Equivalence Between Nested Gibbs Measures and Log-Linear Combinations of Gibbs Measures

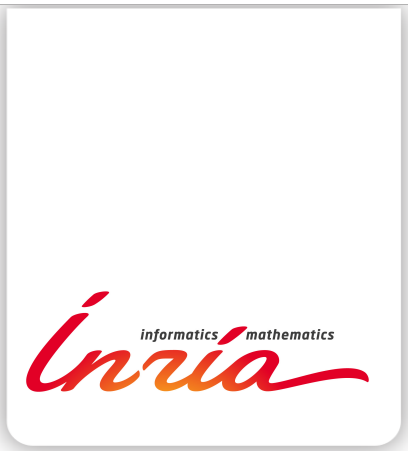
Yaiza Bermudez, Samir M. Perlaza, Iñaki Esnaola

Equivalence Between Nested Gibbs Measures and Log-Linear Combinations of Gibbs Measures

Yaiza Bermudez, Samir M. Perlaza, and Iñaki Esnaola

RESEARCH REPORT
N° 9613
May 2026
Project-Team AT-SOP

ISSN 0249-6399 ISRN INRIA/RR--9613--FR+ENG



Equivalence Between Nested Gibbs Measures and Log-Linear Combinations of Gibbs Measures

Yaiza Bermudez, Samir M. Perlaza, and Iñaki Esnaola

Project-Team AT-SOP

Research Report n° 9613 — May 2026 — 27 pages

Abstract: In this report, three operations on Gibbs probability measures are studied. The first operation, which takes as argument one Gibbs probability measure and is often referred to as renormalization, consists in generating a new Gibbs measure by normalizing a power of the density of the given measure. Such a normalization has a twofold effect: first, it changes the regularization factor; and second, it concentrates the support within a subset of the original support. Interestingly, it is shown that these effects can be independently controlled by different parameters. The second operation, which takes as argument two Gibbs probability measures, consists of changing the reference measure of the latter by the former. Hence, the former is said to be "nested" within the latter, yielding a new Gibbs probability measure. The third operation consists of a normalized log-linear combination of the densities of Gibbs probability measures. The resulting probability measures, from both second and third operations, which are also Gibbs probability measures, are shown, respectively, to be the solutions to optimization problems of the expectations of linear combinations of the objective functions of the given measures, subject to a relative entropy regularization. Such optimization problems differ exclusively in the coefficients of the linear combinations. This observation leads the conclusion that there exists a set of parameters for which nesting one Gibbs probability measure into another has the same effect as log-linearly combining them. These operations are shown to have relevant applications in statistical learning. As an example, a one-shot federated learning system in which clients send to the server their locally trained Gibbs algorithms for being log-linearly combined by the server, is shown to achieve the same performance as a Gibbs algorithm trained upon the aggregation of all local training datasets.

Key-words: Empirical Risk Minimization, Relative Entropy Regularization, Gibbs Probability Measures, Gibbs Algorithms, Renormalization, Nested Gibbs Measures, Log-Linear Combinations.

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 are also 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

Équivalence entre mesures de Gibbs imbriquées et combinaisons log-linéaires de mesures de Gibbs.

Résumé : Dans ce rapport, trois opérations sur des mesures de probabilité de Gibbs sont étudiées. La première opération, qui prend comme argument une mesure de probabilité de Gibbs et est souvent appelée renormalisation, consiste à générer une nouvelle mesure de Gibbs en normalisant une puissance de la densité de la mesure donnée. Une telle normalisation a un double effet : premièrement, elle modifie le facteur de régularisation ; et deuxièmement, elle concentre le support dans un sous-ensemble du support original. De manière intéressante, il est montré que ces effets peuvent être contrôlés indépendamment par différents paramètres. La deuxième opération, qui prend comme arguments deux mesures de probabilité de Gibbs, consiste à remplacer la mesure de référence de la seconde par la première. Ainsi, la première est dite « imbriquée » dans la seconde, ce qui donne lieu à une nouvelle mesure de probabilité de Gibbs. La troisième opération consiste en une combinaison log-linéaire normalisée des densités de mesures de probabilité de Gibbs. Les mesures de probabilité résultantes, issues des deuxième et troisième opérations, qui sont également des mesures de probabilité de Gibbs, sont montrées, respectivement, comme étant les solutions de problèmes d'optimisation des espérances de combinaisons linéaires des fonctions objectives des mesures données, soumises à une régularisation par entropie relative. De tels problèmes d'optimisation diffèrent exclusivement par les coefficients des combinaisons linéaires. Cette observation conduit à conclure qu'il existe un ensemble de paramètres pour lequel imbriquer une mesure de probabilité de Gibbs dans une autre a le même effet que les combiner de manière log-linéaire.

Mots-clés : Minimisation du Risque Empirique, Régularisation par Entropie Relative, Mesures de Probabilité de Gibbs, Algorithmes de Gibbs, Renormalisation, Measures de Gibbs Imbriquées, Combinaison Log-Linéaire.

Contents

1 Introduction

Gibbs probability measures have gained significant attention in statistical learning theory and information theory, in part due to their numerous properties and the fact that they represent the well-known Gibbs algorithm [1–10]. More importantly, as shown in [11], Gibbs measures form a benchmark to which algorithms can be compared for assessing their generalization error [9, 12–14]. Gibbs measures have also been shown to describe the long-run behavior of stochastic-gradient-based learning algorithms with time-invariant learning rate [15]; and to represent the worst-case data-generating probability distributions for a fixed model in supervised learning [16].

The main contributions of this work concern three operations on Gibbs probability measures. The first operation consists of constructing a Gibbs probability measure by normalizing (to one) a power of the density of a given Gibbs probability measure. This operation is classical in statistical physics and information theory, and is not limited to Gibbs measures. See for instance [17–26] and references therein. While some authors coined such operation as "renormalization", others coined as "escort distribution" the resulting probability distribution. In this work, renormalization is shown to lead to two independent effects on the original Gibbs measure: first, trimming the regularization factor; and second, concentrating the support to a subset of its original support (Theorem 1). Interestingly, both effects are shown to be independently controlled by different parameters of this operation. This new result complements existing evidence on the benefits of trimming the regularization factor in [4, 6, 12], as well as on the benefits of strategically choosing the reference measure as shown in [9, 27, 28]. Given two Gibbs probability measures, the second operation consists of constructing a new Gibbs probability measure by using the first measure as the reference measure of the second. The first measure is then said to be nested within the second, yielding a *nested Gibbs probability measure*. Nested Gibbs measures form a special class of Gibbs measures with important applications in statistical learning theory, including decentralized learning and machine unlearning; see, for instance, [27] and [28]. In this work, nested Gibbs measures are shown to be solutions to a minimization of the expectation of a linear combination of two original objective functions subject to a regularization by a relative entropy with respect to a given reference measure (Theorem 6). The third operation is the normalized log-linear combination of the densities of Gibbs probability measures. This operation has been studied beyond Gibbs probability measures in [17, 19, 29] and references therein. In this work, it is shown that the normalized log-linear combination of two Gibbs measures is itself a Gibbs measure, and that it is the solution to a minimization of the expectation of a linear combination of the objective functions of the original Gibbs measures subject to a regularization by a relative entropy with respect to a given reference measure (Theorem 8). Hence, nesting Gibbs probability measures or log-linearly combining them, leads to new Gibbs probability measures that are solutions to optimization problems whose difference is only on the coefficients of the linear combination of the objective functions. Using this observation, the central result is a formal proof

that there exists a choice of parameters such that both operations yield the same Gibbs probability measure (Corollary 12). Finally, this equivalence between such operations is put at play to design a one-shot federated learning [30] that achieves at all clients, the same performance of a centralized system that trains a Gibbs algorithm upon the aggregation of the training datasets of all clients (Corollary 13). This result is reminiscent of the decentralized learning construction in [28], which also achieves the same performance as a centralized system. While the construction in [28] is based on sequentially nesting the Gibbs algorithms of all clients, the construction in this report is based on simultaneously log-linearly combining the clients' Gibbs algorithms at a central server.

2 Preliminaries

This section describes the notation used in this work and formally introduces conditional Gibbs probability measures.

2.1 Notation

Given a set $\mathcal{X} \subseteq \mathbb{R}^d$, for some $d \in \mathbb{N}$, the Borel sigma-field defined on $\mathcal{X}$ is denoted by $\mathscr{B}(\mathcal{X})$. The set of probability measures on the measurable space $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ is denoted by $\triangle(\mathcal{X})$. Given a product measurable space $(\mathcal{X} \times \mathcal{Y}, \mathscr{B}(\mathcal{X} \times \mathcal{Y},))$. 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 $\triangle(\mathcal{Y})$ indexed by $\mathcal{X}$ is said to be a conditional probability measure if, for all measurable sets $\mathcal{B} \in \mathscr{B}(\mathcal{Y})$, 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 $\triangle(\mathcal{Y} \mid \mathcal{X})$.

Given a σ-finite measure Q on $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$, the set of probability measures in $\triangle(\mathcal{X})$ that are absolutely continuous with respect to Q is denoted by $\triangle_Q(\mathcal{X})$. 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". The Radon-Nikodym derivative of P with respect to Q is denoted by $\dfrac{\mathrm{d}P}{\mathrm{d}Q}$. 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*

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

2.2 Gibbs Conditional Probability Measures

Consider two sets $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$, for some given integers d_1 and d_2. Let $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ and $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$ be two Borel measurable spaces. Consider also the product measurable space $(\mathcal{X} \times \mathcal{Y}, \mathscr{B}(\mathcal{X} \times \mathcal{Y}))$. Using this notation, Gibbs conditional probability measures in $\triangle(\mathcal{Y}|\mathcal{X})$ are parametrized by a Borel measurable function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$; a σ-finite measure Q on the measurable space $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$; and a real $\lambda \in \mathbb{R} \setminus \{0\}$. In order to define such conditional measures, consider a fixed $x \in \mathcal{X}$ and the following function:

$$\mathsf{K}_{h,Q,x} : \left\{ \begin{array}{l} \mathbb{R} \longrightarrow \mathbb{R} \\ t \longmapsto \log\left(\int \exp\left(t\, h\left(x, y\right)\right) \mathrm{d}Q\left(y\right)\right). \end{array} \right. \tag{3}$$

Under the assumption that the reference measure Q is a probability measure, the function $\mathsf{K}_{h,Q,x}$ in (3) is the cumulant generating function of $h(x, y)$, when $x \in \mathcal{X}$ is kept fixed and y is sampled from Q. Using this notation, the definition of Gibbs conditional probability measures is presented hereunder.

Definition 3 (Gibbs Conditional Probability Measures)**.** *Given a Borel measurable function $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$; a σ-finite measure Q on the measurable space $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$; and a real $\lambda \in \mathbb{R} \setminus \{0\}$, the probability measure $P_{Y|X}^{(h,Q,\lambda)} \in \triangle(\mathcal{Y} \mid \mathcal{X})$ is said to be an (h, Q, λ)-Gibbs conditional probability measure if*

$$\forall x \in \mathcal{X},\ \mathsf{K}_{h,Q,x}\left(-\frac{1}{\lambda}\right) < +\infty; \tag{4}$$

and for all $(x, y) \in \mathcal{X} \times \operatorname{supp} Q$,

$$\frac{\mathrm{d}P_{Y|X=x}^{(h,Q,\lambda)}}{\mathrm{d}Q}(y) = \exp\left(-\frac{1}{\lambda} h\left(x, y\right) - \mathsf{K}_{h,Q,x}\left(-\frac{1}{\lambda}\right)\right), \tag{5}$$

where the function $\mathsf{K}_{h,Q,x}$ is defined in (3)*.*

In Definition 3, under the assumptions that Q is a probability measure and $\lambda > 0$, the condition in (4) is always met. Thus, under such assumptions, a conditional measure that satisfies (5) is always a Gibbs conditional probability measure. Alternatively, when $\lambda < 0$, the condition in (4) is more restrictive and might impose particular constraints on the function h.

In general, Gibbs conditional probability measures exhibit numerous properties, see for instance, [9] and [31]. The following section introduces a class of Gibbs conditional probability measures that exhibit properties that are shown to be central in certain applications in statistical learning.

3 Main Results

In this section three operations on Gibbs probability measures are thoroughly studied: renormalization, nesting and log-linear combinations.

3.1 Renormalized Gibbs Probability Measures

The normalization of a power of a probability measure, which is often referred to as *renormalization*, is defined as follows.

Definition 4 (Renormalization)**.** *Consider $\alpha \in \mathbb{R} \setminus \{0\}$, two σ-finite measures Q_1 and Q_2, and a probability measure P, all on the same measurable space with $P \ll Q_1$ and $Q_2 \ll Q_1$. The (α, Q_1, Q_2)-renormalization of P is a probability measure, denoted by R, such that for all $y \in \operatorname{supp} Q_2$,*

$$\frac{\mathrm{d}R}{\mathrm{d}Q_2}(y) = \frac{\left(\frac{\mathrm{d}P}{\mathrm{d}Q_1}(y)\right)^{\alpha}}{\int \left(\frac{\mathrm{d}P}{\mathrm{d}Q_1}(\tilde{y})\right)^{\alpha} \mathrm{d}Q_2(\tilde{y})}, \tag{6}$$

provided that $0 < \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q_1}(y)\right)^{\alpha} \mathrm{d}Q_2(y) < \infty$.

The renormalization of a Gibbs probability measure is also a Gibbs probability measure, as shown hereunder. More specifically, given the (h, Q, λ)-Gibbs probability measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5), denoted by R_1, its (α, Q, Q_1)-renormalization, satisfies, for all $y \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}R_1}{\mathrm{d}Q_1}(y) = \frac{\left(\frac{\mathrm{d}P_{Y|X=x}^{(h,Q,\lambda)}}{\mathrm{d}Q}(y)\right)^{\alpha}}{\int \left(\frac{\mathrm{d}P_{Y|X=x}^{(h,Q,\lambda)}}{\mathrm{d}Q}(\widehat{y})\right)^{\alpha} \mathrm{d}Q_1(\widehat{y})}. \tag{7}$$

The following theorem shows that the measure R_1 in (7) is a Gibbs probability measure.

Theorem 1. *The (α, Q, Q_1)-renormalization of the Gibbs probability measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5), namely the measure R_1 in (7), satisfies for all $y \in \operatorname{supp} Q_1$,*

$$\frac{\mathrm{d}R_1}{\mathrm{d}Q_1}(y) = \frac{\exp\left(-\frac{\alpha}{\lambda} h(x,y)\right)}{\int \exp\left(-\frac{\alpha}{\lambda} h(x,\widehat{y})\right) \mathrm{d}Q_1(\widehat{y})} = \frac{\mathrm{d}P_{Y|X=x}^{(h,Q_1,\frac{\lambda}{\alpha})}}{\mathrm{d}Q_1}(y). \tag{8}$$

Proof: The proof is presented in Appendix A. ∎

Theorem 1 shows that the (α, Q, Q_1)-renormalization of the (h, Q, λ)-Gibbs probability measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5) is identical to the $(h, Q_1, \frac{\lambda}{\alpha})$-Gibbs probability measure $P_{Y|X=x}^{(h,Q_1,\frac{\lambda}{\alpha})}$ in (8). More specifically, $D\left(R_1 \,\middle\|\, P_{Y|X=x}^{(h,Q_1,\frac{\lambda}{\alpha})}\right) = 0$. Hence, such a renormalization has a twofold effect on the Gibbs measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5): (i) the regularization factor is changed from λ to λ/α; and (ii) the reference measure is changed from Q to Q_1. The second effect might imply a concentration of

the measure given that $Q_1 \ll Q$. More specifically if $\operatorname{supp} Q_1 \subset \operatorname{supp} Q$, then, $\operatorname{supp} P_{Y|X=x}^{(h,Q_1,\frac{\lambda}{\alpha})} = \operatorname{supp} Q_1 \subset \operatorname{supp} Q = \operatorname{supp} P_{Y|X=x}^{(h,Q,\lambda)}$. Interestingly, both effects are not necessarily simultaneously observed, which follows from the fact that α and Q_1 can be chosen independently. For instance, if Q_1 is chosen to be identical to Q, then only the change of the regularization factor is observed. Alternatively, if $\alpha = 1$, only the change of reference measure is observed. This observation reveals a structural property of Gibbs probability measures through the optimization problems stated in the following corollaries. In particular, [31, Lemma 1] yields the following result.

Corollary 2. *When $\frac{\lambda}{\alpha} > 0$, the (α, Q, Q_1)-renormalization of the Gibbs probability measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5), namely the measure R_1 in (7), is the unique solution, if it exists, to the following optimization problem*

$$\min_{P \in \triangle_{Q_1}(\mathcal{Y})} \int h(x,y) \, \mathrm{d}P(y) \; + \; \frac{\lambda}{\alpha} D(P \parallel Q_1) . \tag{9}$$

Alternatively, when $\frac{\lambda}{\alpha} < 0$, the (α, Q, Q_1)-renormalization of the Gibbs probability measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5), namely the measure R_1 in (7), is the unique solution, if it exists, to the following optimization problem

$$\max_{P \in \triangle_{Q_1}(\mathcal{Y})} \int h(x,y) \, \mathrm{d}P(y) \; + \; \frac{\lambda}{\alpha} D(P \parallel Q_1) . \tag{10}$$

Similarly, [31, Lemma 2] leads to the following corollary.

Corollary 3. *When $\frac{\lambda}{\alpha} > 0$, the (α, Q, Q_1)-renormalization of the Gibbs probability measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5), namely the measure R_1 in (7), is the unique solution, if it exists, to the following optimization problem*

$$\begin{aligned} \min_{P \in \triangle_{Q_1}(\mathcal{Y})} \quad & \int h(x,y) \, \mathrm{d}P(y) \\ s.t. \quad & D(P \parallel Q_1) \leqslant D\left(P_{Y|X=x}^{(h,Q_1,\frac{\lambda}{\alpha})} \,\middle\|\, Q_1\right) . \end{aligned} \tag{11}$$

Alternatively, when $\frac{\lambda}{\alpha} < 0$, the (α, Q, Q_1)-renormalization of the Gibbs probability measure $P_{Y|X=x}^{(h,Q,\lambda)}$ in (5), namely the measure R_1 in (7), is the unique solution, if it exists, to the following optimization problem

$$\begin{aligned} \max_{P \in \triangle_{Q_1}(\mathcal{Y})} \quad & \int h(x,y) \, \mathrm{d}P(y) \\ s.t. \quad & D(P \parallel Q_1) \leqslant D\left(P_{Y|X=x}^{(h,Q_1,\frac{\lambda}{\alpha})} \,\middle\|\, Q_1\right) . \end{aligned} \tag{12}$$

These structural properties shed light on the open question concerning the choice of the reference measure Q in regularized optimization problems involving Gibbs probability measures, which is a central question in statistical learning theory. See for instance, [6, 9, 27, 28], and references therein.

3.2 Nested Gibbs Probability Measures

Nested Gibbs probability measures can be formally defined as follows.

Definition 5 (Nested Gibbs Probability Measures)**.** *An (h, Q_1, λ)-Gibbs probability measure $P^{(h,Q_1,\lambda)}_{Y|X=x} \in \triangle(\mathcal{Y})$, as the one in* (5)*, is said to be a nested Gibbs probability measure if Q_1 is a Gibbs probability measure.*

Consider two Borel measurable functions

$$h_1 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \text{ and} \tag{13}$$

$$h_2 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}. \tag{14}$$

Given $x_1 \in \mathcal{X}$, consider an (h_1, Q_1, λ_1)-Gibbs probability measure, denoted by

$$P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1} \in \triangle(\mathcal{Y}). \tag{15}$$

Given $x_2 \in \mathcal{X}$, consider an (h_2, Q_2, λ_2)-Gibbs probability measure, denoted by

$$P^{(h_2,Q_2,\lambda_2)}_{Y_2|X_2=x_2} \in \triangle(\mathcal{Y}). \tag{16}$$

In the following, the measure $P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}$ is said to be nested within $P^{(h_2,Q_2,\lambda_2)}_{Y_2|X_2=x_2} \in \triangle(\mathcal{Y})$, when Q_2 is chosen to be identical to $P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}$, which yields a $\left(h_2, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}, \lambda_2\right)$-Gibbs probability measure, denoted by

$$P^{\left(h_2, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}, \lambda_2\right)}_{Y_2|X_2=x_2} \in \triangle(\mathcal{Y}). \tag{17}$$

Such a measure is a nested Gibbs probability measure (Definition 5). It is also characterized by several optimization problems, as shown in the following corollaries. In particular, [31, Lemma 1] yields the first characterization.

Corollary 4. *When $\lambda_2 > 0$, the nested Gibbs probability measure $P^{\left(h_2, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}, \lambda_2\right)}_{Y_2|X_2=x_2}$ in* (17) *is the unique solution, if it exists, to the following optimization problem*

$$\min_{P \in \triangle_{Q_1}(\mathcal{Y})} \int h_2(x_2, y)\, \mathrm{d}P(y) + \lambda_2 D\left(P \,\middle\|\, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}\right). \tag{18}$$

Alternatively, when $\lambda_2 < 0$, the nested Gibbs probability measure $P^{\left(h_2, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}, \lambda_2\right)}_{Y_2|X_2=x_2}$ in (17) *is the unique solution, if it exists, to the following optimization problem*

$$\max_{P \in \triangle_{Q_1}(\mathcal{Y})} \int h_2(x_2, y)\, \mathrm{d}P(y) + \lambda_2 D\left(P \,\middle\|\, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}\right). \tag{19}$$

Similarly, [31, Lemma 4] leads to the following corollary.

Corollary 5. *When $\lambda_2 > 0$, the nested Gibbs probability measure $P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)}$ in (17) is the unique solution, if it exists, to the following optimization problem*

$$\min_{P \in \triangle_{Q_1}(\mathcal{Y})} \quad \int h_2(x_2, y)\, \mathrm{d}P(y)$$
$$s.t. \quad D\left(P \,\middle\|\, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}\right) \leqslant D\left(P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)} \,\middle\|\, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}\right). \quad (20)$$

Alternatively, when $\lambda_2 < 0$, the nested Gibbs probability measure $P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)}$ in (17) is the unique solution, if it exists, to the following optimization problem

$$\max_{P \in \triangle_{Q_1}(\mathcal{Y})} \quad \int h_2(x_2, y)\, \mathrm{d}P(y)$$
$$s.t. \quad D\left(P \,\middle\|\, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}\right) \leqslant D\left(P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)} \,\middle\|\, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}\right). \quad (21)$$

The following theorem provides an explicit expression for the Radon-Nikodym derivative of $P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)}$ with respect to the measure Q_1. For doing so, consider the following function

$$\widehat{h}_{\beta_1,\beta_2} : \begin{cases} \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R} \\ (x_1, x_2, y) \longmapsto \beta_1 h_1(x_1, y) + \beta_2 h_2(x_2, y), \end{cases} \quad (22)$$

where (β_1, β_2) is a pair of given parameters.

Theorem 6. *The nested Gibbs probability measure $P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)}$ in (17) satisfies for all $y \in \operatorname{supp} Q_1$,*

$$\frac{\mathrm{d}P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)}}{\mathrm{d}Q_1}(y) = \frac{\exp\left(-\frac{1}{\lambda_1} h_1(x_1, y) - \frac{1}{\lambda_2} h_2(x_2, y)\right)}{\int \exp\left(-\frac{1}{\lambda_1} h_1(x_1, \tilde{y}) - \frac{1}{\lambda_2} h_2(x_2, \tilde{y})\right) \mathrm{d}Q_1(\tilde{y})} \quad (23)$$
$$= \frac{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{1}{\lambda_1}, \frac{1}{\lambda_2}}, Q_1, 1\right)}}{\mathrm{d}Q_1}(y). \quad (24)$$

Proof: The proof is presented in Appendix B. ∎

Theorem 6 shows that the nested Gibbs probability measure $P_{Y_2|X_2=x_2}^{\left(h_2, P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}, \lambda_2\right)}$ in (17) is identical to the $\left(\widehat{h}_{\frac{1}{\lambda_1}, \frac{1}{\lambda_2}}, Q_1, 1\right)$-Gibbs probability measure $P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{1}{\lambda_1}, \frac{1}{\lambda_2}}, Q_1, 1\right)}$ in (24), which leads to the following implications.

Theorem 7. *Consider the* $\left(h_2, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}, \lambda_2\right)$*-Gibbs probability measure* $P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}$ *in* (17) *and the* $\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}}, Q_1, 1\right)$*-Gibbs probability measure, denoted by* $P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$ *in* (24)*. Then, the following statements are equivalent:*

(i) *For all* $y \in \operatorname{supp} Q_1$*,*

$$\frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}(y) = 1. \tag{25}$$

(ii)

$$D\left(P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2} \,\middle\|\, P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}\right) = 0. \tag{26}$$

(iii) *The probability measure* $P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}$ *is the unique solution, if it exists, to the following optimization problem*

$$\min_{P\in\triangle_{Q_1}(\mathcal{Y})} \int \widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}}(x_1,x_2,y)\mathrm{d}P(y) + D\left(P \parallel Q_1\right). \tag{27}$$

(iv) *The measure* $P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}$ *is the unique solution, if it exists, to the following optimization problem*

$$\begin{aligned} \min_{P\in\triangle_{Q_1}(\mathcal{Y})} \quad & \int \widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}}(x_1,x_2,y)\,\mathrm{d}P(y) \\ s.t. \quad & D\left(P \parallel Q_1\right) \leqslant D\left(P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2} \,\middle\|\, Q_1\right). \end{aligned} \tag{28}$$

Proof: The proof is presented in Appendix C. ∎

Theorem 7 establishes that the measures $P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}$ and $P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$, both in $\triangle(\mathcal{Y})$, are identical. This observation implies that the probability measure $P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$ is also the solution to the optimization problems in (27) and (28), if such solutions exist. Together with Corollary 4 and Corollary 5, it follows that if $\lambda_2 > 0$, it is also the solution to the optimization problems in (18) and (20), if such solutions exist. Alternatively, if $\lambda_2 < 0$, it is also the solution to the optimization problems in (19) and (21), if such solutions exist.

3.3 Normalized Log-Linear Combinations of Gibbs Probability Measures

The normalization of the product of powers of probability measures is formalized as follows.

Definition 6 (Normalized Log-Linear Combination of Measures)**.** *Consider $\alpha_1 \in \mathbb{R} \setminus \{0\}$, $\alpha_2 \in \mathbb{R} \setminus \{0\}$, two probability measures P_1 and P_2; and three σ-finite measures Q, Q_1 and Q_2, all on the same measurable space, such that $P_1 \ll Q_1$, $Q \ll Q_1$, $P_2 \ll Q_2$ and $Q \ll Q_2$. The $(\alpha_1, \alpha_2, Q_1, Q_2, Q)$-normalized log-linear combination of P_1 and P_2 is a probability measure, denoted by S, such that for all $y \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}S}{\mathrm{d}Q}(y) = \frac{\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(y)\right)^{\alpha_1}\left(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(y)\right)^{\alpha_2}}{\int \left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(\widehat{y})\right)^{\alpha_1}\left(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(\widehat{y})\right)^{\alpha_2} \mathrm{d}Q(\widehat{y})}, \tag{29}$$

provided that $0 < \int \left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_1}(\widehat{y})\right)^{\alpha_1}\left(\frac{\mathrm{d}P_2}{\mathrm{d}Q_2}(\widehat{y})\right)^{\alpha_2} \mathrm{d}Q(\widehat{y}) < \infty$.

Definition 6, which is restricted to the normalized log-linear combination of only two probability measures, can be generalized to normalized log-linear combinations of multiple probability measures. Nonetheless, for pedagogical purposes, this work focuses exclusively on the case of two Gibbs probability measures. The $(\alpha_1, \alpha_2, Q_1, Q_2, Q)$-normalized log-linear combination of the Gibbs probability measures $P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}$ in (15) and $P^{(h_2,Q_2,\lambda_2)}_{Y_2|X_2=x_2}$ in (16), denoted by $S_1 \in \triangle_Q(\mathcal{Y})$, satisfies for all $y \in \operatorname{supp} Q$

$$\frac{\mathrm{d}S_1}{\mathrm{d}Q}(y) = \frac{\left(\frac{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}{\mathrm{d}Q_1}(y)\right)^{\alpha_1}\left(\frac{\mathrm{d}P^{(h_2,Q_2,\lambda_2)}_{Y_2|X_2=x_2}}{\mathrm{d}Q_2}(y)\right)^{\alpha_2}}{\int \left(\frac{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}{\mathrm{d}Q_1}(\tilde{y})\right)^{\alpha_1}\left(\frac{\mathrm{d}P^{(h_2,Q_2,\lambda_2)}_{Y_2|X_2=x_2}}{\mathrm{d}Q_2}(\tilde{y})\right)^{\alpha_2} \mathrm{d}Q(\tilde{y})}. \tag{30}$$

The following theorem provides an explicit expression for the Radon-Nikodym derivative of S_1 with respect to the measure Q.

Theorem 8. *The $(\alpha_1, \alpha_2, Q_1, Q_2, Q)$-normalized log-linear combination of $P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}$ in (15) and $P^{(h_2,Q_2,\lambda_2)}_{Y_2|X_2=x_2}$ in (16), namely the probability measure S_1 in (30), satisfies for all $y \in \operatorname{supp} Q$,*

$$\frac{\mathrm{d}S_1}{\mathrm{d}Q}(y) = \frac{\exp\left(-\frac{\alpha_1}{\lambda_1}h_1(x_1, y) - \frac{\alpha_2}{\lambda_2}h_2(x_2, y)\right)}{\int \exp\left(-\frac{\alpha_1}{\lambda_1}h_1(x_1, \widehat{y}) - \frac{\alpha_2}{\lambda_2}h_2(x_2, \widehat{y})\right) \mathrm{d}Q(\widehat{y})} \tag{31}$$

$$= \frac{\mathrm{d}P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}{\mathrm{d}Q}(y), \tag{32}$$

where $P_{Y|(X_1,X_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}$ *is an* $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)$*-Gibbs conditional probability measure in* $\triangle(\mathcal{Y}|\mathcal{X}\times\mathcal{X})$*; and the function* $\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}}$ *is defined in* (22).

Proof: The proof is presented in Appendix D. ■

Theorem 8 shows that the $(\alpha_1,\alpha_2,Q_1,Q_2,Q)$-normalized log-linear combination of $P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}$ in (15) and $P_{Y_2|X_2=x_2}^{(h_2,Q_2,\lambda_2)}$ in (16) is identical to the $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)$-Gibbs probability measure $P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}$ in (32). The normalized log-linear combination S_1 exhibits all properties of Gibbs probability measures described in [31]. In particular, [31, Lemma 1] leads to the following corollary.

Corollary 9. *The* $(\alpha_1,\alpha_2,Q_1,Q_2,Q)$*-normalized log-linear combination* S_1 *of* $P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}$ *and* $P_{Y_2|X_2=x_2}^{(h_2,Q_2,\lambda_2)}$ *is the unique solution, if it exists, to the following optimization problem*

$$\min_{P\in\triangle_Q(\mathcal{Y})}\int \widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}}(x_1,x_2,y)\mathrm{d}P(y)+D\left(P\parallel Q\right). \tag{33}$$

Similarly, [31, Lemma 4] leads to the following corollary.

Corollary 10. *The* $(\alpha_1,\alpha_2,Q_1,Q_2,Q)$*-normalized log-linear combination* S_1 *of* $P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}$ *and* $P_{Y_2|X_2=x_2}^{(h_2,Q_2,\lambda_2)}$ *is the unique solution, if it exists, to the following optimization problem*

$$\begin{aligned}\min_{P\in\triangle_Q(\mathcal{Y})}\quad & \int \widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}}(x_1,x_2,y)\mathrm{d}P(y)\\ s.t.\quad & D\left(P\parallel Q\right)\leqslant D\left(S_1\parallel Q\right).\end{aligned} \tag{34}$$

The following theorem collects the equivalent formulations associated with the normalized log-linear combination S_1.

Theorem 11. *Consider the* $(\alpha_1,\alpha_2,Q_1,Q_2,Q)$*-normalized log-linear combination* S_1 *of* $P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}$ *and* $P_{Y_2|X_2=x_2}^{(h_2,Q_2,\lambda_2)}$ *in* (30)*, and the* $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)$*-Gibbs probability measure, denoted by* $P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}$ *in* (32)*. Then, the following statements are equivalent:*

(i) *For all* $y\in\operatorname{supp}Q$*,*

$$\frac{\mathrm{d}S_1}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}(y)=1. \tag{35}$$

(ii)

$$D\left(S_1\left\|\, P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}\right.\right)=0. \tag{36}$$

(iii) *The probability measure S_1 is the unique solution, if it exists, to the optimization problem in* (33).

(iv) *The probability measure S_1 is the unique solution, if it exists, to the optimization problem in* (34).

Proof: The proof is presented in Appendix E. ∎

Theorem 11 establishes that the normalized log-linear combination S_1 and the Gibbs probability measure $P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}\,\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2},Q,1\right)}$, both in $\triangle(\mathcal{Y})$, are identical. This observation implies that the normalized log-linear combination S_1 is also the solution to the optimization problems in (33) and (34), if such solutions exist. In particular, when $\alpha_1 = \alpha_2 = 1$ and Q and Q_1 are identical, the optimization problems in (33) and (34) coincide with the optimization problems in (27) and (28), respectively. More specifically, denote by $S_2 \in \triangle_{Q_1}(\mathcal{Y})$ the $(1, 1, Q_1, Q_2, Q_1)$-normalized log-linear combination of $P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}$ and $P_{Y_2|X_2=x_2}^{(h_2,Q_2,\lambda_2)}$. Hence, for all $y \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}S_2}{\mathrm{d}Q_1}(y) = \frac{\exp\left(-\frac{1}{\lambda_1}h_1(x_1,y) - \frac{1}{\lambda_2}h_2(x_2,y)\right)}{\int \exp\left(-\frac{1}{\lambda_1}h_1(x_1,\widehat{y}) - \frac{1}{\lambda_2}h_2(x_2,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})} \tag{37}$$

$$= \frac{\mathrm{d}P_{Y_2|X_2=x_2}^{\left(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2\right)}}{\mathrm{d}Q_1}(y), \tag{38}$$

where the equality in (37) follows from Theorem 8; and the equality in (38) follows from Theorem 6. Using this notation, the following holds.

Corollary 12. *The nested Gibbs probability measure* $P_{Y_2|X_2=x_2}^{(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2)}$ *in* (17) *and the Gibbs probability measure S_2 in* (37) *are identical. That is,*

$$D\left(S_2 \,\middle\|\, P_{Y_2|X_2=x_2}^{(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2)}\right)=0. \tag{39}$$

4 Applications in One-Shot Federated Learning

Consider a federated learning system in which K clients collaboratively tune their local learning algorithms by communicating with a common server. Let the sets $\mathcal{M} \subseteq \mathbb{R}^d$, with $d \in \mathbb{N}$; $\mathcal{X}$; and $\mathcal{Y}$, denote some sets of *models*, *patterns*, and *labels*, respectively. For all $k \in \{1, 2, \ldots, K\}$, client k has n_k training data points $(x_{k,1}, y_{k,1}), \ldots, (x_{k,n_k}, y_{k,n_k})$, which are elements of the set $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$, and form the dataset

$$z_k \triangleq \left((x_{k,1}, y_{k,1}), \ldots, (x_{k,n_k}, y_{k,n_k})\right) \in \mathcal{Z}^{n_k}. \tag{40}$$

The aggregation of all local training datasets is denoted by

$$z_0 \triangleq (z_1, \ldots, z_K) \in \mathcal{Z}^{n_0}, \tag{41}$$

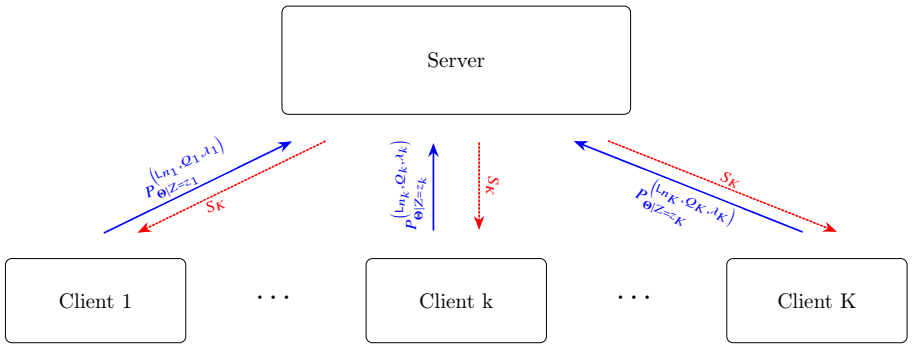


Figure 1: One-shot client–server communication system. For all $k \in \{1, 2, \ldots, K\}$, the Gibbs algorithm of client k, $P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(\mathsf{L}_{n_k}, Q_k, \lambda_k)}$, is defined in (43), and the measure representing the aggregation at the server, S_K, is defined in (44).

with $n_0 \triangleq \sum_{k=1}^{K} n_k$. Given a model $\boldsymbol{\theta} \in \mathcal{M}$, the loss induced by such a model on a data point $(x, y) \in \mathcal{Z}$ is denoted by $\ell(x, y, \boldsymbol{\theta})$, where $\ell : \mathcal{Z} \times \mathcal{M} \to [0, +\infty)$ is referred to as the *loss function*, which is assumed to be Borel measurable.

The *empirical risk* induced by a model with respect to an m-length dataset is determined by the function

$$\mathsf{L}_m : \begin{cases} \mathcal{Z}^m \times \mathcal{M} \longrightarrow [0, +\infty) \\ \quad (\boldsymbol{z}, \boldsymbol{\theta}) \longmapsto \frac{1}{m} \sum_{i=1}^{m} \ell\left(x_{k,i}, y_{k,i}, \boldsymbol{\theta}\right). \end{cases} \tag{42}$$

A supervised learning algorithm trained upon n_k-length datasets is represented by a conditional probability measure in $\triangle\left(\mathcal{M} \mid \mathcal{Z}^{n_k}\right)$. A class of algorithms that is central in this section is that of Gibbs algorithms. For all $k \in \{1, 2, \ldots, K\}$, given a σ-finite measure Q_k defined on $\mathcal{M}$, a regularization factor $\lambda_k \in (0, +\infty)$, and a fixed dataset $\boldsymbol{z}_k \in \mathcal{Z}^{n_k}$, the instance of the Gibbs algorithm at client k, trained upon the dataset $\boldsymbol{z}_k$, is represented by the Gibbs probability measure

$$P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(\mathsf{L}_{n_k}, Q_k, \lambda_k)} \in \triangle_{Q_k}\left(\mathcal{M}\right). \tag{43}$$

Consider the one-shot federated learning system (two communication phases) described in Figure 1. In the first phase, clients transmit their local Gibbs algorithms to the server. During the second phase, the server aggregates these local algorithms by means of a normalized log-linear combination and broadcasts the resulting probability measure to all clients. More specifically, the server

defines the measure $S_K \in \triangle_Q(\mathcal{M})$ such that, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}S_K}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\displaystyle\prod_{k=1}^{K}\left(\frac{\mathrm{d}P^{(\mathsf{L}_{n_k},Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\theta})\right)^{\alpha_k}}{\displaystyle\int\prod_{k=1}^{K}\left(\frac{\mathrm{d}P^{(\mathsf{L}_{n_k},Q_k,\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\boldsymbol{z}_k}}{\mathrm{d}Q_k}(\boldsymbol{\nu})\right)^{\alpha_k} dQ(\boldsymbol{\nu})} \tag{44}$$

$$= \frac{\exp\left(-\displaystyle\sum_{k=1}^{K}\frac{\alpha_k}{\lambda_k}\mathsf{L}_{n_k}(\boldsymbol{z}_k,\boldsymbol{\theta})\right)}{\displaystyle\int\exp\left(-\sum_{k=1}^{K}\frac{\alpha_k}{\lambda_k}\mathsf{L}_{n_k}(\boldsymbol{z}_k,\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu})}, \tag{45}$$

where, for all $k \in \{1,2,\ldots,K\}$, it has been assumed that $Q \ll Q_k$. The equality in (45) follows from iteratively using Theorem 8. Moreover, from [31, Lemma 1], it follows that the measure S_K is the unique solution to

$$\min_{P\in\triangle_Q(\mathcal{M})}\int\left(\sum_{k=1}^{K}\frac{\alpha_k}{\lambda_k}\mathsf{L}_{n_k}(\boldsymbol{z}_k,\boldsymbol{\theta})\right)\mathrm{d}P(\boldsymbol{\theta}) + D\left(P \parallel Q\right). \tag{46}$$

The following corollary leverages this observation for introducing a choice of parameters in (46) that guarantees the achievability of centralized-performance for the federated learning system depicted above.

Corollary 13. *Assume that for all $k \in \{1,2,\ldots,K\}$,*

$$\lambda_k = \frac{1}{n_k};\ \textit{and}\ \alpha_k = \frac{1}{n_0\lambda_0}, \tag{47}$$

for some $\lambda_0 > 0$ in (44). Then, for all $\boldsymbol{\theta} \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}S_K}{\mathrm{d}Q}(\boldsymbol{\theta}) = \frac{\exp\left(\frac{-1}{\lambda_0}\,\mathsf{L}_{n_0}(\boldsymbol{z}_0,\boldsymbol{\theta})\right)}{\displaystyle\int\exp\left(\frac{-1}{\lambda_0}\,\mathsf{L}_{n_0}(\boldsymbol{z}_0,\boldsymbol{\nu})\right)\mathrm{d}Q(\boldsymbol{\nu})} = \frac{\mathrm{d}P^{(\mathsf{L}_{n_0},Q,\lambda_0)}_{\boldsymbol{\Theta}|\mathbf{Z}=\boldsymbol{z}_0}}{\mathrm{d}Q}(\boldsymbol{\theta}). \tag{48}$$

Corollary 13 shows that the normalized log-linear combination of the K algorithms independently obtained by the clients by training Gibbs algorithms upon their local training datasets, namely the measure S_K in (45), is identical to a Gibbs algorithm trained on the aggregated dataset $\boldsymbol{z}_0$. The main assumption in Corollary 13 is that, for all $k \in \{1,2,\ldots,K\}$,

$$\frac{\alpha_k}{\lambda_k} = \frac{n_k}{n_0\lambda_0}. \tag{49}$$

This condition can be satisfied through several choices of the pairs (λ_k, α_k). Interstingly, λ_k is chosen locally by client k, whereas α_k is chosen by the server. In particular, the choice $\lambda_k = 1/n_k$ depends only on the size of the local dataset at client k. Under this choice, the condition reduces to $\alpha_k = \frac{1}{n_0 \lambda_0}$, which is independent of k. Hence, the server does not need any client-specific information about the local datasets; it only requires the total number of samples n_0 and the centralized regularization parameter λ_0. In other words, in a one-shot federated learning setting, this choice of parameters allows the server to achieve the same performance than a centralized Gibbs algorithm trained on the aggregation of all training datasets, by forming a normalized log-linear combination of the locally trained Gibbs algorithms. In particular, this occurs by exclusively sharing probability distributions on the set of models and not the actual data. The feasibility of such a learning system is constrained by the possibility of transmitting the corresponding Gibbs measures without distortion, which is not possible in practice. This limitation is related to the practical issue mentioned in [28]. The effect of such a distortion remains to be formally studied.

5 Conclusion and Final Discussion

In this report, three operations on Gibbs probability measures have been presented: renormalization, nesting, and normalized log-linear combinations. A first result is that Gibbs probability measures are closed under renormalization, nesting, or normalized log-linear combinations. In the case of nesting and normalized log-linear combinations, the resulting probability measure is the solution to an optimization of the expectation of a linear combination of the original objective functions (one for each measure) subject to a relative entropy regularization. This observation gives a common interpretation to two operations that are defined differently: nesting operates through the reference measure, whereas normalized log-linear combination operates directly on the densities. More specifically, the same Gibbs probability measure can be obtained either through a sequential construction (nesting) or through an aggregate construction (normalized log-linear combination). This point of view is useful in statistical learning problems in which objective functions, often represent empirical risks. In such settings, the results show that combining several Gibbs probability measures, via nesting or normalized log-linear combinations, can be understood as combining the corresponding learning objectives. Applications of the nesting operation in this context include decentralized learning and machine unlearning [27, 28], while an application of normalized log-linear combinations in federated learning is described in Section 4.

In general, the equivalence between nesting Gibbs probability measures and log-linearly combining them (subject to normalization) further clarifies this interpretation. More specifically, for a suitable choice of parameters, the sequential construction obtained by nesting one Gibbs probability measure within another yields the same probability measure as the aggregate construction obtained by log-linearly combining them. This equivalence identifies the precise

conditions under which both operations lead to the same measure. Overall, these results provide a unified view of renormalized Gibbs probability measures, nested Gibbs probability measures, and normalized log-linear combinations of Gibbs probability measures. These results also highlight the importance of the reference measure and the regularization factor in the design and interpretation of Gibbs-based methods, particularly in applications where several objective functions must be combined within a single regularized optimization problem.

References

[1] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 2521–2530.

[2] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, vol. 51, Cadiz, Spain, May 2016, pp. 1232–1240.

[3] M. A. Medina, J. L. M. Olea, C. Rush, and A. Velez, "On the robustness to misspecification of α-posteriors and their variational approximations," *Journal of Machine Learning Research*, vol. 23, no. 147, pp. 1–51, 2022.

[4] Y. Bu, "Towards optimal inverse temperature in the Gibbs algorithm," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Athens, Greece, Jul. 2024, pp. 2257–2262.

[5] Y. Bu, H. V. Tetali, G. Aminian, M. Rodrigues, and G. W. Wornell, "On the generalization error of meta learning for the Gibbs algorithm," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT)*, Taipei, Taiwan, Jun. 2023, pp. 2488–2493.

[6] 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.

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

[8] R. Ray, M. A. Medina, and C. Rush, "Asymptotics for power posterior mean estimation," in *Proceedings of the 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, USA, Sep. 2023.

[9] 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.

[10] B. Rodríguez-Gálvez, "An information-theoretic approach to generalization theory," Ph.D. dissertation, KTH Royal Institute of Technology, 2024.

[11] S. M. Perlaza and X. Zou, "The generalization error of supervised machine learning algorithms," *arXiv preprint arXiv:2411.12030*, 2024.

[12] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 34, Virtual Event, Dec. 2021, pp. 8106–8118.

[13] G. Aminian, Y. Bu, L. Toni, M. R. D. Rodrigues, and G. W. Wornell, "Information-theoretic characterizations of generalization error for the Gibbs algorithm," *IEEE Transactions on Information Theory*, vol. 70, no. 1, pp. 632–655, Jan. 2024.

[14] X. Zou, S. M. Perlaza, I. Esnaola, and E. Altman, "Generalization analysis of machine learning algorithms via the worst-case data-generating probability measure," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 15, Vancouver, Canada, Feb. 2024, pp. 17 271–17 279.

[15] 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.

[16] X. Zou, S. M. Perlaza, I. Esnaola, E. Altman, and H. V. Poor, "The worst-case data-generating probability measure in statistical learning," *IEEE Journal on Selected Areas in Information Theory*, vol. 5, pp. 175–189, Apr. 2024.

[17] J.-F. Bercher, "A simple probabilistic construction yielding generalized entropies and divergences, escort distributions and q-Gaussians," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 19, pp. 4460–4469, Oct. 2012.

[18] K. G. Wilson, "The renormalization group and critical phenomena," *Reviews of Modern Physics*, vol. 55, no. 3, pp. 583–600, Jul. 1983.

[19] A. R. Asadi, "Hierarchical maximum entropy via the renormalization group," *arXiv preprint arXiv:2509.01424*, Sep. 2025.

[20] A. Chhabra and R. V. Jensen, "Direct determination of the f(α) singularity spectrum," *Physical Review Letters*, vol. 62, no. 12, pp. 1327–1330, Mar. 1989.

[21] C. Beck and F. Schlögl, *Thermodynamics of Chaotic Systems: An Introduction*, ser. Cambridge Nonlinear Science Series. Cambridge, UK: Cambridge University Press, 1993, vol. 4.

[22] S. Abe, "Geometry of escort distributions," *Physical Review E*, vol. 68, no. 3, p. 031101, Sep. 2003.

[23] A. Ohara, H. Matsuzoe, and S. Amari, "A dually flat structure on the space of escort distributions," *Journal of Physics: Conference Series*, vol. 201, no. 1, p. 012012, Dec. 2010.

[24] J.-F. Bercher, "Source coding with escort distributions and Rényi entropy bounds," *Physics Letters A*, vol. 373, no. 36, pp. 3235–3238, Aug. 2009.

[25] C. Tsallis, *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World*, 1st ed. New York, NY, USA: Springer, 2009.

[26] S. Abe and G. B. Bagci, "Necessity of q-expectation value in nonextensive statistical mechanics," *Physical Review E*, vol. 71, no. 1, p. 016139, Jan. 2005.

[27] 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.

[28] ——, "Decentralized machine learning with centralized performance guarantees via Gibbs algorithms," in *Proceedings of the International Symposium on Information Theory (ISIT)*, Guangzhou, China, Jun. 2026.

[29] A. Lalitha, T. Javidi, and A. D. Sarwate, "Social learning and distributed hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 9, pp. 6161–6179, Sep. 2018.

[30] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, ser. Proceedings of Machine Learning Research, vol. 54, Apr. 2017, pp. 1273–1282.

[31] 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.

[32] 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.

Appendices

A Proof of Theorem 1

From Definition 4, R_1 in (7), the (α, Q, Q_1)-renormalization of (h, Q, λ)-Gibbs probability measure $P^{(h,Q,\lambda)}_{Y|X=x}$ in (5), satisfies, for all $y \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}R_1}{\mathrm{d}Q_1}(y) = \frac{\left(\frac{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}{\mathrm{d}Q}(y)\right)^{\alpha}}{\int \left(\frac{\mathrm{d}P^{(h,Q,\lambda)}_{Y|X=x}}{\mathrm{d}Q}(\widehat{y})\right)^{\alpha} \mathrm{d}Q_1(\widehat{y})} \tag{50}$$

$$= \frac{\left(\exp\left(-\frac{1}{\lambda}h(x,y) - \mathsf{K}_{h,Q,x}\left(-\frac{1}{\lambda}\right)\right)\right)^{\alpha}}{\int \left(\exp\left(-\frac{1}{\lambda}h(x,\widehat{y}) - \mathsf{K}_{h,Q,x}\left(-\frac{1}{\lambda}\right)\right)\right)^{\alpha} \mathrm{d}Q_1(\widehat{y})} \tag{51}$$

$$= \frac{\exp\left(-\frac{\alpha}{\lambda}h(x,y)\right)}{\int \exp\left(-\frac{\alpha}{\lambda}h(x,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})} \frac{\exp\left(\alpha\mathsf{K}_{h,Q,x}\left(-\frac{1}{\lambda}\right)\right)}{\exp\left(\alpha\mathsf{K}_{h,Q,x}\left(-\frac{1}{\lambda}\right)\right)} \tag{52}$$

$$= \frac{\exp\left(-\frac{\alpha}{\lambda}h(x,y)\right)}{\int \exp\left(-\frac{\alpha}{\lambda}h(x,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})} \tag{53}$$

$$= \frac{\mathrm{d}P^{(h,Q_1,\frac{\lambda}{\alpha})}_{Y|X=x}}{\mathrm{d}Q_1}(y), \tag{54}$$

where the equalities in (51) and (54) follow from Definition 3. This completes the proof.

B Proof of Theorem 6

From Definition 3, for all $y \in \operatorname{supp} Q_1$, it follows that

$$\frac{\mathrm{d}P^{\left(h_2, P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}, \lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}(y) = \frac{\exp\left(-\frac{1}{\lambda_2}h_2(x_2,y)\right)}{\int \exp\left(-\frac{1}{\lambda_2}h_2(x_2,\tilde{y})\right)\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}(\tilde{y})} \tag{55}$$

$$= \frac{\exp\left(-\frac{1}{\lambda_2}h_2(x_2,y)\right)}{\int \exp\left(-\frac{1}{\lambda_2}h_2(x_2,\tilde{y})\right)\frac{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}{\mathrm{d}Q_1}(\tilde{y})\mathrm{d}Q_1(\tilde{y})} \tag{56}$$

$$= \frac{\exp\left(-\frac{1}{\lambda_2}h_2(x_2,y)\right)\int \exp\left(-\frac{1}{\lambda_1}h_1(x_1,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})}{\int \exp\left(-\frac{1}{\lambda_2}h_2(x_2,\tilde{y})\right)\exp\left(-\frac{1}{\lambda_1}h_1(x_1,\tilde{y})\right)\mathrm{d}Q_1(\tilde{y})} \tag{57}$$

From [9, Lemma 3] and [32, Theorem 2] the Radon–Nikodym derivative $\frac{\mathrm{d}Q_1}{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}$ is well defined and from [32, Theorem 4], for all $y \in \operatorname{supp} Q_1$, it follows that

$$\frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}(y) = \frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}Q_1}(y)\frac{\mathrm{d}Q_1}{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}(y) \tag{58}$$

Combining (57) and (58) yields for all $y \in \operatorname{supp} Q_1$

$$\begin{aligned}
&\frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}Q_1}(y)\\
&= \left(\frac{\mathrm{d}Q_1}{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}(y)\right)^{-1}\frac{\exp\left(-\frac{1}{\lambda_2}h_2(x_2,y)\right)\int\exp\left(-\frac{1}{\lambda_1}h_1(x_1,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})}{\int\exp\left(-\frac{1}{\lambda_2}h_2(x_2,\tilde{y})-\frac{1}{\lambda_1}h_1(x_1,\tilde{y})\right)\mathrm{d}Q_1(\tilde{y})} &(59)\\
&= \frac{\mathrm{d}P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1}}{\mathrm{d}Q_1}(y)\frac{\exp\left(-\frac{1}{\lambda_2}h_2(x_2,y)\right)\int\exp\left(-\frac{1}{\lambda_1}h_1(x_1,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})}{\int\exp\left(-\frac{1}{\lambda_2}h_2(x_2,\tilde{y})-\frac{1}{\lambda_1}h_1(x_1,\tilde{y})\right)\mathrm{d}Q_1(\tilde{y})} &(60)\\
&= \frac{\exp\left(-\frac{1}{\lambda_1}h_1(x_1,y)\right)\quad\exp\left(-\frac{1}{\lambda_2}h_2(x_2,y)\right)\int\exp\left(-\frac{1}{\lambda_1}h_1(x_1,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})}{\int\exp\left(-\frac{1}{\lambda_1}h_1(x_1,\widehat{y})\right)\mathrm{d}Q_1(\widehat{y})\quad\int\exp\left(-\frac{1}{\lambda_2}h_2(x_2,\tilde{y})-\frac{1}{\lambda_1}h_1(x_1,\tilde{y})\right)\mathrm{d}Q_1(\tilde{y})} &(61)\\
&= \frac{\exp\left(-\frac{1}{\lambda_1}h_1(x_1,y)-\frac{1}{\lambda_2}h_2(x_2,y)\right)}{\int\exp\left(-\frac{1}{\lambda_1}h_1(x_1,\tilde{y})-\frac{1}{\lambda_2}h_2(x_2,\tilde{y})\right)\mathrm{d}Q_1(\tilde{y})} &(62)\\
&= \frac{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}{\mathrm{d}Q_1}(y), &(63)
\end{aligned}$$

where the equality in (60) follows from [32, Theorem 5]; the equality in (61) follows from Definition 3; and the equality in (63) follows from the Definition 3 and (22). This completes the proof.

C Proof of Theorem 7

From [9, Lemma 3], the following absolute continuity condition holds

$$P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2} \ll P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1} \ll Q_1. \tag{64}$$

Moreover, since $P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$ is a Gibbs probability measure with reference measure Q_1, from [9, Lemma 3], it also follows that

$$Q_1 \ll P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)} \ll Q_1. \tag{65}$$

Thus, from [32, Theorem 2] the Radon–Nikodym derivative $\dfrac{\mathrm{d}Q_1}{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}$ is well defined and from the multiplicative inverse of the Radon–Nikodym derivative [32, Theorem 5], for all $y \in \operatorname{supp} Q_1$, it follows that

$$\frac{\mathrm{d}Q_1}{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}(y) = \left(\frac{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}{\mathrm{d}Q_1}(y)\right)^{-1}. \tag{66}$$

Therefore, from Theorem 6 and the equality in (66) , for all $y \in \operatorname{supp} Q_1$,

$$\frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}(y) = \frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}Q_1}(y)\frac{\mathrm{d}Q_1}{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}(y) \tag{67}$$

$$= 1, \tag{68}$$

where the equality in(67) follows from the chain rule of the Radon–Nikodym derivative [32, Theorem 4] ; and the equality in (68) follows from Theorem 6. This proves *(i)*.

Assume now that *(i)* holds. Then,

$$D\left(P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2} \,\middle\|\, P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}\right)$$

$$= \int \log\left(\frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}(y)\right)\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}(y) \tag{69}$$

$$= 0, \tag{70}$$

which proves *(ii)*. Conversely, if *(ii)* holds, then from [9, Theorem 1] it follows that

$$\frac{\mathrm{d}P^{\left(h_2,P^{(h_1,Q_1,\lambda_1)}_{Y_1|X_1=x_1},\lambda_2\right)}_{Y_2|X_2=x_2}}{\mathrm{d}P^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}(y) = 1. \tag{71}$$

Therefore, *(i)* and *(ii)* are equivalent.

Assume next that *(i)* holds. Then $P_{Y_2|X_2=x_2}^{\left(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2\right)}$ is identical to the $\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)$-Gibbs probability measure in (24). Hence, from [31, Lemma 1], it is the unique solution, if it exists, to the optimization problem in (27). This proves that *(i)* implies *(iii)*.

Assume now that *(iii)* holds. From [31, Lemma 1], the unique solution, if it exists, to the optimization problem in (27) is precisely the $\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)$-Gibbs probability measure $P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}$ in (24). Hence,

$$P_{Y_2|X_2=x_2}^{\left(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2\right)}=P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}. \tag{72}$$

Thus,

$$D\left(P_{Y_2|X_2=x_2}^{\left(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2\right)}\,\middle\|\,Q_1\right)=D\left(P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}\,\middle\|\,Q_1\right). \tag{73}$$

Since $P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}$ is the $\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)$-Gibbs probability measure, [31, Lemma 4] implies that it is the unique solution, if it exists, to the optimization problem in (28). Therefore, $P_{Y_2|X_2=x_2}^{\left(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2\right)}$ is the unique solution, if it exists, to the same optimization problem. This proves that *(iii)* implies *(iv)*.

Assume finally that *(iv)* holds. From (67) and (68), which follow from Theorem 6, [32, Theorem 5], and [32, Theorem 4], it follows that, for all $y\in\operatorname{supp}Q_1$,

$$\frac{\mathrm{d}P_{Y_2|X_2=x_2}^{\left(h_2,P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)},\lambda_2\right)}}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{1}{\lambda_1},\frac{1}{\lambda_2}},Q_1,1\right)}}(y)=1. \tag{74}$$

This proves that *(iv)* implies *(i)*.

Therefore, *(i)*, *(ii)*, *(iii)*, and *(iv)* are equivalent. This concludes the proof.

D Proof of Theorem 8

From the definition of Gibbs probability measures in (5), for all $y\in\operatorname{supp}Q$,

$$\left(\frac{\mathrm{d}P_{Y_1|X_1=x_1}^{(h_1,Q_1,\lambda_1)}}{\mathrm{d}Q_1}(y)\right)^{\alpha_1}\left(\frac{\mathrm{d}P_{Y_2|X_2=x_2}^{(h_2,Q_2,\lambda_2)}}{\mathrm{d}Q_2}(y)\right)^{\alpha_2}$$

$$= \left(\exp\left(-\frac{1}{\lambda_1}h_1\left(x_1,y\right) - \mathsf{K}_{h_1,Q_1,x_1}\left(-\frac{1}{\lambda_1}\right)\right)\right)^{\alpha_1} \left(\exp\left(-\frac{1}{\lambda_2}h_2\left(x_2,y\right) - \mathsf{K}_{h_2,Q_2,x_2}\left(-\frac{1}{\lambda_2}\right)\right)\right)^{\alpha_2} \tag{75}$$

$$= \exp\left(-\frac{\alpha_1}{\lambda_1}h_1\left(x_1,y\right) - \frac{\alpha_2}{\lambda_2}h_2\left(x_2,y\right)\right) \exp\left(-\alpha_1\mathsf{K}_{h_1,Q_1,x_1}\left(-\frac{1}{\lambda_1}\right) - \alpha_2\mathsf{K}_{h_2,Q_2,x_2}\left(-\frac{1}{\lambda_2}\right)\right). \tag{76}$$

Substituting the equality in (76) into the normalized log-linear combination of Gibbs measures in (30) yields

$$\frac{\mathrm{d}S1}{\mathrm{d}Q}\left(y\right) = \frac{\exp\left(-\frac{\alpha_1}{\lambda_1}h_1\left(x_1,y\right) - \frac{\alpha_2}{\lambda_2}h_2\left(x_2,y\right)\right)}{\int \exp\left(-\frac{\alpha_1}{\lambda_1}h_1\left(x_1,\widehat{y}\right) - \frac{\alpha_2}{\lambda_2}h_2\left(x_2,\widehat{y}\right)\right)\mathrm{d}Q\left(\widehat{y}\right)} \frac{\exp\left(-\alpha_1\mathsf{K}_{h_1,Q_1,x_1}\left(-\frac{1}{\lambda_1}\right) - \alpha_2\mathsf{K}_{h_2,Q_2,x_2}\left(-\frac{1}{\lambda_2}\right)\right)}{\exp\left(-\alpha_1\mathsf{K}_{h_1,Q_1,x_1}\left(-\frac{1}{\lambda_1}\right) - \alpha_2\mathsf{K}_{h_2,Q_2,x_2}\left(-\frac{1}{\lambda_2}\right)\right)} \tag{77}$$

$$= \frac{\exp\left(-\frac{\alpha_1}{\lambda_1}h_1\left(x_1,y\right) - \frac{\alpha_2}{\lambda_2}h_2\left(x_2,y\right)\right)}{\int \exp\left(-\frac{\alpha_1}{\lambda_1}h_1\left(x_1,\widehat{y}\right) - \frac{\alpha_2}{\lambda_2}h_2\left(x_2,\widehat{y}\right)\right)\mathrm{d}Q\left(\widehat{y}\right)} \tag{78}$$

$$= \frac{\mathrm{d}P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}{\mathrm{d}Q}(y), \tag{79}$$

where the equality in (79) follows from Definition 3. This completes the proof of (31).

E Proof of Theorem 11

From Theorem 8, the normalized log-linear combination S_1 is identical to the $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)$-Gibbs probability measure $P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$. Moreover, since $P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$ is a Gibbs probability measure with reference measure Q, from [9, Lemma 3], it follows that

$$Q \ll P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)} \ll Q. \tag{80}$$

Thus, from [32, Theorem 2], the Radon–Nikodym derivative

$$\frac{\mathrm{d}Q}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}$$

is well defined and, from the multiplicative inverse of the Radon–Nikodym derivative [32, Theorem 5], for all $y \in \operatorname{supp} Q$, it follows that

$$\frac{\mathrm{d}Q}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}(y) = \left(\frac{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}{\mathrm{d}Q}(y) \right)^{-1}. \tag{81}$$

Therefore, from Theorem 8 and the equality in (81), for all $y \in \operatorname{supp} Q$,

$$\begin{aligned}
\frac{\mathrm{d}S_1}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}(y) &= \frac{\mathrm{d}S_1}{\mathrm{d}Q}(y) \frac{\mathrm{d}Q}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}(y) && (82)\\
&= \frac{\mathrm{d}S_1}{\mathrm{d}Q}(y) \left(\frac{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}{\mathrm{d}Q}(y) \right)^{-1} && (83)\\
&= 1, && (84)
\end{aligned}$$

where the equality in (82) follows from the chain rule of the Radon–Nikodym derivative [32, Theorem 4]; the equality in (83) follows from (81); and the equality in (84) follows from Theorem 8. This proves *(i)*.

Assume now that *(i)* holds. Then,

$$\begin{aligned}
D\left(S_1 \,\middle\|\, P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}\right) &= \int \log\left(\frac{\mathrm{d}S_1}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}(y) \right) \mathrm{d}S_1(y) && (85)\\
&= 0, && (86)
\end{aligned}$$

which proves *(ii)*. Conversely, if *(ii)* holds, then from [9, Theorem 1] it follows that

$$\frac{\mathrm{d}S_1}{\mathrm{d}P_{Y|(X_1,X_2)=(x_1,x_2)}^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}}(y) = 1. \tag{87}$$

Therefore, *(i)* and *(ii)* are equivalent.

Assume next that *(i)* holds. Then S_1 is identical to the $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}}, Q, 1\right)$-Gibbs probability measure $P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$. Hence, from [31, Lemma 1], it is the unique solution, if it exists, to the optimization problem in (33). This proves that *(i)* implies *(iii)*.

Assume now that *(iii)* holds. From [31, Lemma 1], the unique solution, if it exists, to the optimization problem in (33) is precisely the $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}}, Q, 1\right)$-Gibbs probability measure $P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$. Hence,

$$S_1 = P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}. \tag{88}$$

Thus,

$$D\left(S_1 \parallel Q\right) = D\left(P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)} \middle\| Q\right). \tag{89}$$

Since $P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$ is the $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}}, Q, 1\right)$-Gibbs probability measure, [31, Lemma 4] implies that it is the unique solution, if it exists, to the optimization problem in (34). Therefore, S_1 is the unique solution, if it exists, to the same optimization problem. This proves that *(iii)* implies *(iv)*.

Assume finally that *(iv)* holds. From [31, Lemma 4], the unique solution, if it exists, to the optimization problem in (34) is precisely the $\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}}, Q, 1\right)$-Gibbs probability measure $P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}$. Hence,

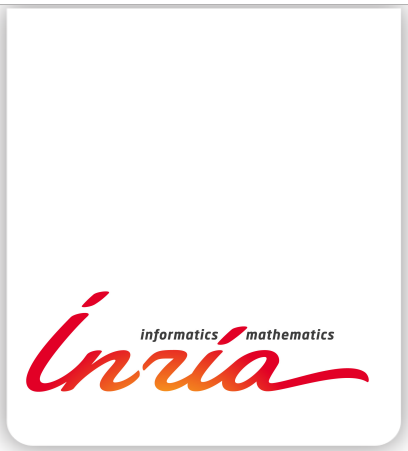
$$S_1 = P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}. \tag{90}$$

Together with the absolute-continuity relation established above and the explicit Radon–Nikodym derivatives in Theorem 8, it follows that, for all $y \in \operatorname{supp} Q$,

$$\frac{\mathrm{d}S_1}{\mathrm{d}P^{\left(\widehat{h}_{\frac{\alpha_1}{\lambda_1},\frac{\alpha_2}{\lambda_2}},Q,1\right)}_{Y|(X_1,X_2)=(x_1,x_2)}}(y) = 1. \tag{91}$$

This proves that *(iv)* implies *(i)*.

Therefore, *(i)*, *(ii)*, *(iii)*, and *(iv)* are equivalent. This concludes the proof.



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