Q_{k+1}}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q_1,\lambda_{k+1})}_{\boldsymbol{\Theta}_{k+1}|\boldsymbol{Z}_{k+1}=\boldsymbol{z}_{k+1}}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\exp\left(C_{k+1}\right), \tag{58}$$

which corresponds to (32) in the case of $k+1$ and concludes the proof. ∎

B Proof of Theorem 4

From Lemma 7, for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$, it follows that,

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\exp\left(C_k\right), \tag{59}$$

with

$$C_k = \log\left(\frac{\int \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \tilde{\nu}\right)\right) \mathrm{d}Q_1(\tilde{\nu}) \left(\exp\left(\mathsf{K}_{k,Q_1,z_k}\left(\frac{-1}{\lambda_k}\right)\right)\right)}{\int \exp\left(\sum_{j=1}^{k} \frac{-1}{\lambda_j} \mathsf{L}_j\left(z_j, \nu\right)\right) \mathrm{d}Q_1(\nu)}\right). \quad (60)$$

From [6, Lemma 3] and [20, Theorem 2] the Radon–Nikodym derivatives $\frac{\mathrm{d}Q_1}{\mathrm{d}Q_k}$ and $\frac{\mathrm{d}Q_k}{\mathrm{d}Q_1}$ are well defined. From the chain rule of the Radon-Nikodym derivative [20, Theorem 4], for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$, it follows that,

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \frac{\mathrm{d}Q_1}{\mathrm{d}Q_k}(\boldsymbol{\theta}). \quad (61)$$

Moreover, from the multiplicative inverse of the Radon-Nikodym derivative [20, Theorem 5], by combining (59) and (61), for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$, it follows that,

$$\begin{aligned}
&\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \\
&= \frac{\mathrm{d}Q_k}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \exp(C_k) && (62)\\
&= \frac{\mathrm{d}Q_k}{\mathrm{d}Q_{k-1}}(\boldsymbol{\theta}) \frac{\mathrm{d}Q_{k-1}}{\mathrm{d}Q_{k-2}}(\boldsymbol{\theta}) \ldots \frac{\mathrm{d}Q_3}{\mathrm{d}Q_2}(\boldsymbol{\theta}) \frac{\mathrm{d}Q_2}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \exp(C_k) && (63)\\
&= \prod_{j=2}^{k} \left(\frac{\mathrm{d}Q_j}{\mathrm{d}Q_{j-1}}(\boldsymbol{\theta})\right) \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \exp(C_k), && (64)
\end{aligned}$$

where (63) follows from [20, Theorem 4]. From Lemma 7 and (52), for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$ it follows that

$$\prod_{j=2}^{k} \left(\frac{\mathrm{d}Q_j}{\mathrm{d}Q_{j-1}}(\boldsymbol{\theta})\right) = \frac{\exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \nu\right)\right) \mathrm{d}Q_1(\nu)}. \quad (65)$$

Plugging (65) into (64) yield

$$\begin{aligned}
&\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \\
&= \frac{\exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \nu\right)\right) \mathrm{d}Q_1(\nu)} \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \exp(C_k) && (66)\\
&= \frac{\exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \nu\right)\right) \mathrm{d}Q_1(\nu)} \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})
\end{aligned}$$

$$\frac{\int \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \tilde{\boldsymbol{\nu}}\right)\right) \mathrm{d}Q_1(\tilde{\boldsymbol{\nu}}) \left(\exp\left(\mathsf{K}_{k,Q_1,\boldsymbol{z}_k}\left(\frac{-1}{\lambda_k}\right)\right)\right)}{\int \exp\left(\sum_{i=1}^{k} \frac{-1}{\lambda_i} \mathsf{L}_i\left(z_i, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})} \tag{67}$$

$$= \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \boldsymbol{\theta}\right)\right) \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \frac{\exp\left(\mathsf{K}_{k,Q_1,\boldsymbol{z}_k}\left(\frac{-1}{\lambda_k}\right)\right)}{\int \exp\left(\sum_{i=1}^{k} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})} \tag{68}$$

$$= \frac{\exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \boldsymbol{\theta}\right)\right) \exp\left(\frac{-1}{\lambda_k} \mathsf{L}_k\left(\boldsymbol{z}_k, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{i=1}^{k} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})} \tag{69}$$

$$= \frac{\exp\left(\sum_{i=1}^{k} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{i=1}^{k} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})}, \tag{70}$$

where (67) follows from (60); and (69) follows from Definition 4. This completes the proof.

C Proof of Theorem 6

For all $k \in \{1, \ldots, K\}$, under the assumption in (25), it holds that $\lambda_k = \frac{n_0}{n_k}\lambda_0$. From Theorem 4, it follows that for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\exp\left(\sum_{k=1}^{K} \frac{-n_k}{n_0\lambda_0} \mathsf{L}_k\left(\boldsymbol{z}_k, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{k=1}^{K} \frac{-n_k}{n_0\lambda_0} \mathsf{L}_k\left(\boldsymbol{z}_k, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})} \tag{71}$$

$$= \frac{\exp\left(\sum_{k=1}^{K} \frac{-n_k}{n_0\lambda_0} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k\left(x_{k,i}, y_{k,i}, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\sum_{k=1}^{K} \frac{-n_k}{n_0\lambda_0} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k\left(x_{k,i}, y_{k,i}, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})} \tag{72}$$

$$= \frac{\exp\left(\frac{-1}{n_0\lambda_0} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \ell_k\left(x_{k,i}, y_{k,i}, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\frac{-1}{n_0\lambda_0} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \ell_k\left(x_{k,i}, y_{k,i}, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})}. \tag{73}$$

where the equality in (72) follows from (12).

Under the assumption on the loss functions, *i.e.*, $\ell_1 = \ell_2 = \cdots = \ell_K = \ell$ and the fact that $\boldsymbol{z}_0$ in (8) is the aggregation of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K$, and $n_0 = \sum_{k=1}^{K} n_k$, the equality in (73) yields for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) = \frac{\exp\left(\frac{-1}{n_0\lambda_0} \sum_{i=1}^{n_0} \ell\left(x_{0,i}, y_{0,i}, \boldsymbol{\theta}\right)\right)}{\int \exp\left(\frac{-1}{n_0\lambda_0} \sum_{i=1}^{n_0} \ell\left(x_{0,i}, y_{0,i}, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})} \tag{74}$$

$$= \frac{\mathrm{d}P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}}{\mathrm{d}Q_1}(\boldsymbol{\theta}), \tag{75}$$

where the equality in (75) follows from (27).

Finally, from [20, Theorem 5] and (75), for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$,

$$1 = \frac{\mathrm{d}P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \; \frac{\mathrm{d}Q_1}{\mathrm{d}P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}}(\boldsymbol{\theta}) \tag{76}$$

$$= \frac{\mathrm{d}P^{(Q_K,\lambda_K)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_K=\boldsymbol{z}_K}}{\mathrm{d}P^{(Q_1,\lambda_0)}_{\boldsymbol{\Theta}_K|\boldsymbol{Z}_0=\boldsymbol{z}_0}}(\boldsymbol{\theta}), \tag{77}$$

where the equality in (77) follows from [20, Theorem 4] and completes the proof.

D Complementary Results

The following lemma provides an information-theoretic expression for C_k in (33) in terms of relative entropies.

Lemma 8. *The value C_k in* (33) *satisfies*

$$C_k = D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q_k\right) - D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q_{k+1}\right) - D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q_1\right) \tag{78}$$

$$= D\left(Q_1 \,\|\, Q_k\right) - D\left(Q_1 \,\|\, Q_{k+1}\right) + D\left(Q_1 \,\middle\|\, P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right), \tag{79}$$

where the measure $P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$, is an $(\mathsf{L}_k, Q_1, \lambda_k)$-Gibbs probability measure; and the measure Q_k is defined in (21b).

Proof: From Lemma 7, for all $\boldsymbol{\theta} \in \operatorname{supp} Q_1$, it follows that

$$\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta}) = \frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta}) \exp\left(C_k\right), \tag{80}$$

with

$$C_k = \log\left(\frac{\int \exp\left(\sum_{i=1}^{k-1} \frac{-1}{\lambda_i} \mathsf{L}_i\left(\boldsymbol{z}_i, \tilde{\boldsymbol{\nu}}\right)\right) \mathrm{d}Q_1(\tilde{\boldsymbol{\nu}}) \left(\exp\left(\mathsf{K}_{k,Q_1,\boldsymbol{z}_k}\left(\frac{-1}{\lambda_k}\right)\right)\right)}{\int \exp\left(\sum_{j=1}^{k} \frac{-1}{\lambda_j} \mathsf{L}_j\left(\boldsymbol{z}_j, \boldsymbol{\nu}\right)\right) \mathrm{d}Q_1(\boldsymbol{\nu})}\right). \tag{81}$$

By taking a logarithm and integrating with respect to $P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}$ in (80), it follows that

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

$$= \int \log\left(\frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\right) \mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}(\boldsymbol{\theta}) + C_k \tag{82}$$

$$= D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q_1\right) + C_k. \tag{83}$$

On the other hand, from Lemma 1 it follows that

$$\int \log\left(\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta})\right)\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}(\boldsymbol{\theta})$$
$$= D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_k\right) - D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right). \tag{84}$$

Combining (83) and (84) yields

$$C_k = D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_k\right) - D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_{k+1}\right) - D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_1\right), \tag{85}$$

which completes the proof of (78).

Following the same method, if the integral in (80) is taken with respect to Q_1 it then follows that,

$$\int \log\left(\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta})\right)\mathrm{d}Q_1(\boldsymbol{\theta})$$
$$= \int \log\left(\frac{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}{\mathrm{d}Q_1}(\boldsymbol{\theta})\right)\mathrm{d}Q_1(\boldsymbol{\theta}) + C_k \tag{86}$$
$$= \int \log\left(\left(\frac{\mathrm{d}Q_1}{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta})\right)^{-1}\right)\mathrm{d}Q_1(\boldsymbol{\theta}) + C_k \tag{87}$$
$$= -\int \log\left(\frac{\mathrm{d}Q_1}{\mathrm{d}P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta})\right)\mathrm{d}Q_1(\boldsymbol{\theta}) + C_k \tag{88}$$
$$= -D\left(Q_1 \,\middle\|\, P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right) + C_k, \tag{89}$$

where (87) follows from [6, Lemma 3], [20, Theorem 2] and [20, Theorem 5].

On the other hand, from Lemma 1 it follows that

$$\int \log\left(\frac{\mathrm{d}P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta})\right)\mathrm{d}Q_1(\boldsymbol{\theta})$$
$$= D(Q_1 \,\|\, Q_k) - D\left(Q_1 \,\middle\|\, P^{(Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right). \tag{90}$$

Combining (89) and (90) yields

$$C_k = D(Q_1 \,\|\, Q_k) - D(Q_1 \,\|\, Q_{k+1}) + D\left(Q_1 \,\middle\|\, P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right), \tag{91}$$

which completes the proof of (79) and concludes the proof. ∎

The explicit characterization of this constant makes clear that the nested choice of reference measures is not innocuous: it determines how the information contained in previous clients is incorporated through a client's reference measure

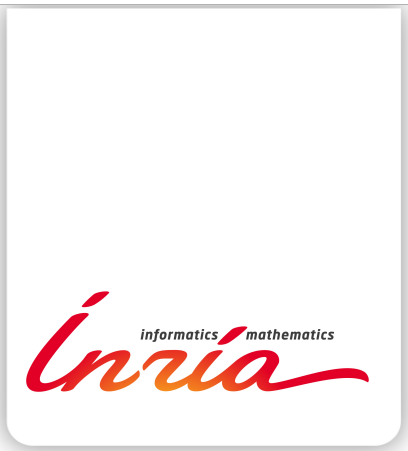
Q_k in (21b). This observation motivates comparing choices of reference measures in (21b) Q_1 or Q_k.

Lemma 8 implies a characterization of the Jeffreys divergence [22] between the $(\mathsf{L}_k, Q_1, \lambda_k)$-Gibbs probability measure $P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (21) and Q_k in (21b) , as shown by the following corollary.

Corollary 9. *The* $(\mathsf{L}_k, Q_1, \lambda_k)$*-Gibbs probability measure* $P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ *in* (21)*, satisfies*

$$\begin{aligned} & D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_1\right) + D\left(Q_1 \,\middle\|\, P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right) \\ = & D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_k\right) - D\left(Q_1 \,\|\, Q_k\right) \\ & + D\left(Q_1 \,\|\, Q_{k+1}\right) - D\left(P^{(Q_1,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k} \,\middle\|\, Q_{k+1}\right), \end{aligned} \tag{92}$$

where the measure Q_k *is defined in* (21b)*.*



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399