Statistical Federated Learning and Generalization

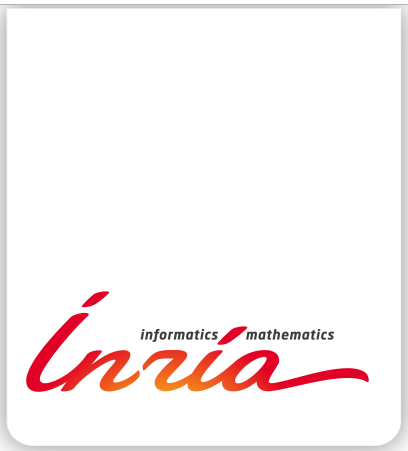
Yaiza Bermudez, Samir M. Perlaza, Iñaki Esnaola, H. Vincent Poor

Statistical Federated Learning and Generalization

Yaiza Bermudez, Samir M. Perlaza, Iñaki Esnaola and H. Vincent Poor

RESEARCH REPORT
N° 9599
October 2025
Project-Team NEO

ISSN 0249-6399 ISRN INRIA/RR--9599--FR+ENG



Statistical Federated Learning and Generalization

Yaiza Bermudez, Samir M. Perlaza, Iñaki Esnaola and H. Vincent Poor

Project-Team NEO

Research Report n° 9599 — October 2025 — 67 pages

Abstract: In this report, the generalization error of federated learning (FL) systems is characterized through a novel statistical framework. Central to this framework is the concept of a meta-federated learning algorithm, defined as a probability measure over a client's local models conditioned on the datasets of all participating clients. By means of this abstraction, several fundamental properties of FL systems are stated and closed-form expressions for the generalization error are derived. More specifically, the method of gaps, originally introduced for non-federated settings, is extended to FL, and closed-form expressions for the generalization error are obtained in terms of classical information measures, including relative entropy, mutual information, and lautum information. A central role in these new expressions is played by some specific Gibbs probability measures (Gibbs algorithms). More importantly, it is revealed that the challenge of evaluating the generalization error in FL is reduced to two distinct tasks: (*a*) measuring the dependence of client model choices on the datasets of all clients; and (*b*) distinguishing the meta-federated learning algorithm from a Gibbs algorithm trained solely on local data. These findings establish new links between generalization in FL, mismatched hypothesis testing, Shannon's information measures, and Pythagorean identities for the generalization error.

Key-words: Federated Learning, Generalization, Gibbs Measures, and Method of Gaps

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. H. Vincent Poor, Samir M. Perlaza, and Iñaki Esnaola 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 Fédéré et Généralisation

Résumé : Dans ce rapport, l'erreur de généralisation des systèmes d'apprentissage fédéré (FL) est caractérisée au moyen d'un nouveau cadre statistique. Au cœur de ce cadre se trouve la notion d'algorithme d'apprentissage méta-fédéré, défini comme une mesure de probabilité sur les modèles locaux d'un client, conditionnée par les ensembles de données de l'ensemble des clients participants. Grâce à cette abstraction, plusieurs propriétés fondamentales des systèmes de FL sont énoncées et des expressions explicites (en forme fermée) de l'erreur de généralisation sont dérivées. Plus précisément, la méthode des écarts (« method of gaps »), initialement introduite pour des contextes non fédérés, est étendue au FL, et des expressions en forme fermée de l'erreur de généralisation sont obtenues en termes de mesures d'information classiques, notamment l'entropie relative (divergence de Kullback-Leibler), l'information mutuelle et l'« information de lautum ». Un rôle central dans ces nouvelles expressions est joué par certaines mesures de probabilité de Gibbs (algorithmes de Gibbs). Plus important encore, il apparaît que la difficulté d'évaluer l'erreur de généralisation en FL se ramène à deux tâches distinctes : (*a*) mesurer la dépendance des choix de modèles des clients à l'égard des ensembles de données de tous les clients; et (*b*) distinguer l'algorithme d'apprentissage méta-fédéré d'un algorithme de Gibbs entraîné uniquement sur des données locales. Ces résultats établissent de nouveaux liens entre la généralisation en FL, les tests d'hypothèses « mismatched » (non adaptés), les mesures d'information de Shannon et des identités pythagoriciennes pour l'erreur de généralisation.

Mots-clés : Apprentissage Fédéré, Généralisation, Mesures de Gibbs et Méthode des Ecarts

Contents

1 Introduction

Federated learning (FL), first introduced in [1], harnesses training across a population of clients by repeatedly exchanging information with a coordinating server [2, 3]. This setting departs from classical centralized learning in two crucial ways: (*i*) data are statistically heterogeneous across clients [4–7] and often severely unbalanced in terms of integrity and dataset sizes [1, 2]; and (*ii*) the set of participating clients varies over time due to system constraints, communication costs, and availability [8–10].

From this perspective, the analysis of generalization capabilities of FL has been out of the reach of classical methods based on VC-dimension or Rademacher complexity [11–15], despite important contributions, such as [4, 16, 17]. This contrasts with non-federated learning for which a rich variety of bounds [18–27] and exact expressions are known for broad classes of statistical learning algorithms. Most of the existing exact expressions for the generalization error were separately introduced by different authors using specific methods of proofs for particular learning algorithms [26, 28, 29]. Interestingly, the method of gaps, introduced in [30], establishes a unified framework from which all existing exact expressions can be obtained. More importantly, such a method establishes strong connections among generalization and several areas in statistics, e.g., mismatched hypothesis testing [31], classical information measures [32–34], Pythagorean identities of variations of expectations [35], and I-projections [36].

1.1 Related Work

Optimization and statistical methods have led to insightful lower and upper bounds on the generalization error of FL systems [9, 17, 37–50]. Some contributions focus on a *participation* (client-sampling) generalization gap, i.e., the discrepancy between the generalization performance obtained when training involves a random subset of clients and the performance exclusively involving local training datasets. Along the participation generalization gap, also an *out-of-sample* generalization gap has been studied in [51,52], where out-of-sample generalization must be understood in the sense of [53]. The interest in these gaps lies in the fact that their sum equals the generalization error of the FL system under study. These gaps are typically studied via optimization-centric convergence analyses for specific algorithms, including FedAvg [1], FedProx [5], SCAFFOLD [6], FedOpt [54], among others, under heterogeneous data and partial participation [55, 56]. Despite the numerous lower and upper bounds on the generalization error of federated learning algorithms, exact expressions are uncharted territory.

1.2 Contributions

This report develops a statistical model that explicitly characterizes, in statistical terms, each of the components of an FL system. For a fixed set of participating clients, this model reproduces as special cases, most standard FL scenarios,

e.g., FedAvg [1], FedProx [5], among others. Central to this formulation is a new notion, referred to as *meta-federated learning algorithm*, which consists of a probability measure over each client's local models conditioned on the joint training datasets of all clients. This conditional probability measure arises naturally from the client-server communication process and serves as a stand-alone mathematical abstraction of such a client within the federation. Each client selects its models by sampling such a probability measure. Leveraging this flexible and general abstraction, important conclusions are drawn for all FL systems. First, it is shown that, the meta-federated learning algorithm of each client is a convex combination of all instances of the local algorithm, which is represented by a probability measure on the models conditioned on the local training dataset and all messages exchanged with the server. Hence, the set of all meta-federated learning algorithms in FL is a strict subset of all probability measures that can be defined over the set of models. Second, it is shown that relevant performance metrics for a given client, such as the expected test empirical risk, the expected training empirical risk, and the expected generalization error, can be expressed in terms of the corresponding meta-federated learning algorithm. This observation enables an extension of the method of gaps, introduced in [30] for the case of non-federated systems, to the federated setting. As a result, closed-form expressions for the generalization error of FL systems are obtained in terms of classical information measures (relative entropy, mutual information [32, 33], and lautum information [34]). In such expressions, a Gibbs probability measure on the models, which represents a Gibbs algorithm [57], plays a central role. In particular, the difficulty of computing a client's generalization error in an FL system is shown to be decomposed into two separated challenges: (i) quantifying the dependence of the client's model selection on the training datasets of all participating clients via the mutual and lautum information induced by the meta-federated learning algorithm; and (ii) reliably distinguishing the meta-federated learning algorithm from a Gibbs algorithm exclusively trained on the individual training dataset. These challenges unveil exciting connections among generalization in FL, mismatched hypothesis testing, Shannon's information measures and Pythagorean identities for the generalization error.

2 System Model

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

$$z_k \triangleq \big((x_{k,1}, y_{k,1}), (x_{k,2}, y_{k,2}), \ldots, (x_{k,n_k}, y_{k,n_k})\big) . \tag{1}$$

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

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

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

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

$$n_0 \triangleq \sum_{k=1}^{K} n_k. \tag{4}$$

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

$$\ell_k : \mathcal{Z}_k \times \mathcal{M}_k \to [0, +\infty), \tag{5}$$

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 z_k in (1), is determined by the function

$$\mathsf{L}_k : \begin{cases} \mathcal{Z}_k^{n_k} \times \mathcal{M}_k \longrightarrow [0, +\infty) \\ \quad (z_k, \boldsymbol{\theta}) \longmapsto \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k \left(x_{k,i}, y_{k,i}, \boldsymbol{\theta}\right), \end{cases} \tag{6}$$

where the function ℓ_k is defined in (5).

The underlying assumption of federated learning is that each client sends to and receives messages from a server. Such messages, which circulate through a communication network, help clients to cooperatively search for models. A typical formulation of a federated learning system consists in distributively solving the following optimization problem via messages exchanges between the server and the clients,

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \sum_{k=1}^{K} \frac{n_k}{n_0} \mathsf{L}_k \left(z_k, \boldsymbol{\theta}\right), \tag{7}$$

where $\mathcal{M} \triangleq \bigcap_{i=1}^{K} \mathcal{M}_k$ is the set of common models. The motivation of clients for adopting models that are solutions to the problem in (7) becomes evident in the case in which all agents use the same loss function, i.e., $\ell_1 = \ell_2 = \ldots = \ell_K = \ell$. In such a case, the problem in (7) becomes

$$\min_{\boldsymbol{\theta} \in \mathcal{M}} \frac{1}{n_0} \sum_{t=1}^{n_0} \ell \left(\boldsymbol{\theta}, x_{0,t}, y_{0,t}\right), \tag{8}$$

where the pair $(x_{0,t}, y_{0,t})$ is a data point in the dataset z_0 in (3). Note such a dataset z_0 aggregates all local datasets and thus, the solution to (8) can be seen as a model that minimizes the empirical risk with respect to the aggregation of all local datasets.

In the following, the communication network is described. Using such a network, the stochastic description of the federated learning system under study is presented. Finally, the mathematical formulation of the problem studied in this work is thoroughly described.

2.1 Communication System

Server and clients engage on a synchronous communication over bidirectional, noiseless channels with information transmission rate limitations. This communication consists of m communication rounds. In each round, all clients send a message to the server, which subsequently sends a message to each client. At the end of these m communication rounds, all clients locally choose their corresponding models. Previous to this communication, a preamble is conducted to configure both local and global parameters for the clients and the server.

2.1.1 Preamble

The preamble precedes the federated learning process and serves to initialize the system. During this period, clients acquire reference probability measures, typically in the form of priors, over the set of models and/or the set of datasets. Such reference measures may be transmitted by the server or derived locally by each client using available data, e.g., training datasets.

2.1.2 Communication Protocol

During communication round t, with $t \in \{1, 2, \ldots, m\}$ and $k \in \{1, 2, \ldots, K\}$, the message received by client k from the server and the message sent by client k to the server are denoted by

$$u_{k,t} \in \mathcal{U}_k, \text{ and} \tag{9}$$

$$v_{k,t} \in \mathcal{V}_k, \tag{10}$$

respectively, with $\mathcal{U}_k$ and $\mathcal{V}_k$ some given finite sets. Communication round t starts with all clients sending their corresponding messages $v_{1,t}$, $v_{2,t}$, …, $v_{K,t}$ to the server. Once all messages have been received by the server, it sends the messages $u_{1,t}$, $u_{2,t}$, …, $u_{K,t}$ to the corresponding clients. Using this notation, Figure 1 depicts the exchange of messages during two communication rounds, namely $t-1$ and t, in the case in which $K = 2$ and $t > 1$.

For all $t \in \{1, 2, \ldots, m\}$, messages $v_{1,t}$, $v_{2,t}$, …, $v_{K,t}$ and $u_{1,t}$, $u_{2,t}$, …, $u_{K,t}$ are subject to information transmission rate constraints. More specifically, at each communication round, at most $\overline{R}_k \in (0, +\infty)$ bits are sent by client k to the server; and at most $\underline{R}_k \in (0, +\infty)$ bits are received by client k from the server. Hence, $\overline{R}_k$ and $\underline{R}_k$ are the maximum information transmission rates in bits per

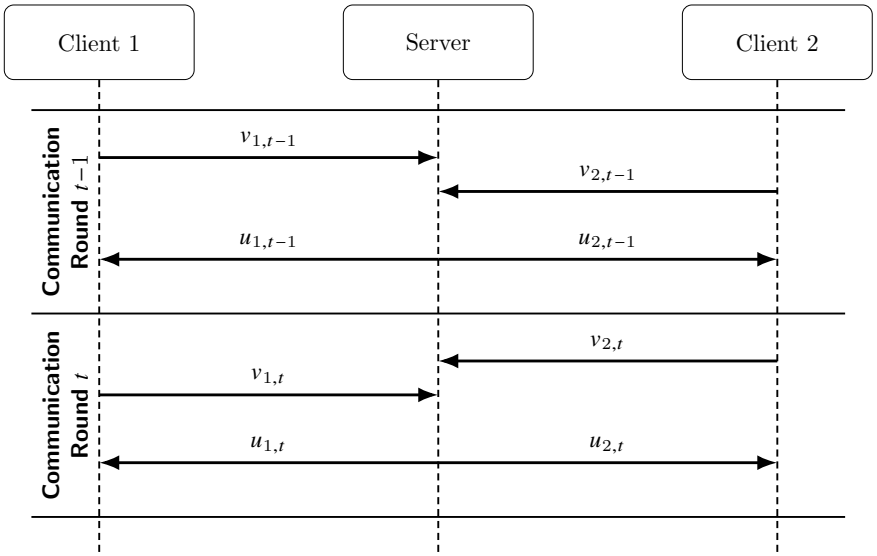


Figure 1: Two consecutive communication rounds, namely $t-1$ and t, with $t \in \{2, 3, \ldots, m\}$, in the case in which $K = 2$.

communication-round duration. These constraints translate into the following inequalities, for all $k \in \{1, 2, \ldots, K\}$,

$$|\mathcal{U}_k| \leqslant 2^{\underline{R}_k} \text{ and} \tag{11}$$

$$|\mathcal{V}_k| \leqslant 2^{\bar{R}_k}. \tag{12}$$

This is due to the fact that for establishing a one-to-one relation between the elements of the set $\mathcal{U}_k$ (respectively $\mathcal{V}_k$) and some binary sequences, it requires having $|\mathcal{U}_k|$ (respectively $|\mathcal{V}_k|$) sequences of at least $\log_2(|\mathcal{U}_k|)$ bits (respectively $\log_2(|\mathcal{V}_k|)$ bits). This implies that the rate $\underline{R}_k$ (respectively $\bar{R}_k$) shall be at least $\log_2(|\mathcal{U}_k|)$ (respectively $\log_2(|\mathcal{V}_k|)$) bits per communication round from the server to client k (respectively from client k to the server). This justifies the inequalities (11) and (12). In the following, the uplink and downlink data rates, $\bar{R}_k$ and $\underline{R}_k$ respectively, are fixed parameters in this analysis.

2.2 Stochastic Modelling

For the ease of notation, all messages sent to the server by the clients during round t are denoted by

$$\boldsymbol{v}^{(t)} = \left(v_{1,t}, v_{2,t}, \ldots, v_{K,t}\right)^\top \in \mathcal{V}_1 \times \mathcal{V}_2 \times \ldots \times \mathcal{V}_K; \tag{13}$$

and all messages sent by client k during all previous rounds, including round t, are denoted by

$$\boldsymbol{v}_k^{(t)} = \left(v_{k,1}, v_{k,2}, \ldots, v_{k,t}\right)^\top \in \mathcal{V}_k^t. \tag{14}$$

The messages sent by the server to the clients, during such a communication round t, are denoted by

$$\boldsymbol{u}^{(t)} = \left(u_{1,t}, u_{2,t}, \ldots, u_{K,t}\right)^\top \in \mathcal{U}_1 \times \mathcal{U}_2 \times \ldots \times \mathcal{U}_K; \tag{15}$$

and all the messages sent by the server to client k during all previous rounds, including round t, are denoted by

$$\boldsymbol{u}_k^{(t)} = \left(u_{k,1}, u_{k,2}, \ldots, u_{k,t}\right)^\top \in \mathcal{U}_k^t. \tag{16}$$

Note that the tuples in (13) and (14) share the following structure,

$$\left(\boldsymbol{v}_1^{(t)}\ \boldsymbol{v}_2^{(t)}\ \cdots\ \boldsymbol{v}_K^{(t)}\right)^\top = \underbrace{\begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,t} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ v_{K,1} & v_{K,2} & \cdots & v_{K,t} \end{bmatrix}}_{\triangleq \underline{\boldsymbol{v}}^{(t)} \in (\mathcal{V}_1 \times \mathcal{V}_2 \times \ldots \times \mathcal{V}_K)^t} = \left(\boldsymbol{v}^{(1)}\ \boldsymbol{v}^{(2)}\ \cdots\ \boldsymbol{v}^{(t)}\right). \tag{17}$$

Similarly, the tuples in (15) and (16) share the following structure,

$$\left(\boldsymbol{u}_1^{(t)}\ \boldsymbol{u}_2^{(t)}\ \cdots\ \boldsymbol{u}_K^{(t)}\right)^\top = \underbrace{\begin{bmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,t} \\ u_{2,1} & u_{2,2} & \cdots & u_{2,t} \\ \vdots & \vdots & \ddots & \vdots \\ u_{K,1} & u_{K,2} & \cdots & u_{K,t} \end{bmatrix}}_{\triangleq \underline{\boldsymbol{u}}^{(t)} \in (\mathcal{U}_1 \times \mathcal{U}_2 \times \ldots \times \mathcal{U}_K)^t} = \left(\boldsymbol{u}^{(1)}\ \boldsymbol{u}^{(2)}\ \cdots\ \boldsymbol{u}^{(t)}\right). \tag{18}$$

For the ease of notation, let the sets $\mathcal{M}_0$, $\mathcal{U}_0$, $\mathcal{V}_0$, and $\mathcal{Z}_0$ be respectively defined as follows:

$$\mathcal{M}_0 \triangleq \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_K \subset \mathbb{R}^{d \times K}; \tag{19}$$

$$\mathcal{U}_0 \triangleq \mathcal{U}_1 \times \mathcal{U}_2 \times \ldots \times \mathcal{U}_K; \tag{20}$$

$$\mathcal{V}_0 \triangleq \mathcal{V}_1 \times \mathcal{V}_2 \times \ldots \times \mathcal{V}_K; \tag{21}$$

$$\mathcal{Z}_0 \triangleq \mathcal{Z}_1^{n_1} \times \mathcal{Z}_2^{n_2} \times \ldots \times \mathcal{Z}_K^{n_K}. \tag{22}$$

After m communication rounds, the model chosen by client k is denoted by $\boldsymbol{\theta}_k \in \mathcal{M}_k$. The tuple containing the models chosen by all clients is denoted by

$$\underline{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K) \in \mathcal{M}_0. \tag{23}$$

In this analysis, at the end of m communication rounds, the dataset, $\boldsymbol{z}_0$ in (2); the messages sent and received throughout the communication process, $\underline{\boldsymbol{v}}^{(m)}$

in (17) and $\underline{\boldsymbol{u}}^{(m)}$ and (18); and the models chosen by all clients, $\underline{\boldsymbol{\theta}}$ in (23), are obtained by sampling the joint probability measure

$$P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)},\boldsymbol{Z}_0} \in \triangle\left(\mathcal{M}_0 \times \mathcal{U}_0^m \times \mathcal{V}_0^m \times \mathcal{Z}_0\right), \tag{24}$$

where the sets $\mathcal{M}_0$, $\mathcal{U}_0$, $\mathcal{V}_0$, and $\mathcal{Z}_0$ are defined in (19), (20), (21), and (22), respectively. The joint probability measure in (24) can be factorized as

$$P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)},\boldsymbol{Z}_0} = P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}\, P_{\boldsymbol{Z}_0}, \tag{25}$$

with,

$$P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0} \in \triangle\left(\mathcal{M}_0 \times \mathcal{U}_0^m \times \mathcal{V}_0^m \mid \mathcal{Z}_0\right); \text{ and} \tag{26}$$

$$P_{\boldsymbol{Z}_0} \in \triangle\left(\mathcal{Z}_0\right). \tag{27}$$

The conditional probability measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}$ in (26) is a statistical description of the federated learning system. Alternatively, the probability measure $P_{\boldsymbol{Z}_0}$ in (27) is a statistical description of the data source.

The analysis continues by highlighting some of the constraints derived from the federated learning system implementation.

Assumption 1. *The model chosen by client k, with $k \in \{1, 2, \ldots, K\}$, is obtained by sampling a probability measure exclusively conditioned on*

- *The dataset $\boldsymbol{z}_k$ in (1);*
- *The m messages sent to the server, $\boldsymbol{v}_k^{(m)}$ in (14); and*
- *The m messages received from the server, $\boldsymbol{u}_k^{(m)}$ in (16).*

The conditional probability measure described in Assumption 1 is denoted by

$$P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{U}_k^m \times \mathcal{V}_k^m \times \mathcal{Z}_k^{n_k}\right), \tag{28}$$

and can be interpreted as the *machine learning algorithm* used by client k, whose inputs are the dataset $\boldsymbol{z}_k$ in (1); the m messages sent to the server, $\boldsymbol{v}_k^{(m)}$ in (14); and the m messages received from the server, $\boldsymbol{u}_k^{(m)}$ in (16).

Assumption 2. *The message sent by client k to the server at communication round t, with $k \in \{1, 2, \ldots, K\}$ and $t \in \{1, 2, \ldots, m\}$, is obtained by sampling a probability measure exclusively conditioned on*

- *The dataset $\boldsymbol{z}_k$ in (1);*
- *The $t-1$ messages previously sent to the server, $\boldsymbol{v}_k^{(t-1)}$ in (14); and*
- *The $t-1$ messages previously received from the server, $\boldsymbol{u}_k^{(t-1)}$ in (16).*

The conditional probability measure described in Assumption 2, at communication round t, with $t \in \{1, 2, \ldots, m\}$, is denoted by

$$P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k} \in \triangle\left(\mathcal{V}_k|\mathcal{U}_k^{t-1} \times \mathcal{V}_k^{t-1} \times \mathcal{Z}_k^{n_k}\right), \tag{29}$$

and can be interpreted as a mathematical abstraction of the part of client k that determines the message to be sent to the server during communication round t.

Assumption 3. *The messages sent by the server at communication round t to all clients, $\boldsymbol{u}^{(t)}$ in (15), with $t \in \{1, 2, \ldots, m\}$, are obtained by sampling a probability measure exclusively conditioned on*

- *The t messages received from each client, $\underline{\boldsymbol{v}}^{(t)}$ in (17); and*
- *The $t-1$ messages previously sent to the clients, $\underline{\boldsymbol{u}}^{(t-1)}$ in (18).*

The conditional probability measure described in Assumption 3, at communication round t, with $t \in \{1, 2, \ldots, m\}$, is denoted by

$$P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}} \in \triangle\left(\mathcal{U}_0|\mathcal{U}_0^{t-1} \times \mathcal{V}_0^{t}\right), \tag{30}$$

and can be seen as a mathematical abstraction of the server during communication round t. This follows by observing that the messages sent to the clients during communication round t, $\boldsymbol{u}^{(t)}$ in (15), are chosen exclusively based on all messages previously sent to the clients $\boldsymbol{u}^{(t-1)}$; and on all messages previously received from the clients, including those of the current communication round, $\boldsymbol{v}^{(t)}$ in (13), which is all the data available at the server.

The following lemma puts the above assumptions at play to factorize the conditional probability measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}$ in (26).

Lemma 1. *Under Assumptions 1, 2, and 3, the joint probability measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}$ in (26) satisfies,*

$$\begin{aligned} P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0} = &\left(\prod_{k=1}^{K} P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}\right) \\ &\prod_{t=1}^{m}\left(P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}} \prod_{k=1}^{K} P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}\right), \end{aligned} \tag{31}$$

where the conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ is defined in (28); the conditional probability measure $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}$ is defined in (29); and the conditional probability measure $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}}$ is defined in (30).

Proof: The proof is presented in Appendix A. ∎

Essentially, given the m exchanged messages, $\underline{\boldsymbol{v}}^{(m)}$ and $\underline{\boldsymbol{u}}^{(m)}$ in (13) and (15); and the local training dataset $\boldsymbol{z}_k$, client k chooses its models by sampling the probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\boldsymbol{Z}_k=\boldsymbol{z}_k} \in \triangle(\mathcal{M}_k)$. The message sent by

client k to the server at communication round t, $v_{k,t}$ in (10), is obtained by sampling the probability measure $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)}=\boldsymbol{u}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)}=\boldsymbol{v}_k^{(t-1)},\boldsymbol{Z}_k=\boldsymbol{z}_k} \in \triangle(\mathcal{V}_k)$. Alternatively, the messages sent by the server at communication round t to the clients can be obtained by sampling the probability measure $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)}=\underline{\boldsymbol{u}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}=\underline{\boldsymbol{v}}^{(t)}} \in \triangle(\mathcal{U}_0)$. Using these measures, it is possible to establish mathematical representations for all objects intervening in the learning process, namely, the data source, the server, and the clients. See for instance, Figure 2, which describes a *one-shot* federated learning, as is often referred to the case in which $m = 1$.

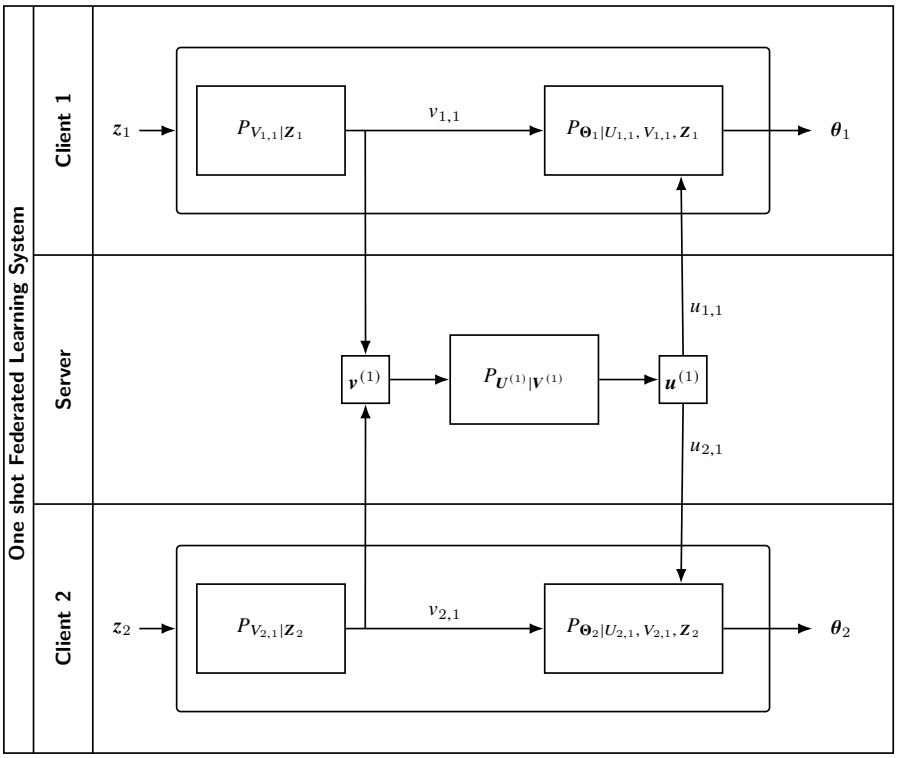


Figure 2: System model with $K = 2$ and $m = 1$ (one-shot federated learning).

2.2.1 Model of the Data Source

The data source is modeled by the probability measure $P_{\boldsymbol{Z}_0}$ in (27). That is, the datasets $\boldsymbol{z}_0 = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_K)$ in (2) are obtained by sampling such a measure $P_{\boldsymbol{Z}_0}$. The marginal probability measures of $P_{\boldsymbol{Z}_0}$ in $\triangle\left(\mathcal{Z}_k^{n_k}\right)$, with $k \in \{1, 2, \ldots, K\}$, are denoted by $P_{\boldsymbol{Z}_k}$ and are defined as follows. For all $k \in \{1, 2, \ldots, K\}$ and for

all measurable sets $\mathcal{A} \subset \mathcal{Z}_k^{n_k}$,

$$P_{\mathbf{Z}_k}(\mathcal{A}) = P_{\mathbf{Z}_0}\left(\prod_{i=1}^{k-1} \mathcal{Z}_i^{n_i} \times \mathcal{A} \times \prod_{i=k+1}^{K} \mathcal{Z}_i^{n_i}\right). \tag{32}$$

A common assumption on data sources is that they generate independent and identically distributed datasets. The following definition formalizes these assumptions.

Definition 1 (Independent and Identically Distributed Data Sets)**.** *A data source represented by the probability measure $P_{\mathbf{Z}_0}$ in* (27) *is said to generate independent datasets if for all measurable sets $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \ldots \times \mathcal{A}_K$, with $\mathcal{A}_k \subset \mathcal{Z}_k^{n_k}$,*

$$P_{\mathbf{Z}_0}(\mathcal{A}) = \prod_{k=1}^{K} P_{\mathbf{Z}_k}(\mathcal{A}_k), \tag{33}$$

where the probability measures $P_{\mathbf{Z}_1}, \ldots, P_{\mathbf{Z}_K}$ are defined in (32)*. Moreover, such datasets are said to be identically distributed if*

$$P_{\mathbf{Z}_1} = P_{\mathbf{Z}_2} = \ldots = P_{\mathbf{Z}_K}. \tag{34}$$

Another common assumption consists in considering that datasets are formed by independent datapoints. The following definition formalizes such an assumption.

Definition 2 (Independent and Identically Distributed Data Points)**.** *A data source represented by the probability measure $P_{\mathbf{Z}_0}$ in* (27) *is said to generate datasets formed by independent and identically distributed data points, if for all $k \in \{1, 2, \ldots, K\}$, the marginal measure $P_{\mathbf{Z}_k}$ in* (32) *is a product measure formed by a probability measure $P_{Z_k} \in \triangle(\mathcal{Z}_k)$. That is, for all measurable sets $\mathcal{A} = \mathcal{A}_1 \times \ldots \times \mathcal{A}_{n_k}$, with $\mathcal{A}_j \subset \mathcal{Z}_k$, and $j \in \{1, 2, \ldots, n_k\}$, it holds that*

$$P_{\mathbf{Z}_k}(\mathcal{A}) = \prod_{j=1}^{n_k} P_{Z_k}(\mathcal{A}_j). \tag{35}$$

2.2.2 Model of the Server

At each communication round, the server is modeled by a conditional probability measure, from which the messages sent to the clients are sampled. During communication round t, with $t \in \{1, 2, \ldots, m\}$, the server is modeled by the conditional probability measure $P_{\mathbf{U}^{(t)}|\underline{\mathbf{V}}^{(t)},\underline{\mathbf{U}}^{(t-1)}}$ in (30). The following definitions introduce the most common assumptions on the server.

Definition 3 (Deterministic servers)**.** *A server is said to be* deterministic *if there exists a family of functions*

$$g_t : \mathcal{V}_0^t \times \mathcal{U}_0^{t-1} \to \mathcal{U}_0, \quad \text{with } t \in \{1, 2, \ldots, m\}, \tag{36}$$

such that for all $\underline{\boldsymbol{u}}^{(t-1)} \in \mathcal{U}_0^{t-1}$*, for all* $\underline{\boldsymbol{v}}^{(t)} \in \mathcal{V}_0^t$*, and for all measurable sets* $\mathcal{A} \subseteq \mathcal{U}_0$*, the conditional probability measure* $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{V}}^{(t)},\underline{\boldsymbol{U}}^{(t-1)}}$ *in* (30) *satisfies*

$$P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{V}}^{(t)}=\underline{\boldsymbol{v}}^{(t)},\underline{\boldsymbol{U}}^{(t-1)}=\underline{\boldsymbol{u}}^{(t-1)}}(\mathcal{A}) = \begin{cases} 1 & \text{if } g_t(\underline{\boldsymbol{v}}^{(t)}, \underline{\boldsymbol{u}}^{(t-1)}) \in \mathcal{A}, \\ 0 & \text{otherwise.} \end{cases} \tag{37}$$

Essentially, the messages sent to the clients by a deterministic server are functions of all messages previously received from the clients, including those received during the current communication round, as well as, of all messages previously sent to the clients. That is, the following holds with probability one (w.r.t. $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{V}}^{(t)}=\underline{\boldsymbol{v}}^{(t)},\underline{\boldsymbol{U}}^{(t-1)}=\underline{\boldsymbol{u}}^{(t-1)}}$ in (37), with $(\underline{\boldsymbol{v}}^{(t)}, \underline{\boldsymbol{u}}^{(t-1)}) \in \mathcal{V}_0^t \times \mathcal{U}_0^{t-1}$):

$$\boldsymbol{u}^{(1)} = g_1\left(\boldsymbol{v}^{(1)}\right) \text{ and} \tag{38}$$

$$\boldsymbol{u}^{(t)} = g_t\left(\underline{\boldsymbol{v}}^{(t)}, \underline{\boldsymbol{u}}^{(t-1)}\right), \text{with } t \in \{2, 3, \ldots, m\}. \tag{39}$$

Definition 4 (Memoryless server)**.** *A server is said to be* memoryless *if for all* $t \in \{2, 3, \ldots, m\}$*, the conditional probability measure* $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{V}}^{(t)},\underline{\boldsymbol{U}}^{(t-1)}}$ *in* (30) *satisfies*

$$P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{V}}^{(t)},\underline{\boldsymbol{U}}^{(t-1)}} = P_{\boldsymbol{U}^{(t)}|\boldsymbol{V}^{(t)}}, \tag{40}$$

for some given conditional probability measures

$P_{\boldsymbol{U}^{(1)}|\boldsymbol{V}^{(1)}}, \quad P_{\boldsymbol{U}^{(2)}|\boldsymbol{V}^{(2)}}, \quad \ldots, \quad P_{\boldsymbol{U}^{(m)}|\boldsymbol{V}^{(m)}}$ *in* $\triangle(\mathcal{U}_0|\mathcal{V}_0)$.

At each communication round t, with $t \in \{1, 2, 3, \ldots, m\}$, the messages sent by a memoryless server depend exclusively on the messages received by the server at the beginning of the current communication round. Servers that are not memoryless, are said to be servers *with memory*.

Definition 5 (Time-Invariant server)**.** *A server is said to be* time-invariant *if for all* $t \in \{2, 3, \ldots, m\}$*, the conditional probability measure* $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{V}}^{(t)},\underline{\boldsymbol{U}}^{(t-1)}}$ *in* (30) *satisfies*

$$P_{\boldsymbol{U}^{(1)}|\boldsymbol{V}^{(1)}} = P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{V}}^{(t)},\underline{\boldsymbol{U}}^{(t-1)}} = P_{\boldsymbol{U}|\boldsymbol{V}}, \tag{41}$$

for some fixed conditional probability measure $P_{\boldsymbol{U}|\boldsymbol{V}} \in \triangle(\mathcal{U}_0|\mathcal{V}_0)$.

Servers that are time-invariant are also memoryless. Servers that are not time-invariant are referred to as *time-varying* servers.

Definition 6 (Uniform server)**.** *A server is said to be* uniform *if, at each communication round, it sends identical messages to all clients. That is, for all* $t \in \{1, 2, \ldots, m\}$*,*

$$u_{1,t} = u_{2,t} = \ldots = u_{K,t}, \tag{42}$$

where $u_{1,t}, u_{2,t}, \ldots, u_{K,t}$ *are the individual messages sent to the clients during round* t*, as defined in* (9).

Uniform servers are also referred to as *homogeneous*, *broadcast*, or *global* servers. Servers that are not uniform are often called *heterogeneous* or *personalized* servers.

2.2.3 Model of the Clients

A client is modeled by two objects, namely the *machine learning algorithm* and the *message generator* or *messenger*. At Client k, with $k \in \{1, 2, \ldots, K\}$, the former is modeled by the conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ in (28). The latter is modeled by the measures $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}$ in (29), with $t \in \{1, 2, \ldots, m\}$. The following definitions describe the main assumptions on a client expressed in terms of such a client's machine learning algorithm and messenger.

Definition 7 (Deterministic Client)**.** *Client k, with $k \in \{1, 2, \ldots, K\}$, is said to be* deterministic *if there exists a family of functions*

$$g_{k,t} : \mathcal{U}_k^{t-1} \times \mathcal{V}_k^{t-1} \times \mathcal{Z}_k^{n_k} \to \mathcal{V}_k, \text{ with } t \in \{1, 2, \ldots, m\}, \tag{43}$$

such that for all $\boldsymbol{u}_k^{(t-1)} \in \boldsymbol{U}_k^{(t-1)}$, for all $\boldsymbol{v}_k^{(t-1)} \in \boldsymbol{V}_k^{(t-1)}$, for all $\boldsymbol{z}_k \in \boldsymbol{Z}_k$, and for all measurable sets $\mathcal{B} \in \mathcal{V}_k$, the conditional probability measure $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}$ in (29) satisfies

$$P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)}=\boldsymbol{u}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)}=\boldsymbol{v}_k^{(t-1)},\boldsymbol{Z}_k=\boldsymbol{z}_k}(\mathcal{B}) = \begin{cases} 1 & \text{if } g_{k,t}(\boldsymbol{u}_k^{(t-1)}, \boldsymbol{V}_k^{(t-1)}, z_k) \in \mathcal{B}, \\ 0 & \text{otherwise.} \end{cases} \tag{44}$$

Equation (44) implies, that the messenger is deterministic, *i.e.* operates without randomness in the generation of the transmitted messages. A client is therefore deterministic when its messenger is deterministic.

Definition 8 (Conforming Client)**.** *Client k, with $k \in \{1, 2, \ldots, K\}$, is said to be* conforming *if*

1. *the conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ in (28) satisfies*

$$P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k} = P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{Z}_k}, \tag{45}$$

   for some conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{U}_k^m \times \mathcal{Z}_k^{n_k}\right)$; and

2. *for all $t \in \{1, 2, \ldots, m\}$, the conditional probability measure $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}$ in (29) satisfies*

$$P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k} = P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{Z}_k}, \tag{46}$$

   for some conditional probability measure $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{Z}_k} \in \triangle\left(\mathcal{V}_k|\mathcal{U}_k^{t-1} \times \mathcal{Z}_k^{n_k}\right)$.

Equations (45) and (46) imply, respectively, that the machine-learning algorithm and the messenger are both conforming.

Conforming algorithms are represented by probability measures conditioned on the m messages received from the server and on the local training dataset. There is no conditioning on the m messages sent to the server. The algorithms are therefore driven by the server's guidance. The same principle holds for conforming messengers. A conforming messenger is represented during communication round t by a probability measure that is conditioned on the $(t-1)$ messages received from server and the local dataset. There is no dependence on the $(t-1)$ messages transmitted. By contrast, non-conforming algorithms and messengers are represented by probability measures conditioned not only on the messages received and the local training data but also on some of messages they sent to the server. Hence, model selection reflects both the server's instructions and their own prior transmissions.

A special case of *conforming clients* is referred to as *myopic clients*.

Definition 9 (Myopic Client)**.** *Client k, with $k \in \{1, 2, \ldots, K\}$, is said to be* myopic *if it satisfies that*

1. *the conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ in* (28) *satisfies*

$$P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k} = P_{\boldsymbol{\Theta}_k|U_{k,m},\boldsymbol{Z}_k}, \tag{47}$$

   for some conditional probability measure $P_{\boldsymbol{\Theta}_k|U_{k,m},\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{U}_k \times \mathcal{Z}_k^{n_k}\right)$; and

2. *the conditional probability measure $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}$ in* (29) *satisfies, for all $t \in \{1, 2, \ldots, m\}$,*

$$P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k} = P_{V_{k,t}|U_{k,t},\boldsymbol{Z}_k}, \tag{48}$$

   for some conditional probability measure $P_{V_{k,t}|U_{k,t},\boldsymbol{Z}_k} \in \triangle\left(\mathcal{V}_k|\mathcal{U}_k \times \mathcal{Z}_k^{n_k}\right)$.

Equations (47) and (48) imply, respectively, that the machine-learning algorithm and the messenger are myopic. A myopic client is, by construction, conforming, whereas the converse does not hold in general. Myopic servers choose both the messages to be transmitted to the server and the models by sampling probability measures that are conditioned solely on the last message received from the server and the local training dataset.

The following definitions describe the main assumptions on the machine learning algorithm. The most basic classification of algorithms is between *federated algorithms* and *non-federated algorithms*.

Definition 10 (Non-federated Algorithms)**.** *The machine learning algorithm of client k, with $k \in \{1, 2, \ldots, K\}$, is said to be* nonfederated *if the conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ in* (28) *satisfies*

$$P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k} = P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k}, \tag{49}$$

for some conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$.

Non-federated learning algorithms are represented by probability measures exclusively conditioned on the local training datasets. That is, regardless of the messages sent to or received from the server. Alternatively, *federated algorithms* choose models by sampling a probability measure that is conditioned on the local training dataset and some (or all) messages sent to and received from the server.

A special case of non-federated algorithm is often referred to as Gibbs algorithms and play a central role in this work. Section 4 thoroughly describes such algorithms and their fundamental properties.

3 Generalization Error

The key feature of federated learning is that the local model chosen by client k, $\boldsymbol{\theta}_k \in \mathcal{M}_k$ with $k \in \{1, 2, \ldots, K\}$, is not determined solely by the client's own training dataset $\mathbf{z}_k \in \mathcal{Z}_k^{n_k}$. Instead, due to the server-mediated communication, it is fundamentally influenced by the collection of all training datasets $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K$ across all clients in the federation. See the factorization in (31). A statistical description of how clients choose their models depending on the training datasets of all clients is referred to as the *meta-federated learning algorithm* (or meta-algorithm) of client k.

3.1 Meta-Federated Learning Algorithm

The meta-federated learning algorithm of client k consists of the marginal (conditional) probability measure in $\triangle(\mathcal{M}_k \mid \mathcal{Z}_0)$ derived from the (conditional) probability measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}$ in (26), which describes the whole federated learning system. In the following, such a conditional probability measure is denoted by $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0} \in \triangle(\mathcal{M}_k \mid \mathcal{Z}_0)$. The superindex (m) is for highlighting the number of communication rounds. The following lemma provides an explicit expression of this conditional probability measure.

Lemma 2. *The marginal probability measure in $\triangle(\mathcal{M}_k \mid \mathcal{Z}_0)$ of the probability measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}$ in (26), denoted $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0} \in \triangle(\mathcal{M}_k \mid \mathcal{Z}_0)$, is such that for all measurable sets $\mathcal{A} \subseteq \mathcal{M}_k$ and for all $\mathbf{z}_0 = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K) \in \mathcal{Z}_0$,*

$$P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\mathcal{A}) = \sum_{(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}) \in \mathcal{U}_0^m \times \mathcal{V}_0^m} a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\mathbf{z}_0} P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\mathbf{Z}_k=\mathbf{z}_k}(\mathcal{A}), \quad (50)$$

where the conditional probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\mathbf{Z}_k}$ is in (31). Moreover, $a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\mathbf{z}_0} \geqslant 0$ and $\sum_{(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}) \in \mathcal{U}_0^m \times \mathcal{V}_0^m} a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\mathbf{z}_0} = 1$.

Proof: The proof is presented in Appendix B. ∎

The main observation of Lemma 2 is that the *meta-algorithm* of client k, $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}$ in (50), is a convex combination of all possible instances of client k's

algorithm $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ in (31). More specifically, given the local training dataset $\boldsymbol{z}_k$ in (1); and the sequence of messages $\underline{\boldsymbol{v}}^{(m)}$ and $\underline{\boldsymbol{u}}^{(m)}$ in (13) and (15), respectively; an instance of client k's algorithm is the probability measure $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\boldsymbol{Z}_k=\boldsymbol{z}_k} \in \triangle(\mathcal{M}_k)$. Given the aggregated training dataset $\boldsymbol{z}_0$ in (2), there are at most $2^{m\sum_{k=1}^{K}\bar{R}_k+\underline{R}_k}$ possible instances, where $\bar{R}_k$ and $\underline{R}_k$ are the uplink and downlink information transmission rates of client k described in (11) and (12).The key observation here is that the meta-federated learning algorithm of client k depends on the training datasets of all clients via the weights $a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\boldsymbol{z}_0}$ with which the different instances of the algorithm $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ are combined. Interestingly, such weights are determined by the client's messenger and the server. This is clearer from (31) and the following expression

$$a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\boldsymbol{z}_0} \triangleq \prod_{t=1}^{m} P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)}=\underline{\boldsymbol{u}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}=\underline{\boldsymbol{v}}^{(t)}}\left(\{\boldsymbol{u}^{(t)}\}\right) \prod_{j=1}^{K} P_{V_{j,t}|\boldsymbol{U}_j^{(t-1)}=\boldsymbol{u}_j^{(t-1)},\boldsymbol{V}_j^{(t-1)}=\boldsymbol{v}_j^{(t-1)},\boldsymbol{Z}_j=\boldsymbol{z}_j}\left(\{v_{j,t}\}\right), \tag{51}$$

which appears also in the proof Lemma 2, in Appendix B, Equation (123).

An additional observation from Lemma 2 is that the probability measure from which client k draws its models, *i.e.*, its meta-algorithm $P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}$, is not arbitrary in $\triangle(\mathcal{M}_k)$. In fact, the only probability distributions (meta-algorithms) achievable by client k are mixtures of the instances of its algorithm $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ in (50). Consequently, the realizable set of meta-algorithms is a subset of $\triangle(\mathcal{M}_k)$. In particular, while a meta-algorithm, induced by the aggregated training dataset, must lie in this subset, an algorithm trained using the aggregated dataset may, in principle, range over the entire set $\triangle(\mathcal{M}_k)$. More importantly, if the individual target learning algorithm is not a mixture of the instances of the local learning algorithm $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ in (50), then it is not a meta-algorithm and thus not achievable by client k. This highlights an important weakness of federated learning that often goes uncommented in existing literature.

Interestingly, the definition of generalization error in federated learning can be expressed in terms of such a statistical description, as shown in the remaining of this section.

3.2 Expected Empirical Risks

In a practical federated setting, the generalization capability of a client-side learning algorithm is evaluated via *testing datasets*, which differ from the *training dataset* [57]. To formally define the expectation of both local testing and local training empirical risks, consider the following functionals, for some fixed dataset

$\boldsymbol{z}_k \in \mathcal{Z}_k^{n_k}$ and some fixed model $\boldsymbol{\theta}_k \in \mathcal{M}_k$,

$$\mathsf{R}_{k,\boldsymbol{z}_k} : \begin{cases} \triangle(\mathcal{M}_k) \longrightarrow [0, +\infty) \\ P \longmapsto \int \mathsf{L}_k(\boldsymbol{z}_k, \boldsymbol{\theta}) \, \mathrm{d}P(\boldsymbol{\theta}) \end{cases} \tag{52}$$

and

$$\mathsf{R}_{k,\boldsymbol{\theta}_k} : \begin{cases} \triangle\left(\mathcal{Z}_k^{n_k}\right) \longrightarrow [0, +\infty) \\ P \longmapsto \int \mathsf{L}_k(\boldsymbol{z}, \boldsymbol{\theta}_k) \, \mathrm{d}P(\boldsymbol{z}), \end{cases} \tag{53}$$

where the function L_k is defined in (6).

Using this notation, the expected testing and training empirical risks are defined hereunder.

Definition 11 (Expected Testing and Training Empirical Risks)**.** *Consider the aggregated training datasets $\boldsymbol{z}_0 = (\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_K) \in \mathcal{Z}_0$ and a testing dataset $\widehat{\boldsymbol{z}}_k \in \mathcal{Z}_k^{n_k}$, with $k \in \{1, 2, \ldots, K\}$. In the federated learning system described by the measure $P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)} | \mathbf{Z}_0}$ in (26), the expected training empirical risk at client k is $\mathsf{R}_{k,\boldsymbol{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\boldsymbol{z}_0}\right)$, where the probability measure $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\boldsymbol{z}_0}$ is defined in (50) and the functional $\mathsf{R}_{k,\boldsymbol{z}_k}$ is defined in (52). Alternatively, the expected testing empirical risk at client k is $\mathsf{R}_{k,\widehat{\boldsymbol{z}}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\boldsymbol{z}_0}\right)$.*

Consider the conditional probability measure $P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k} \in \triangle(\mathcal{Z}_0|\mathcal{M}_k)$ to be such that for all $(\boldsymbol{\theta}_k, \boldsymbol{z}_0) \in \mathcal{M}_k \times \mathcal{Z}_0$,

$$\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_0}}(\boldsymbol{z}_0) = \frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\boldsymbol{z}_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k), \tag{54}$$

where the conditional probability measure $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}$ is defined in (50) and the probability measure $P^{(m)}_{\boldsymbol{\Theta}_k}$ is such that for all measurable sets $\mathcal{A} \subseteq \mathcal{M}_k$

$$P^{(m)}_{\boldsymbol{\Theta}_k}(\mathcal{A}) = \int P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\boldsymbol{z}_0}(\mathcal{A}) \, \mathrm{d}P_{\mathbf{Z}_0}(\boldsymbol{z}_0). \tag{55}$$

The existence of the conditional probability measure $P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k}$ in (54) is guaranteed by [58, Theorem 11]. Moreover, consider the conditional probability measure

$$P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right), \tag{56}$$

which is the marginal in $\triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right)$ of the conditional probability measure $P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k}$ in (54).

The following lemmas, establish a connection between the functionals $\mathsf{R}_{k,\boldsymbol{z}_k}$ in (52) and $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ in (53).

Lemma 3. *Assume that for all* $(z_0, \boldsymbol{\theta}_k) \in (\mathcal{Z}_0 \times \mathcal{M}_k)$*, the measures* $P_{\mathbf{Z}_0}$ *in* (27)*;* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ *in* (50)*;* $P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *in* (54)*; and* $P^{(m)}_{\boldsymbol{\Theta}_k}$ *in* (55)*, satisfy* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \ll P^{(m)}_{\boldsymbol{\Theta}_k}$ *and* $P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \ll P_{\mathbf{Z}_0}$*. Then, it follows that*

$$\int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}(z_0) = \int \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k), \quad (57)$$

where the functionals R_{k,z_k} *and* $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ *are defined in* (52) *and* (53)*, respectively; and the conditional probability measure* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}$ *is defined in* (56)*.*

Proof: The proof is presented in Appendix C ■

Lemma 4. *Assume that for all* $(z_0, \boldsymbol{\theta}_k) \in (\mathcal{Z}_0 \times \mathcal{M}_k)$*, the measures* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ *in* (50)*;* $P^{(m)}_{\boldsymbol{\Theta}_k}$ *in* (55)*;* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *in* (56)*; and* $P_{\mathbf{Z}_k}$ *in* (32)*, satisfy* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \ll P^{(m)}_{\boldsymbol{\Theta}_k}$ *and* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \ll P_{\mathbf{Z}_k}$*. Then, it follows that*

$$\begin{aligned}
&\iint \mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}(z_0)\, \mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) \\
&= \int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k}\right) \mathrm{d}P_{\mathbf{Z}_k}(z_k) && (58)\\
&= \int \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P_{\mathbf{Z}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) && (59)\\
&= \iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\, \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right), && (60)
\end{aligned}$$

where the functionals R_{k,z_k} *and* $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ *are defined in* (52) *and* (53)*, respectively; and the probability measure* $P_{\mathbf{Z}_0}$ *is defined in* (27)*.*

Proof: The proof is presented in Appendix D ■

3.3 Definition of Generalization Error

The generalization error of client k, with $k \in \{1, 2, \ldots, K\}$, is the expectation with respect to the joint probability measure of local models and all clients' training datasets of the difference between the local testing and local training empirical risks. In particular, for client k, let the meta-algorithm be $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0} \in \triangle(\mathcal{M}_k \mid \mathcal{Z}_0)$, let $\widehat{z}_k \in (\mathcal{X} \times \mathcal{Y})^{n_k}$ be a testing dataset, and $z_k \in (\mathcal{X} \times \mathcal{Y})^{n_k}$ a training dataset, where z_k and $\widehat{z}_k$ are components of z_0 and $\widehat{z}_0$ as in (2). A typical evaluation of the generalization capability of the instance $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ consists in the difference

$$\mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) - \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right), \quad (61)$$

where the functionals $\mathsf{R}_{k,\widehat{z}_k}$ and R_{k,z_k} are defined in (52). In a nutshell, the evaluation measures the variation of the expected empirical risk induced by $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ when the dataset changes from z_k (training) to $\widehat{z}_k$ (testing). A small variation indicates good generalization for client k. While a small value of (61)

on specific testing datasets does not guarantee generalization to all datasets, this is a widely used practical criterion.

The analysis can be deepened under additional assumptions. A common one and the one adopted here to define the *generalization error* is that, for each client k, training and testing datasets are independently drawn from the same probability measure. In our federated setting, models are produced conditionally on the joint dataset $\mathbf{z}_0 \sim P_{\mathbf{Z}_0}$, while the client-$k$ testing dataset $\widehat{\mathbf{z}}_k$ is drawn independently from the marginal $P_{\mathbf{Z}_k}$.

Definition 12 (Generalization Error)**.** *In the federated learning system described by the measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}$ in* (26)*, under the assumption that datasets $\mathbf{z}_0 = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K) \in \mathcal{Z}_0$ are obtained by sampling a probability measure $P_{\mathbf{Z}_0} \in \triangle(\mathcal{Z}_0)$, the generalization error at client k is*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) \triangleq \iint \left(\mathsf{R}_{k,\widehat{\mathbf{z}}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\right) - \mathsf{R}_{k,\mathbf{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\right)\right) \mathrm{d}P_{\mathbf{Z}_k}(\widehat{\mathbf{z}}_k)\mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0), \tag{62}$$

where the functionals $\mathsf{R}_{k,\mathbf{z}_k}$ and $\mathsf{R}_{k,\widehat{\mathbf{z}}_k}$ are defined in (52)*; the probability measure $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}$ is defined in* (50)*; and the measure $P_{\mathbf{Z}_k}$ is defined in* (32)*.*

From Definition 12, it follows that the generalization error is the expectation of the difference between the expected testing empirical error $\mathsf{R}_{k,\widehat{\mathbf{z}}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\right)$ and the expected training empirical error $\mathsf{R}_{k,\mathbf{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\right)$ induced by the meta-algorithm $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}$ of client k under the assumption that the training dataset $\mathbf{z}_0 = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K)$ and the testing dataset $\widehat{\mathbf{z}}_k$ are independently sampled from the measures $P_{\mathbf{Z}_0}$ and $P_{\mathbf{Z}_k}$, respectively. Note that this assumption concerns only the training and testing datasets; it does not require clients' datasets to be independent or identically distributed. This definition, while written explicitly using the meta-algorithm, is identical to the one typically used in federated learning literature, cf. [37], [44], and [45].

The generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ in (62) can also be expressed in terms of the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ in (53), using Lemma 3 and Lemma 4.

Lemma 5. *The generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ in* (62) *satisfies*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) = \iint \left(\mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k), \tag{63}$$

where the functionals $\mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}$ *and* $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ *are defined in* (53)*; the probability measure* $P^{(m)}_{\boldsymbol{\Theta}_k} \in \triangle(\mathcal{M}_k)$ *is defined in* (55)*; and the conditional probability measure* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k} \mid \mathcal{M}_k\right)$ *is defined in* (56)*.*

Proof: The proof is presented in Appendix E ∎

The expression of the generalization error in (63) is, from a system-implementation perspective, less intuitive than the form in (62). Nevertheless, it is pivotal for the conclusions drawn in the following sections. A statistical intuition for (63) can be obtained as follows. Given the meta algorithm $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}$ in (50); the probability measure $P_{\mathbf{Z}_0}$ in (27); and a model $\boldsymbol{\theta}_k$ sampled from $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ for some aggregated training dataset z_0, the probability measure $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ can be interpreted as the posterior of $P_{\mathbf{Z}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\right)$ after the observation of the model $\boldsymbol{\theta}_k$.

4 Gibbs Measures

4.1 Data-Dependent Gibbs Measures

The Gibbs conditional probability measure used in this work is parametrized by the empirical risk function L_k; a σ-finite measure $Q_{\boldsymbol{\Theta}_k} \in \triangle(\mathcal{M}_k)$; and a dataset $z_k \in \mathcal{Z}_k^{n_k}$, with $k \in \{1, 2, \ldots, K\}$. In order to define such Gibbs measure consider the following function:

$$\mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},z_k}: \left\{ \begin{array}{l} \mathbb{R} \longrightarrow \mathbb{R} \\ t \longmapsto \log\left(\int \exp\left(t\,\mathsf{L}_k\left(z_k,\boldsymbol{\theta}_k\right)\right) \mathrm{d}Q_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right)\right). \end{array} \right. \tag{64}$$

Under the assumption that the reference measure $Q_{\boldsymbol{\Theta}_k}$ is a probability measure, the function $\mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},z_k}$ in (64) 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 $Q_{\boldsymbol{\Theta}_k}$. Using this notation, the definition of Gibbs conditional probability measures on the models is presented hereunder.

Definition 13. *Given the function* L_k *in* (6)*, with* $k \in \{1, 2, \ldots, K\}$*; a* σ*-finite measure* $Q_{\boldsymbol{\Theta}_k} \in \triangle(\mathcal{M}_k)$*; and a* $\lambda_k \in \mathbb{R} \setminus \{0\}$*, the probability measure* $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$ *is said to be an* $(\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k)$*-Gibbs conditional probability measure if*

$$\forall z_k \in \mathcal{S}_k,\ \mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},z_k}\left(-\frac{1}{\lambda_k}\right) < +\infty; \tag{65}$$

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

$$\frac{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}\left(\boldsymbol{\theta}_k\right) = \exp\left(-\frac{1}{\lambda_k}\mathsf{L}_k\left(z_k,\boldsymbol{\theta}_k\right) - \mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},z_k}\left(-\frac{1}{\lambda_k}\right)\right), \tag{66}$$

where the function $\mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},z_k}$ *is defined in* (64)*.*

Note that, while $P_{\mathbf{\Theta}_k|\mathbf{Z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}$ in (66) is referred to as a Gibbs conditional probability measure, the measure $P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}$, obtained by conditioning upon a given dataset $\mathbf{z}_k \in \mathcal{Z}_k^{n_k}$, is referred to as a Gibbs probability measure.

On another note, the Gibbs probability measure $P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}$ in (66) can be related to the following optimization problem, under the assumption that $\lambda_k > 0$:

$$\min_{P \in \triangle_{Q_{\mathbf{\Theta}_k}}(\mathcal{M}_k)} \mathsf{R}_{k,\mathbf{z}_k}(P) + \lambda_k D\left(P \,\middle\|\, Q_{\mathbf{\Theta}_k}\right), \tag{67}$$

where the functional $\mathsf{R}_{k,\mathbf{z}_k}$ is defined in (52). Alternatively, when $\lambda_k < 0$, such a Gibbs probability measure can be related to the following optimization problem:

$$\max_{P \in \triangle_{Q_{\mathbf{\Theta}_k}}(\mathcal{M}_k)} \mathsf{R}_{k,\mathbf{z}_k}(P) + \lambda_k D\left(P \,\middle\|\, Q_{\mathbf{\Theta}_k}\right). \tag{68}$$

The following lemma shows the connections between the Gibbs probability measure $P_{\mathbf{\Theta}_k|\mathbf{Z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}$ in (66) and the optimization problems mentioned above.

Lemma 6. *Assume that the optimization problem in* (67) *(respectively, in* (68)*) admits a solution. Then, if $\lambda_k > 0$ (respectively, if $\lambda_k < 0$) the probability measure $P_{\mathbf{\Theta}_k|\mathbf{Z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}$ in* (66) *is the unique solution.*

Proof: The proof is immediate from [35, Lemma 1]. ∎

This result has also been reported in a more general context in [59, 60]. In particular, when $\lambda_k > 0$, the conditional probability measure $P_{\mathbf{\Theta}_k|\mathbf{Z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}$ in (66) represents the long-run distribution of a stochastic gradient descent algorithm [61] that aims to minimize the expected empirical risk. In statistical learning, such a distribution is often referred to as the *Gibbs algorithm* [57]. Alternatively, when $\lambda_k < 0$, such a conditional measure represents an adversarial algorithm that aims to maximize the expected empirical risk.

The following lemma introduces a connection between the function $\mathsf{K}_{k,Q_{\mathbf{\Theta}_k},\mathbf{z}_k}$ in (64) and the optimization problems mentioned above.

Lemma 7. *Given an $(\mathsf{L}_k, Q_{\mathbf{\Theta}_k}, \lambda_k)$-Gibbs probability measure, denoted by $P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}$, with $\mathbf{z}_k \in \triangle\left(\mathcal{Z}_k^{n_k}\right)$, the following holds,*

$$-\lambda_k \mathsf{K}_{k,Q_{\mathbf{\Theta}_k},\mathbf{z}_k}\left(-\frac{1}{\lambda_k}\right) = \mathsf{R}_{k,\mathbf{z}_k}\left(P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}\right) + \lambda_k D\left(P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)} \,\middle\|\, Q_{\mathbf{\Theta}_k}\right) \tag{69}$$

$$= \mathsf{R}_{k,\mathbf{z}_k}\left(Q_{\mathbf{\Theta}_k}\right) - \lambda_k D\left(Q_{\mathbf{\Theta}_k} \,\middle\|\, P_{\mathbf{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}^{(Q_{\mathbf{\Theta}_k},\lambda_k)}\right), \tag{70}$$

where the function $\mathsf{K}_{k,Q_{\mathbf{\Theta}_k},\mathbf{z}_k}$ is defined in (64)*; and the functional $\mathsf{R}_{k,\mathbf{z}_k}$ is defined in* (52)*. Moreover, if $\lambda_k > 0$,*

$$-\lambda_k \mathsf{K}_{k,Q_{\mathbf{\Theta}_k},\mathbf{z}_k}\left(-\frac{1}{\lambda_k}\right) = \min_{P \in \triangle_{Q_{\mathbf{\Theta}_k}}(\mathcal{M}_k)} \mathsf{R}_{k,\mathbf{z}_k}(P) + \lambda_k D\left(P \,\middle\|\, Q_{\mathbf{\Theta}_k}\right). \tag{71}$$

Alternatively, if $\lambda_k < 0$,

$$-\lambda_k \mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},\boldsymbol{z}_k}\left(-\frac{1}{\lambda_k}\right) = \max_{P \in \triangle_{Q_{\boldsymbol{\Theta}_k}}(\mathcal{M}_k)} \mathsf{R}_{k,\boldsymbol{z}_k}(P) + \lambda_k D\left(P \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right). \tag{72}$$

Proof: The proof follows immediately from [35, Lemma 2]. ∎

4.2 Model-Dependent Gibbs Measures

In this section the conditional Gibbs measure is in $\triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right)$ and is parametrized by the empirical risk function L_k; a σ-finite measure $Q_{\mathbf{Z}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\right)$; and a model $\boldsymbol{\theta}_k \in \mathcal{M}_k$. Using this notation, consider the following function:

$$\mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k} : \begin{cases} \mathbb{R} \longrightarrow \mathbb{R} \\ t \longmapsto \log\left(\int \exp\left(t\,\mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\right) \mathrm{d}Q_{\mathbf{Z}_k}(\boldsymbol{z}_k)\right). \end{cases} \tag{73}$$

Under the assumption that the reference measure $Q_{\mathbf{Z}_k}$ is a probability measure, the function $\mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}$ in (73) is the cumulant generating function of the random variable $\mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k)$, for some fixed model $\boldsymbol{\theta}_k \in \mathcal{M}_k$ when the dataset $\boldsymbol{z}_k$ is sampled from $Q_{\mathbf{Z}_k}$. Using this notation, the definition of the Gibbs conditional probability measures on the datasets is presented hereunder.

Definition 14. *Given the empirical risk function* L_k *in* (6) *of client* k*, with* $k \in \{1, 2, \ldots, K\}$*; a* σ*-finite measure* $Q_{\mathbf{Z}_k}$ *on* $\triangle\left(\mathcal{Z}_k^{n_k}\right)$*; and a* $\alpha_k \in \mathbb{R} \setminus \{0\}$*, the probability measure* $P_{\widetilde{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k}^{(Q_{\mathbf{Z}_k},\alpha_k)} \in \triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right)$ *is said to be an* $(\mathsf{L}_k, Q_{\mathbf{Z}_k}, \alpha_k)$*-Gibbs conditional probability measure if*

$$\forall \boldsymbol{\theta}_k \in \mathcal{B}_k,\ \mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}\left(-\frac{1}{\alpha_k}\right) < +\infty; \tag{74}$$

for some set $\mathcal{B}_k \subseteq \mathcal{M}_k$*; and for all* $(\boldsymbol{z}_k,\boldsymbol{\theta}_k) \in \operatorname{supp} Q_{\mathbf{Z}_k} \times \mathcal{B}_k$*,*

$$\frac{\mathrm{d}P_{\widetilde{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}^{(Q_{\mathbf{Z}_k},\alpha_k)}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k) = \exp\left(-\frac{1}{\alpha_k}\mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k) - \mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}\left(-\frac{1}{\alpha_k}\right)\right), \tag{75}$$

where the function $\mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}$ *is defined in* (73)*.*

Note that, while $P_{\widetilde{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k}^{(Q_{\mathbf{Z}_k},\alpha_k)}$ in (75) is referred to as Gibbs conditional probability measure, the measure $P_{\widetilde{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}^{(Q_{\mathbf{Z}_k},\alpha_k)}$, obtained by conditioning upon a given model $\boldsymbol{\theta}_k \in \mathcal{M}_k$, is referred to as Gibbs probability measure.

Using this notation, the Gibbs probability measures $P_{\widetilde{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k}^{(Q_{\mathbf{Z}_k},\alpha_k)}$ in (75) can be related to the following optimization problems, under the assumption that $\alpha_k > 0$:

$$\min_{P \in \triangle_{Q_{\mathbf{Z}_k}}\left(\mathcal{Z}_k^{n_k}\right)} \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P\right) + \alpha_k D\left(P \,\middle\|\, Q_{\mathbf{Z}_k}\right), \tag{76}$$

where the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ is defined in (53). Alternatively, when $\alpha_k < 0$, such a Gibbs probablity measure can be related to the following optimization problem:

$$\max_{P \in \triangle_{Q_{\mathbf{Z}_k}}\left(\mathcal{Z}_k^{n_k}\right)} \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P\right) + \alpha_k D\left(P \,\middle\|\, Q_{\mathbf{Z}_k}\right). \tag{77}$$

The following lemma shows the connections between the Gibbs probability measure $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k}$ in (75) and the optimization problems mentioned above.

Lemma 8. *Assume that the optimization problem in* (76) *(respectively, in* (77)*) admits a solution. Then, if $\alpha_k > 0$ (respectively, $\alpha_k < 0$), the probability measure $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k}$ in* (75) *is the unique solution.*

Proof: The proof is immediate from [35, Lemma 1]. ∎

In particular, when $\alpha_k > 0$ and $Q_{\mathbf{Z}_k}$ is a probability measure, the objective functional in (76) is the sum of two positive functionals, namely the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ and a relative entropy. While the former is linear, the latter is strictly convex. Hence, the objective functional is strictly convex and thus, such an optimization problem admits a unique minimizer. Alternatively, when $\alpha_k < 0$ and $Q_{\mathbf{Z}_k}$ is a probability measure, the objective functional in (77) is strictly concave. Thus, this justifies the existence of a unique maximizer.

The following lemma introduces a connection between the function $\mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}$ in (73) and the optimization problems mentioned above.

Lemma 9. *Given an $(\mathsf{L}_k, Q_{\mathbf{Z}_k}, \alpha_k)$-Gibbs probability measure, denoted by $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ with $\boldsymbol{\theta}_k \in \triangle\left(\mathcal{M}_k\right)$, the following holds,*

$$-\alpha_k \mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}\left(-\frac{1}{\alpha_k}\right) = \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) + \alpha_k D\left(P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) \tag{78}$$

$$= \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(Q_{\mathbf{Z}_k}\right) - \alpha_k D\left(Q_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right), \tag{79}$$

where the function $\mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}$ is defined in (73)*; and the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ is defined in* (53)*. Moreover, if $\alpha_k > 0$,*

$$-\alpha_k \mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}\left(-\frac{1}{\alpha_k}\right) = \min_{P \in \triangle_{Q_{\mathbf{Z}_k}}\left(\mathcal{Z}_k^{n_k}\right)} \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P\right) + \alpha_k D\left(P \,\middle\|\, Q_{\mathbf{Z}_k}\right). \tag{80}$$

Alternatively, if $\alpha_k < 0$,

$$-\alpha_k \mathsf{K}_{k,Q_{\mathbf{Z}_k},\boldsymbol{\theta}_k}\left(-\frac{1}{\alpha_k}\right) = \max_{P \in \triangle_{Q_{\mathbf{Z}_k}}\left(\mathcal{Z}_k^{n_k}\right)} \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P\right) + \alpha_k D\left(P \,\middle\|\, Q_{\mathbf{Z}_k}\right). \tag{81}$$

Proof: The proof follows immediately from [35, Lemma 2]. ∎

5 Method of Gaps

The method of gaps was introduced in [30] to obtain exact expressions for the generalization error of non-federated machine learning algorithms. The method consists in two steps, such steps are retained in our analysis for clarity of presentation. Interestingly, as shown hereunder, the same method applies to federated learning systems by leveraging the notion of meta-federated learning introduced above. The key observation is that the generalization error $\mathsf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0};P_{\boldsymbol{Z}_0}\right)$ in (62) can also be expressed in terms of gaps [30, Section IV-A].

5.1 Expected Empirical Risk Gaps

For all $k \in \{1,2,\ldots,K\}$, the functional $\mathsf{R}_{k,\boldsymbol{z}_k}$ in (52), for some fixed dataset $\boldsymbol{z}_k \in \mathcal{Z}_k^{n_k}$; and those of the variations of the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ in (53), for some fixed model $\boldsymbol{\theta}_k \in \mathcal{M}_k$, due to changes in their arguments, are referred to as *gaps*. These gaps are studied via the following functionals:

$$\mathsf{G}_k : \begin{cases} \mathcal{Z}_k^{n_k} \times \triangle\left(\mathcal{M}_k\right) \times \triangle\left(\mathcal{M}_k\right) \longrightarrow \mathbb{R} \\ \qquad\qquad (\boldsymbol{z}_k, P_1, P_2) \longmapsto \mathsf{R}_{k,\boldsymbol{z}_k}\left(P_1\right) - \mathsf{R}_{k,\boldsymbol{z}_k}\left(P_2\right) \end{cases}, \tag{82}$$

and

$$\mathsf{G}_k : \begin{cases} \mathcal{M}_k \times \triangle\left(\mathcal{Z}_k^{n_k}\right) \times \triangle\left(\mathcal{Z}_k^{n_k}\right) \longrightarrow \mathbb{R} \\ \qquad\qquad (\boldsymbol{\theta}_k, P_1, P_2) \longmapsto \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P_1\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P_2\right) \end{cases}, \tag{83}$$

where the functionals $\mathsf{R}_{k,\boldsymbol{z}_k}$ and $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ are defined in (52) and (53), respectively. Although both functionals in (82) and (83) are denoted by G_k, there is no ambiguity as they can be distinguished from their arguments.

The value $\mathsf{G}_k\left(\boldsymbol{z}_k, P_1, P_2\right)$, with G_k in (82), represents the variation of the functional $\mathsf{R}_{k,\boldsymbol{z}_k}$ in (52), when its argument changes from P_2 to P_1. Such a value is often referred to as a *data-dependent gap* as the functional G_k in (82) is parametrized on a fixed dataset $\boldsymbol{z}_k \in \mathcal{Z}_k^{n_k}$. Conversely, the value $\mathsf{G}_k\left(\boldsymbol{\theta}_k, P_1, P_2\right)$, with G_k in (83), represents the variation of the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ in (53), when its argument changes from P_2 to P_1. Such a value is often referred to as a *model-dependent gap* as the functional G_k in (83) is parametrized on a fixed model $\boldsymbol{\theta}_k \in \mathcal{M}_k$. This viewpoint aligns with the perspective of variations of an expectation due to changes in the probability measure presented in [35].

The following lemmas introduce a closed-form expression for the gap $\mathsf{G}_k\left(\boldsymbol{z}_k, P_{\boldsymbol{\Theta}_k}^{(m)}, P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}^{(m)}\right)$ in (88) and the gap $\mathsf{G}_k\left(\boldsymbol{\theta}_k, P_{\boldsymbol{Z}_k}, P_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}^{(m)}\right)$ in (89).

Lemma 10. *Given an* $(\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k)$*-Gibbs probability measure, denoted by* $P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)} \in \triangle\left(\mathcal{M}_k\right)$*, with* $\boldsymbol{z}_k \in \triangle\left(\mathcal{Z}_k^{n_k}\right)$*. For all* $P \in \triangle_{Q_{\boldsymbol{\Theta}_k}}\left(\mathcal{M}_k\right)$*,*

$$\mathsf{G}_k\left(\boldsymbol{z}_k, P, P_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}\right)$$

$$= \lambda_k \Big(D\Big(P \,\Big\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\Big) + D\Big(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\Big\|\, Q_{\boldsymbol{\Theta}_k}\Big) - D\left(P \,\|\, Q_{\boldsymbol{\Theta}_k}\right)\Big). \tag{84}$$

Proof: The proof follows immediately from [35, Lemma 3] ∎

Lemma 11. *Given an* $(\mathsf{L}_k, Q_{\boldsymbol{Z}_k}, \alpha_k)$*-Gibbs probability measure, denoted by* $P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\right)$*, with* $\boldsymbol{\theta}_k \in \triangle\left(\mathcal{M}_k\right)$*. For all* $P \in \triangle_{Q_{\boldsymbol{Z}_k}}\left(\mathcal{Z}_k^{n_k}\right)$*,*

$$\begin{aligned}&\mathsf{G}_k\left(\boldsymbol{\theta}_k, P, P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\\ &= \alpha_k \Big(D\Big(P \,\Big\|\, P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\Big) + D\Big(P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\Big\|\, Q_{\boldsymbol{Z}_k}\Big) - D\left(P \,\|\, Q_{\boldsymbol{Z}_k}\right)\Big).\end{aligned} \tag{85}$$

Proof: The proof follows immediately from [35, Lemma 3] ∎

An important observation is that a gap $\mathsf{G}_k\left(\boldsymbol{z}_k, P_1, P_2\right)$ (or $\mathsf{G}_k\left(\boldsymbol{\theta}_k, P_1, P_2\right)$) is a variation of the expectation of the functions $\mathsf{L}_k\left(\boldsymbol{z}_k, \cdot\right) : \mathcal{M}_k \to [0, +\infty)$ (or $\mathsf{L}_k\left(\cdot, \boldsymbol{\theta}_k\right) : \mathcal{Z}_k^{n_k} \to [0, +\infty)$), with L_k in (6) for some fixed $\boldsymbol{z}_k \in \mathcal{Z}_k^{n_k}$ (or $\boldsymbol{\theta}_k \in \mathcal{M}_k$) due to a change of the probability measure from P_2 to P_1, both in $\triangle\left(\mathcal{M}_k\right)$ (or both in $\triangle\left(\mathcal{Z}_k^{n_k}\right)$). From this perspective, the following lemmas introduce a closed-form expression for the gap $\mathsf{G}_k\left(\boldsymbol{z}_k, P^{(m)}_{\boldsymbol{\Theta}_k}, P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}\right)$ in (88) and the gap $\mathsf{G}_k\left(\boldsymbol{\theta}_k, P_{\boldsymbol{Z}_k}, P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)$ in (89).

Lemma 12. *Consider the probability measures* P_1 *and* P_2 *both in* $\triangle_{Q_{\boldsymbol{\Theta}_k}}\left(\mathcal{M}_k\right)$ *and a dataset* $\boldsymbol{z}_k \in \mathcal{Z}_k^{n_k}$*, with* $k \in \{1, 2, \ldots, K\}$*. Then,*

$$\begin{aligned}\mathsf{G}_k\left(\boldsymbol{z}_k, P_1, P_2\right) = \lambda_k \Bigg(& D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right) - D\left(P_2 \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\\ & + D\left(P_2 \,\|\, Q_{\boldsymbol{\Theta}_k}\right) - D\left(P_1 \,\|\, Q_{\boldsymbol{\Theta}_k}\right)\Bigg),\end{aligned} \tag{86}$$

where $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$ *is the* $(\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k)$*-Gibbs conditional probability measure defined in* (66)*; and the functional* G_k *is defined in* (82)*.*

Proof: The proof is presented in Appendix F. ∎

Lemma 13. *Consider the probability measures* P_1 *and* P_2 *both in* $\triangle_{Q_{\boldsymbol{Z}_k}}\left(\mathcal{Z}_k^{n_k}\right)$ *and a model* $\boldsymbol{\theta}_k \in \mathcal{M}_k$*, with* $k \in \{1, 2, \ldots, K\}$*. Then,*

$$\begin{aligned}\mathsf{G}_k\left(\boldsymbol{\theta}_k, P_1, P_2\right) = \alpha_k \Bigg(& D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) - D\left(P_2 \,\middle\|\, P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\\ & + D\left(P_2 \,\|\, Q_{\boldsymbol{Z}_k}\right) - D\left(P_1 \,\|\, Q_{\boldsymbol{Z}_k}\right)\Bigg),\end{aligned} \tag{87}$$

where $P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right)$ *is the* $(\mathsf{L}_k, Q_{\boldsymbol{Z}_k}, \alpha_k)$*-Gibbs conditional probability measure defined in* (75)*; and the functional* G_k *is defined in* (83)*.*

Proof: The proof is presented in Appendix G. ∎

5.2 Step One

The *step one* of the method of gaps consists in writing the generalization error as the expectation of a gap. The following lemmas can be assimilated to such a step.

Lemma 14. *The generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *satisfies*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int \mathsf{G}_k\Big(z_k,P^{(m)}_{\boldsymbol{\Theta}_k},P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\Big)\mathrm{d}P_{\mathbf{Z}_0}(z_0), \tag{88}$$

where the probability measure $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ *is defined in* (50)*; the probability measure* $P^{(m)}_{\boldsymbol{\Theta}_k}\in\triangle\left(\mathcal{M}_k\right)$ *is defined in* (55)*; and the functional* G_k *is defined in* (82)*.*

Proof: The proof is presented in Appendix H ■

Lemma 15. *The generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *satisfies*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int \mathsf{G}_k\Big(\boldsymbol{\theta}_k,P_{\mathbf{Z}_k},P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\Big)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k), \tag{89}$$

where the probability measure $P_{\mathbf{Z}_k}\in\triangle\left(\mathcal{Z}_k^{n_k}\right)$ *is defined in* (27)*; the probability measure* $P^{(m)}_{\boldsymbol{\Theta}_k}\in\triangle\left(\mathcal{M}_k\right)$ *is defined in* (55)*; the conditional probability measure* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}\in\triangle\left(\mathcal{Z}_k^{n_k}\mid\mathcal{M}_k\right)$ *is defined in* (56)*; and the functional* G_k *is defined in* (83)*.*

Proof: The proof is presented in Appendix I ■

Observe that Lemma 14 and Lemma 15 express the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62) as the expectation of a data-dependent gap and a model-dependent gap, respectively. The significance of this observation is that these gaps admit closed-form expressions in terms of information measures namely, relative entropy.

5.3 Step Two

The *step two* of the method of gaps consists in leveraging the properties of such gaps for expressing the expectations mentioned above in terms of information measures. Using Lemma 12 and Lemma 14 leads to the following expression for the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62), which can be assimilated to the *step two* of the data-dependent method of gaps.

Lemma 16. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that, for all* $z_0\in\mathcal{Z}_0$*, the probability measures* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ *in* (50)

and $P^{(m)}_{\boldsymbol{\Theta}_k}$ *in* (55) *are both absolutely continuous with respect to the* σ*-finite measure* $Q_{\boldsymbol{\Theta}_k}$ *in* (66)*. Then,*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int\lambda_k\Bigg(D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\,\middle\|\,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right)$$

$$-D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\,\middle\|\,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right)+D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\,\middle\|\,Q_{\boldsymbol{\Theta}_k}\right)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\,\middle\|\,Q_{\boldsymbol{\Theta}_k}\right)\Bigg)\mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0),\quad(90)$$

where $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k}\in\triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$ *is the* $(\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k)$*-Gibbs conditional probability measure defined in* (66)*.*

Proof: The proof follows directly from Lemma 12 and Lemma 14. ∎

Similarly, using Lemma 13 and Lemma 15 leads to the following expression for the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62), which can be assimilated to the *step two* of the model-dependent method of gaps.

Lemma 17. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that, for all* $\boldsymbol{\theta}_k\in\mathcal{M}_k$*, the probability measures* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *in* (56) *and* $P_{\mathbf{Z}_k}$ *in* (32)*, are both absolutely continuous with respect to the* σ*-finite measure* $Q_{\mathbf{Z}_k}$ *in* (75)*. Then,*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int\alpha_k\Bigg(D\left(P_{\mathbf{Z}_k}\,\middle\|\,P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)$$

$$-D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\,\middle\|\,P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)+D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\,\middle\|\,Q_{\mathbf{Z}_k}\right)-D\left(P_{\mathbf{Z}_k}\,\middle\|\,Q_{\mathbf{Z}_k}\right)\Bigg)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k),\quad(91)$$

where the probability measure $P^{(m)}_{\boldsymbol{\Theta}_k}$ *is defined in* (55)*; and* $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k}\in\triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right)$ *is the* $(\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k)$*-Gibbs conditional probability measure defined in* (75)*.*

Proof: The proof follows directly from Lemma 13 and Lemma 15. ∎

Lemma 16 and Lemma 17 are the fundamental results from which *data-dependent* and *model-dependent* expressions are obtained for the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62).

The following sections introduce closed-form expressions for the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62) using the method of gaps. The expressions for the generalization error derived from (62) are referred to as *data-dependent expressions*. Those derived from (63) are referred to as *model-dependent expressions*.

6 Data-dependent Closed-Form Expressions

The exact expressions for the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62) presented in this work involve an $(\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k)$-Gibbs conditional probability measure defined in (66).

6.1 The First Closed-Form Expression

The first closed-form expression of $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62) involves a mutual information, a lautum information and a new information measure parametrized by the same parameters of the $(\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k)$-Gibbs probability measure $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ in (66). Such a new information measure is denoted by $J_{\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k}$ and satisfies

$$J_{\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)$$
$$-\int\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\,, \quad (92)$$

where $P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\in\triangle(\mathcal{M}_k\times\mathcal{Z}_0)$ is a product measure formed by $P_{\mathbf{Z}_0}$ in (27) and $P^{(m)}_{\boldsymbol{\Theta}_k}$ in (55); $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\in\triangle(\mathcal{M}_k\times\mathcal{Z}_0)$ is the joint probability measure induced by the conditional measure $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}$ in (50) and $P_{\mathbf{Z}_0}$; and $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ is the $(\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k)$-Gibbs probability measure in (66).

The following theorem consolidates the hypothesis testing nature of the new information measure introduced in (92).

Theorem 18. *Consider a free parameter $\gamma>0$; the following constants*

$$\underline{\gamma}\triangleq\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{M}_k\times\mathcal{Z}_0}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right)\ \textit{and} \quad (93)$$

$$\bar{\gamma}\triangleq\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{M}_k\times\mathcal{Z}_0}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right); \quad (94)$$

and the set $\mathcal{A}_\gamma=\left\{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{M}_k\times\mathcal{Z}_0:\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right)\geqslant\gamma\right\}$, *where the probability measures* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}$ *is defined in* (50) *and* $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}$ *is the* $(\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k)$*-Gibbs conditional probability measure defined in* (66).

Then, the value $J_{\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (92), *satisfies*

$$(\bar{\gamma}-\gamma)\,P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}(\mathcal{A}_\gamma)+\left(\gamma-\underline{\gamma}\right)P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)$$

$$\geqslant J_{\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) \geqslant$$
$$\left(\underline{\gamma} - \gamma\right) P^{(m)}_{\boldsymbol{\Theta}_k} P_{\mathbf{Z}_0}\left(\mathcal{A}^{\mathsf{c}}_{\gamma}\right) + \left(\gamma - \bar{\gamma}\right) P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0} P_{\mathbf{Z}_0}\left(\mathcal{A}_{\gamma}\right), \tag{95}$$

where $P^{(m)}_{\boldsymbol{\Theta}_k} P_{\mathbf{Z}_0} \in \triangle\left(\mathcal{M}_k \times \mathcal{Z}_0\right)$ *is a product measure formed by* $P_{\mathbf{Z}_0}$ *in* (27) *and* $P^{(m)}_{\boldsymbol{\Theta}_k}$ *in* (55)*; and* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0} P_{\mathbf{Z}_0} \in \triangle\left(\mathcal{M}_k \times \mathcal{Z}_0\right)$ *is the joint probability measure induced by the conditional measure* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}$ *in* (50) *and* $P_{\mathbf{Z}_0}$.

Proof: The proof is presented in Appendix J. ∎

Using mutual information, lautum information and this new information measure, the first data-dependent closed-form expression of the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ in (62) is introduced hereunder.

Theorem 19. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that for all* $\mathbf{z}_0 \in \mathcal{Z}_0$*, the measures* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}$ *in* (50)*;* $P^{(m)}_{\boldsymbol{\Theta}_k}$ *in* (55)*; and* $Q_{\boldsymbol{\Theta}_k}$ *in* (66)*, satisfy that* $P^{(m)}_{\boldsymbol{\Theta}_k} \ll Q_{\boldsymbol{\Theta}_k} \ll P^{(m)}_{\boldsymbol{\Theta}_k} \ll P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0} \ll Q_{\boldsymbol{\Theta}_k}$. *Then,*

$$\begin{aligned}&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)\\ &= \lambda_k\Big(I\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) + L\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) + J_{\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)\Big),\end{aligned} \tag{96}$$

where $J_{\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ *is defined in* (92).

Proof: The proof is presented in Appendix K. ∎

In most practical settings, the absolute-continuity assumptions in Theorem 19 are satisfied; exceptions typically arise only in specialized academic constructions [62]. Together, Theorem 18 and Theorem 19 lead to a new bound on the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ in (62), which beyond its tightness, reveals the entanglement between the analysis of the generalization error and the study of mismatched hypothesis tests.

6.2 The Second Closed-Form Expression

The second data-dependent exact expression for the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ in (62) involves exclusively relative entropies.

Theorem 20. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that for all* $\mathbf{z}_0 \in \mathcal{Z}_0$*,* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0} \ll Q_{\boldsymbol{\Theta}_k}$*, with* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}$ *in* (50) *and* $Q_{\boldsymbol{\Theta}_k}$ *in* (66)*. Then,*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) = \lambda_k \iint \left(D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k}, \lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{\mathbf{z}}_k}\right)\right.$$

$$-D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=z_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)\Big)\mathrm{d}P_{\boldsymbol{Z}_k}(\widehat{z}_k)\mathrm{d}P_{\boldsymbol{Z}_0}(z_0), \text{ (97)}$$

where $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k} \in \triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$ *is the* $(\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k)$*-Gibbs conditional probability measure defined in* (66)*.*

Proof: The proof is presented in Appendix L. ■

6.3 A Pythagorean Identity of Generalization Error

The following theorem, together with the converse of the Pythagorean theorem [63, Book I, Proposition 48] leads to a Pythagorean identity, shown in Figure 3, which is reminiscent of the information projections studied in [35,64,65].

Theorem 21. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}; P_{\boldsymbol{Z}_0}\right)$ *in* (62)*. Assume that the conditional measure* $P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0}$ *in* (50)*; the measure* $P^{(m)}_{\boldsymbol{\Theta}_k}$ *in* (55)*; and* σ*-finite measure* $Q_{\boldsymbol{\Theta}_k}$ *in* (66)*, satisfy for all* $z_0 \in \mathcal{Z}_0$*,* $P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=z_0} \ll P^{(m)}_{\boldsymbol{\Theta}_k} \ll Q_{\boldsymbol{\Theta}_k}$*. Then,*

$$\begin{aligned}
&\frac{1}{\lambda_k}\,\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}; P_{\boldsymbol{Z}_0}\right) + \int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=z_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)\mathrm{d}P_{\boldsymbol{Z}_0}(z_0)\\
&= \int D\left(P^{(m)}_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{z}_k}\right)\mathrm{d}P_{\boldsymbol{Z}_k}(\widehat{z}_k) + \int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=z_0} \,\middle\|\, P^{(m)}_{\boldsymbol{\Theta}_k}\right)\mathrm{d}P_{\boldsymbol{Z}_0}(z_0), \text{ (98)}
\end{aligned}$$

where the measure $P_{\boldsymbol{Z}_k}$ *is defined in* (32) *and* $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}$ *is the* $(\mathsf{L}_k, Q_{\boldsymbol{\Theta}_k}, \lambda_k)$*-Gibbs probability measure in* (66)*.*

Proof: The proof is presented in Appendix M. ■

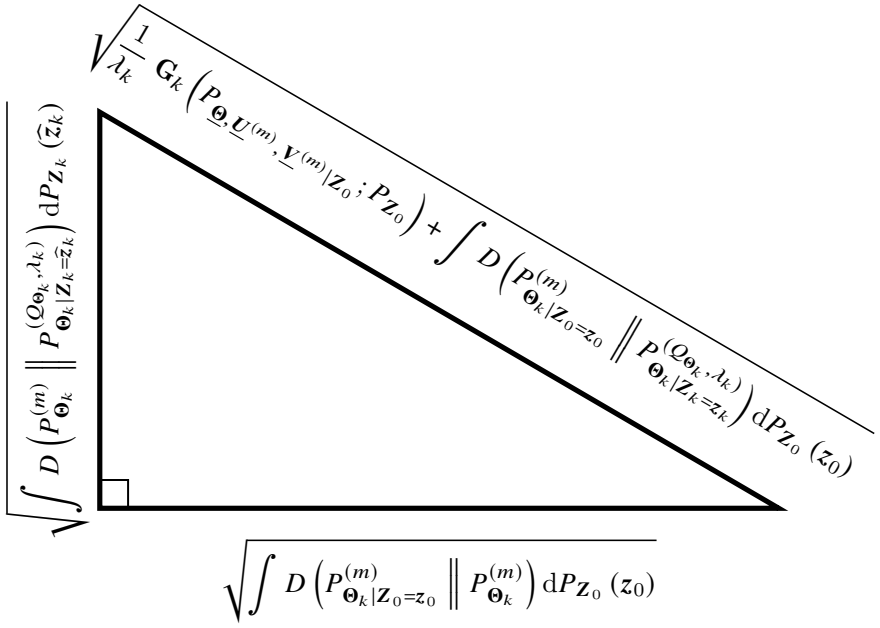


Figure 3: Geometric interpretation of Theorem 21

7 Model-dependent Closed-Form Expressions

7.1 The First Closed-Form Expression

The first closed-form expression of $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62) involves a mutual information, a lautum information and a new information measure parametrized by the same parameters of the $(\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k)$-Gibbs probability measure $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ in (75). Such a new information measure is denoted by $J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}$ and satisfies

$$J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right)=\int\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)\mathrm{d}P_{\mathbf{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(z_k,\boldsymbol{\theta}_k)$$
$$-\int\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(z_k,\boldsymbol{\theta}_k)\,,\quad(99)$$

where $P_{\mathbf{Z}_k}$ is the marginal probability measure in $\triangle\left(\mathcal{Z}_k^{n_k}\right)$ of $P_{\mathbf{Z}_0}$ in (27), which satisfies

$$P_{\mathbf{Z}_0}(\mathcal{A}) = \int P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(\mathcal{A})\, \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k), \tag{100}$$

with $P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ defined in (54); $P_{\mathbf{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\times\mathcal{M}_k\right)$ is a product measure formed by $P_{\mathbf{Z}_k}$ and $P^{(m)}_{\boldsymbol{\Theta}_k}$ in (55); and $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\times\mathcal{M}_k\right)$ is the joint probability measure induced by the conditional measure $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}$ in (56) and $P^{(m)}_{\boldsymbol{\Theta}_k}$; and $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ is the $(\mathsf{L}_k, Q_{\mathbf{Z}_k}, \alpha_k)$-Gibbs probability measure in (75). Note that the equality in (100) is a consequence of the equality in (54).

Using this notation the following Theorem is introduced and follows along the same lines as Theorem (18).

Theorem 22. *Consider a free parameter $\gamma > 0$; the following constants*

$$\underline{\gamma} \triangleq \min_{(\mathbf{z}_k,\boldsymbol{\theta}_k)\in\mathcal{Z}_k^{n_k}\times\mathcal{M}_k} \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}\right) \text{ and} \tag{101}$$

$$\bar{\gamma} \triangleq \max_{(\mathbf{z}_k,\boldsymbol{\theta}_k)\in\mathcal{Z}_k^{n_k}\times\mathcal{M}_k} \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}\right); \tag{102}$$

and the set $\mathcal{A}_\gamma = \left\{(\mathbf{z}_k,\boldsymbol{\theta}_k)\in\mathcal{Z}_k^{n_k}\times\mathcal{M}_k : \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}\right) \geqslant \gamma\right\}$, *where the probability measure $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ defined in (56) and $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ is the $(\mathsf{L}_k, Q_{\mathbf{Z}_k}, \alpha_k)$-Gibbs conditional probability measure defined in (75). Then, the value $J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}; P^{(m)}_{\boldsymbol{\Theta}_k}\right)$ in (99), satisfies*

$$\begin{aligned}(\bar{\gamma}-\gamma)\,P_{\mathbf{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(\mathcal{A}_\gamma) + \left(\gamma-\underline{\gamma}\right)P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)\\ \geqslant J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}; P^{(m)}_{\boldsymbol{\Theta}_k}\right) \geqslant\\ \left(\underline{\gamma}-\gamma\right)P_{\mathbf{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right) + (\gamma-\bar{\gamma})\,P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(\mathcal{A}_\gamma),\end{aligned} \tag{103}$$

where $P_{\mathbf{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\times\mathcal{M}_k\right)$ is a product measure formed by $P_{\mathbf{Z}_k}$ in (32) and $P^{(m)}_{\boldsymbol{\Theta}_k}$ in (55); and $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}\times\mathcal{M}_k\right)$ is the joint probability measure induced by the conditional measure $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}$ in (56) and $P^{(m)}_{\boldsymbol{\Theta}_k}$.

Proof: The proof is presented in Appendix N. ∎

Using mutual information, lautum information and this new information measure, the first model-dependent closed-form expression of the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ in (62) is introduced below.

Theorem 23. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that for all* $\boldsymbol{\theta}_k \in \mathcal{M}_k$*, the measures* $P_{\mathbf{Z}_k}$ *in* (32)*;* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *in* (56)*; and* $Q_{\mathbf{Z}_k}$ *in* (75)*; satisfy that* $P_{\mathbf{Z}_k} \ll Q_{\mathbf{Z}_k} \ll P_{\mathbf{Z}_k} \ll P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \ll Q_{\mathbf{Z}_k}$*. Then,*

$$\begin{aligned}&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)\\&=\alpha_k\Big(I\Big(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\Big)+L\Big(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\Big)+J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\Big(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\Big)\Big),\end{aligned}\tag{104}$$

where $J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right)$ *is defined in* (99)*.*

Proof: The proof is presented in Appendix O. ∎

Together, Theorem 22 and Theorem 23 lead to a new bound on the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62).

7.2 The Second Closed-Form Expression

The second model-dependent exact expression for the generalization error $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (62) involves exclusively relative entropies.

Theorem 24. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that, for all* $\boldsymbol{\theta}_k \in \mathcal{M}_k$*, the measures* $P_{\mathbf{Z}_k}$ *in* (32)*;* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *in* (56)*; and* $Q_{\mathbf{Z}_k}$ *in* (75)*; satisfy that* $P_{\mathbf{Z}_k} \ll Q_{\mathbf{Z}_k}$ *and* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \ll Q_{\mathbf{Z}_k}$*. Then,*

$$\begin{aligned}&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\alpha_k\iint D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\,\middle\|\,P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right)\\&-D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\,\middle\|\,P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k),\end{aligned}\tag{105}$$

where the probability measures $P^{(m)}_{\boldsymbol{\Theta}_k}$ *is defined in* (55) *and* $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right)$ *is an* $(\mathsf{L}_k, Q_{\mathbf{Z}_k}, \alpha_k)$*-Gibbs conditional probability measure defined in* (75)*.*

Proof: The proof is presented in Appendix P. ∎

7.3 A Pythagorean Identity of Generalization Error

The following theorem, together with the converse of the Pythagorean theorem [63, Book I, Proposition 48] leads to a Pythagorean identity, shown in Figure 4.

Theorem 25. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that, for all* $\boldsymbol{\theta}_k \in \mathcal{M}_k$*, the measures* $P_{\mathbf{Z}_k}$ *in* (32) *;* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *in* (56)*; and* $Q_{\mathbf{Z}_k}$ *in* (75)*; satisfy that* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \ll P_{\mathbf{Z}_k} \ll Q_{\mathbf{Z}_k}$*. Then,*

$$\begin{aligned}&\frac{1}{\alpha_k}\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)+\int D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\,\middle\|\,P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\\ &=\int D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\,\middle\|\,P_{\mathbf{Z}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)+\int D\left(P_{\mathbf{Z}_k}\,\middle\|\,P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right),\end{aligned}\quad (106)$$

where the measure $P^{(m)}_{\boldsymbol{\Theta}_k}$ *is defined in* (55) *and* $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *is the* $(\mathsf{L}_k, Q_{\mathbf{Z}_k}, \alpha_k)$*-Gibbs probability measure in* (75)*.*

Proof: The proof is presented in Appendix Q ■

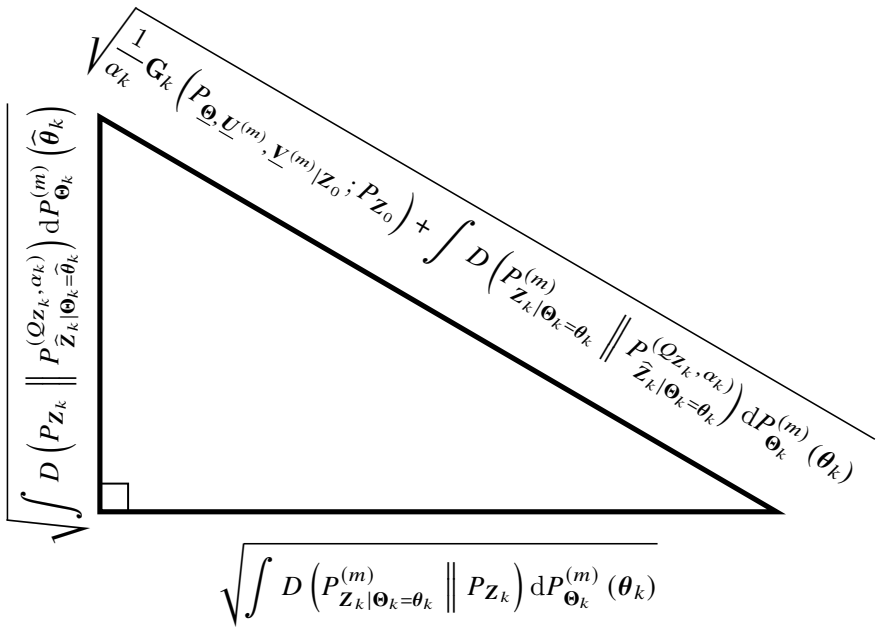


Figure 4: Geometric interpretation of Theorem 25

References

[1] 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. PMLR, Apr. 2017, pp. 1273–1282.

[2] 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.

[3] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, B. Agüera y Arcas, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, S. Diggavi, H. Eichner, A. Gadhikar, Z. Garrett, A. M. Girgis, F. Hanzely, A. Hard, C. He, S. Horvath, Z. Huo, A. Ingerman, M. Jaggi, T. Javidi, P. Kairouz, S. Kale, S. P. Karimireddy, J. Konečný, S. Koyejo, T. Li, L. Liu, M. Mohri, H. Qi, S. J. Reddi, P. Richtárik, K. Singhal, V. Smith, M. Soltanolkotabi, W. Song, A. T. Suresh, S. U. Stich, A. Talwalkar, H. Wang, B. Woodworth, S. Wu, F. X. Yu, H. Yuan, M. Zaheer, M. Zhang, T. Zhang, C. Zheng, C. Zhu, and W. Zhu, *A Field Guide to Federated Optimization*, 1st ed. Ithaca, NY, USA: arXiv, 2021.

[4] M. Mohri, G. Sivek, and A. T. Suresh, "Agnostic federated learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 4615–4625.

[5] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine Learning and Systems*, vol. 2, pp. 429–450, Mar. 2020.

[6] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2020.

[7] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local SGD on identical and heterogeneous data," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 108. PMLR, 2020.

[8] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv*, Oct. 2016.

[9] S. U. Stich, "Local sgd converges fast and communicates little," *International Conference on Learning Representations (ICLR)*, May 2019.

[10] E. Diao, J. Ding, and V. Tarokh, "HeteroFL: Computation and communication efficient federated learning for heterogeneous clients," in *International Conference on Learning Representations (ICLR)*, 2021.

[11] V. Vapnik, *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[12] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," *Journal of the ACM (JACM)*, vol. 44, no. 4, pp. 615–631, 1997.

[13] M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-iid processes," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 21, 2008.

[14] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. New York, NY, USA: Cambridge University Press, 2014.

[15] V. Cherkassky, X. Shao, F. M. Mulier, and V. N. Vapnik, "Model complexity control for regression using VC generalization bounds," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 1075–1089, Sep. 1999.

[16] B. Wei, J. Li, Y. Liu, and W. Wang, "Non-iid federated learning with sharper risk bound," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 6906–6917, 2024.

[17] X. Hu, S. Li, and Y. Liu, "Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses," in *International Conference on Learning Representations (ICLR)*, 2023.

[18] O. Bousquet and A. Elisseeff, "Stability and generalization," *The Journal of Machine Learning Research*, vol. 2, no. 1, pp. 499–526, Mar. 2002.

[19] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," in *Proceedings of the 33rd International Conference on Machine Learning*, vol. 48, 2016, pp. 1225–1234.

[20] 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.

[21] G. K. Dziugaite and D. M. Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," in *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

[22] 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, Dec. 2017, pp. 1–10.

[23] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, Jan. 2020.

[24] J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, "Information-theoretic generalization bounds for SGLD via data-dependent estimates," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 11 013–11 023.

[25] 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.

[26] 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, Dec. 2021, pp. 8106–8118.

[27] 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, 2024.

[28] 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*, Vancouver, Canada, Feb. 2024.

[29] 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.

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

[31] P. Boroumand and A. G. i Fàbregas, "Mismatched binary hypothesis testing: Error exponent sensitivity," *IEEE Transactions on Information Theory*, vol. 68, no. 10, pp. 6738–6761, 2022.

[32] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, Jul. 1948.

[33] ——, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 623–656, Oct. 1948.

[34] D. P. Palomar and S. Verdú, "Lautum information," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 964–975, Mar. 2008.

[35] 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.

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

[37] S. Yagli, A. Dytso, and H. V. Poor, "Information-theoretic bounds on the generalization error and privacy leakage in federated learning," in *Proceedings of the International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, May 2020, pp. 1–5.

[38] R. Pathak and M. J. Wainwright, "Fedsplit: An algorithmic framework for fast federated optimization," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[39] M. Sefidgaran, R. Chor, A. Zaidi, and Y. Wan, "Lessons from generalization error analysis of federated learning: You may communicate less often!" in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 235. PMLR, Jul. 2024, pp. 44 093–44 135.

[40] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. Song, "A principled approach to data valuation for federated learning," in *Federated Learning: Privacy and Incentive*, ser. Lecture Notes in Computer Science, Q. Yang, L. Fan, and H. Yu, Eds. Springer, 2020, vol. 12500, pp. 153–167.

[41] J. Wang, V. Tantia, N. Ballas, and M. Rabbat, "Slowmo: Improving communication-efficient distributed SGD with slow momentum," in *International Conference on Learning Representations (ICLR)*, 2020.

[42] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 97. PMLR, 2019.

[43] S. Horváth and P. Richtárik, "A better alternative to error feedback for communication-efficient distributed learning," *arXiv*, Jun. 2020.

[44] Z. Sun, X. Niu, and E. Wei, "Understanding generalization of federated learning via stability: Heterogeneity matters," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 238. PMLR, May 2024, pp. 676–684.

[45] L. P. Barnes, A. Dytso, and H. V. Poor, "Improved information theoretic generalization bounds for distributed and federated learning," *arXiv*, Feb. 2022.

[46] A. Ramezani-Kebrya, F. Liu, T. Pethick, G. Chrysos, and V. Cevher, "Federated learning under covariate shifts with generalization guarantees," *Transactions on Machine Learning Research*, Jun. 2023.

[47] R. Zhang, Q. Xu, J. Yao, Y. Zhang, Q. Tian, and Y. Wang, "Federated domain generalization with generalization adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2023, pp. 3954–3963.

[48] M. Sefidgaran, R. Chor, and A. Zaidi, "Rate-distortion theoretic bounds on generalization error for distributed learning," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[49] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, "Federated meta-learning with fast convergence and efficient communication," *arXiv*, Feb. 2019.

[50] H. Zhang, C. Li, N. Kan, Z. Zheng, W. Dai, J. Zou, and H. Xiong, "Improving generalization in federated learning with model-data mutual information regularization: A posterior inference approach," in *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[51] H. Yuan, W. Morningstar, L. Ning, and K. Singhal, "What do we mean by generalization in federated learning?" *arXiv*, Oct. 2022.

[52] Z. Wang, C. Long, and Y. Mao, "Generalization in federated learning: A conditional mutual information framework," *arXiv*, Mar. 2025.

[53] W. Liu, G. Yu, L. Wang, and R. Liao, "An information-theoretic framework for out-of-distribution generalization with applications to stochastic gradient Langevin dynamics," *IEEE Transactions on Information Theory*, vol. 71, no. 11, pp. 8798–8817, 2025.

[54] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," *International Conference on Learning Representations*, May 2021.

[55] B. Ying, Z. Li, and H. Yang, "Exact and linear convergence for federated learning under arbitrary client participation is attainable," Jun. 2025.

[56] Y. J. Cho, P. Sharma, G. Joshi, Z. Xu, S. Kale, and T. Zhang, "On the convergence of federated averaging with cyclic client participation," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5677–5721.

[57] 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 *International Symposium on Information Theory (ISIT)*, June 2023, pp. 328–333.

[58] 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.

[59] 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 *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2024, pp. 759–764.

[60] ——, "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.

[61] 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*, July 2024, pp. 2168 – 2229.

[62] Y. Polyanskiy and Y. Wu, *Information Theory*, 1st ed. Cambridge, UK: Cambridge University Press, 2023.

[63] Euclid, *The Thirteen Books of Euclid's Elements*, 2nd ed., T. L. Heath, Ed. Dover Publications, Inc., 1956.

[64] N. N. Chentsov, "Nonsymmetrical distance between probability distributions, entropy and the theorem of pythagoras," *Mathematical notes of the Academy of Sciences of the USSR*, vol. 4, no. 3, pp. 686–691, 1968.

[65] I. Csiszár and F. Matus, "Information projections revisited," *IEEE Transactions on Information Theory*, vol. 49, no. 6, pp. 1474–1490, 2003.

[66] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Biometrika*, vol. 45, no. 3-4, pp. 296–315, 1958.

[67] R. B. Ash and C. A. Doleans-Dade, *Probability and Measure Theory*, 2nd ed. Burlington, MA, USA: Academic Press, 2000.

A Proof of Lemma 1

From Bayes' theorem [66], the joint probability measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)},\boldsymbol{Z}_0}$ in (24) can be factorized as,

$$P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0} = P_{\underline{\boldsymbol{\Theta}}|\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)},\boldsymbol{Z}_0} P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0} \tag{107a}$$

with,

$$P_{\underline{\boldsymbol{\Theta}}|\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)},\boldsymbol{Z}_0} \in \triangle\left(\mathcal{M}_0|\mathcal{U}_0^m \times \mathcal{V}_0^m \times \mathcal{Z}_0\right), \text{ and} \tag{107b}$$

$$P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0} \in \triangle\left(\mathcal{U}_0^m \times \mathcal{V}_0^m \mid \mathcal{Z}_0\right). \tag{107c}$$

Furthermore, under Assumption 1, the conditional probability measure $P_{\underline{\boldsymbol{\Theta}}|\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)},\boldsymbol{Z}_0}$ in (107b) can be factorized as

$$P_{\underline{\boldsymbol{\Theta}}|\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)},\boldsymbol{Z}_0} = \prod_{k=1}^{K} P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k} \tag{108a}$$

with, $P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)},\boldsymbol{V}_k^{(m)},\boldsymbol{Z}_k}$ defined in (28).

Furthermore, from Bayes' theorem [66], the joint probability measure $P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}$ in (107c) can be rewritten as,

$$P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0} = \prod_{t=1}^{m} P_{\boldsymbol{U}^{(t)},\boldsymbol{V}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t-1)},\boldsymbol{Z}_0} \tag{109}$$

$$= \prod_{t=1}^{m} P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)},\boldsymbol{Z}_0} P_{\boldsymbol{V}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t-1)},\boldsymbol{Z}_0} \tag{110}$$

$$= \prod_{t=1}^{m} \left(P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}} \prod_{k=1}^{K} P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k} \right). \tag{111}$$

In the case in which $t = 1$, the notations $P_{\boldsymbol{U}^{(1)},\boldsymbol{V}^{(1)}|\underline{\boldsymbol{U}}^{(0)},\underline{\boldsymbol{V}}^{(0)},\boldsymbol{Z}_0}$, $P_{\boldsymbol{U}^{(1)}|\underline{\boldsymbol{U}}^{(0)},\underline{\boldsymbol{V}}^{(1)},\boldsymbol{Z}_0}$, $P_{\boldsymbol{V}^{(1)}|\underline{\boldsymbol{U}}^{(0)},\underline{\boldsymbol{V}}^{(0)},\boldsymbol{Z}_0}$, $P_{\boldsymbol{U}^{(1)}|\underline{\boldsymbol{U}}^{(0)},\underline{\boldsymbol{V}}^{(1)}}$, and $P_{V_{k,1}|\boldsymbol{U}_k^{(0)},\boldsymbol{V}_k^{(0)},\boldsymbol{Z}_k}$ are longhands for the measures

$$P_{\boldsymbol{U}^{(1)},\boldsymbol{V}^{(1)}|\boldsymbol{Z}_0} \in \triangle\left(\mathcal{U}_0 \times \mathcal{V}_0|\mathcal{Z}_0\right) \tag{112a}$$

$$P_{\boldsymbol{U}^{(1)}|\underline{\boldsymbol{V}}^{(1)},\boldsymbol{Z}_0} \in \triangle\left(\mathcal{U}_0|\mathcal{V}_0 \times \mathcal{Z}_0\right), \tag{112b}$$

$$P_{\boldsymbol{V}^{(1)}|\boldsymbol{Z}_0} \in \triangle\left(\mathcal{V}_0|\mathcal{Z}_0\right), \tag{112c}$$

$$P_{\boldsymbol{U}^{(1)}|\underline{\boldsymbol{V}}^{(1)}} \in \triangle\left(\mathcal{U}_0|\mathcal{V}_0\right) \text{ and} \tag{112d}$$

$$P_{V_{k,1}|\boldsymbol{Z}_k} \in \triangle\left(\mathcal{V}_k|\mathcal{Z}_k^{n_k}\right), \tag{112e}$$

respectively, and for all $t \in \{2, 3, \ldots, m\}$,

$$P_{\boldsymbol{U}^{(t)},\boldsymbol{V}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t-1)},\boldsymbol{Z}_0} \in \triangle\left(\mathcal{U}_0 \times \mathcal{V}_0|\mathcal{U}_0^{t-1} \times \mathcal{V}_0^{t-1} \times \mathcal{Z}_0\right), \tag{112f}$$

$$P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)},\boldsymbol{Z}_0} \in \triangle\left(\mathcal{U}_0|\mathcal{U}_0^{t-1}\times\mathcal{V}_0^{t}\times\mathcal{Z}_0\right), \tag{112g}$$

$$P_{\boldsymbol{V}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t-1)},\boldsymbol{Z}_0} \in \triangle\left(\mathcal{V}_0|\mathcal{U}_0^{t-1}\times\mathcal{V}_0^{t-1}\times\mathcal{Z}_0\right). \tag{112h}$$

The conditional probability measure $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)},\boldsymbol{Z}_0}$ in (110) simplifies to $P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}}$ in (30) under Assumption 3. This is because the server does not have access to the training datasets $\boldsymbol{z}_0$ in (2). Similarly, under Assumption 2, the joint conditional probability $P_{\boldsymbol{V}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t-1)},\boldsymbol{Z}_0}$ in (112h) satisfies

$$P_{\boldsymbol{V}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)},\underline{\boldsymbol{V}}^{(t-1)},\boldsymbol{Z}_0} = \prod_{k=1}^{K} P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}, \tag{113}$$

with the conditional probability measure $P_{V_{k,t}|\boldsymbol{U}_k^{(t-1)},\boldsymbol{V}_k^{(t-1)},\boldsymbol{Z}_k}$ defined in (29). This completes the proof.

B Proof of Lemma 2

Consider the measurable set $\mathcal{A}\subseteq\mathcal{M}_k$, for some $k\in\{1,2,\ldots,K\}$, and define the set $\mathcal{A}\times\mathcal{M}_{-k}\subseteq\mathcal{M}_0$ as follows:

$$\mathcal{A}\times\mathcal{M}_{-k} \triangleq \mathcal{M}_1\times\cdots\times\mathcal{M}_{k-1}\times\mathcal{A}\times\mathcal{M}_{k+1}\times\cdots\times\mathcal{M}_K. \tag{114}$$

Then, for all $\underline{\boldsymbol{u}}^{(m)}\in\mathcal{U}_0^m$, for all $\underline{\boldsymbol{v}}^{(m)}\in\mathcal{V}_0^m$ and for all $\boldsymbol{z}_0\in\mathcal{Z}_0$

$$P_{\underline{\boldsymbol{\Theta}}|\underline{\boldsymbol{U}}^{(m)}=\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{V}}^{(m)}=\underline{\boldsymbol{v}}^{(m)},\boldsymbol{Z}_0=\boldsymbol{z}_0}(\mathcal{A}\times\mathcal{M}_{-k}) = \prod_{t=1}^{k-1}\left(P_{\boldsymbol{\Theta}_t|\boldsymbol{U}_t^{(m)}=\boldsymbol{u}_t^{(m)},\boldsymbol{V}_t^{(m)}=\boldsymbol{v}_t^{(m)},\boldsymbol{Z}_t=\boldsymbol{z}_t}(\mathcal{M}_t)\right)$$

$$P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\boldsymbol{Z}_k=\boldsymbol{z}_k}(\mathcal{A}) \prod_{t=k+1}^{K}\left(P_{\boldsymbol{\Theta}_t|\boldsymbol{U}_t^{(m)}=\boldsymbol{u}_t^{(m)},\boldsymbol{V}_t^{(m)}=\boldsymbol{v}_t^{(m)},\boldsymbol{Z}_t=\boldsymbol{z}_t}(\mathcal{M}_t)\right) \tag{115}$$

$$= P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\boldsymbol{Z}_k=\boldsymbol{z}_k}(\mathcal{A}), \tag{116}$$

where the equality in (115) follows from Assumption 1. Moreover, for all $i\in\{1,\ldots,K\}$, the conditional measure $P_{\boldsymbol{\Theta}_i|\boldsymbol{U}_i^{(m)},\boldsymbol{V}_i^{(m)},\boldsymbol{Z}_i}$ in (31) satisfies, for all $\boldsymbol{u}_i^{(m)}\in\mathcal{U}_i^m$, for all $\boldsymbol{v}_i^{(m)}\in\mathcal{V}_i^m$ and for all $\boldsymbol{z}_i\in\mathcal{Z}_i^{n_i}$,

$$P_{\boldsymbol{\Theta}_i|\boldsymbol{U}_i^{(m)}=\boldsymbol{u}_i^{(m)},\boldsymbol{V}_i^{(m)}=\boldsymbol{v}_i^{(m)},\boldsymbol{Z}_i=\boldsymbol{z}_i}(\mathcal{M}_i) = 1, \tag{117}$$

which justifies (116).

On the other hand, for all $\boldsymbol{z}_0\in\mathcal{Z}_0$, it follows that

$$\int P_{\underline{\boldsymbol{\Theta}}|\underline{\boldsymbol{U}}^{(m)}=\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{V}}^{(m)}=\underline{\boldsymbol{v}}^{(m)},\boldsymbol{Z}_0=\boldsymbol{z}_0}(\mathcal{A}\times\mathcal{M}_{-k})\,\mathrm{d}P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0=\boldsymbol{z}_0}\left(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}\right)$$

$$= \int P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\boldsymbol{Z}_k=\boldsymbol{z}_k}(\mathcal{A})\,\mathrm{d}P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0=\boldsymbol{z}_0}\left(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}\right) \tag{118}$$

$$= P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}(\mathcal{A}), \tag{119}$$

where the equality in (118) follows from (116); and the equality in (119) represents the marginal probability measure in $\triangle(\mathcal{M}_k \mid \mathcal{Z}_0)$ of the probability measure $P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}$ in (26).

Finally, given that the sets $\mathcal{U}_0$ and $\mathcal{V}_0$ are finite, the integral in (118) reduces to a finite sum,

$$\begin{aligned} & P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}(\mathcal{A}) \\ &= \int P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\boldsymbol{Z}_k=\boldsymbol{z}_k}(\mathcal{A})\, \mathrm{d}P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0=\boldsymbol{z}_0}\left(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}\right) \quad (120) \\ &= \sum_{\left(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}\right)\in\mathcal{U}_0^m\times\mathcal{V}_0^m} a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\boldsymbol{z}_0} P_{\boldsymbol{\Theta}_k|\boldsymbol{U}_k^{(m)}=\boldsymbol{u}_k^{(m)},\boldsymbol{V}_k^{(m)}=\boldsymbol{v}_k^{(m)},\boldsymbol{Z}_k=\boldsymbol{z}_k}(\mathcal{A}), \quad (121) \end{aligned}$$

where the coefficients $a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\boldsymbol{z}_0}$ satisfy

$$\begin{aligned} & a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\boldsymbol{z}_0} \triangleq P_{\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0=\boldsymbol{z}_0}\left(\left\{\left(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}\right)\right\}\right) \quad (122) \\ =& \prod_{t=1}^{m} P_{\boldsymbol{U}^{(t)}|\underline{\boldsymbol{U}}^{(t-1)}=\underline{\boldsymbol{u}}^{(t-1)},\underline{\boldsymbol{V}}^{(t)}=\underline{\boldsymbol{v}}^{(t)}}\left(\{\boldsymbol{u}^{(t)}\}\right) \\ & \prod_{j=1}^{K} P_{V_{j,t}|\boldsymbol{U}_j^{(t-1)}=\boldsymbol{u}_j^{(t-1)},\boldsymbol{V}_j^{(t-1)}=\boldsymbol{v}_j^{(t-1)},\boldsymbol{Z}_j=\boldsymbol{z}_j}\left(\{v_{j,t}\}\right), \quad (123) \end{aligned}$$

where (123) follows from (111). Note that from (122), it follows that $a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\boldsymbol{z}_0} \geqslant 0$; and

$$\sum_{\left(\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)}\right)\in\mathcal{U}_0^m\times\mathcal{V}_0^m} a_{\underline{\boldsymbol{u}}^{(m)},\underline{\boldsymbol{v}}^{(m)},\boldsymbol{z}_0}=1, \tag{124}$$

which completes the proof.

C Proof of Lemma 3

The proof of (57) is primarily algebraic and starts as follows

$$\begin{aligned} & \int \mathsf{R}_{k,\boldsymbol{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}\right) \mathrm{d}P_{\boldsymbol{Z}_0}(\boldsymbol{z}_0) \\ &= \iint \mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\, \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\boldsymbol{Z}_0}(\boldsymbol{z}_0) \quad (125) \\ &= \iint \mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k) \frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}} \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\boldsymbol{Z}_0}(\boldsymbol{z}_0) \quad (126) \\ &= \iint \mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k) \frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_0)}{\mathrm{d}P_{\boldsymbol{Z}_0}} \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\boldsymbol{Z}_0}(\boldsymbol{z}_0) \quad (127) \end{aligned}$$

$$= \iint \mathsf{L}_k\left(z_k, \boldsymbol{\theta}_k\right) \frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_0}}\left(z_0\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \tag{128}$$

$$= \iint \mathsf{L}_k\left(z_k, \boldsymbol{\theta}_k\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_0|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\left(z_0\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \tag{129}$$

$$= \iint \mathsf{L}_k\left(z_k, \boldsymbol{\theta}_k\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\left(z_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \tag{130}$$

$$= \int \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right), \tag{131}$$

where (126) follows from [58, Theorem 2]; (127) follows from [58, Theorem 11]; (128) follows from follows from [67, Theorem 2.6.6]; (129) follows from [58, Theorem 2]; (130) from the fact that the functional L_k depends exclusively on z_k (instead of z_0); and (131) follows from the definition of the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ in (53). This completes the proof.

D Proof of Lemma 4

The proof of Lemma 4 is primarily algebraic and proceeds as follows

$$\iint \mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right)$$

$$= \iiint \mathsf{L}_k\left(\widehat{z}_k, \boldsymbol{\theta}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right) \tag{132}$$

$$= \iiint \mathsf{L}_k\left(\widehat{z}_k, \boldsymbol{\theta}_k\right) \frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right) \tag{133}$$

$$= \iint \mathsf{L}_k\left(\widehat{z}_k, \boldsymbol{\theta}_k\right) \int \frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right) \tag{134}$$

$$= \iint \mathsf{L}_k\left(\widehat{z}_k, \boldsymbol{\theta}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right) \tag{135}$$

$$= \int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k}\right) \mathrm{d}P_{\mathbf{Z}_k}\left(z_k\right) \tag{136}$$

$$= \int \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P_{\mathbf{Z}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right), \tag{137}$$

where (133) follows from a change of measure [58, Theorem 2]; (134) follows from [67, Theorem 2.6.6]; (135) follows from [58, Theorem 10]; and (136) follows from the definition of the functional R_{k,z_k} in (52) and concludes the proof of (58). The equality in (137) follows from [67, Theorem 2.6.6] and the definition of the functional $\mathsf{R}_{k,\boldsymbol{\theta}_k}$ in (53). This concludes the proof of (59). The proof continues from (136) as follows

$$\int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k}\right) \mathrm{d}P_{\mathbf{Z}_k}\left(z_k\right)$$

$$= \iint \mathsf{L}_k\left(\widehat{z}_k, \boldsymbol{\theta}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right) \tag{138}$$

$$= \iint \mathsf{L}_k\left(z_k, \widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \int \frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}\left(z_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P_{\mathbf{Z}_k}\left(z_k\right) \tag{139}$$

$$= \iiint \mathsf{L}_k\left(z_k, \widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}\left(z_k\right) \mathrm{d}P_{\mathbf{Z}_k}\left(z_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \tag{140}$$

$$= \iiint \mathsf{L}_k\left(z_k, \widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\left(z_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \tag{141}$$

$$= \iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right), \tag{142}$$

where (139) follows from [58, Theorem 10]; (140) from [67, Theorem 2.6.6]; and (141) from [58, Theorem 2]. The equality in (142) concludes the proof of (60).

E Proof of Lemma 5

The proof follows from Definition 12, from which it holds that

$$\begin{aligned}
&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) \\
&= \iint \left(\mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) - \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)\right) \mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) \mathrm{d}P_{\mathbf{Z}_0}(z_0) \qquad (143)\\
&= \iint \mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) \mathrm{d}P_{\mathbf{Z}_0}(z_0) - \int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}(z_0). \qquad (144)
\end{aligned}$$

The proof focuses on the term $\iint \mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right)$ in (144). From Lemma 4, it follows that

$$\begin{aligned}
&\iint \mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_k}\left(\widehat{z}_k\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right) \\
&= \iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right). \qquad (145)
\end{aligned}$$

Consider now the term $\int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right)$ in (144). From Lemma 3, it follows that

$$\int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}\left(z_0\right) = \int \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right). \tag{146}$$

The proof is completed by using (145) and (146) in (144).

F Proof of Lemma 12

The proof of (86) is primarily algebraic and proceeds by analyzing the KL divergence $D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)$. More precisely,

$$D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)$$

$$= \int \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}(\boldsymbol{\theta})\right)\mathrm{d}P_1(\boldsymbol{\theta}) \tag{147}$$

$$= \int \log\left(\frac{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}(\boldsymbol{\theta}_k)\,\frac{\mathrm{d}P_1}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P_1(\boldsymbol{\theta}_k) \tag{148}$$

$$= \int \log\left(\left(\frac{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)^{-1}\frac{\mathrm{d}P_1}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P_1(\boldsymbol{\theta}_k) \tag{149}$$

$$= \int -\log\left(\frac{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)+\log\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P_1(\boldsymbol{\theta}_k) \tag{150}$$

$$= -\int \log\left(\frac{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P_1(\boldsymbol{\theta}_k)+D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right) \tag{151}$$

$$= -\int\left(-\frac{1}{\lambda_k}\mathsf{L}_k(z_k,\boldsymbol{\theta}_k)-\mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},z_k}\left(-\frac{1}{\lambda_k}\right)\right)\mathrm{d}P_1(\boldsymbol{\theta}_k)+D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right) \tag{152}$$

$$= \int \frac{1}{\lambda_k}\mathsf{L}_k(z_k,\boldsymbol{\theta}_k)\,\mathrm{d}P_1(\boldsymbol{\theta}_k)+\mathsf{K}_{k,Q_{\boldsymbol{\Theta}_k},z_k}\left(-\frac{1}{\lambda_k}\right)+D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right) \tag{153}$$

$$\begin{aligned}= &\int \frac{1}{\lambda_k}\mathsf{L}_k(z_k,\boldsymbol{\theta}_k)\,\mathrm{d}P_1(\boldsymbol{\theta}_k)-\frac{1}{\lambda_k}\int \mathsf{L}_k(z_k,\boldsymbol{\theta}_k)\,\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}(\boldsymbol{\theta}_k)\\ &-D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)+D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)\end{aligned} \tag{154}$$

$$= \frac{1}{\lambda_k}\left(\mathsf{R}_{k,z_k}(P_1)-\mathsf{R}_{k,z_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)\right)-D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)+D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right) \tag{155}$$

$$= \frac{1}{\lambda_k}\mathsf{G}_k\left(z_k,P_1,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)-D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)+D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right), \tag{156}$$

where equality (148) follows from [58, Theorem 4]; (149) from [58, Theorem 5]; (152) from (66); (154) from Lemma 7; and (156) from (82).

Hence, from the equality in (156), it follows that

$$\begin{aligned}&\mathsf{G}_k\left(z_k,P_1,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)\\ &=\lambda_k\left(D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)+D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)-D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)\right).\end{aligned} \tag{157}$$

The proof proceeds from the definition of the functional G_k in (82), from which it holds that

$$\begin{aligned}&\mathsf{G}_k(z_k,P_1,P_2)\\ &=\mathsf{R}_{k,z_k}(P_1)-\mathsf{R}_{k,z_k}(P_2)+\mathsf{R}_{k,z_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)-\mathsf{R}_{k,z_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)\end{aligned} \tag{158}$$

$$= \mathsf{G}_k\left(z_k,P_1,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)-\mathsf{G}_k\left(z_k,P_2,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right) \tag{159}$$

$$= \lambda_k\left(D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k}\right)+D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=z_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)-D\left(P_1 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)\right)$$

$$-\lambda_k \left(D\left(P_2 \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right) + D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right) - D\left(P_2 \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)\right) \tag{160}$$

$$= \lambda_k \Big(D\Big(P_1 \,\Big\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\Big) - D\Big(P_2 \,\Big\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\Big) + D\big(P_2 \,\big\|\, Q_{\boldsymbol{\Theta}_k}\big) - D\big(P_1 \,\big\|\, Q_{\boldsymbol{\Theta}_k}\big)\Big), \tag{161}$$

where the equality in (159) follows from the definition of the functional G_k in (82); and the equality in (160) follows from (157). This completes the proof.

G Proof of Lemma 13

The proof follows along the same lines as the proof of Lemma 12. It is included entirely hereunder for the sake of completeness.

The proof of (87) is primarily algebraic and proceeds by analyzing the KL divergence $D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)$. More precisely,

$$D\left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)$$

$$= \int \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P_1(\boldsymbol{z}_k) \tag{162}$$

$$= \int \log\left(\frac{\mathrm{d}Q_{\boldsymbol{Z}_k}}{\mathrm{d}P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\,\frac{\mathrm{d}P_1}{\mathrm{d}Q_{\boldsymbol{Z}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P_1(\boldsymbol{z}_k) \tag{163}$$

$$= \int \log\left(\left(\frac{\mathrm{d}P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\boldsymbol{Z}_k}}(\boldsymbol{z}_k)\right)^{-1}\frac{\mathrm{d}P_1}{\mathrm{d}Q_{\boldsymbol{Z}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P_1(\boldsymbol{z}_k) \tag{164}$$

$$= \int -\log\left(\frac{\mathrm{d}P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\boldsymbol{Z}_k}}(\boldsymbol{z}_k)\right) + \log\left(\frac{\mathrm{d}P_1}{\mathrm{d}Q_{\boldsymbol{Z}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P_1(\boldsymbol{z}_k) \tag{165}$$

$$= -\int \log\left(\frac{\mathrm{d}P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\boldsymbol{Z}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P_1(\boldsymbol{z}_k) + D\left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k}\right) \tag{166}$$

$$= -\int \left(-\frac{1}{\alpha_k}\mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k) - \mathsf{K}_{k,Q_{\boldsymbol{Z}_k},\boldsymbol{\theta}_k}\left(-\frac{1}{\alpha_k}\right)\right)\mathrm{d}P_1(\boldsymbol{z}_k) + D\left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k}\right) \tag{167}$$

$$= \mathsf{K}_{k,Q_{\boldsymbol{Z}_k},\boldsymbol{\theta}_k}\left(-\frac{1}{\alpha_k}\right) + \int \frac{1}{\alpha_k}\mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\,\mathrm{d}P_1(\boldsymbol{z}_k) + D\left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k}\right) \tag{168}$$

$$= -\frac{1}{\alpha_k}\int \mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\,\mathrm{d}P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_k) - D\left(P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\boldsymbol{Z}_k}\right)$$
$$+ \int \frac{1}{\alpha_k}\mathsf{L}_k(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\,\mathrm{d}P_1(\boldsymbol{z}_k) + D\left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k}\right) \tag{169}$$

$$= \frac{1}{\alpha_k}\Big(\mathsf{R}_{k,\boldsymbol{\theta}_k}(P_1) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\Big(P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\Big)\Big) - D\Big(P^{(Q_{\boldsymbol{Z}_k},\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\Big\|\, Q_{\boldsymbol{Z}_k}\Big) + D\big(P_1 \,\big\|\, Q_{\boldsymbol{Z}_k}\big) \tag{170}$$

$$= \frac{1}{\alpha_k} \mathsf{G}_k \left(\boldsymbol{\theta}_k, P_1, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) - D \left(P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) + D \left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k} \right), \tag{171}$$

where equality (163) follows from [58, Theorem 4]; (164) from [58, Theorem 5]; (167) from (66); (169) from Lemma 9; and (171) from (82).

Hence, from the equality in (171) it follows that

$$\begin{aligned} &\mathsf{G}_k \left(\boldsymbol{\theta}_k, P_1, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) \\ &= \alpha_k \Big(D \Big(P_1 \,\Big\|\, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \Big) + D \Big(P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \,\Big\|\, Q_{\boldsymbol{Z}_k} \Big) - D \left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) \Big). \end{aligned} \tag{172}$$

The proof proceeds from the definition of the functional G_k in (83), from which it holds that

$$\begin{aligned} &\mathsf{G}_k \left(\boldsymbol{\theta}_k, P_1, P_2 \right) \\ &= \mathsf{R}_{k, \boldsymbol{\theta}_k} \left(P_1 \right) - \mathsf{R}_{k, \boldsymbol{\theta}_k} \left(P_2 \right) + \mathsf{R}_{k, \boldsymbol{\theta}_k} \left(P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) - \mathsf{R}_{k, \boldsymbol{\theta}_k} \left(P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) && (173) \\ &= \mathsf{G}_k \left(\boldsymbol{\theta}_k, P_1, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) - \mathsf{G}_k \left(\boldsymbol{\theta}_k, P_2, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) && (174) \\ &= \alpha_k \left(D \left(P_1 \,\middle\|\, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) + D \left(P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) - D \left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) \right) \\ &\quad - \alpha_k \left(D \left(P_2 \,\middle\|\, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \right) + D \left(P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) - D \left(P_2 \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) \right) && (175) \\ &= \alpha_k \Big(D \Big(P_1 \,\Big\|\, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \Big) - D \left(P_1 \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) - D \Big(P_2 \,\Big\|\, P^{(Q_{\boldsymbol{Z}_k}, \alpha_k)}_{\widehat{\boldsymbol{Z}}_k | \boldsymbol{\Theta}_k = \boldsymbol{\theta}_k} \Big) + D \left(P_2 \,\middle\|\, Q_{\boldsymbol{Z}_k} \right) \Big), && (176) \end{aligned}$$

where the equality in (174) follows from the definition of the functional G_k in (83); and the equality in (175) follows from (172). This completes the proof.

H Proof of Lemma 14

The proof follows from Definition 12, which yields

$$\begin{aligned} &\mathbf{G}_k \left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)} | \boldsymbol{Z}_0}; P_{\boldsymbol{Z}_0} \right) \\ &= \iint \left(\mathsf{R}_{k, \widehat{z}_k} \Big(P^{(m)}_{\boldsymbol{\Theta}_k | \boldsymbol{Z}_0 = z_0} \Big) - \mathsf{R}_{k, z_k} \Big(P^{(m)}_{\boldsymbol{\Theta}_k | \boldsymbol{Z}_0 = z_0} \Big) \right) \mathrm{d} P_{\boldsymbol{Z}_k} (\widehat{z}_k) \mathrm{d} P_{\boldsymbol{Z}_0} (z_0) && (177) \\ &= \iint \mathsf{R}_{k, \widehat{z}_k} \Big(P^{(m)}_{\boldsymbol{\Theta}_k | \boldsymbol{Z}_0 = z_0} \Big) \mathrm{d} P_{\boldsymbol{Z}_k} (\widehat{z}_k) \mathrm{d} P_{\boldsymbol{Z}_0} (z_0) - \int \mathsf{R}_{k, z_k} \Big(P^{(m)}_{\boldsymbol{\Theta}_k | \boldsymbol{Z}_0 = z_0} \Big) \mathrm{d} P_{\boldsymbol{Z}_0} (z_0). && (178) \end{aligned}$$

The proof focuses on the term $\iint \mathsf{R}_{k, \widehat{z}_k} \left(P^{(m)}_{\boldsymbol{\Theta}_k | \boldsymbol{Z}_0 = z_0} \right) \mathrm{d} P_{\boldsymbol{Z}_k} (\widehat{z}_k) \, \mathrm{d} P_{\boldsymbol{Z}_0} (z_0)$ in (144). From Lemma 4, it follows that

$$\iint \mathsf{R}_{k, \widehat{z}_k} \left(P^{(m)}_{\boldsymbol{\Theta}_k | \boldsymbol{Z}_0 = z_0} \right) \mathrm{d} P_{\boldsymbol{Z}_k} (\widehat{z}_k) \, \mathrm{d} P_{\boldsymbol{Z}_0} (z_0) = \int \mathsf{R}_{k, z_k} \Big(P^{(m)}_{\boldsymbol{\Theta}_k} \Big) \mathrm{d} P_{\boldsymbol{Z}_k} (z_k). \tag{179}$$

Using (179) in (178) yields

$$\mathbf{G}_k \left(P_{\underline{\boldsymbol{\Theta}}, \underline{\boldsymbol{U}}^{(m)}, \underline{\boldsymbol{V}}^{(m)} | \boldsymbol{Z}_0}; P_{\boldsymbol{Z}_0} \right)$$

$$= \int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k}\right) \mathrm{d}P_{\mathbf{Z}_k}(z_k) - \int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}(z_0) \tag{180}$$

$$= \int \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k}\right) - \mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}(z_0) \tag{181}$$

$$= \int \mathsf{G}_k\left(z_k; P^{(m)}_{\boldsymbol{\Theta}_k}; P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right) \mathrm{d}P_{\mathbf{Z}_0}(z_0), \tag{182}$$

which completes the proof of (88).

I Proof of Lemma 15

The proof follows along the same lines as the proof of Lemma 14. It is included entirely hereunder for the sake of completeness.

The proof follows from Lemma 5, which yields

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$$

$$= \iint \left(\mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \tag{183}$$

$$= \iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) - \int \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k). \tag{184}$$

The proof focuses on the term $\iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)$ in (144). From Lemma 4, it follows that

$$\iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) = \int \mathsf{R}_{k,\boldsymbol{\theta}_k}(P_{\mathbf{Z}_k}) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k). \tag{185}$$

Using (185) in (184) yields

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$$

$$= \int \mathsf{R}_{k,\boldsymbol{\theta}_k}(P_{\mathbf{Z}_k}) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) - \int \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \tag{186}$$

$$= \int \mathsf{R}_{k,\boldsymbol{\theta}_k}(P_{\mathbf{Z}_k}) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \tag{187}$$

$$= \int \mathsf{G}_k\left(\boldsymbol{\theta}_k, P_{\mathbf{Z}_k}, P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k), \tag{188}$$

which completes the proof of (89).

J Proof of Theorem 18

The proof starts by considering the information measure $J_{\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (92), from which it follows that

$$
\begin{aligned}
&J_{\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int_{\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&+\int_{\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&-\int_{\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&-\int_{\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0) \qquad (189)\\
&\leqslant\gamma P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)-\gamma P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right)\\
&+\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right)\\
&-\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right) \qquad (190)\\
&=-\gamma P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right)+\gamma P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)\\
&+\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right)\\
&-\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right) \qquad (191)\\
&=P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right)\left(\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)-\gamma\right)\\
&+P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)\left(\gamma-\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\right), \qquad (192)
\end{aligned}
$$

where the inequality in (190) follows from the definition of the set $\mathcal{A}_\gamma$ and the extrema of the log-likelihood ratio.

A lower bounds on the information measure $J_{\mathsf{L}_k,\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (92) can also be obtained as follows,

$$
\begin{aligned}
&J_{\mathsf{L}_k,\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int_{\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&+\int_{\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&-\int_{\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&-\int_{\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0) \qquad (193)\\
&\geqslant\gamma P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right)-\gamma P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)\\
&+\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)\\
&-\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right) \qquad (194)\\
&=-\gamma P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)+\gamma P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right)\\
&+\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)\\
&-\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma\right) \qquad (195)\\
&=P^{(m)}_{\boldsymbol{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_\gamma^{\mathsf{c}}\right)\left(\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma^{\mathsf{c}}}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)-\gamma\right)\\
&+P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\mathcal{A}_\gamma)\left(\gamma-\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)}{\mathrm{d}P^{(\mathcal{Q}_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}\right)\right), \qquad (196)
\end{aligned}
$$

where the equality in (194) follows from the definition of the set $\mathcal{A}_\gamma$ and the extrema of the log-likelihood ratio.

K Proof of Theorem 19

The following lemma is introduced as part of this proof.

Lemma 26. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that for all* $z_0\in\mathcal{Z}_0$*, the probability measures* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}$ *in* (50)*;* $P^{(m)}_{\boldsymbol{\Theta}_k}$ *in* (55)*; and the* σ*-finite measure* $Q_{\boldsymbol{\Theta}_k}$ *in* (66) *satisfy* $P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\ll Q_{\boldsymbol{\Theta}_k}\ll P^{(m)}_{\boldsymbol{\Theta}_k}\ll Q_{\boldsymbol{\Theta}_k}$*. Then,*

$$\begin{aligned}&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\lambda_k\ I\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)\\&+\lambda_k\int\left(D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\ \middle\|\ P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\ \middle\|\ P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0),\end{aligned}\tag{197}$$

where the conditional probability measure $P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k}\in\triangle\left(\mathcal{M}_k|\mathcal{Z}_k^{n_k}\right)$ *is an* $\left(\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k\right)$*-Gibbs conditional probability measure.*

Proof:

From lemma 16, it follows that

$$\begin{aligned}&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)=\int\lambda_k\Bigg(D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\ \middle\|\ P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)\\&-D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\ \middle\|\ P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)+D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)\Bigg)\mathrm{d}P_{\mathbf{Z}_0}(z_0).\end{aligned}\tag{198}$$

The proof proceeds by focusing on the expression $\int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0)$, from which it holds that

$$\begin{aligned}&\int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0)\\&=\int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k}\ \middle\|\ Q_{\boldsymbol{\Theta}_k}\right)\qquad(199)\\&=\iint\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\\&\quad-\int\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\qquad(200)\\&=\iint\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\\&\quad-\int\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\left(\int\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\qquad(201)\\&=\iint\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\end{aligned}$$

$$
\begin{aligned}
&- \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right) \frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\, \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (202)\\
&= \iint \left(\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right) - \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (203)\\
&= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\, \frac{\mathrm{d}Q_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (204)\\
&= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (205)\\
&= I\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}; P_{\mathbf{Z}_0}\right), && (206)
\end{aligned}
$$

where the equality in (201) follows from [58, Theorem 10]; (202) follows from [67, Theorem 2.6.6]; (203) follows from a change of measure [58, Theorem 2]; (204) follows from [58, Theorem 5]; and (205) follows from [58, Theorem 4]. This completes the proof. ∎

The proof proceeds by focusing on the term $\int D\left(P^{(m)}_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right) \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0)$ in (197), from which it follows that

$$
\begin{aligned}
&\int D\left(P^{(m)}_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}\right) \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0)\\
&= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (207)\\
&= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\, \frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (208)\\
&= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0)\\
&\quad + \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\, \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (209)\\
&= \int D\left(P^{(m)}_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}\right) \mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0)\\
&\quad + \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0) && (210)\\
&= L\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}; P_{\mathbf{Z}_0}\right) + \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=\mathbf{z}_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\mathbf{z}_k}}(\boldsymbol{\theta}_k)\right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_0}(\mathbf{z}_0), && (211)
\end{aligned}
$$

where the equality in (209) follows from [58, Theorem 8].

From Lemma 26, by using (211) in (197), it holds that

$$\begin{aligned}
&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)\\
&=\lambda_k\, I\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)+\lambda_k\Bigg(L\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0};P_{\mathbf{Z}_0}\right)\\
&\quad+\iint\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\\
&\quad-\int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\,\middle\|\,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0)\Bigg) \qquad (212)\\
&=\lambda_k\Big(I\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)+L\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)+J_{\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)\Big), \qquad (213)
\end{aligned}$$

where (213) follows from the definition of the information measure $J_{\mathsf{L}_k,Q_{\boldsymbol{\Theta}_k},\lambda_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$ in (92). This completes the proof.

L Proof of Theorem 20

From the definition of the generalization error in (62), it follows that,

$$\begin{aligned}
&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)\\
&=\iint\mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)-\mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k)\mathrm{d}P_{\mathbf{Z}_0}(z_0). \qquad (214)
\end{aligned}$$

The proof proceeds by focusing on $\mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)-\mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)$, from which it follows that

$$\begin{aligned}
&\mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)-\mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)\\
&=\mathsf{G}_k\left(\widehat{z}_k,P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0},P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\right)+\mathsf{R}_{k,\widehat{z}_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\right)\\
&\quad-\mathsf{G}_k\left(z_k,P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0},P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)-\mathsf{R}_{k,z_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right) \qquad (215)\\
&=\mathsf{R}_{k,\widehat{z}_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)-\mathsf{R}_{k,\widehat{z}_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\right)+\mathsf{R}_{k,\widehat{z}_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\right)\\
&\quad-\mathsf{R}_{k,z_k}\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\right)+\mathsf{R}_{k,z_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)-\mathsf{R}_{k,z_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right) \qquad (216)\\
&=\mathsf{R}_{k,\widehat{z}_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\right)-\mathsf{R}_{k,z_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)\\
&\quad+\lambda_k\left(D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\,\middle\|\,Q_{\boldsymbol{\Theta}_k}\right)+D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\,\middle\|\,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\right)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\,\middle\|\,Q_{\boldsymbol{\Theta}_k}\right)\right)\\
&\quad-\lambda_k\left(D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\,\middle\|\,Q_{\boldsymbol{\Theta}_k}\right)+D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\,\middle\|\,P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\right)-D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}\,\middle\|\,Q_{\boldsymbol{\Theta}_k}\right)\right) \qquad (217)
\end{aligned}$$

$$
\begin{aligned}
&= \mathsf{R}_{k,\widehat{\boldsymbol{z}}_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right) + \lambda_k D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)\\
&\quad -\mathsf{R}_{k,\boldsymbol{z}_k}\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right) - \lambda_k D\left(P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k} \,\middle\|\, Q_{\boldsymbol{\Theta}_k}\right)\\
&\quad +\lambda_k\left(D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right) - D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\right) \qquad (218)\\
&= \mathsf{R}_{k,\widehat{\boldsymbol{z}}_k}\left(Q_{\boldsymbol{\Theta}_k}\right) + \lambda_k D\left(Q_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right) - \mathsf{R}_{k,\boldsymbol{z}_k}\left(Q_{\boldsymbol{\Theta}_k}\right) + \lambda_k D\left(Q_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\\
&\quad +\lambda_k\left(D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right) - D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\right), \qquad (219)
\end{aligned}
$$

where the equality in (217) follows from Lemma 12; and the equality in (219) follows from Lemma 7. By plugging (219) into (214) it follows that

$$
\begin{aligned}
&\mathsf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\boldsymbol{Z}_0}; P_{\boldsymbol{Z}_0}\right) = \iint \mathsf{R}_{k,\widehat{\boldsymbol{z}}_k}\left(Q_{\boldsymbol{\Theta}_k}\right) + \lambda_k D\left(Q_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right)\\
&-\mathsf{R}_{k,\boldsymbol{z}_k}\left(Q_{\boldsymbol{\Theta}_k}\right) + \lambda_k D\left(Q_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\\
&+\lambda_k\Bigg(D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right)\\
&-D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\Bigg)\mathrm{d}P_{\boldsymbol{Z}_k}(\widehat{\boldsymbol{z}}_k)\mathrm{d}P_{\boldsymbol{Z}_0}(\boldsymbol{z}_0) \qquad (220)\\
&= \iint \mathsf{R}_{k,\widehat{\boldsymbol{z}}_k}\left(Q_{\boldsymbol{\Theta}_k}\right) + \lambda_k D\left(Q_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right)\mathrm{d}P_{\boldsymbol{Z}_k}\left(\widehat{\boldsymbol{z}}_k\right)\mathrm{d}P_{\boldsymbol{Z}_0}\left(\boldsymbol{z}_0\right)\\
&\quad - \iint \mathsf{R}_{k,\boldsymbol{z}_k}\left(Q_{\boldsymbol{\Theta}_k}\right) + \lambda_k D\left(Q_{\boldsymbol{\Theta}_k} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\mathrm{d}P_{\boldsymbol{Z}_k}\left(\widehat{\boldsymbol{z}}_k\right)\mathrm{d}P_{\boldsymbol{Z}_0}\left(\boldsymbol{z}_0\right)\\
&\quad + \iint \lambda_k\Bigg(D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right)\\
&\quad -D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\Bigg)\mathrm{d}P_{\boldsymbol{Z}_k}(\widehat{\boldsymbol{z}}_k)\mathrm{d}P_{\boldsymbol{Z}_0}(\boldsymbol{z}_0) \qquad (221)\\
&= \lambda_k \iint \Bigg(D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right)\\
&\quad -D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\boldsymbol{z}_k}\right)\Bigg)\mathrm{d}P_{\boldsymbol{Z}_k}(\widehat{\boldsymbol{z}}_k)\mathrm{d}P_{\boldsymbol{Z}_0}(\boldsymbol{z}_0). \qquad (222)
\end{aligned}
$$

This completes the proof.

M Proof of Theorem 21

From Theorem 20, it holds that

$$
\iint D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_0=\boldsymbol{z}_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\boldsymbol{Z}_k=\widehat{\boldsymbol{z}}_k}\right)\mathrm{d}P_{\boldsymbol{Z}_k}\left(\widehat{\boldsymbol{z}}_k\right)\mathrm{d}P_{\boldsymbol{Z}_0}\left(\boldsymbol{z}_0\right)
$$

$$= \frac{1}{\lambda_k} \mathbf{G}_k\Big(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\Big) + \int D\Big(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \,\Big\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\Big)\mathrm{d}P_{\mathbf{Z}_0}(z_0). \quad (223)$$

Moreover, the term on the left side of the equality in (223) can be expressed as follows

$$\begin{aligned}
&\iint D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \,\middle\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0)\,\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k)\\
&= \iiint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_0}(z_0)\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) \quad (224)\\
&= \iiint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k k}}(\boldsymbol{\theta}_k)\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_0}(z_0)\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) \quad (225)\\
&= \iiint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\,\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k)\\
&\quad + \iiint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\,\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) \quad (226)\\
&= \iiint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}}(\boldsymbol{\theta}_k)\right)\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\,\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\,\mathrm{d}P_{\mathbf{Z}_0}(z_0)\,\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k)\\
&\quad + \int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \,\middle\|\, P^{(m)}_{\boldsymbol{\Theta}_k}\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0) \quad (227)\\
&= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}{\mathrm{d}P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}}(\boldsymbol{\theta}_k)\right)\left(\int \frac{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0}}{\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_0}(z_0)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k)\\
&\quad + \int D\left(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \,\middle\|\, P^{(m)}_{\boldsymbol{\Theta}_k}\right)\mathrm{d}P_{\mathbf{Z}_0}(z_0) \quad (228)\\
&= \int D\Big(P^{(m)}_{\boldsymbol{\Theta}_k} \,\Big\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\Big)\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) + \int D\Big(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \,\Big\|\, P^{(m)}_{\boldsymbol{\Theta}_k}\Big)\mathrm{d}P_{\mathbf{Z}_0}(z_0), \quad (229)
\end{aligned}$$

where (225) follows from [58, Theorem 3]; (226) follows from [58, Theorem 4]; (227) follows from [58, Theorem 2]; and (228) follows from [67, Theorem 2.6.6] and [58, Theorem 10].

Using (229) in (223) yields

$$\begin{aligned}
&\int D\Big(P^{(m)}_{\boldsymbol{\Theta}_k} \,\Big\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=\widehat{z}_k}\Big)\mathrm{d}P_{\mathbf{Z}_k}(\widehat{z}_k) + \int D\Big(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \,\Big\|\, P^{(m)}_{\boldsymbol{\Theta}_k}\Big)\mathrm{d}P_{\mathbf{Z}_0}(z_0)\\
&= \frac{1}{\lambda_k} \mathbf{G}_k\Big(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\Big) + \int D\Big(P^{(m)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_0=z_0} \,\Big\|\, P^{(Q_{\boldsymbol{\Theta}_k},\lambda_k)}_{\boldsymbol{\Theta}_k|\mathbf{Z}_k=z_k}\Big)\mathrm{d}P_{\mathbf{Z}_0}(z_0), \quad (230)
\end{aligned}$$

which completes the proof.

N Proof of Theorem 22

The proof follows along the same lines as the proof of Theorem 18. It is included entirely hereunder for the sake of completeness.

The proof starts by considering $J_{\mathsf{L}_k,Q\boldsymbol{Z}_k,\alpha_k}\left(P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right)$ in (99), from which it follows that

$$
\begin{aligned}
&J_{\mathsf{L}_k,Q\boldsymbol{Z}_k,\alpha_k}\left(P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right)\\
&=\int_{\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P_{\boldsymbol{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\\
&\quad+\int_{\mathcal{A}^{\mathsf{c}}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P_{\boldsymbol{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\\
&\quad-\int_{\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{z}_k,\boldsymbol{\theta}_k)\\
&\quad-\int_{\mathcal{A}^{\mathsf{c}}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{z}_k,\boldsymbol{\theta}_k) \qquad (231)\\
&\leqslant\gamma P_{\boldsymbol{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}^{\mathsf{c}}_\gamma\right)-\gamma P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_\gamma\right)\\
&\quad+\max_{(\boldsymbol{\theta}_k,\boldsymbol{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)P_{\boldsymbol{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_\gamma\right)\\
&\quad-\min_{(\boldsymbol{\theta}_k,\boldsymbol{z}_0)\in\mathcal{A}^{\mathsf{c}}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}^{\mathsf{c}}_\gamma\right) \qquad (232)\\
&=-\gamma P_{\boldsymbol{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_\gamma\right)+\gamma P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}^{\mathsf{c}}_\gamma\right)\\
&\quad+\max_{(\boldsymbol{\theta}_k,\boldsymbol{z}_0)\in\mathcal{A}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)P_{\boldsymbol{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_\gamma\right)\\
&\quad-\min_{(\boldsymbol{\theta}_k,\boldsymbol{z}_0)\in\mathcal{A}^{\mathsf{c}}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}^{\mathsf{c}}_\gamma\right) \qquad (233)\\
&=P_{\boldsymbol{Z}_k}P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_\gamma\right)\left(\max_{(\boldsymbol{\theta}_k,\boldsymbol{z}_0)\in\mathcal{A}^{\mathsf{c}}_\gamma}\log\left(\frac{\mathrm{d}P^{(m)}_{\boldsymbol{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q\boldsymbol{Z}_k,\alpha_k)}_{\widehat{\boldsymbol{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right)-\gamma\right)
\end{aligned}
$$

$$+P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k}P^{(m)}_{\mathbf{\Theta}_k}\left(\mathcal{A}^{\mathsf{c}}_{\gamma}\right)\left(\gamma-\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)\right), \tag{234}$$

where the inequality in (232) follows from the definition of the set $\mathcal{A}_{\gamma}$ and the extrema of the log-likelihood ratio.

A lower bounds on the information measure $J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k};P^{(m)}_{\mathbf{\Theta}_k}\right)$ in (99) can also be obtained as follows,

$$\begin{aligned}
&J_{\mathsf{L}_k,Q_{\mathbf{\Theta}_k},\lambda_k}\left(P^{(m)}_{\mathbf{\Theta}_k|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)\\
=&\int_{\mathcal{A}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&+\int_{\mathcal{A}^{\mathsf{c}}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&-\int_{\mathcal{A}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0)\\
&-\int_{\mathcal{A}^{\mathsf{c}}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k|\mathbf{Z}_0}P_{\mathbf{Z}_0}(\boldsymbol{\theta}_k,\mathbf{z}_0) \qquad (235)\\
\geqslant&\gamma P^{(m)}_{\mathbf{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}_{\gamma}\right)-\gamma P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k}P^{(m)}_{\mathbf{\Theta}_k}\left(\mathcal{A}^{\mathsf{c}}_{\gamma}\right)\\
&+\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}^{\mathsf{c}}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)P^{(m)}_{\mathbf{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}^{\mathsf{c}}_{\gamma}\right)\\
&-\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k}P^{(m)}_{\mathbf{\Theta}_k}\left(\mathcal{A}_{\gamma}\right) \qquad (236)\\
=&-\gamma P^{(m)}_{\mathbf{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}^{\mathsf{c}}_{\gamma}\right)+\gamma P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k}P^{(m)}_{\mathbf{\Theta}_k}\left(\mathcal{A}_{\gamma}\right)\\
&+\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}^{\mathsf{c}}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)P^{(m)}_{\mathbf{\Theta}_k}P_{\mathbf{Z}_0}\left(\mathcal{A}^{\mathsf{c}}_{\gamma}\right)\\
&-\max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_{\gamma}}\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\mathbf{z}_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}\right)P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k}P^{(m)}_{\mathbf{\Theta}_k}\left(\mathcal{A}_{\gamma}\right) \qquad (237)
\end{aligned}$$

$$= P^{(m)}_{\boldsymbol{\Theta}_k} P_{\mathbf{Z}_0}\left(\mathcal{A}^{\mathsf{c}}_{\gamma}\right)\left(\min_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}^{\mathsf{c}}_{\gamma}} \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\mathbf{z}_k)\right) - \gamma\right)$$

$$+ P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k} P^{(m)}_{\boldsymbol{\Theta}_k}\left(\mathcal{A}_{\gamma}\right)\left(\gamma - \max_{(\boldsymbol{\theta}_k,\mathbf{z}_0)\in\mathcal{A}_{\gamma}} \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(\mathbf{z}_k)\right)\right), \tag{238}$$

where the equality in (236) follows from the definition of the set $\mathcal{A}_{\gamma}$ and the extrema of the log-likelihood ratio.

O Proof of Theorem 23

The proof follows along the same lines as the proof of Theorem 19. It is included entirely hereunder for the sake of completeness.

The following lemma is introduced as part of this proof.

Lemma 27. *Consider the generalization error* $\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\mathbf{U}}^{(m)},\underline{\mathbf{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right)$ *in* (62) *and assume that for all* $\boldsymbol{\theta}_k \in \mathcal{M}_k$*, the probability measures* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}$ *in* (56)*;* $P_{\mathbf{Z}_k}$ *in* (32)*; and the* σ*-finite measure* $Q_{\mathbf{Z}_k}$ *in* (75) *satisfy* $P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \ll Q_{\mathbf{Z}_k} \ll P_{\mathbf{Z}_k} \ll Q_{\mathbf{Z}_k}$*. Then,*

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\mathbf{U}}^{(m)},\underline{\mathbf{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) = \alpha_k\, I\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k}; P^{(m)}_{\boldsymbol{\Theta}_k}\right)$$
$$+\alpha_k \int \left(D\left(P_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) - D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k), \tag{239}$$

where the conditional probability measure $P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k} \in \triangle\left(\mathcal{Z}_k^{n_k}|\mathcal{M}_k\right)$ *is an* $(\mathsf{L}_k, Q_{\mathbf{Z}_k}, \alpha_k)$*-Gibbs conditional probability measure; and the probability measures* $P^{(m)}_{\boldsymbol{\Theta}_k}$ *is defined in* (55)*.*

Proof: From lemma 17 it follows that

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\mathbf{U}}^{(m)},\underline{\mathbf{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) = \int \alpha_k \Bigg(D\left(P_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)$$
$$-D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) + D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) - D\left(P_{\mathbf{Z}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right)\Bigg)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k). \tag{240}$$

The proof proceeds by focusing on the expression $\int D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) - D\left(P_{\mathbf{Z}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)$, from which it holds that

$$\int D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) - D\left(P_{\mathbf{Z}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)$$

$$= \int D\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) \mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) - D\left(P_{\mathbf{Z}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) \tag{241}$$

$$= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{242}$$

$$- \int \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) \mathrm{d}P_{\mathbf{Z}_k}(\boldsymbol{z}_k)$$

$$= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)$$

$$- \int \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right)\left(\int \frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)\right)\mathrm{d}P_{\mathbf{Z}_k}(\boldsymbol{z}_k) \tag{243}$$

$$= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)$$

$$- \iint \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) \frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\,\mathrm{d}P_{\mathbf{Z}_k}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{244}$$

$$= \iint \left(\log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) - \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right)\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{245}$$

$$= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}Q_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\,\frac{\mathrm{d}Q_{\mathbf{Z}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{246}$$

$$= \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}(\boldsymbol{z}_k)\right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(\boldsymbol{z}_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{247}$$

$$= I\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}; P^{(m)}_{\mathbf{\Theta}_k}\right), \tag{248}$$

where the equality (243) follows from [58, Theorem 10]; (244) follows from [67, Theorem 2.6.6]; (245) follows from a change of measure [58, Theorem 2]; (246) follows from [58, Theorem 5]; and (247) follows from [58, Theorem 4]. This completes the proof. ■

The proof proceeds by focusing on $\int D\left(P_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)$ in (239), from which it follows that,

$$\int D\left(P_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}\right) \mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)$$

$$= \iint \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}(\boldsymbol{z}_k)\right) \mathrm{d}P_{\mathbf{Z}_k}(\boldsymbol{z}_k)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{249}$$

$$= \iint \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\,\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)\mathrm{d}P_{\mathbf{Z}_k}(z_k)\,\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \tag{250}$$

$$= \iint \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)(z_k)\,\mathrm{d}P_{\mathbf{Z}_k}(z_k)\,\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)$$
$$+ \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)\mathrm{d}P_{\mathbf{Z}_k}(z_k)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \tag{251}$$

$$= \int D\left(P_{\mathbf{Z}_k}\,\middle\|\,P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \tag{252}$$
$$+ \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)\mathrm{d}P_{\mathbf{Z}_k}(z_k)\,\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)$$

$$= L\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right) + \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)\mathrm{d}P_{\mathbf{Z}_k}(z_k)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k), \tag{253}$$

where the equality in (251) follows from [58, Theorem 8].

From Lemma 27, by using (253) in (239), it holds that

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$$
$$= \alpha_k\; I\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right) + \alpha_k\Bigg(L\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right)$$
$$+ \iint \log\left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}(z_k)\right)\mathrm{d}P_{\mathbf{Z}_k}(z_k)\,\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k) \tag{254}$$
$$- \int D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\,\middle\|\,P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k)\Bigg) \tag{255}$$
$$= \alpha_k\Big(I\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right) + L\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right) + J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right)\Big), \tag{256}$$

where (256) follows from the definition of the information measure $J_{\mathsf{L}_k,Q_{\mathbf{Z}_k},\alpha_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k};P^{(m)}_{\boldsymbol{\Theta}_k}\right)$ in (99). This completes the proof.

P Proof of Theorem 24

The proof follows along the same lines as the proof of Theorem 20. It is included entirely hereunder for the sake of completeness.

From lemma 5, it holds that

$$\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right)$$

$$= \iint \left(\mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right) \mathrm{d}P_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P_{\boldsymbol{\Theta}_k}(\boldsymbol{\theta}_k). \tag{257}$$

The proof proceeds by focusing on $\mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)$, from which it follows that

$$\begin{aligned}
&\mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \\
&= \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) - \mathsf{G}_k\left(\widehat{\boldsymbol{\theta}}_k, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}, P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \\
&\quad + \mathsf{G}_k\left(\boldsymbol{\theta}_k, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}, P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) && (258) \\
&= \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) - \mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \\
&\quad -\alpha_k\Bigg(- D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) + D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) \\
&\quad -D\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right)\Bigg) + \alpha_k\Bigg(- D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) \\
&\quad +D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) - D\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right)\Bigg) && (259) \\
&= \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) + \alpha_k D\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right) \\
&\quad - \left(\mathsf{R}_{k,\boldsymbol{\theta}_k}\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right) + \alpha_k D\left(P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, Q_{\mathbf{Z}_k}\right)\right) \\
&\quad +\alpha_k\left(D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) - D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right) && (260) \\
&= \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}(Q_{\mathbf{Z}_k}) - \alpha_k D\left(Q_{\mathbf{Z}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) - \left(\mathsf{R}_{k,\boldsymbol{\theta}_k}(Q_{\mathbf{Z}_k}) - \alpha_k D\left(Q_{\mathbf{Z}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right) \\
&\quad +\alpha_k\left(D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) - D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right), && (261)
\end{aligned}$$

where the equality in (259) follows from Lemma 13; and the equality in (261) follows from Lemma 9. By plugging (261) into (257), it follows that

$$\begin{aligned}
&\mathbf{G}_k\left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0}\right) = \iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(Q_{\mathbf{Z}_k}\right) - \alpha_k D\left(Q_{\mathbf{Z}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) \\
&-\left(\mathsf{R}_{k,\boldsymbol{\theta}_k}(Q_{\mathbf{Z}_k}) - \alpha_k D\left(Q_{\mathbf{Z}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\right) + \alpha_k\Bigg(D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) \\
&-D\left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}\right)\Bigg) \mathrm{d}P_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right) && (262) \\
&= \iint \mathsf{R}_{k,\widehat{\boldsymbol{\theta}}_k}\left(Q_{\mathbf{Z}_k}\right) - \alpha_k D\left(Q_{\mathbf{Z}_k} \,\middle\|\, P^{(Q\mathbf{Z}_k,\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right) \mathrm{d}P_{\boldsymbol{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \mathrm{d}P_{\boldsymbol{\Theta}_k}\left(\boldsymbol{\theta}_k\right)
\end{aligned}$$

$$
\begin{aligned}
&- \iint \left(\mathsf{R}_{k,\boldsymbol{\theta}_k} \left(Q_{\mathbf{Z}_k} \right) - \alpha_k D \left(Q_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \right) \right) \mathrm{d}P_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P_{\boldsymbol{\Theta}_k} \left(\boldsymbol{\theta}_k \right) \\
&+ \iint \alpha_k \Bigg(D \left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k} \right) \\
&- D \left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \right) \Bigg) \mathrm{d}P_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P_{\boldsymbol{\Theta}_k} \left(\boldsymbol{\theta}_k \right) \qquad (263) \\
&= \iint \alpha_k \Bigg(D \left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k} \right) \\
&- D \left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \right) \Bigg) \mathrm{d}P_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P_{\boldsymbol{\Theta}_k} \left(\boldsymbol{\theta}_k \right) . \qquad (264)
\end{aligned}
$$

This completes the proof.

Q Proof of Theorem 25

The proof follows along the same lines as the proof of Theorem 21. It is included entirely hereunder for the sake of completeness.

From Theorem 24, it holds that

$$
\begin{aligned}
&\iint D \left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k} \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} \left(\boldsymbol{\theta}_k \right) \\
&= \frac{1}{\alpha_k} \mathbf{G}_k \left(P_{\underline{\boldsymbol{\Theta}},\underline{\boldsymbol{U}}^{(m)},\underline{\boldsymbol{V}}^{(m)}|\mathbf{Z}_0}; P_{\mathbf{Z}_0} \right) + \int D \left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} (\boldsymbol{\theta}_k) . \qquad (265)
\end{aligned}
$$

Moreover, the term on the left of the equality in (265) can be expressed as follows

$$
\begin{aligned}
&\iint D \left(P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k} \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} \left(\boldsymbol{\theta}_k \right) \\
&= \iiint \log \left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(z_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}} \right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(z_k) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} (\boldsymbol{\theta}_k) \qquad (266) \\
&= \iiint \log \left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(z_k)}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}} \frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}(z_k) \right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(z_k) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} (\boldsymbol{\theta}_k) \qquad (267) \\
&= \iiint \log \left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(z_k)}{\mathrm{d}P_{\mathbf{Z}_k}} \frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\boldsymbol{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}}(z_k) \right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}(z_k) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} \left(\widehat{\boldsymbol{\theta}}_k \right) \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} (\boldsymbol{\theta}_k) \qquad (268) \\
&= \iint \log \left(\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}} (z_k) \right) \mathrm{d}P^{(m)}_{\mathbf{Z}_k|\boldsymbol{\Theta}_k=\boldsymbol{\theta}_k} (z_k) \, \mathrm{d}P^{(m)}_{\boldsymbol{\Theta}_k} (\boldsymbol{\theta}_k)
\end{aligned}
$$

$$+\iiint \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}}(z_k)\right)\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}(z_k)\,\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{269}$$

$$= \int D\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P_{\mathbf{Z}_k}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)$$
$$+\iiint \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}}(z_k)\right)\frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}(z_k)\mathrm{d}P_{\mathbf{Z}_k}(z_k)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) \tag{270}$$

$$= \int D\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P_{\mathbf{Z}_k}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)$$
$$+\iint \log\left(\frac{\mathrm{d}P_{\mathbf{Z}_k}}{\mathrm{d}P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}}(z_k)\right)\left(\int \frac{\mathrm{d}P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}}{\mathrm{d}P_{\mathbf{Z}_k}}(z_k)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k)\right)\mathrm{d}P_{\mathbf{Z}_k}(z_k)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right) \tag{271}$$

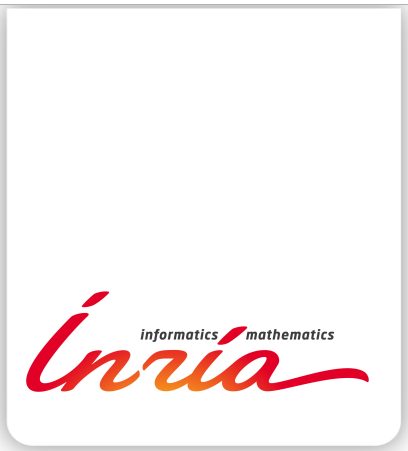
$$= \int D\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P_{\mathbf{Z}_k}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) + \int D\left(P_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right), \tag{272}$$

where (267) follows from [58, Theorem 3]; in (268) follows from [58, Theorem 4]; (270) follows from [58, Theorem 2]; and (272) follows from [67, Theorem 2.6.6] and [58, Theorem 10].

Using (272) in (265) yields

$$\int D\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P_{\mathbf{Z}_k}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k) + \int D\left(P_{\mathbf{Z}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\widehat{\boldsymbol{\theta}}_k}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}\left(\widehat{\boldsymbol{\theta}}_k\right)$$
$$= \frac{1}{\alpha_k}\,\mathbf{G}_k\left(P_{\underline{\mathbf{\Theta}},\underline{\mathbf{U}}^{(m)},\underline{\mathbf{V}}^{(m)}|\mathbf{Z}_0};P_{\mathbf{Z}_0}\right) + \int D\left(P^{(m)}_{\mathbf{Z}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k} \,\middle\|\, P^{(Q_{\mathbf{Z}_k},\alpha_k)}_{\widehat{\mathbf{Z}}_k|\mathbf{\Theta}_k=\boldsymbol{\theta}_k}\right)\mathrm{d}P^{(m)}_{\mathbf{\Theta}_k}(\boldsymbol{\theta}_k), \tag{273}$$

which completes 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