Bayesian data fusion with shared priors

Peng Wu[†], Tales Imbiriba[†], Víctor Elvira[*] *Senior, IEEE*, Pau Closas[†] *Senior, IEEE.*
[†]Electrical and Computer Engineering Department, Northeastern University, Boston, MA
{wu.p,talesim,closas}@northeastern.edu
[*]School of Mathematics, University of Edinburgh, UK
victor.elvira@ed.ac.uk

Abstract—The integration of data and knowledge from several sources is known as data fusion. When data is only available in a distributed fashion or when different sensors are used to infer a quantity of interest, data fusion becomes essential. In Bayesian settings, a priori information of the unknown quantities is available and, possibly, present among the different distributed estimators. When the local estimates are fused, the prior knowledge used to construct several local posteriors might be overused unless the fusion node accounts for this and corrects it. In this paper, we analyze the effects of shared priors in Bayesian data fusion contexts. Depending on different common fusion rules, our analysis helps to understand the performance behavior as a function of the number of collaborative agents and as a consequence of different types of priors. The analysis is performed by using two divergences which are common in Bayesian inference, and the generality of the results allows to analyze very generic distributions. These theoretical results are corroborated through experiments in a variety of estimation and classification problems, including linear and nonlinear models, and federated learning schemes.

Index Terms—Data fusion, Bayesian inference, Distributed Machine Learning, Federated learning.

I. INTRODUCTION AND RELATED WORKS

Distributed data fusion (DDF) is a process in which a group of agents senses their immediate environment, communicates with other agents, and aims at inferring knowledge about a specific process collectively. Some applications include cooperative robots mapping a room [1], multisensor data fusion in internet of things[2], navigation systems [3], medical diagnosis, or pattern recognition to name a few[4].

There are several data fusion architectures which have been proposed. Arguably, the joint directors of laboratories (JDL) Data Fusion Group is the most widely-used taxonomy method for data fusion-related functions [5], [6]. It defines data fusion as a "multilevel, multifaceted process handling the automatic detection, association, correlation, estimation, and combination of data and information from several sources." Two recent fusion frameworks were proposed in [7], [8], both based on category theory and are claimed to be sufficiently general to capture all kinds of fusion techniques [9], including raw data fusion, feature fusion, and decision fusion.

Distributed data fusion research can be categorized into Bayesian or consensus-based approaches [10]. Bayesian methods focus on preserving the full distribution of the unknowns given the data, called posterior, over the estimated process at each agent, so that sensor data can be easily and recursively merged with prior knowledge and does not need to be stored. On the other hand, consensus algorithms are designed in such a way that agents can continue to exchange information until they agree on certain parameters or quantities of interest. In this context, many data fusion methods focus on Bayesian methodologies, such as naive Bayesian fusion, federated Kalman filtering [11], [12], and other different fusion methods that may leverage data-driven models [13], [14], [15]. Recently, there has been increasing interest in the consensus area, particularly in the machine learning community, where a plethora of distributed learning or federated learning methods have been proposed [16], [17], [18]. In this paper, we focus on a Bayesian perspective to the data fusion problem.

There are many data fusion challenges [5], [19], being the data correlation problem one of the most prominent cases. In general, the performance of DDF solutions cannot outperform a centralized scheme, which is typically considered as a benchmark. Although on the other side, DDF is intrinsically more adaptive and resilient to failures. When dealing with distributed agents, some of the challenges involve both dealing with observations impacted by the same process noise [20] and also non-independence of local estimates due to multiple counting of data [21], [22], [14], [10], which essentially means that in a distributed architecture local agent estimates may be correlated. To maintain optimality and consistency, a distributed fusion method should account for such cross-correlation issues. The multiple-counting problem occurs when data is utilized numerous times without the user's knowledge.[1] This might be due to recirculation of data through cyclic channels or the same data traveling through several paths from another agent to the fusion node. To avoid multiple counting of data, two popular solutions are typically adopted [23]: arithmetic average (AA) and geometric average (GA). These two methods can be employed by most fusion methods that *average* the parameters of interest across many agents. Sometimes the GA solution are referred to as Chernoff fusion approaches [22], also known as covariance intersection [24], [21] under Gaussian assumptions. Related to the latter are the works on *mixture of experts* [25], robust *product of experts* [15], fusion of Monte Carlo approximations [26], and the

[1]Note that certain data fusion literature refers to this issue as the *double-counting* problem. We will refer to *multiple-counting* problem in this paper, since it is a more appropriate term in the considered fusion context.

weighted fusion of Kalman filters [27], [28], as well as the linear fusion of partial estimators [29], [30].

DDF is a field that connects to other disciplines, for that reason related works can be found also under the umbrella of model fusion, estimator fusion, distributed estimation, and distributed learning. Particularly, in the area of machine learning, the interest of distributed learning is growing rapidly under the so-called federated learning (FL) paradigm. Most of the research in FL focuses on frequentist approaches to inference, where data-driven neural network (NN) parameters are aggregated [31], [18], [32] based on the AA method. Besides the fusion methodology, privacy and communication efficiency are also a key part of FL research [16]. Recently, some works explored Bayesian FL schemes [33], [34], [35], [36], [37] where the model fusion problem is addressed from a Bayesian perspective, an approach that is thoroughly reviewed in [38]. While this paper described the different fusion strategies including the CIP and CIL we discussed, the present paper provides additional theoretical analyses and connections to different application examples that help provide further practical insights. Also within the Bayesian approach, Bayesian committee machine (BCM) [13], [15], [39], [40] is another framework that relates to DDF, in this case focused on Gaussian process models.

In this paper, we analyse the use of *a priori* information in distributed fusion problems, where data is not directly shared and thus kept private. Particularly, we focus on schemes where a parameter $\boldsymbol{\theta}$ is inferred locally by M agents in a Bayesian setting. In this scheme, $\boldsymbol{\theta}$ takes values in an arbitrary d-dimensional space Θ. The quantity of interest is therefore treated as a random variable for which an a priori distribution is available, $p(\boldsymbol{\theta})$. This prior is shared by all participating agents and updated using local data $\mathcal{D}_m$ to produce the a posteriori distribution of the unknown given the local information, $p(\boldsymbol{\theta}|\mathcal{D}_m)$. In this context, we analyse two conditional independent (exact and approximate) Bayesian fusion methodologies that employ the M local posteriors and, in particular, connect those to the case where the full posterior is computed at a central node accessing all data, $p(\boldsymbol{\theta}|\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M)$. One of the main challenges of Bayesian fusion rules is the repeated use of prior data, which is generally not properly dealt with. We discuss corrections to avoid systematic bias, as well as provide a discussion on how the number of clients affects the performance of the fusion rule when $p(\boldsymbol{\theta})$ is used in excess without corrections. In practical problems of distributed Bayesian inference, not dealing properly with the multiple counting problem can yield to very sub-optimal solutions. In this paper, we discuss practical situations where this problem occurs, particularly when either the number of participating agents M grows or when the prior distribution is very informative. The challenge is often that details on how prior information was used are not always available at the fusion center. Also, in other systems, efficiency is preferred over accuracy due to computational limitations, which can endanger the quality of the final posterior. An illustrative scenario is in distributed target tracking [28], where multiple sensors are combined to enhance accuracy and coverage. However, due to band-limited communications the priors are not always
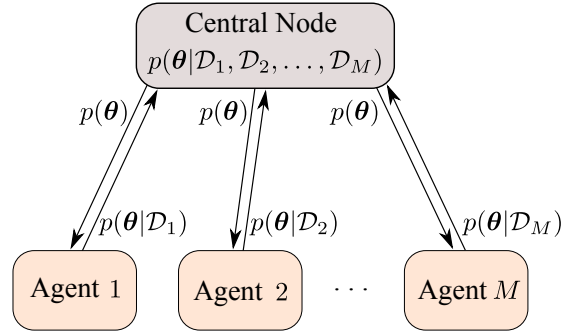

Fig. 1: Distributed framework of M agents sharing a priori information on a quantity, $p(\boldsymbol{\theta})$, and returning local posterior updates, $p(\boldsymbol{\theta}|\mathcal{D}_m)$. A central node is in charge of fusing the local updates with the goal of approximating the full posterior $p(\boldsymbol{\theta}|\mathcal{D}_1, \ldots, \mathcal{D}_M)$.

available at the central node, in which case fusion typically occurs at the decision level without considering the shared prior influence, ultimately yielding to potentially erroneous tracking results.

The novelty of this work is to present a rigorous analysis of distributed Bayesian fusion strategies, which advances our understanding on how prior information impacts their performance. The case of Gaussian posteriors, of interest in a Bayesian inference context due to its tractability, is investigated in detail. In particular, we provide analytic results on the impact of both the quality of the shared prior information and the number of collaborative clients. These results are shown to be valid for arbitrary distributions, going beyond the Gaussian assumption. The theoretical findings are validated on a set of representative inference problems such as distributed linear and neural network based regression and classification, and on a federated learning scheme. This work paves the way for further methodological advances in distributed Bayesian fusion, equipped with a solid theoretical justification.

The remainder of the paper is organized as follows. Section II introduces the optimal (global) Bayesian solution as well as two fusion rules using local updates and assumptions. Section III provides insights for the important case where inferring $\boldsymbol{\theta}$ can be seen as a regression problem where the relevant distributions are approximated by a Gaussian distributions. Section IV provides an analysis of the two fusion rules, showing results in terms of the number of participating agents and the informativeness of the a priori distribution, under arbitrary distributions and not necessarily restricted to the Gaussian case. Different experimental results, featuring various situations, are discussed in Section V, validating the theoretical results on earlier sections in relevant cases. The paper is concluded by Section VI with final remarks.

II. BAYESIAN DATA FUSION SCHEMES

This paper considers a framework (see Fig. 1) where observations related to an unknown quantity of interest $\boldsymbol{\theta} \in \Theta^d$ are obtained by M distributed agents, such that the m-th agent has only access to its own partial data $\mathcal{D}_m$. The data is assumed to be conditionally independent and identically distributed

TABLE I: Table of relevant notation conventions.

Notation	Definition
$\boldsymbol{\theta}$	Parameter of interest.
$p(\boldsymbol{\theta})$	Prior distribution for $\boldsymbol{\theta}$.
m, M	m-th agent and total of M agents.
$\mathcal{D}_m$, $\mathcal{D}$	Data available to the m-th agent and to all agents.
$p(\boldsymbol{\theta}\mid\mathcal{D}_m)$	m-th agent local posterior of $\boldsymbol{\theta}$ given $\mathcal{D}_m$.
$p(\boldsymbol{\theta}\mid\mathcal{D})$	Centralized posterior of $\boldsymbol{\theta}$ given $\mathcal{D}$.
q_0	Scalar value representing the variance of $p(\boldsymbol{\theta})$.
$\boldsymbol{\theta}_0$, $\mathbf{C}_0$	Mean and covariance of Gaussian prior.
$\boldsymbol{\theta}_m$, $\mathbf{C}_m$	Mean and covariance of local Gaussian posterior.
$\boldsymbol{\mu}$, $\boldsymbol{\Lambda}$	Mean and covariance of CIL's Gaussian posterior.
$\widetilde{\boldsymbol{\mu}}$, $\widetilde{\boldsymbol{\Lambda}}$	Mean and covariance of CIP's Gaussian posterior.

(*i.i.d.*) across agents. In addition, a Bayesian approach to the inference of $\boldsymbol{\theta}$ is considered, therefore a prior distribution is assumed $p(\boldsymbol{\theta})$ before data is observed. The main goal of Bayesian data fusion is to evaluate the posterior of the quantity given all the data $p(\boldsymbol{\theta}\mid\mathcal{D})$, however this might be unpractical due to the need to transmit local data to a central node in charge of processing it. Alternatively, distributed Bayesian data fusion provides a framework to combine the local inference results in a manner that approximates the optimal Bayesian solution. For the sake of clarity, Table I provides a summary of the notation used throughout the paper.

This section formulates the optimal Bayesian fusion strategy, as it would be implemented at a central node using all available data $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$ to compute $p(\boldsymbol{\theta}\mid\mathcal{D})$. Then, we establish connections to different distributed Bayesian data fusion strategies where we observe that, unless properly accounted for, the a priori information may be used multiple times. This situation can potentially cause inconsistent approximations of the optimal Bayesian solution [21], [13], [14]. In particular, two approximate Bayesian solutions are discussed in the remainder of this section.

CIL fusion rule. Under the assumption that datasets $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$ are *i.i.d.* given $\boldsymbol{\theta}$, it is possible to express the optimal Bayesian fusion rule in terms of the local likelihoods and the globally used prior distribution. We refer to this solution [38] as the optimal Bayesian fusion rule under *conditionally independent likelihoods* (CIL). The posterior $p(\boldsymbol{\theta}\mid\mathcal{D})$ is then

$$
\begin{aligned}
p(\boldsymbol{\theta}\mid\mathcal{D}) &= \frac{p(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M)} \\
&= \frac{p(\boldsymbol{\theta})\prod_{m=1}^{M} p(\mathcal{D}_m\mid\boldsymbol{\theta})}{p(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M)} ,
\end{aligned}
\tag{1}
$$

where we assumed that the joint likelihood can be factorized into local likelihoods by the conditional independence assumption, which can be further manipulated using Bayes' rule as

$$
p(\boldsymbol{\theta}\mid\mathcal{D}) = Z\frac{\prod_{m=1}^{M} p(\boldsymbol{\theta}\mid\mathcal{D}_m)}{p^{M-1}(\boldsymbol{\theta})} ,
\tag{2}
$$

where

$$
Z = \frac{\prod_{m=1}^{M} p(\mathcal{D}_m)}{p(\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M)} ,
\tag{3}
$$

is the normalization constant and $p(\boldsymbol{\theta}\mid\mathcal{D}_m)$ represents the local posterior for the m-th agent.

This approach is optimal under the *i.i.d.* data assumption, which enables the likelihood factorization in (2). It presents a Bayesian approach for data fusion when the same prior is used by the M users. This solution enables for a distributed implementation, where prior is sent to each agent who returns their local posteriors in exchange. It is easy to see that this Bayesian fusion rule for the case where each agent uses different local prior information $p_m(\boldsymbol{\theta})$ is

$$
p(\boldsymbol{\theta}\mid\mathcal{D}) = Zp(\boldsymbol{\theta}) \prod_{m=1}^{M} \frac{p(\boldsymbol{\theta}\mid\mathcal{D}_m)}{p_m(\boldsymbol{\theta})} ,
\tag{4}
$$

where the global prior $p(\boldsymbol{\theta})$ might differ from the local prior distributions.

CIP fusion rule. There are situations where distributing the priors across the collaborative agents is not possible and, instead, the data fusion node has access to the local posteriors only. These agents might have used the same priors for the quantity or different across agents. A popular fusion rule is the so-called product of experts (PoE) [15], [25] by which the local posteriors are multiplied together in order to produce a global posterior estimate, which is necessarily an approximation of the optimal solution in (2). Following the nomenclature used in this paper, we refer to this approach [15], [38] as the approximate Bayesian fusion rule under *conditionally independent posterior* (CIP), where the main assumption is that global posterior can be factorized into a product of the local posteriors. The resulting posterior $\widetilde{p}(\boldsymbol{\theta}\mid\mathcal{D})$ is an approximation of the *true* posterior $p(\boldsymbol{\theta}\mid\mathcal{D})$ which is accurate when, indeed, the assumption holds. More precisely,

$$
\widetilde{p}(\boldsymbol{\theta}\mid\mathcal{D}) = \widetilde{Z} \prod_{m=1}^{M} p(\boldsymbol{\theta}\mid\mathcal{D}_m) ,
\tag{5}
$$

where $\widetilde{Z}$ is a normalization term and $p(\boldsymbol{\theta}\mid\mathcal{D}_m)$ is m-th local posterior as computed by each of the M agents. Notice that the prior $p(\boldsymbol{\theta})$ is reused multiple times in CIP. In contrast, CIL rule in (2) is optimally dealing with that issue by virtue of the dividing factor $p^{M-1}(\boldsymbol{\theta})$.

The results in Section V will provide a comparison between CIP and CIL approaches. In addition, we are interested in developing analytical tools to gain understanding about their differences. Mainly, CIP in (5) differs from CIL in (2) in its underlying assumption that the global posterior can be factorized into local posteriors. We propose to quantify this deviation using particular cases of the f-divergence between CIL and CIP solutions, such that we can understand theoretically when it is worth accounting for the multiple counting of the prior which is discussed in the Section IV.

III. FUSION OF GAUSSIAN ESTIMATORS

Gaussian assumptions are often considered in order to allow for inferential tractability [41]. Furthermore, it is also common to approximate the posterior by a Gaussian distribution, for instance in the context the so-called Gaussian filters [42] or Laplace approximations (see for instance INLA [43]).

In this section, we analyze the relevant case where $\Theta = \mathbb{R}^d$ in a Gaussian context, such that inferring $\boldsymbol{\theta} \in \mathbb{R}^d$ can be

interpreted as a regression problem or a classification task [44]. In this context, we assume that the prior on the parameter is normally distributed, $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{C}_0)$, and that the distribution of the local likelihood for the m-th agent is

$$\mathbf{y}_{n,m}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{f}(\boldsymbol{\theta}, \mathbf{x}_{n,m}), \mathbf{R}_m) \qquad (6)$$

where $\mathbf{x}_{n,m} \in \mathbb{R}^{d_x}$ is the n-th feature input vector, $\mathbf{f}(\boldsymbol{\theta}, \mathbf{x}_{n,m})$ is a mapping function from those inputs to observed data $\mathbf{y}_{n,m} \in \mathbb{R}^{d_y}$. $n = 1, \ldots, N_m$ denotes the sample index for the m-th agent. The local dataset, as described in Section II, is then composed of the N_m pairs $\mathcal{D}_m = \{\mathbf{y}_{n,m}, \mathbf{x}_{n,m}\}_{n=1}^{N_m}$, and $\mathbf{R}_m$ denotes the observation covariance matrix, which potentially can be different across agents. Note that when the mapping is a linear function on $\boldsymbol{\theta}$, then the posterior is also normally distributed and its parameters can be optimally computed [44]. In general problems, if the function $\mathbf{f}(\cdot, \mathbf{x}_{n,m})$ is non-linear in the parameter $\boldsymbol{\theta}$, one can still make an approximation that the posterior would be approximately Gaussian and compute its mean/covariance leveraging for instance a Laplace approximation (e.g. an approach discussed in [44]) to linearize the model.

Under the Gaussian approximation, the resulting local posterior $p(\boldsymbol{\theta}|\mathcal{D}_m) \approx \mathcal{N}(\boldsymbol{\theta}_m, \mathbf{C}_m)$ with

$$\boldsymbol{\theta}_m = \mathbf{C}_m \left(\mathbf{C}_0^{-1} \boldsymbol{\theta}_0 + \sum_{n=1}^{N_m} \boldsymbol{F}_{n,m}^\top \mathbf{R}_m^{-1} \mathbf{y}_{n,m} \right) , \qquad (7)$$

$$\mathbf{C}_m^{-1} = \mathbf{C}_0^{-1} + \sum_{n=1}^{N_m} \boldsymbol{F}_{n,m}^\top \mathbf{R}^{-1} \boldsymbol{F}_{n,m}, \qquad (8)$$

where $\boldsymbol{F} = \partial \mathbf{f}/\partial \boldsymbol{\theta}$ is the Jacobian of the model, resulting from its linearization (or the coefficients of the model in case it is already linear).

CIL fusion rule. The optimal Bayesian fusion rule under conditionally independent likelihoods, or CIL for short, under (approximately) local Gaussian posteriors leads to a Gaussian posterior given by

$$p(\boldsymbol{\theta}|\mathcal{D}_1, \ldots, \mathcal{D}_M) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) , \qquad (9)$$

where (see Appendix A for details) the posterior precision matrix can be obtained as

$$\boldsymbol{\Lambda} = \sum_{m=1}^{M} \mathbf{C}_m^{-1} - \mathbf{C}_0^{-1}(M-1) , \qquad (10)$$

which depends on the local and a priori covariance matrices. The posterior mean is then

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1} \left(\sum_{m=1}^{M} \mathbf{C}_m^{-1} \boldsymbol{\theta}_m - (M-1)\mathbf{C}_0^{-1} \boldsymbol{\theta}_0 \right) , \qquad (11)$$

which can be rearranged as

$$\boldsymbol{\mu} = \sum_{m=1}^{M} \boldsymbol{\Xi}_m \boldsymbol{\theta}_m + \boldsymbol{\Xi}_0 \boldsymbol{\theta}_0 = \sum_{m=0}^{M} \boldsymbol{\Xi}_m \boldsymbol{\theta}_m , \qquad (12)$$

with the weight matrices as

$$\boldsymbol{\Xi}_m = \begin{cases} \boldsymbol{\Lambda}^{-1} \mathbf{C}_m^{-1}, & m \neq 0, \\ (1-M)\boldsymbol{\Lambda}^{-1}\mathbf{C}_0^{-1}, & m = 0. \end{cases} \qquad (13)$$

CIP fusion rule. Under the Gaussian assumption, the approximate fusion rule in (5), or CIP in short, results in

$$\widetilde{p}(\boldsymbol{\theta}|\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M) = \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Lambda}}^{-1}) , \qquad (14)$$

where

$$\widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\Lambda}}^{-1} \sum_{m=1}^{M} \mathbf{C}_m^{-1} \boldsymbol{\theta}_m, \qquad (15)$$

$$\widetilde{\boldsymbol{\Lambda}} = \sum_{m=1}^{M} \mathbf{C}_m^{-1} , \qquad (16)$$

as detailed in [29] and which can also be interpreted as a particular case of the Gaussian CIL rule (9) when no a priori information is considered in the fusion stage. That is, intuitively, that $\mathbf{C}_0$ takes very large values such that $\boldsymbol{\mu} \rightarrow \widetilde{\boldsymbol{\mu}}$ and $\boldsymbol{\Lambda} \rightarrow \widetilde{\boldsymbol{\Lambda}}$.

IV. CONVERGENCE ANALYSIS OF CIL AND CIP FUSION STRATEGIES.

We are interested in understanding when the two fusion rules are equivalent. That is, when accounting for the multiple use of the a priori information makes a difference and when, on the contrary, can be neglected thus simplifying the calculus. To that aim, we use two divergences that belong to the family of f-divergences to quantify the similarities of the posteriors resulting from both methods. Namely, we consider the Kullback-Leibler (KL) and the χ^2 divergences. While KL is widely used in probability and information theory [45], χ^2-divergence is highly relevant in Bayesian inference and importance sampling [46], since it relates to the variance of the Monte Carlo estimators that attempt to approximate moments of the posterior (see for instance [47], [48], [49]). The two divergences are introduced here and particularized for the CIL and CIP expression of interest in this work.

The KL divergence between CIL and CIP, for a general model where M agents are fused, results in

$$\mathrm{KL}_M \left(p(\boldsymbol{\theta}|\mathcal{D}) \parallel \widetilde{p}(\boldsymbol{\theta}|\mathcal{D}) \right) = \int p(\boldsymbol{\theta}|\mathcal{D}) \log \left(\frac{p(\boldsymbol{\theta}|\mathcal{D})}{\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})} \right) d\boldsymbol{\theta} \qquad (17)$$

$$= \log \left(\frac{p_M(\mathcal{D})}{p(\mathcal{D})} \right) - (M-1) \int \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \log(p(\boldsymbol{\theta})) d\boldsymbol{\theta},$$

where we explicitly note the number of agents as a subindex in the KL for clarity. We notice the dependence of the divergence with M and the parameter's prior distribution (see Appendix C for additional details on how this expression was obtained):

The χ^2 divergence between the fusion rules can be generally expressed as

$$\chi_M^2(p(\boldsymbol{\theta}|\mathcal{D}) \parallel \widetilde{p}(\boldsymbol{\theta}|\mathcal{D})) = \int \frac{p^2(\boldsymbol{\theta}|\mathcal{D})}{\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} - 1 \qquad (18)$$

$$= \int p(\boldsymbol{\theta}|\mathcal{D})p^{M-1}(\boldsymbol{\theta}) d\boldsymbol{\theta} \int p(\boldsymbol{\theta}|\mathcal{D})p^{1-M}(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1,$$

where we again notice the dependence on M and $p(\boldsymbol{\theta})$. Additional details regarding this expression can be consulted in Appendix D.

Based on these divergences we provide an analysis of convergence between CIL and CIP fusion rules with respect to the design parameters related to the number of agents and the informativeness of the prior on $\boldsymbol{\theta}$.

Before reaching our main results we define the fusion setup as follows. Let $\mathcal{D}$ be a global dataset of *i.i.d.* observations, which are related to a parameter of interest $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$. Let $\mathcal{D}$ be processed by M local agents, each observing mutually exclusive sets $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$ and sharing the same a priori information from the quantity of interest $p(\boldsymbol{\theta})$. The m-th agent updates the prior to compute its local posterior, $p(\boldsymbol{\theta}|\mathcal{D}_m)$, which are fused at a central node either using the CIL Bayesian fusion rule $p(\boldsymbol{\theta}|\mathcal{D})$ or the CIP Bayesian fusion rule $\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})$. Finally, let q_0 be a scalar defining the variance of the prior $p(\boldsymbol{\theta})$.

Theorem 1. *For $\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})$ and $p(\boldsymbol{\theta}|\mathcal{D})$ belonging to arbitrary distribution families, $\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})$ asymptotically tends to $p(\boldsymbol{\theta}|\mathcal{D})$ as the shared prior becomes non-informative, that is, their divergence is such that*

$$\lim_{q_0 \to \infty} d(p(\boldsymbol{\theta}|\mathcal{D}) \| \widetilde{p}(\boldsymbol{\theta}|\mathcal{D})) = 0 , \quad \forall M \geqslant 1$$

where q_0 is a parameter that models the a priori uncertainty.

Proof. Details can be found in Appendix B for the general case, as well as a particularization to the Gaussian posterior assumption. □

Theorem 2. *For $\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})$ and $p(\boldsymbol{\theta}|\mathcal{D})$ belonging to arbitrary distribution families, their divergence increases with M, that is,*

$$d_{M+1}(p(\boldsymbol{\theta} \| \mathcal{D})|\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})) \geq d_M(p(\boldsymbol{\theta}|\mathcal{D}) \| \widetilde{p}(\boldsymbol{\theta}|\mathcal{D}))$$

with $d_M(\cdot \| \cdot)$ representing either the KL or χ^2 divergences.

Proof. Details in Appendices C and D for the KL and χ^2 divergences, respectively. □

Theorem 1 states that CIL converges to CIP as the a priori distribution on $\boldsymbol{\theta}$ becomes less informative. Notice that the result shown in Appendix B is general for any assumed posterior distribution, although we complementary provide a result under the Gaussian case, which is widely used in the context of Bayesian inference since, under mild conditions, the posterior converges in distribution to a normal according to the Bernstein–von Mises Theorem [50].

Intuitively, the result in Theorem 2 shows that for a fixed budget of N samples, increasing the number of agents M increases the separation between the two global fusion solutions. That is, as more agents participate less data per agent is available, in which case the a priori distribution on $\boldsymbol{\theta}$ becomes more relevant and the issue of reusing the prior emerges. This result is valid for both KL and χ^2 divergences under arbitrary distribution families.

V. EXPERIMENTS

We validate the obtained results on various relevant inference problems. Namely, (1) an estimation problem, where a set of M linear regression models are employed to update the prior with local data, which are subsequently fused to produce global estimates; (2) a binary classification problem in which M models are used locally to produce classification results that are then fused to produce a global result; (3) another general classification problem, where M neural networks are locally used to produce classification solutions that are then aggregated into a global Bayesian classifier; and (4) an application in the context of federated learning, which in this case updates the prior recursively instead of just once, unlike the previous experiment.

A. Distributed estimation of linear models

For the experiments presented in this subsection we generated a synthetic data set following a linear model, with respect to the parameters $\boldsymbol{\theta}$, embedded in noise:

$$\mathbf{y}_m = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\mathbf{x}_m) + \mathbf{r}_m, \tag{19}$$

where $\mathbf{r}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$, with $\mathbf{R} = r\mathbf{I}$, and $\boldsymbol{\phi} : \mathbb{R}^{d_x} \to \mathbb{R}^d, \mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x})$ is an arbitrary nonlinear function. In our experiment, we set $r = 4$, the observation dimension is 1, and $\boldsymbol{\phi}$ is identity function. The $\boldsymbol{\theta}$ is a real random integer vector generated uniformly between -10 and 20 with $d_x = 6$ dimensions. The total number of points generated is around 700 for training and around 300 for testing. Then, we split the dataset into several parts, each corresponding to one of the local clients' training data. We estimated local parameters following equation (7) and fused using the approaches as discussed in Section III. Furthermore, we computed the evolution of the KL divergence, and test MSE with both prior variance $q_0 \in \mathbb{R}_+$, and total number of clients M. The results are depicted in Figures 2-4. All results are obtained by average of 400 independent Monte Carlo simulations.

Figure 2 shows the evolution of the KL divergence with respect to the prior variance q_0 (top panel for $M = 6$ and $M = 26$) and the number of clients M (bottom panel for $q_0 = 1$, $q_0 = 3$). In this figure, the KL decreases wrt prior variance q_0. Which can also be shown in the bottom panel for $q_0 = 1$, $q_0 = 3$, where the KL increase wrt the number of clients. These results are consistent with Theorem 2 and 1, which was already predicting such behavior.

Figure 3 shows the test MSE wrt the prior variance q_0 for a fixed number of collaborating clients of 6 and 26. This figure shows that all tested configurations converge to similar results when the prior variance is large enough, that is when the a priori information on $\boldsymbol{\theta}$ is non-informative. However, when given $p(\boldsymbol{\theta})$ is informative, represented by smaller variance values, the impact of properly accounting for it would be more apparent. Independently of the number of users, CIL seems to exhibit stable results, whereas CIP's performance degrades as the prior becomes narrower.

Figure 4 shows the MSE of test wrt the number of clients M, as well as for the two representative values prior variance
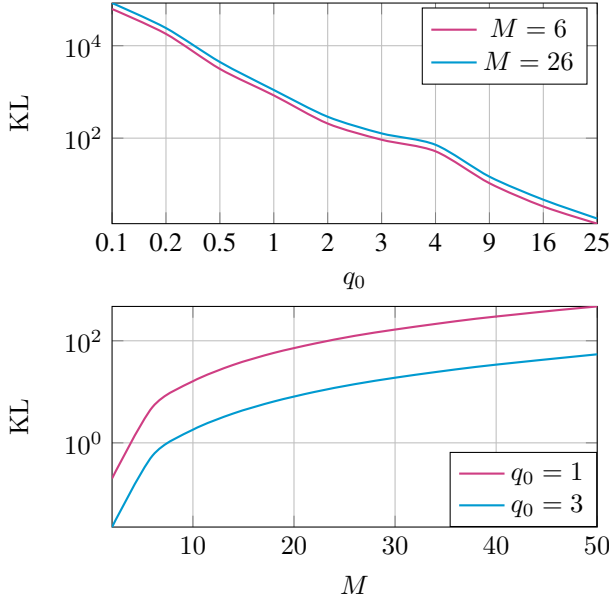
Fig. 2: Evolution of the KL divergence wrt q_0 (top) and number of client M (bottom).
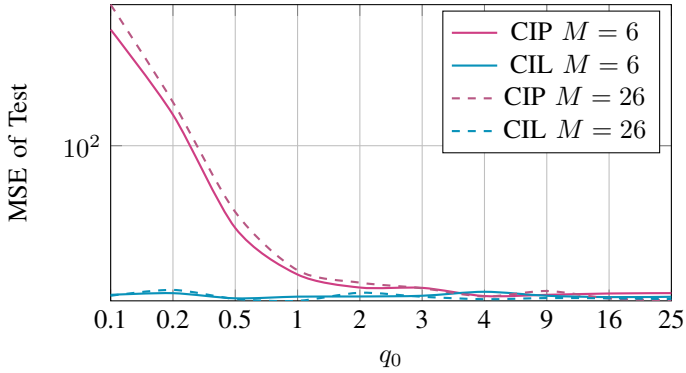

Fig. 3: Test MSE with respect to q_0 for the distributed linear estimation in Section V-A.

$q_0 = 1$ and $q_0 = 3$. The results suggest that CIL can outperform CIP as we increasing the number of clients. In that situation, for growing M, CIP uses the prior information repeatedly, causing the results to be biased unless this fact is taken into consideration as in the CIL appoach. This effect is even more notorious when the a priori information is not accurate (e.g., the mean value is far from the true value of $\boldsymbol{\theta}$).

B. Fusion of local class posterior probabilities

We are now interested in analyzing the performance of the Bayesian data fusion approaches when priors are shared across agents in a classification problem. Similarly to BCM [13], we consider the fusion of classification results as the fusion of probability functions. Analogous to (2), the CIL fusion, given input $\mathbf{x}$, results in a discrete posterior probability distribution:

$$P(C = c|\mathcal{D}_1, \ldots, \mathcal{D}_M, \mathbf{x}) \propto \frac{\Pi_{m=1}^M P(C = c|\mathcal{D}_m, \mathbf{x})}{P^{M-1}(C = c)}, \quad (20)$$
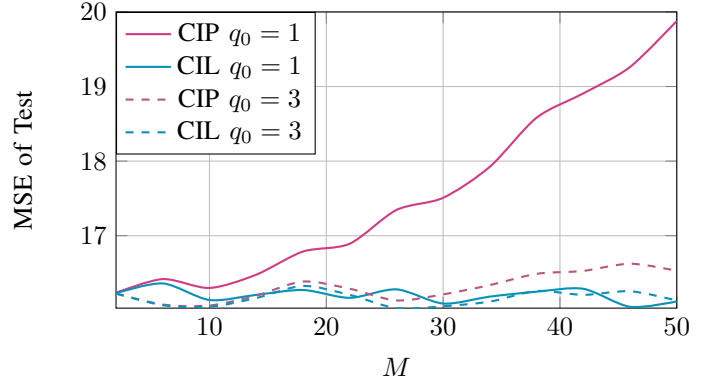

Fig. 4: The test MSE wrt M under different prior information when prior variance at 1 and 3.

where c is the classification label with $c \in \{1, \ldots, L\}$. The a priori probability of the classes, shared among agents, is denoted by $P(C = c)$. To obtain the local class posterior probabilities $P(C|\mathcal{D}_m, \mathbf{x})$, we consider an linear discriminant analysis (LDA) model, which optimally utilizes the shared prior, then fuse the local posteriors to obtain the final result in (20).

In this experiment, for the sake of clarity, we are considering an $L = 2$ classes classification problem. The objective is thus to estimate their class posterior probabilities $P(C = i|\mathcal{D}_1, \ldots, \mathcal{D}_M, \mathbf{x})$ with $i \in \{1, 2\}$. We generated a synthetic dataset, consisting of 1000 data points, by sampling from two normal distributions (for class i, the conditional likelihood of the generated data is $p(\mathcal{D}_m|\mathbf{x}, C = i) = \mathcal{N}(\mathbf{1}_i, \mathbf{I})$, where $\mathbf{1}_i$ is an all-zeroes vector except for a 1 at the i-th element. The dimension of the data $\mathbf{y}$ is 10. The prior class probabilities were 0.6 and 0.4 for classes 1 and 2, respectively. The resulting dataset was divided into M subsets of the same dimension $1000/M$, such that these are *i.i.d.* distributed.

Figure 5 shows the evolution of the KL divergence between class posteriors of CIP and CIL as function of different systems parameters. The top panel depicts the KL divergence as a function of the assumed (misspecified) prior $P_1 \triangleq P(C = 1)$ for $M \in \{6, 12\}$. The panel shows that the KL divergence is minimized when the assumed prior probabilities are $P_1 = P_2 = 1/2$, such that non-informative priors were employed. Additionally, the KL divergence for $M = 12$ clients is larger than the one for $M = 6$ clients, which is also shown in the bottom panel where the KL divergence increases with M. The bottom panel depicts the KL as a function of M while fixing the assumed priors as $P_1 = 0.1$ and $P_1 = 0.4$. Again the plot shows that the KL is smaller for non-informative priors while the KL consistently increases with the number of clients.

Figure 6 shows the accuracy performance of posterior class distribution of both CIP and CIL with respect to the assumed prior P_1. The accuracy is a metric that measures the percentage of correct predictions over the total test samples observed. Similarly to the KL analysis in Figure 5, the accuracy of both approaches are similar for non-informative class priors $P_1 = P_2 = 1/2$. On the other hand, when the prior probability is far from the actual class prior probability, CIP's accuracy
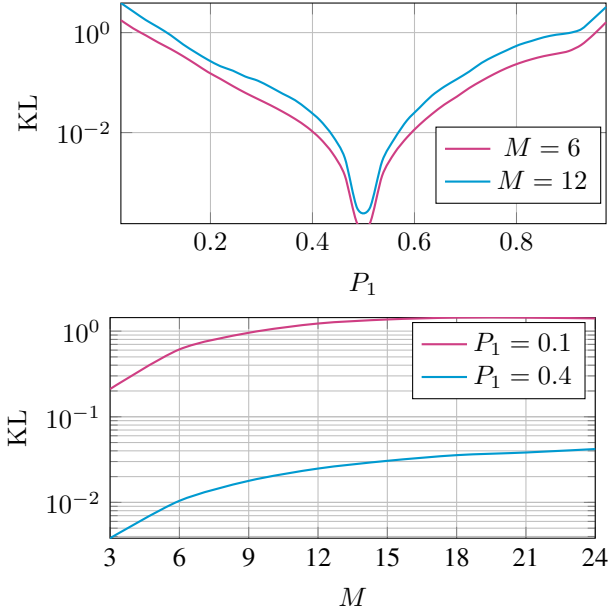
Fig. 5: KL divergence between CIL and CIP class posteriors as a function of (top) the class 1 prior probability P_1 and (bottom) number of clients M. Distributed classification problem from Section V-B.
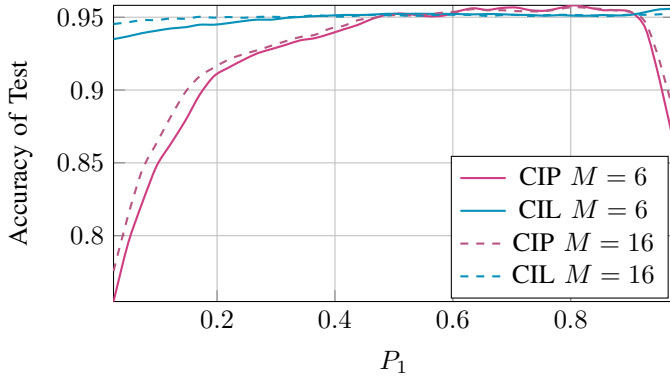

Fig. 6: The accuracy with respect to the class 1 prior probability for various numbers of clients M and fixed amount of total data.

is degraded since that mismatched prior is overused, whereas for CIL the performance does not drop dramatically. When the prior probability is close to real class distribution (that is, $P_1 = 0.6$ and $P_2 = 0.4$), both of these two methods achieve relatively decent results. Figure 7 shows the classification accuracy, this time in terms of the number of clients M and two values for P_1. The results show that CIL outperforms CIP, similarly to Figure 6.

C. Distributed learning of neural network classifiers

In this test we consider a set of M neural networks with the same structure and network parameters $\boldsymbol{\theta}$, each trained on observed local data. The objective is therefore to fuse the trained NNs into a global NN that can then be used for classification purposes, accounting for all local data through
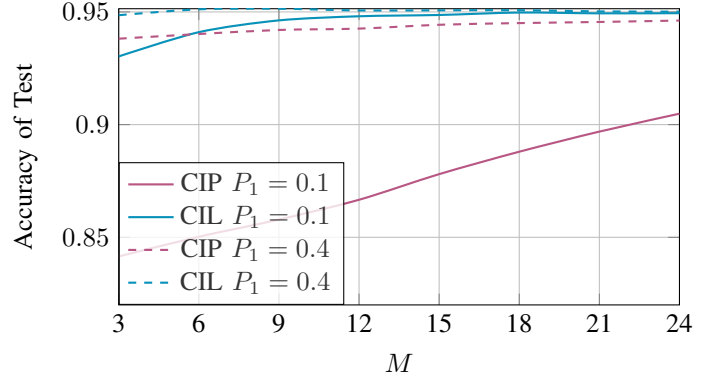

Fig. 7: Testing accuracy with respect to M under class 1 prior probabilities 0.1 and 0.4, for the classification problem in Section V-B.

a distributed training process. More precisely, the fusion is among the locally learned $\boldsymbol{\theta}$, as opposed to fusing the local class posterior as done in the experiments in Section V-B. Additionally, since the focus of this paper is on Bayesian data fusion, those are Bayesian NNs (BNN) in that their weights are treated as random variables. Therefore, training of a BNN involves inferring the joint posterior of its parameters given available data. The objective of this experiment is to locally train the BNNs, then fuse the local posteriors $p(\boldsymbol{\theta}|\mathcal{D}_m)$ to compute a global posterior $p(\boldsymbol{\theta}|\mathcal{D}_1, \ldots, \mathcal{D}_M)$. The use of Bayesian approaches within the context of data-driven models has been investigated previously and is given attention more recently [51], where for the sake of simplicity the weights in the BNN are typically assumed independent and normally distributed such that mean and covariance can characterize their posterior.

In this experiment, we used a similar classification dataset as in Section V-B, sampling from a normal distribution for each class $p(\mathcal{D}_m|\mathbf{x}, C = i) = \mathcal{N}(\mathbf{1}_i, \mathbf{I})$, where in this case the dimension of the problem was increased to 10 classes, $i = \{1, \ldots, 10\}$, while size of the data was kept to 10 as in Section V-B. The prior was normally distributed as $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, q_0\mathbf{I})$, with q_0 adjusting its variance. The BNN trained by the M local agents had a hidden layer of 64 neurons, the training epoch was set to 100, and the learning rate was 0.05. It is worth mentioning that the total amount of data is fixed, such that increasing M would have the impact of reducing the amount of local data available at each client. Given a local dataset, there are several methods that can be used to calculate the posterior distribution of the BNN parameters. Without loss of generality, we considered here a Laplace approximation of the BNN parameters [44], assuming the parameters are normally distributed.

Figure 8 shows the KL divergence between the parameters' posterior of both CIP and CIL solutions, as a function of the a priori variance q_0 (top panel for $M = 6$ and $M = 16$ clients) and the number of clients M (bottom panel for $q_0 = 4$ and $q_0 = 1.6$). From the top panel of this figure, we can see that behavior of the KL divergence with respect to q_0 or M are similar to those from V-A.
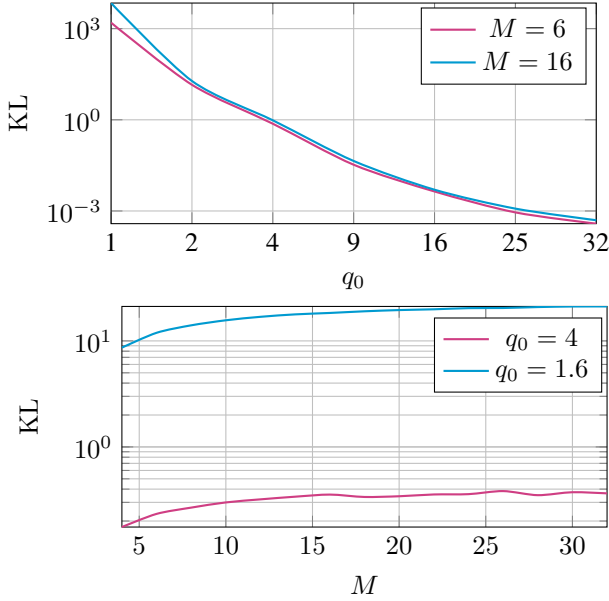
Fig. 8: KL divergence between CIL and CIP class posteriors as a function of (top) q_0 and (bottom) M for the distributed classification training in Section V-C.
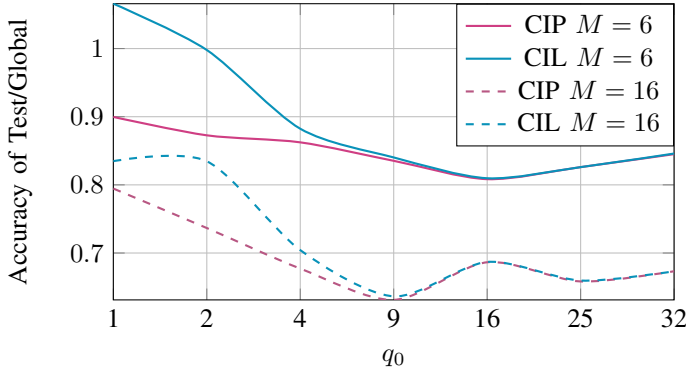

Fig. 9: Testing accuracy with respect to the prior variance for a fixed amount of total data across the M clients.

Figure 9 shows the fusion center NN testing accuracy with respect the prior variance q_0 for $M = 6$ and $M = 16$ clients. For the sake of clarity, and to facilitate the identification of the trend behavior, the reported metric is test accuracy of CIL (and CIP) divided by global accuracy (considered as a benchmark, using the global model for testing, which is trained by using all datasets jointly). As expected, as the prior becomes more informative, CIL outperforms CIP, and as M increases the local data becomes more scarce and the performance degrades. Analogously, Figure 10 shows the testing accuracy with respect to the number of clients M for different values of q_0, drawing similar conclusions.

D. Recursive update of classification problem using M neural networks

Results presented in the previous subsection only considered the case where classifier's parameters are fused once, after learned using local data. In many applications, however, one
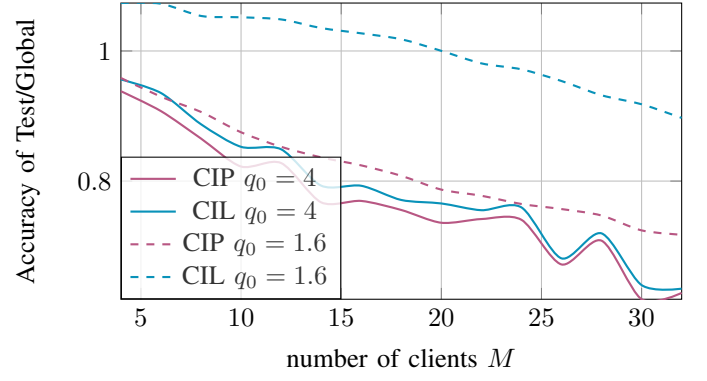

Fig. 10: Testing accuracy with respect to M for a different values of q_0.

aims at fusing models over time in an iterative procedure. For instance, this is the case of many distributed and/or federated methods where a common prior can be shared with several nodes and fused back recursively [36]. In this experiment, we consider the case of Bayesian federated model learning where learned distributions from distributed clients are fused by a central node in the vein of Figure 1. Locally, at the t-th communication round each client obtains their local posterior of the BNN model parameters using Bayes theorem as

$$p_t(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{p_{t-1}(\boldsymbol{\theta})p(\mathcal{D}_m|\boldsymbol{\theta})}{P(\mathcal{D}_m)}, \tag{21}$$

where $p_{t-1}(\boldsymbol{\theta})$ denotes the prior distribution of the parameters at t and the log-posterior can be written as

$$\ln p_t(\boldsymbol{\theta}|\mathcal{D}_m) = \ln p_{t-1}(\boldsymbol{\theta}) + \ln p(\mathcal{D}_m|\boldsymbol{\theta}) - \ln p(\mathcal{D}_m). \tag{22}$$

For simplicity, we consider the Gaussian assumption on the posterior $p_t(\boldsymbol{\theta}|\mathcal{D}_m) \approx \mathcal{N}(\boldsymbol{\theta}_{m,t}, \mathbf{C}_{m,t})$, similarly as in (7). Defining the loss for the m-th client as the log-posterior in (22), i.e., $J_{m,t}(\boldsymbol{\theta}) = \ln p_t(\boldsymbol{\theta}|\mathcal{D}_m)$, and using the Laplace approximation for its second term, the loss can be approximated as

$$J_{m,t}(\boldsymbol{\theta}) \approx \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{t-1})^\top \boldsymbol{\Lambda}_{t-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{t-1}) \tag{23}$$
$$+ \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{m,t}^{\mathrm{ML}})^\top \mathbf{H}_{m,t}(\boldsymbol{\theta} - \boldsymbol{\theta}_{m,t}^{\mathrm{ML}}) + \kappa,$$

where $p_{t-1}(\boldsymbol{\theta}) \approx \mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Lambda}_{t-1}^{-1})$ is the parameter prior at t, resulting from the CIL (or CIP) fusion of local posteriors at $k - 1$. The Laplace approximation of the likelihood $p(\mathcal{D}_m|\boldsymbol{\theta}) \approx \mathcal{N}(\boldsymbol{\theta}_{m,t}^{\mathrm{ML}}, \mathbf{H}_{m,t}^{-1})$ requires maximum likelihood (ML) estimation of the parameter, $\boldsymbol{\theta}_{m,t}^{\mathrm{ML}}$, possibly through a gradient method or other numerical optimization approach. A suitable estimate of the inverse covariance $\mathbf{H}_{m,t}^{-1}$ in the Laplacian approximation is known to be [44] the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$, which can be approximated [36] by $\mathcal{I}_m(\boldsymbol{\theta}) \approx \frac{1}{|\mathcal{D}_m|}\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_m} \nabla_\theta \log p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})\nabla_\theta \log p(\mathbf{y}|\mathbf{x},\boldsymbol{\theta})^\top$ for the m-th client. The κ gathers constant terms that are not related to $\boldsymbol{\theta}$ and thus do not contribute to the loss minimization. Then, the local parameter posterior mean for the m-th client can be estimated by solving the following optimization problem:

$$\boldsymbol{\theta}_{m,t} = \arg\max_{\boldsymbol{\theta}} J_{m,t}(\boldsymbol{\theta}), \tag{24}$$
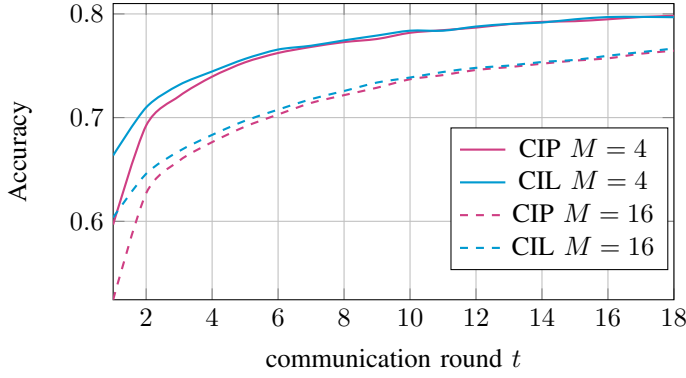
Fig. 11: Testing accuracy evolution with communication round for the recursive classification learning in Section V-D.
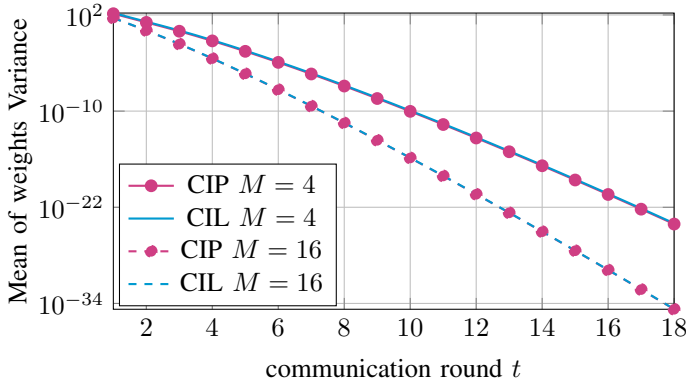

Fig. 12: Evolution over t of the mean variance of model parameters.

and the associated local covariance is [36]

$$\mathbf{C}_{m,t}^{-1} = \frac{\partial^2 J_m(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \approx \mathcal{I}_m(\boldsymbol{\theta}_{m,t}^{\text{ML}}) + \boldsymbol{\Lambda}_{t-1} \,. \tag{25}$$

After obtaining the local posterior parameters at t, $\boldsymbol{\theta}_{m,t}$ and $\mathbf{C}_{m,t}$, for $m = 1, \ldots, M$, these can be fused using either CIL and CIP approaches. This fused posterior would become $p_t(\boldsymbol{\theta}) \approx \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t^{-1})$ the new prior used at the next communication round $t + 1$.

Compared to previous classification experiments, this experiment considers a more complex synthetic dataset generated from a Gaussian mixture with a maximum of 4 Gaussian components, with 3 possible classes and 10 features dimension. A total of 600 *i.i.d.* training points and 300 testing points were generated. The considered BNN model is a multi-layer perceptron (MLP) with two hidden layers, containing 32 and 8 neurons respectively, with a learning rate of 0.01 at training.

The results in Figures 11 and 12 show the evolution of accuracy and the mean value of the variance of all parameters, as a function of the communication round t at which fusion happens. $M = 4$ and $M = 16$ are tested. In the experiments, the initial round is given with non-informative prior information. After several communicate rounds occur, the variance becomes very informative, as shown in Figure 12.

It is seen from Figure 11 that accuracy generally increases with the number of communication rounds, for both CIP and CIL. Also, the accuracy decreases inversely proportional to the number of clients M. The latter is a consequence of having less local data samples at each node, given that the overall amount is kept constant regardless of M. Another conclusion from the experiment is that CIL outperforms CIP, approaching each other when enough communication rounds are run. On the other hand, the average variance of the parameters produced by (25) is negligible, both approaches becoming over-confident.

VI. CONCLUSIONS

This paper investigates the impact of sharing a priori information of parameters of interest in a Bayesian data fusion context. The focus is on distributed learning of models for regression and classification purposes. We analyze two fusion possibilities in a distributed system, so-called conditionally independent posterior (CIP) and conditionally independent likelihoods (CIL). CIL assumes the local datasets are mutually conditionally independent and CIP is based on the approximation that the local posteriors are mutually independent. CIP is sometimes desirable due to its simplicity, however the assumption is compromised when prior information is shared among clients for which CIL is better suited. This article provided an analysis of both fusion approaches and their performance as a function of the uncertainty of the shared prior and the number of agents in the distributed setting. A deeper analysis was provided under the assumption that local posteriors can be approximated as Gaussian distributed. It is shown that both CIL and CIP converge to similar fusion results when the prior becomes non-informative. Another result from the analysis shows that, for the same amount of global data, when the number of collaborative agents increase the CIP solution degrades with respect to CIL. Or in other words, when the number of users decrease, both approaches become equivalent. A comparison of these two methods in different applications of distributed inference was provided. Namely, experimental results were discussed for distributed learning of linear regression and classification models; distributed learning of neural network classifiers; and a practical use case in the context of federated learning, where the prior information is updated recursively over time as opposed to only once. The results consistently show that CIL, which accounts for the shared prior information, generally outperforms CIP, as it was predicted by the analytical results presented in this paper.

APPENDIX A
BAYESIAN DATA FUSION UNDER GAUSSIAN DISTRIBUTIONS

CIL, in (2) is composed of three terms. Namely, i) the Product of Experts term, $\prod_{m=1}^{M} p(\boldsymbol{\theta}|\mathcal{D}_m)$, which we refer to as the CIP; ii) the product of priors, $p^{M-1}(\boldsymbol{\theta})$; and iii) the normalizing constant Z, which has no impact in obtaining the result we are interested in this appendix.

Under the Gaussian assumption, the first term is given by (14) following known results of products of Gaussian distributions [52], such that

$$\prod_{m=1}^{M} p(\boldsymbol{\theta}|\mathcal{D}_m) \propto \mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Lambda}}^{-1}) . \tag{26}$$

The second term is

$$p^{M-1}(\boldsymbol{\theta}) \propto \mathcal{N}(\boldsymbol{\theta}_0, \frac{\mathbf{C}_0}{M-1}) , \tag{27}$$

given that $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{C}_0)$.

The target distribution, $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$, is also a Gaussian with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ which can be computed from

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \propto \frac{\mathcal{N}(\boldsymbol{\theta}_0, \frac{\mathbf{C}_0}{M-1})}{\mathcal{N}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Lambda}}^{-1})}, \tag{28}$$

since the division of two Gaussian distributions is yet another Gaussian distribution up to a normalizing constant, since it belongs to the exponential family.

To calculate the desired mean and covariance, we can use the completing the square method, whereby the exponent in the Gaussian distribution is expanded as

$$-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\boldsymbol{\theta} - \boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} + \text{const.}, \tag{29}$$

where the symmetries of $\boldsymbol{\Lambda}$ are used and the constant represents the terms that are unrelated to $\boldsymbol{\theta}$.

The exponent in the right-hand side of (28) is composed of two quadratic terms of the form in (29), that is , consider the quadratic part:

$$-\frac{1}{2}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\mu}})^\top \widetilde{\boldsymbol{\Lambda}}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\mu}}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top (M-1)\mathbf{C}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0), \tag{30}$$

which can be rearranged in the form of the right-hand side of (29) as:

$$-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Lambda}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} = -\frac{1}{2}\boldsymbol{\theta}^\top \left(\sum_{m=1}^{M} \mathbf{C}_m^{-1} - \mathbf{C}_0^{-1}(M-1) \right) \boldsymbol{\theta}$$
$$+ \boldsymbol{\theta}^\top \left(\sum_{m=1}^{M} \mathbf{C}_m^{-1}\boldsymbol{\theta}_m - (M-1)\mathbf{C}_0^{-1}\boldsymbol{\theta}_0 \right), \tag{31}$$

where we used (15) and (16). From (31) we can identify the desired expressions for $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ in (11) and (10) respectively.

APPENDIX B
PROOF OF THEOREM 1

This appendix provides a general proof for the result by using arbitrary divergences. First, we discuss a general result that is independent of the assumed posterior distribution. Secondly, a complementary result under the Gaussian assumption is provided to provide more intuition.

Given the definitions in (2) and (5), the result is obtained straightforwardly by realizing that when the a priori $p(\boldsymbol{\theta})$ distribution is non-informative, this prior becomes independent of $\boldsymbol{\theta}$. As a consequence, the factor $\frac{p^{M-1}(\boldsymbol{\theta})}{Z}$ in (2) becomes a

normalizing constant that, necessarily, has to be the same as $\widetilde{Z}$ in (5) for it to be a proper distribution. Therefore making CIL and CIP identical, which would indeed make any divergence between both tend to 0.

The results above are valid for any assumed posterior distribution. However we are providing a complementary results here under the Gaussian posterior assumption, with the objective of providing additional intuitions and connecting to the case in Section III. Given the a priori distribution $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{C}_0)$ and under the assumption that $\mathbf{C}_0 = q_0 \mathbf{I}$, we can readily see from the results in Appendix A that the CIL posterior distribution $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ results in a mean of

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1} \left(\sum_{m=1}^{M} \mathbf{C}_m^{-1}\boldsymbol{\theta}_m - \frac{M-1}{q_0}\boldsymbol{\theta}_0 \right), \tag{32}$$

and a precision matrix of

$$\boldsymbol{\Lambda} = \sum_{m=1}^{M} \mathbf{C}_m^{-1} - \frac{M-1}{q_0}\mathbf{I} . \tag{33}$$

The result investigates the asymptotics of having a prior information that becomes non-informative, which for the choice of $p(\boldsymbol{\theta})$ here corresponds to the case where $q_0 \to \infty$ such that the Gaussian is wider. Therefore,

$$\lim_{q_0 \to \infty} \boldsymbol{\mu} = \left(\sum_{m=1}^{M} \mathbf{C}_m^{-1} \right)^{-1} \sum_{m=1}^{M} \mathbf{C}_m^{-1}\boldsymbol{\theta}_m, \tag{34}$$

$$\lim_{q_0 \to \infty} \boldsymbol{\Lambda} = \sum_{m=1}^{M} \mathbf{C}_m^{-1}, \tag{35}$$

and we can easily identify that $\lim_{q_0 \to \infty} \boldsymbol{\mu} = \widetilde{\boldsymbol{\mu}}$ and $\lim_{q_0 \to \infty} \boldsymbol{\Lambda} = \widetilde{\boldsymbol{\Lambda}}$ from (15) and (16), respectively. Since under the Gaussian assumption both CIL and CIP can be parameterized by their means and covariances, we can write

$$\lim_{q_0 \to \infty} p(\boldsymbol{\theta}|\mathcal{D}) = \widetilde{p}(\boldsymbol{\theta}|\mathcal{D}) , \tag{36}$$

since those statistics are asymptotically equivalent as shown in (15) and (16). As a consequence, their divergence tends to zero with $q_0 \to \infty$, independently of M, which proves the result.

APPENDIX C
PROOF OF THEOREM 2 UNDER KULLBACK-LEIBLER DIVERGENCE

Recall that the true posterior (referred to as CIL) was expressed in (2) as

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{\prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}, \tag{37}$$

where $p(\mathcal{D}) = \int \prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ and that the approximate posterior (denoted as CIP) was given in (5) as

$$\widetilde{p}(\boldsymbol{\theta}|\mathcal{D}) = \widetilde{Z} \prod_{m=1}^{M} p(\boldsymbol{\theta}|\mathcal{D}_m) = \frac{\prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p^M(\boldsymbol{\theta})}{p_M(\mathcal{D})}, \tag{38}$$

where $p_M(\mathcal{D}) = \int \prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p^M(\boldsymbol{\theta})d\boldsymbol{\theta}$ and $\widetilde{Z} = \frac{\prod_{m=1}^{M} p(\mathcal{D}_m)}{p_M(\mathcal{D})}$. With those definitions, let us define the KL divergence between both fusion rules as KL_M and manipulate it further

$$\mathrm{KL}_M \triangleq \mathrm{KL}_M(p(\boldsymbol{\theta}|\mathcal{D}) \,\|\, \widetilde{p}(\boldsymbol{\theta}|\mathcal{D})) \qquad (39)$$

$$= \int p(\boldsymbol{\theta}|\mathcal{D}) \log\left(\frac{p(\boldsymbol{\theta}|\mathcal{D})}{\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})}\right) d\boldsymbol{\theta}$$

$$= \int \frac{\prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \log\left(\frac{\frac{\prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}}{\frac{\prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p^M(\boldsymbol{\theta})}{p_M(\mathcal{D})}}\right) d\boldsymbol{\theta}$$

$$= \int \frac{\prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \log\left(\frac{p_M(\mathcal{D})}{p(\mathcal{D})}\frac{1}{p^{M-1}(\boldsymbol{\theta})}\right) d\boldsymbol{\theta}$$

$$= \int \frac{\prod_{m=1}^{M} p(\mathcal{D}_m|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$
$$\cdot \left(\log\left(\frac{p_M(\mathcal{D})}{p(\mathcal{D})}\right) - (M-1)\log(p(\boldsymbol{\theta}))\right) d\boldsymbol{\theta}$$

$$= \log\left(\frac{p_M(\mathcal{D})}{p(\mathcal{D})}\right) + (M-1)H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta})),$$

We also notice that the last term is the cross-entropy of the prior relative to the posterior distribution:

$$H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta})) = -\int \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}\log(p(\boldsymbol{\theta}))d\boldsymbol{\theta}, \quad (40)$$

such that $H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta})) > 0$.

We aim at showing that $\mathrm{KL}_{M+1} > \mathrm{KL}_M$, for which we will show that the following quantity is positive:

$$\mathrm{KL}_{M+1} - \mathrm{KL}_M \qquad (41)$$

$$= \log\left(\frac{p_{M+1}(\mathcal{D})}{p_M(\mathcal{D})}\right) + H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta}))$$

$$= \log\int p(\mathcal{D}|\boldsymbol{\theta})p^{M+1}(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$- \log\int p(\mathcal{D}|\boldsymbol{\theta})p^M(\boldsymbol{\theta})d\boldsymbol{\theta} + H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta}))$$

$$= \log\int \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}p^M(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$- \log\int \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}p^{M-1}(\boldsymbol{\theta})d\boldsymbol{\theta} + H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta}))$$

$$= \log\int p(\boldsymbol{\theta}|\mathcal{D})p^M(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$- \log\int p(\boldsymbol{\theta}|\mathcal{D})p^{M-1}(\boldsymbol{\theta})d\boldsymbol{\theta} + H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta}))$$

$$= \log S_M - \log S_{M-1} + H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta})),$$

where we used Bayes' rule in the last step, and defined $S_M = \int p(\boldsymbol{\theta}|\mathcal{D})p^M(\boldsymbol{\theta})d\boldsymbol{\theta}$. As we will discuss next, if $\log S_M$ is convex with respect to M, then we can show that (41) is positive, thus completing the proof.

Convexity proof for $\log S_M$: To demonstrate the convexity of $\log S_M$ note that it can be rewritten as $S_M = \log \mathbb{E}_p[a^M]$ for $a = p(\boldsymbol{\theta})$, and $p = p(\boldsymbol{\theta}|\mathcal{D})$. Furthermore, also note that a

is non-negative. Now we can show that functions of this type, $f(M) = \log \mathbb{E}_p[a^M]$, $a \geqslant 0$, are convex:

$$f(tM_1 + (1-t)M_2) \stackrel{(a)}{=} \log \mathbb{E}_p[v^t u^{t-1}]$$
$$\stackrel{(b)}{\leqslant} \log\left(\mathbb{E}_p\left[|v^t|^{\frac{1}{t}}\right]\right)^t \left(\mathbb{E}_p\left[|u^{1-t}|^{\frac{1}{(1-t)}}\right]\right)^{1-t}$$
$$\stackrel{(c)}{=} \log \mathbb{E}_p[v]^t \mathbb{E}_p[u]^{1-t}$$
$$= t\log \mathbb{E}_p[v] + (1-t)\log \mathbb{E}_p[u]$$
$$= tf(M_1) + (1-t)f(M_2).$$

In equality (a) we used $v = a^{M_1}$ and $u = a^{M_2}$. Inequality (b) results from the application of the Hölder's inequality for Lebesgue measures, where $\mathbb{E}[XY] \leqslant \mathbb{E}[|X|^z]^{\frac{1}{z}}\mathbb{E}[|Y|^q]^{\frac{1}{q}}$ and we made $t = 1/z$ and $(1-t) = 1/q$, for $z, q \in [1, \infty)$. Equality (c) follows from the fact that the quantities v and u are always positive, concluding the convexity proof for $\log S_M$.

Positiveness of (41): A direct consequence of the convexity of $\log S_M$, is that the difference $\log S_M - \log S_{M-1}$ can be bounded as $\log S_M - \log S_{M-1} \geq \log S_1 - \log S_0$, with equality if $M = 1$. Similarly, we can use the result to show that $\mathrm{KL}_{M+1} - \mathrm{KL}_M \geq \mathrm{KL}_2 - \mathrm{KL}_1$ is bounded when $M = 1$. These results are then used to finally show the positiveness of (41). Based on the convexity of $\log S_M$ and Jensen's inequality, we can show that

$$\log S_M - \log S_{M-1} + H(p(\boldsymbol{\theta}|\mathcal{D}), p(\boldsymbol{\theta})) \qquad (42)$$
$$= \log E_{p(\boldsymbol{\theta}|\mathcal{D})}\left[p^M(\boldsymbol{\theta})\right] - \log E_{p(\boldsymbol{\theta}|\mathcal{D})}\left[p^{M-1}(\boldsymbol{\theta})\right] d\boldsymbol{\theta}$$
$$- E_{p(\boldsymbol{\theta}|\mathcal{D})}\left[\log p(\boldsymbol{\theta})\right]$$
$$\geqslant_{(M=1)} \log E_{p(\boldsymbol{\theta}|\mathcal{D})}\left[p(\boldsymbol{\theta})\right] - E_{p(\boldsymbol{\theta}|\mathcal{D})}\left[\log p(\boldsymbol{\theta})\right]$$
$$\geqslant E_{p(\boldsymbol{\theta}|\mathcal{D})}\left[\log p(\boldsymbol{\theta})\right] - E_{p(\boldsymbol{\theta}|\mathcal{D})}\left[\log p(\boldsymbol{\theta})\right] = 0,$$

which concludes that $\mathrm{KL}_{M+1} \geq \mathrm{KL}_M$, thus proving Theorem 2 under KL divergence and arbitrary distributions. The following subsection in the appendix shows the result under the Gaussian assumption.

A. Proof under Gaussian distributions

A similar, although more tedious, result can be obtained under the more restrictive case of Gaussian distributed posteriors. Given the widespread use of Gaussian models, we provide this complementary result, which uses the Gaussian product results [52]. Similarly, we aim at showing that S_M is convex for the Gaussian case, such that the reasoning in equation (42) follows.

We are interested in the Gaussian assumption for both the prior $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}_0, \mathbf{C}_0)$ and the posterior $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ distributions. The product in S_M can be further manipulated as

$$p(\boldsymbol{\theta}|\mathcal{D})p^M(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right)S_{0,M}\mathcal{N}\left(\boldsymbol{\theta}_0, \frac{\mathbf{C}_0}{M}\right) \qquad (43)$$

$$= S_{0,M}S_{1,M}\mathcal{N}\left((M\mathbf{C}_0^{-1} + \boldsymbol{\Lambda})(\boldsymbol{\Lambda}\boldsymbol{\theta}_0 + \mathbf{C}_0^{-1}M\boldsymbol{\mu}),\right.$$
$$\left.(M\mathbf{C}_0^{-1} + \boldsymbol{\Lambda})^{-1}\right),$$

where $S_{0,M}$ and $S_{1,M}$ are scaling factors [52]

$$S_{0,M} = \frac{1}{(2\pi)^{\frac{M-1}{2}}} \sqrt{\left| \frac{\mathbf{C}_0}{M} (\mathbf{C}_0)^{-M} \right|}$$

$$\cdot \exp[-\frac{1}{2}(M\boldsymbol{\theta}_0^\top \mathbf{C}_0^{-1}\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^\top (\frac{\mathbf{C}_0}{M})^{-1}\boldsymbol{\theta}_0)] \quad (44)$$

$$= \frac{1}{(2\pi)^{\frac{d(M-1)}{2}}} |M\mathbf{C}_0^{M-1}|^{-1/2}$$

$$S_{1,M} = \frac{1}{\sqrt{(2\pi)^d \left| \frac{\mathbf{C}_0}{M} + \boldsymbol{\Lambda}^{-1} \right|}} \quad (45)$$

$$\cdot \exp\left[-\frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\mu})^\top \left(\frac{\mathbf{C}_0}{M} + \boldsymbol{\Lambda}^{-1} \right)^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}) \right],$$

such that computing the integral S_M reduces to a product of those scaling factors

$$S_M = S_{0,M} S_{1,M} \quad (46)$$

$$= (2\pi)^{-\frac{dM}{2}} \left| \frac{\mathbf{C}_0}{M} + \boldsymbol{\Lambda}^{-1} \right|^{-\frac{1}{2}} |M\mathbf{C}_0^{M-1}|^{-\frac{1}{2}}$$

$$\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\mu})^\top \left(\frac{\mathbf{C}_0}{M} + \boldsymbol{\Lambda}^{-1} \right)^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}) \right]$$

$$= (2\pi)^{-\frac{dM}{2}} |\mathbf{C}_0^M + M\mathbf{C}_0^{M-1}\boldsymbol{\Lambda}^{-1}|^{-\frac{1}{2}}$$

$$\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\mu})^\top \left(\frac{\mathbf{C}_0}{M} + \boldsymbol{\Lambda}^{-1} \right)^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}) \right]$$

$$= (2\pi)^{-\frac{d}{2}} |(2\pi)^d \mathbf{C}_0|^{-\frac{M-1}{2}} |\mathbf{C}_0 + M\boldsymbol{\Lambda}^{-1}|^{-\frac{1}{2}}$$

$$\exp\left[-\frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\mu})^\top \left(\frac{\mathbf{C}_0}{M} + \boldsymbol{\Lambda}^{-1} \right)^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}) \right],$$

and

$$\log S_M = -\frac{M-1}{2} \log |2(\pi)^d \mathbf{C}_0| - \frac{1}{2} \log |\mathbf{C}_0 + M\boldsymbol{\Lambda}^{-1}| \quad (47)$$

$$- \frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\mu})^\top \left(\frac{\mathbf{C}_0}{M} + \boldsymbol{\Lambda}^{-1} \right)^{-1} (\boldsymbol{\theta}_0 - \boldsymbol{\mu}) + \text{const} .$$

In order to show that (39) is a positive quantity, we will first show that $\log S_M - \log S_{M-1}$ is bounded. The objective is to show the $\log S_M - \log S_{M-1}$ increases with M, that is to say that $\log S_M$ is a convex function of M, in which case the boundness would follow with its minimum value being $\log S_1 - \log S_0$ when $M = 1$. The following derivation proves the convexity of $\log S_M$. Using basic matrix algebra results, the first derivative of $\log S_M$ is

$$\frac{\partial \log S_M}{\partial M} = -\frac{1}{2} \log |(2\pi)^d \mathbf{C}_0| - \frac{1}{2}\text{Tr}((\mathbf{C}_0 + M\boldsymbol{\Lambda}^{-1})^{-1}\boldsymbol{\Lambda}^{-1}) \quad (48)$$

$$- \frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\mu})^\top ((\mathbf{C}_0 + M\boldsymbol{\Lambda}^{-1})^{-1} \mathbf{C}_0 (\mathbf{C}_0 + M\boldsymbol{\Lambda}^{-1})^{-1})(\boldsymbol{\theta}_0 - \boldsymbol{\mu}),$$

and its second derivative is

$$\frac{\partial^2 \log S_M}{\partial M^2} = \frac{1}{2}\text{Tr}((\mathbf{C}_0\boldsymbol{\Lambda} + M)^{-2} + \quad (49)$$

$$\frac{1}{2}(\boldsymbol{\theta}_0 - \boldsymbol{\mu})^\top [(\mathbf{C}_0 + M\boldsymbol{\Lambda}^{-1})^{-1} \boldsymbol{\Lambda}^{-1} (\mathbf{C}_0 + M\boldsymbol{\Lambda}^{-1})^{-1}](\boldsymbol{\theta}_0 - \boldsymbol{\mu})$$

Since the first term in equation (49) is larger than 0 and the second term is larger or equal than 0, it follows that $\log S_M$ is a convex function and the discussion in equation (42) holds for the Gaussian case. While the result of the theorem for arbitrary distributions is mode general, the above result under Gaussian distributions is complementary and provides consistent results.

APPENDIX D
PROOF OF THEOREM 2 UNDER χ^2 DIVERGENCE

The results in Theorem 2 also holds when the χ^2 divergence [53] is considered. Before showing the result let us first express the χ^2 divergence as a function of the number of clients M:

$$\chi_M^2(p(\boldsymbol{\theta}|\mathcal{D}) \| \widetilde{p}(\boldsymbol{\theta}|\mathcal{D})) = \int \frac{(p(\boldsymbol{\theta}|\mathcal{D}) - \widetilde{p}(\boldsymbol{\theta}|\mathcal{D}))^2}{\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} \quad (50)$$

$$= \int \frac{p^2(\boldsymbol{\theta}|\mathcal{D})}{\widetilde{p}(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} - 1$$

$$= \int \frac{\frac{p^2(\mathcal{D}|\boldsymbol{\theta})p^2(\boldsymbol{\theta})}{p^2(\mathcal{D})}}{\frac{p(\mathcal{D}|\boldsymbol{\theta})p^M(\boldsymbol{\theta})}{p_M(\mathcal{D})}} d\boldsymbol{\theta} - 1$$

$$= \int p(\mathcal{D}|\boldsymbol{\theta})p^{2-M}(\boldsymbol{\theta}) \frac{p_M(\mathcal{D})}{p^2(\mathcal{D})} d\boldsymbol{\theta} - 1$$

$$= \int p(\boldsymbol{\theta}|\mathcal{D})p^{1-M}(\boldsymbol{\theta}) \frac{p_M(\mathcal{D})}{p(\mathcal{D})} d\boldsymbol{\theta} - 1$$

$$= \int p(\boldsymbol{\theta}|\mathcal{D})p^{M-1}(\boldsymbol{\theta}) d\boldsymbol{\theta} \int p(\boldsymbol{\theta}|\mathcal{D})p^{1-M}(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1$$

$$= \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})} [p(\boldsymbol{\theta})^{M-1}] \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})} [p(\boldsymbol{\theta})^{1-M}] .$$

Now, we focus on proving that $\chi_{M+1}^2 - \chi_M^2 \geq 0$ for all $M \geq 1$, where we omit the arguments in (50) for brevity. For this, first note that χ_M^2 is log-convex, that is

$$\log \chi_M^2 = \log \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})} [p(\boldsymbol{\theta})^{M-1}] + \log \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})} [p(\boldsymbol{\theta})^{1-M}]$$

is convex. The convexity of $\log \chi_M^2$ follows from the fact that (i) both $\log \mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})} [p(\boldsymbol{\theta})^{M-1}]$ and $\log [p(\boldsymbol{\theta})^{1-M}]$ fall in the category of functions $f(M) = \log \mathbb{E}_p[a^M]$, $a \geq 0$, for which we proved convexity in Appendix C; and (ii) the fact that the sum of convex functions is also convex [54].

Another necessary result for our proof relies on the fact that the log function is monotonically increasing for non-negative arguments such as the χ^2 divergence. Thus,

$$\log \chi_{M+1}^2 - \log \chi_M^2 \geq 0 \iff \chi_{M+1}^2 - \chi_M^2 \geq 0 .$$

The above fact implies that if one can show that $\log \chi_{M+1}^2 - \log \chi_M^2 \geq 0$, the proof is complete. Now, we can apply these results to complete our proof as follows:

$$\log \chi_{M+1}^2 - \log \chi_M^2 \stackrel{(a)}{=} \log \left(\mathbb{E}_p [a^M] \mathbb{E}_p [a^{-M}] \right) \quad (51)$$

$$- \log \left(\mathbb{E}_p [a^{M-1}] \mathbb{E}_p [a^{-(M-1)}] \right)$$

$$\stackrel{(b)}{\underset{(M=1)}{\geq}} \log \left(\mathbb{E}_p[a] \mathbb{E}_p[a^{-1}] \right)$$

$$= \log \left(\mathbb{E}_p \left[\left(a^{\frac{1}{2}} \right)^2 \right] \mathbb{E}_p \left[\left(a^{-\frac{1}{2}} \right)^2 \right] \right)$$

$$\stackrel{(c)}{\geq} \log \left(\left| \mathbb{E}_p \left[a^{\frac{1}{2}} a^{-\frac{1}{2}} \right] \right|^2 \right) = 0$$

where in Equality (a) we defined $a = p(\boldsymbol{\theta})$ and $p = p(\theta|\mathcal{D})$ for brevity of notation. In Inequality (b) we exploited the

convexity of $\log \chi_M^2$ in the same way done in Appendix C, that is, $\log \chi_M^2 - \log \chi_{M-1}^2 \geqslant \log \chi_1^2 - \log \chi_0^2$. In the last step we applied the Cauchy-Schwarz to obtain Inequality (c). Finally, the result above implies that

$$\chi_{M+1}^2 \geqslant \chi_M^2, \quad \forall M \geqslant 1$$

which concludes the proof.

REFERENCES

[1] D. Dardari, P. Closas, and P. M. Djurić, "Indoor tracking: Theory, methods, and technologies," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 4, pp. 1263–1278, 2015.

[2] W. Ding, X. Jing, Z. Yan, and L. T. Yang, "A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion," *Information Fusion*, vol. 51, pp. 129–144, 2019.

[3] J. Duník, S. K. Biswas, A. G. Dempster, T. Pany, and P. Closas, "State estimation methods in navigation: overview and application," *IEEE Aerospace and Electronic Systems Magazine*, vol. 35, no. 12, pp. 16–31, 2020.

[4] D. L. Hall and S. A. McMullen, *Mathematical techniques in multisensor data fusion*. Artech House, 2004.

[5] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Information Fusion*, vol. 57, pp. 115–129, 2020.

[6] F. White, "Data fusion lexicon," 1991.

[7] M. M. Kokar, J. A. Tomasik, and J. Weyman, "Formalizing classes of information fusion systems," *Information Fusion*, vol. 5, no. 3, pp. 189–202, 2004.

[8] F. Castanedo, "A review of data fusion techniques," *The scientific world journal*, vol. 2013, 2013.

[9] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[10] M. E. Campbell and N. R. Ahmed, "Distributed data fusion: Neighbors, rumors, and the art of collective knowledge," *IEEE Control Systems Magazine*, vol. 36, no. 4, pp. 83–109, 2016.

[11] H. R. Hashemipour, S. Roy, and A. J. Laub, "Decentralized structures for parallel Kalman filtering," *IEEE Transactions on automatic control*, vol. 33, no. 1, pp. 88–94, 1988.

[12] M. S. Mahmoud and H. M. Khalid, "Distributed Kalman filtering: a bibliographic review," *IET Control Theory & Applications*, vol. 7, no. 4, pp. 483–501, 2013.

[13] V. Tresp, "A Bayesian committee machine," *Neural computation*, vol. 12, no. 11, pp. 2719–2741, 2000.

[14] T. Bailey, S. Julier, and G. Agamennoni, "On conservative fusion of information with unknown non-gaussian dependence," in *2012 15th International Conference on Information Fusion*. IEEE, 2012, pp. 1876–1883.

[15] M. Deisenroth and J. W. Ng, "Distributed Gaussian processes," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1481–1490.

[16] H. B. McMahan *et al.*, "Advances and open problems in federated learning," *Foundations and Trends® in Machine Learning*, vol. 14, no. 1, 2021.

[17] S. Ji, T. Saravirta, S. Pan, G. Long, and A. Walid, "Emerging trends in federated learning: From model fusion to federated x learning," *ArXiv*, vol. abs/2102.12920, 2021.

[18] P. Wu, T. Imbiriba, J. Park, S. Kim, and P. Closas, "Personalized Federated Learning over non-IID Data for Indoor Localization," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2021, pp. 421–425.

[19] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[20] Y. Bar-Shalom, "On the track-to-track correlation problem," *IEEE Transactions on Automatic control*, vol. 26, no. 2, pp. 571–572, 1981.

[21] M. A. Bakr and S. Lee, "Distributed multisensor data fusion under unknown correlation and data inconsistency," *Sensors*, vol. 17, no. 11, p. 2472, 2017.

[22] K.-C. Chang, C.-Y. Chong, and S. Mori, "Analytical and computational evaluation of scalable distributed fusion algorithms," *IEEE transactions on Aerospace and Electronic Systems*, vol. 46, no. 4, pp. 2022–2034, 2010.

[23] T. Li, H. Fan, J. García, and J. M. Corchado, "Second-order statistics analysis and comparison between arithmetic and geometric average fusion: Application to multi-sensor target tracking," *Information Fusion*, vol. 51, pp. 233–243, 2019.

[24] S. Julier and J. K. Uhlmann, "General decentralized data fusion with covariance intersection," in *Handbook of multisensor data fusion*. CRC Press, 2017, pp. 339–364.

[25] H. D. Nguyen, L. R. Lloyd-Jones, and G. J. McLachlan, "A universal approximation theorem for mixture-of-experts models," *Neural computation*, vol. 28, no. 12, pp. 2585–2593, 2016.

[26] A. A. Saucan, V. Elvira, P. K. Varshney, and M. Z. Win, "Information fusion via importance sampling," *IEEE Transactions on Signal and Information Processing over Networks*, 2023.

[27] Z. Xing, Y. Xia, L. Yan, K. Lu, and Q. Gong, "Multisensor distributed weighted Kalman filter fusion with network delays, stochastic uncertainties, autocorrelated, and cross-correlated noises," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no. 5, pp. 716–726, 2016.

[28] M. Alimadadi, M. Stojanovic, and P. Closas, "Object tracking in random access networks: A large-scale design," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9784–9792, 2020.

[29] D. Luengo, L. Martino, V. Elvira, and M. Bugallo, "Efficient linear fusion of partial estimators," *Digital Signal Processing*, vol. 78, 03 2018.

[30] ——, "Bias correction for distributed Bayesian estimators," in *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2015, pp. 253–256.

[31] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[32] Q.-V. Pham, K. Dev, P. K. R. Maddikunta, T. R. Gadekallu, T. Huynh-The *et al.*, "Fusion of federated learning and industrial internet of things: a survey," *arXiv preprint arXiv:2101.00798*, 2021.

[33] I. Achituve, A. Shamsian, A. Navon, G. Chechik, and E. Fetaya, "Personalized federated learning with Gaussian processes," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[34] A. Lalitha, O. C. Kilinc, T. Javidi, and F. Koushanfar, "Peer-to-peer federated learning on graphs," *arXiv preprint arXiv:1901.11173*, 2019.

[35] H.-Y. Chen and W.-L. Chao, "Fedbe: Making Bayesian model ensemble applicable to federated learning," *arXiv preprint arXiv:2009.01974*, 2020.

[36] L. Liu, F. Zheng, H. Chen, G.-J. Qi, H. Huang, and L. Shao, "A Bayesian Federated Learning Framework with Online Laplace Approximation," *arXiv preprint arXiv:2102.01936*, 2021.

[37] J. Park, J. Moon, T. Kim, P. Wu, T. Imbiriba, P. Closas, and S. Kim, "Federated learning for indoor localization via model reliability with dropout," *IEEE Communications Letters*, pp. 1–1, 2022.

[38] G. Koliander, Y. El-Laham, P. M. Djurić, and F. Hlawatsch, "Fusion of Probability Density Functions," *Proceedings of the IEEE*, vol. 110, no. 4, pp. 404–453, 2022.

[39] H. Liu, J. Cai, Y. Wang, and Y. S. Ong, "Generalized robust Bayesian committee machine for large-scale Gaussian process regression," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3131–3140.

[40] H. Liu, Y.-S. Ong, X. Shen, and J. Cai, "When Gaussian process meets big data: A review of scalable GPs," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 11, pp. 4405–4423, 2020.

[41] K. Kim and G. Shevlyakov, "Why Gaussianity?" *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 102–113, March 2008.

[42] K. Ito and K. Xiong, "Gaussian Filters for Nonlinear Filtering Problems," *IEEE Trans. Autom. Control*, vol. 45, no. 5, May 2000.

[43] F. Lindgren and H. Rue, "Bayesian spatial modelling with R-INLA," *Journal of statistical software*, vol. 63, pp. 1–25, 2015.

[44] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.

[45] L. Pardo, *Statistical inference based on divergence measures*. CRC press, 2018.

[46] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo, "Generalized multiple importance sampling," *Statistical Science*, vol. 34, no. 1, pp. 129–155, 2019.

[47] S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart, "Importance sampling: Intrinsic dimension and computational cost," *Statistical Science*, pp. 405–431, 2017.

[48] J. Míguez, "On the performance of nonlinear importance samplers and population Monte Carlo schemes," in *2017 22nd International Conference on Digital Signal Processing (DSP)*. IEEE, 2017, pp. 1–5.

[49] V. Elvira and L. Martino, "Advances in importance sampling," *Wiley StatsRef: Statistics Reference Online*, pp. 1–22, 2021.

[50] B. J. Kleijn and A. W. van der Vaart, "The Bernstein-von-Mises theorem under misspecification," *Electronic Journal of Statistics*, vol. 6, pp. 354–381, 2012.

[51] V. Fortuin, A. Garriga-Alonso, F. Wenzel, G. Rätsch, R. Turner, M. van der Wilk, and L. Aitchison, "Bayesian neural network priors revisited," *arXiv preprint arXiv:2102.06571*, 2021.

[52] P. Bromiley, "Products and convolutions of Gaussian probability density functions," *Tina-Vision Memo*, vol. 3, no. 4, p. 1, 2003.

[53] F. Nielsen and R. Nock, "On the chi square and higher-order chi distances for approximating f-divergences," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 10–13, 2013.

[54] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.