Non-asymptotic Concentration Rates in Cooperative Learning Part I: Variational Non-Bayesian Distributed Learning

César A. Uribe, Alex Olshevsky, and Angelia Nedić,

***Abstract*—We study the problem of cooperative inference where a group of agents interact over a network and seeks to estimate a joint parameter that best explains a set o network-wide observations using local information only. Agents do not know the network topology or the observations of other agents. We explore a variational interpretation of the Bayesian posterior and its relation to the stochastic mirror descent algorithm to prove that, under appropriate assumptions, the beliefs generated by the proposed algorithm concentrate around the true parameter exponentially fast. In Part I of this two-part paper series, we focus on providing a variation approach to distributed Bayesian filtering. Moreover, we develop explicit and computationally efficient algorithms for observation models in exponential families. We provide a novel non-asymptotic belief concentration analysis for distributed non-Bayesian learning on finite hypotheses sets. This new analysis method is the basis for the results presented in Part II. We provide the first non-asymptotic belief concentration rate analysis for distributed non-Bayesian learning over networks on compact hypotheses sets in Part II. Additionally, we provide extensive numerical analysis for various distributed inference tasks on networks for observational models in the exponential distributions family.**

***Index Terms*—Distributed Inference, non-Bayesian social learning, estimation over networks, non-asymptotic rates.**

I. Introduction

The increasing amount of data generated by recent applications of distributed systems such as social media, sensor networks, and cloud-based databases has brought considerable attention to distributed data processing, in particular the design of distributed algorithms that take into account the communication constraints and make coordinated decisions in a distributed manner [1]–[11]. In a distributed system, interactions between agents are usually constrained by the network structure and agents can only use locally available information. This contrasts with centralized approaches where all information and computation resources are available at a single location [12]–[15].

One traditional problem in decision-making is that of parameter estimation. Given a set of noisy observations coming from a joint distribution one would like to estimate a parameter or distribution that minimizes a certain loss function. For example, Maximum a Posteriori (MAP) or Minimum Least Squared Error (MLSE) estimators fit a parameter to some model of the observations. Both, MAP and MLSE estimators require some form of Bayesian posterior computation based on models that explain the observations for a given parameter. Computation of such a posteriori distributions depends on having exact models about the likelihood of the corresponding observations. This is one of the main difficulties of using Bayesian approaches in a distributed setting. A fully Bayesian approach is not possible because full knowledge of the network structure, or of other agents' likelihood models, may not be available [16]–[18].

Following the seminal work of Jadbabaie et al. in [1], [19], [20], there have been many studies of distributed non-Bayesian update rules over networks. In this case, agents are assumed to be boundedly rational (i.e., they fail to aggregate information in a fully Bayesian way [21]). Proposed non-Bayesian algorithms involve an aggregation step, typically consisting of weighted geometric or arithmetic average of the received beliefs [7], [22]–[25], and a Bayesian update with the locally available data [18], [26]. Lalitha et al. [27], Qipeng et al. [28], [29], Shahrampour et al. [20], [30], [31] and Rahimian et al. [32] have proposed variations of the non-Bayesian approach and proved consistent, geometric and non-asymptotic convergence rates for a general class of distributed algorithms; from asymptotic analysis to non-asymptotic bounds [33], [34], time-varying directed graphs [35]. Su et al. [36] have also considered adversarial agents and transmission and node failures. Constant elasticity of substitution models [37], minimum operators [38], [39], and uncertain models [] have been also studied. See [40] and [41] for an extended literature review.

We build upon the work in [42] on non-asymptotic behaviors of Bayesian estimators to derive new non-asymptotic concentration results for distributed learning algorithms. In contrast to the existing results which assume a finite hypothesis set, in this paper we extend the framework to compact sets of hypotheses. Our results show that in general, the network structure will induce a transient time after which all agents learn at a network independent rate, and this rate is geometric.

The contributions of this paper (Part I) are as follows:

- We provide a variational analysis of Bayesian posterior and derive an optimization problem for which the posterior is a step of the Stochastic Mirror Descent method.
- We use a variational interpretation to propose a distributed Stochastic Mirror Descent method for distributed learning. Moreover, we specialize the proposed algorithm

A. Nedić is with the School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, 85287 USA e-mail: angelia.nedich@asu.edu.

A. Olshevsky is with the Department of ECE and Division of Systems Engineering, Boston University, Boston, MA, 02215 USA e-mail: alexols@bu.edu.

C.A. Uribe is with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, 77006 USA e-mail: cauribe@rice.edu.

to parametric models of an exponential family which results in especially simple updates.
- We derive novel analysis methods to prove high probability non-asymptotic bounds for the convergence rate for the case of finite hypothesis sets. We show that this distributed learning algorithm concentrates the beliefs of all agents around the true parameter at an exponential rate.

The results in Part I serve as basis for Part II of this paper series where we analyze the case where the parameter spaces are compact. A subset of the problem description and a weaker set of results was presented in [43]. However, in this paper series, we extend such results with a specific treatment of the distributed inference problem for parametric estimation in the exponential family. Theorem statements and proofs have been extended.

The rest of this paper is organized as follows. Section II introduces the problem setup, it describes the networked observation model and the inference task. Section III presents a variational analysis of the Bayesian posterior, shows the implicit representation of the posterior as steps in a stochastic program and extends this program to the distributed setup. Section IV specializes the proposed distributed learning protocol to the case of observation models that are members of the exponential family. Section V shows our main results about the exponential concentration of beliefs around the true parameter. Section V begins by gently introducing our techniques by proving a concentration result in the case of countably many hypotheses, before turning to our main focus: the case when the set of hypotheses is a compact subset of $\mathbb{R}^d$. Finally, conclusions, open problems, and potential future work are discussed.

Notation: Random variables are denoted with upper-case letters, e.g. X, while the corresponding lower-case are used for their realizations, e.g. x. Time indices are denoted by subscripts, and the letter k or t is generally used. Agent indices are denoted by superscripts, and the letters i or j are used. We write $[A]_{ij}$ or a_{ij} to denote the entry of a matrix A in its i-th row and j-th column. We use A' for the transpose of a matrix A, and x' for the transpose of a vector x. The complement of a set B is denoted as B^c.

II. Problem Setup

We begin by introducing the learning problem from a centralized perspective, where all information is available at a single location. Later, we will generalize the setup to the distributed setting where only partial and distributed information is available.

Assume that we observe a sequence of independent random variables $X_1, X_2, \ldots$, all taking values in some measurable space $(\mathcal{X}, \mathcal{A})$ and identically distributed with a common *unknown* distribution P on $\mathcal{X}$, i.e. $X_k \sim P$ for all k. In addition, we have a statistical model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ composed by a parametrized family of probability measures on the sample space $(\mathcal{X}, \mathcal{A})$, where the map $\Theta \to \mathscr{P}$ from parameter to distribution is injective. Moreover, all distributions in the model are dominated[1] by a σ-finite measure λ, with corresponding densities $p_\theta = dP_\theta/d\lambda$[2]. Assume also that the model $\mathscr{P}$ is well-specified, thus *there exists a* θ^* *such that* $P_{\theta^*} = P$. The objective is to estimate θ^* based on the sequence of received observations $x_1, x_2, \ldots$. For example, given a random variable X, the maximum likelihood estimator (MLE) can be defined as

$$\hat{\theta}(X) = \arg\sup_{\theta\in\Theta} p_\theta(X) = \arg\sup_{P\in\mathcal{P}} p(X).$$

Following a Bayesian approach, the parameter is represented as a random variable ϑ on the set Θ is equipped with a σ-algebra $\mathcal{T}$ and a prior probability measure μ_0 on the measurable space $(\Theta, \mathcal{T})$. Moreover, we assume the existence of a probability measure Π on the product space $(\mathcal{X} \times \Theta)$ with σ-algebra $(\mathcal{A} \times \mathcal{T})$. Therefore one can pair the elements of the parametric model with the conditional distributions $\Pi_{X|\vartheta}$. Furthermore, the densities $p_\theta(x)$ are measurable functions of θ for any $x \in \mathcal{X}$. We then define the belief μ_k as the posterior distribution given the sequence of observations up to time k, i.e.,

$$\mu_k(B) = \Pi(\vartheta \in B \mid X_1, \ldots, X_k) = \frac{\int_B \prod_{t=1}^k p_\theta(X_t) d\mu_0(\theta)}{\int_\Theta \prod_{t=1}^k p_\theta(X_t) d\mu_0(\theta)}. \tag{1}$$

for all $B \in \mathcal{T}$ (note that we used the independence of the observations at each time step).

Assuming that all observations, up to time k, are readily available at a centralized location, under appropriate conditions, the recursive Bayesian posterior in Eq. (1) will be consistent in the sense that the beliefs μ_k will concentrate around θ^*; see [44], [45], and [46] for a formal statement. Furthermore, several authors have studied the rate at which this concentration occurs, in both asymptotic and non-asymptotic regimes [42], [47], [48].

Now consider the case where there is a network of n agents observing the process $X_1, X_2, \ldots$, where X_k is now a random vector belonging to the product space $\prod_{i=1}^n \mathcal{X}^i$ and $X_k = [X_k^1, X_k^2, \ldots, X_k^n]'$. Specifically, agent i observes the sequence $X_1^i, X_2^i, \ldots$, where X_k^i is now distributed according to an unknown distributions P^i, effectively making $X_k \sim \boldsymbol{P} = \prod_{i=1}^n P^i$. The statistical model is now distributed, where each agent agent i has a private family of distributions $\mathscr{P}^i = \{P_\theta^i : \theta \in \Theta\}$ it would like to fit to the observations. However, the goal is for *all* agents to agree on a *single* θ that best explains the complete set of observations instead of their local observations only. In other words, the agents collaboratively seek to find θ^* such that $\boldsymbol{P}_{\theta^*} = \prod_{i=1}^n P_{\theta^*}^i = \prod_{i=1}^n P^i = \boldsymbol{P}$.

Agents interact over a network defined by an undirected graph $\mathcal{G} = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of agents and E is a set of undirected edges, i.e., $(i, j) \in E$ if and only if agents i and j can communicate with each

[1] A measure μ is dominated by (or absolutely continuous with respect to) a measure λ if $\lambda(B) = 0$ implies $\mu(B) = 0$ for every measurable set B.

[2] Without loss of generality we will further assume that $\int_{\mathcal{X}} d\lambda(x) = 1$, this will only require our distributions to be absolutely continuous with respect to such measure.

other. We study a simple interaction model where, at each step, agents exchange their beliefs with their neighbors in the graph. Thus at every time step k, agent i will receive the sample x_k^i from X_k^i as well as the beliefs of its neighboring agents, i.e., it will receive μ_{k-1}^j for all j such that $(i,j) \in E$. Applying a fully Bayesian approach runs into some obstacles in this setting, we assume agents know neither the network topology nor the private family of distributions of other agents. Our goal is to design a learning procedure that is both distributed and consistent. That is, we are interested in a belief update algorithm that aggregates information in a non-Bayesian manner and guarantees that the beliefs of all agents will concentrate around θ^*.

As a motivating example, consider the problem of distributed source localization [49], [50]. In this scenario, a network of n agents receives noisy measurements of the distance to a source. The sensing capabilities of each sensor might be limited to a specific region. The group objective is to identify the location of the source jointly. Figure 1 shows a group of 7 agents (circles) seeking to localize a source (star). There is an underlying graph that indicates which nodes can exchange messages. Moreover, each node has a sensing region indicated by the dashed circle around it. Each agent observes signals proportional to the distance to the target. Since a target cannot be localized effectively from a single measure of the distance, agents must cooperate to have any hope of achieving proper localization. For more details on the problem, as well as simulations of the several discrete learning rules, we refer the reader to our earlier paper [33] dealing with the case when the set Θ is finite.

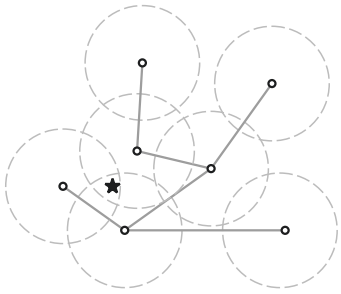

Fig. 1: Distributed source localization example.

III. A VARIATIONAL APPROACH TO DISTRIBUTED BAYESIAN FILTERING

In this section, we make the observation that the posterior in Eq. (1) corresponds to an iteration of a first-order optimization algorithm, namely Stochastic Mirror Descent [51]–[54]. Closely related variational interpretations of Bayes' rule are well-known, and in particular have been given in [55]–[57]. The specific connection to Stochastic Mirror Descent has not been noted, as far as we are aware of. This connection will serve to motivate a distributed learning method which will be the main focus of the paper.

A. Bayes' rule as Stochastic Mirror Descent

Suppose we want to solve the following optimization problem

$$\min_{\theta\in\Theta} F(\theta) \triangleq D_{KL}(P\|P_\theta), \tag{2}$$

where P is an unknown distribution and $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is a parametrized family of distributions. Here, $D_{KL}(P\|Q)$ is the Kullback-Leibler (KL) divergence[3] between distributions P and Q.

First note that we can rewrite the optimization problem in Eq. (2) as

$$\begin{aligned}\min_{\theta\in\Theta} D_{KL}(P\|P_\theta) &= \min_{\pi\in\Delta_\Theta} \mathbb{E}_\pi D_{KL}(P\|P_\vartheta) \qquad \text{where } \vartheta\sim\pi \\ &= \min_{\pi\in\Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P\left[-\log\frac{dP_\vartheta(X)}{dP(X)}\right] \\ &\qquad \text{where } \vartheta\sim\pi, X\sim P,\end{aligned}$$

where Δ_Θ is the set of all possible distributions on the parameter space Θ. Since the distribution P does not depend on ϑ, it follows that

$$\begin{aligned}&\arg\min_{\pi\in\Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P\left[-\log\frac{dP_\vartheta(X)}{dP(X)}\right] \\ &\quad= \arg\min_{\pi\in\Delta_\Theta} \mathbb{E}_\pi \mathbb{E}_P\left[-\log p_\vartheta(X)\right] \\ &\quad= \arg\min_{\pi\in\Delta_\Theta} \mathbb{E}_P \mathbb{E}_\pi\left[-\log p_\vartheta(X)\right].\end{aligned} \tag{3}$$

The equality in Eq. (3), where we exchange the order of the expectations, follows from the Fubini-Tonelli theorem. Clearly, if θ^* minimizes Eq. (2), then a distribution π^* which puts all the mass on θ^* (i.e. $\pi^*(\vartheta=\theta^*)=1$) minimizes Eq. (3).

The difficulty in evaluating the objective function in Eq. (3) lies in the fact that the distribution P is unknown. A generic approach to solving such problems is using algorithms from stochastic approximation methods, where the objective is minimized by constructing a sequence of gradient-based iterates whereby the true gradient of the objective (which is not available) is replaced with a gradient sample that is available at a given time.

A particular method that is relevant for the solution of stochastic programs as in Eq. (3) is the *stochastic mirror descent* method [51], [52], [58], [59]. In particular, recall that the mirror descent method to find the minimum of a function $f(x)$ performs the update

$$x_{k+1} \in \arg\min\left\{\nabla f(x_k)'x + \frac{1}{\alpha_k} D(x, x_k)\right\},$$

where $D(\cdot,\cdot)$ is a specific Bregman divergence. Moreover, note that (3) is linear in π, this the derivative with respect to π is $\mathbb{E}_P\left[-\log p_\vartheta(X)\right]$. Finally, we use the stochastic approximation provided by the current sample x_{k+1} of X. Therefore, the stochastic mirror descent approach constructs a sequence of densities $\{d\mu_k\}$, as follows:

$$d\mu_{k+1} = \arg\min_{\pi\in\Delta_\Theta}\left\{\langle -\log p_\theta(x_{k+1}), \pi\rangle + \frac{1}{\alpha_k} D_w(\pi, d\mu_k)\right\}, \tag{4}$$

[3] $D_{KL}(P\|Q)$ between distributions P and Q (with P dominated by Q) is defined to be $D_{KL}(P\|Q) = -\mathbb{E}_P\left[\log dQ/dP\right]$.

where $\alpha_k > 0$ is the step-size, the inner product is defined as $\langle p, q\rangle = \int_\Theta p(\theta)q(\theta)d\sigma$, and $D_w(x, x_k)$ is a (functional) Bregman distance function associated with a distance-generating function w, i.e.,

$$D_w(x, z) = w(x) - w(z) - \delta w[z; x - z],$$

where $\delta w[z; x - z]$ is the Fréchet derivative of w at z in the direction of $x - z$. If we choose $w(x) = \int x \log x$ as the distance-generating function, then the corresponding Bregman distance is the Kullback-Leibler (KL) divergence D_{KL}. Additionally, by selecting $\alpha_k = 1$, the solution to the optimization problem in Eq. (4) can be computed explicitly, where for each $\theta \in \Theta$,

$$d\mu_{k+1}(\theta) \propto p_\theta(x_{k+1})d\mu_k(\theta),$$

which is the posterior distribution as defined in Eq. (1) (a formal proof of this assertion is a special case of Proposition 1 shown later in the paper).

We have just shown how Bayes rule, i.e., the posterior computation, can be viewed as an instance of mirror descent with an stochastic approximation, for a particular choice of Bregman function; in the next subsection, we show how this interpretation leads to a natural algorithm in the distributed Bayesian posterior.

B. Distributed Stochastic Mirror Descent

Now, consider the distributed problem where the network of agents want to collectively solve the following optimization problem

$$\min_{\theta\in\Theta} F(\theta) \triangleq D_{KL}\left(\boldsymbol{P}\|\boldsymbol{P}_\theta\right) = \sum_{i=1}^n D_{KL}(P^i\|P^i_\theta). \tag{5}$$

Recall that the distribution $\boldsymbol{P}$ is unknown (though, of course, agents gain information about it by observing samples from $X^i_1, X^i_2, \dots$ and interacting with other agents) and that $\mathscr{P}^i$ containing all the distributions P^i_θ is a private family of distributions and is only available to agent i.

We propose the following algorithm as a distributed version of the stochastic mirror descent for the solution of problem Eq. (5):

$$d\mu^i_{k+1} = \arg\min_{\pi\in\Delta_\Theta}\left\{\langle -\log p^i_\theta(x^i_{k+1}), \pi\rangle + \sum_{j=1}^n a_{ij}D_{KL}(\pi\|d\mu^j_k)\right\}$$
$$\text{where } \theta \sim \pi, \tag{6}$$

with $a_{ij} > 0$ denoting the weight that agent i assigns to beliefs coming from its neighbor j. Specifically, $a_{ij} > 0$ if $(i, j) \in E$ or $j = i$, and $a_{ij} = 0$ if $(i, j) \notin E$. The optimization problem in Eq. (6) has a closed form solution. In particular, the posterior density at each $\theta \in \Theta$ is given by

$$d\mu^i_{k+1}(\theta) \propto p^i_\theta(x^i_{k+1}) \prod_{j=1}^n (d\mu^j_k(\theta))^{a_{ij}},$$

or equivalently, the belief on a measurable set B of an agent i at time $k + 1$ is

$$\mu^i_{k+1}(B) \propto \int_B p^i_\theta(x^i_{k+1}) \prod_{j=1}^n (d\mu^j_k(\theta))^{a_{ij}}. \tag{7}$$

We state the correctness of this claim in the following proposition.

Proposition 1. *Assume the weights (a_{ij}) form a doubly stochastic matrix. Then. the probability measure μ^i_{k+1} over the set Θ defined by the update protocol Eq. (7) coincides, almost everywhere, with the update the distributed stochastic mirror descent algorithm applied to the optimization problem in Eq. (5).*

Proof. We need to show that the density $d\mu^i_{k+1}$ associated with the probability measure μ^i_{k+1} defined by Eq. (7) minimizes the problem in Eq. (6). To do so, let $G(\pi)$ be the objective function for the problem in Eq. (6), i.e.,

$$G(\pi) = \langle -\log p^i_\theta(x^i_{k+1}), \pi\rangle + \sum_{j=1}^n a_{ij}D_{KL}(\pi\|d\mu^j_k).$$

Next, we add and subtract the KL divergence between π and the density $d\mu^i_{k+1}$ to obtain

$$\begin{aligned} G(\pi) &= \langle -\log p^i_\theta(x^i_{k+1}), \pi\rangle + \sum_{j=1}^n a_{ij}D_{KL}(\pi\|d\mu^j_k) - \\ &\quad - D_{KL}\left(\pi\|d\mu^i_{k+1}\right) + D_{KL}\left(\pi\|d\mu^i_{k+1}\right) \\ &= \langle -\log p^i_\theta(x^i_{k+1}), \pi\rangle + D_{KL}\left(\pi\|d\mu^i_{k+1}\right) + \\ &\quad + \sum_{j=1}^n a_{ij}\mathbb{E}_\pi \log\frac{d\mu^i_{k+1}}{d\mu^j_k}. \end{aligned}$$

Now, from Eq. (7) it follows that

$$\begin{aligned} G(\pi) &= \langle -\log p^i_\theta(x^i_{k+1}), \pi\rangle + D_{KL}\left(\pi\|d\mu^i_{k+1}\right) + \\ &\quad \sum_{j=1}^n a_{ij}\mathbb{E}_\pi \log\left(\frac{1}{d\mu^j_k}\frac{1}{Z^i_{k+1}}\prod_{l=1}^n \left(d\mu^l_k\right)^{a_{il}} p^i_\theta(x^i_{k+1})\right) \\ &= \langle -\log p^i_\theta(x^i_{k+1}), \pi\rangle + D_{KL}\left(\pi\|d\mu^i_{k+1}\right) \\ &\quad - \log Z^i_{k+1} + \langle \log p^i_\theta(x^i_{k+1}), \pi\rangle \\ &\quad + \sum_{j=1}^n a_{ij}\mathbb{E}_\pi \log\left(\frac{1}{d\mu^j_k}\prod_{l=1}^n \left(d\mu^l_k\right)^{a_{il}}\right) \\ &= -\log Z^i_{k+1} + D_{KL}\left(\pi\|d\mu^i_{k+1}\right) - \sum_{j=1}^n a_{ij}\mathbb{E}_\pi \log d\mu^j_k \\ &\quad + \sum_{l=1}^n a_{il}\mathbb{E}_\pi \log d\mu^l_k \\ &= -\log Z^i_{k+1} + D_{KL}\left(\pi\|d\mu^i_{k+1}\right), \end{aligned} \tag{8}$$

where $Z^i_{k+1} = \int_\theta p^i_\theta(x^i_{k+1}) \prod_{j=1}^n (d\mu^j_k(\theta))^{a_{ij}}$ is the corresponding normalizing constant.

The first term in Eq. (8) does not depend on the distribution π. Thus, we conclude that the solution to the problem in Eq. (6) is the density $\pi^* = d\mu^i_{k+1}$ as defined in Eq. (7) (almost everywhere). □

We remark that the update in Eq. (7) can be viewed as two-step processes: first every agent constructs an aggregate belief using a weighted geometric average of its own belief and the beliefs of its neighbors, and then each agent performs a Bayes' update using the aggregated belief as a prior. We note that similar arguments in the context of distributed optimization have

been proposed in [54], [60] for general Bregman distances. In the case when the number of hypotheses is finite, variations on this update rule were previously analyzed in [27], [30], [33].

C. An example

Example 1. *Consider a group of 4 agents, connected over a network as shown in Figure 2. A set of metropolis weights for this network is given by the following matrix:*

$$A = \begin{bmatrix} 2/3 & 1/6 & 0 & 1/6 \\ 1/6 & 2/3 & 1/6 & 0 \\ 0 & 1/6 & 2/3 & 1/6 \\ 1/6 & 0 & 1/6 & 2/3 \end{bmatrix}.$$

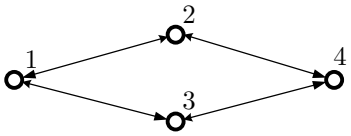

Fig. 2: A network of 4 agents.

Furthermore, assume that each agent is observing a Bernoulli random variable such that $X_k^1 \sim Bern(0.2)$, $X_k^2 \sim Bern(0.4)$, $X_k^3 \sim Bern(0.6)$ and $X_k^4 \sim Bern(0.8)$. In this case, the parameter space is $\Theta = [0, 1]$. Thus, the objective is to collectively find a parameter θ^* that best explains the joint observations in the sense of the problem in Eq. (5)*, i.e.*

$$\begin{aligned} \min_{\theta\in[0,1]} F(\theta) &= \sum_{j=1}^{4} D_{KL}(Bern(\theta^j)\|Bern(\theta)) \\ &= \sum_{j=1}^{4} \left(\theta \log \frac{\theta}{\theta^j} + (1-\theta)\log\frac{1-\theta}{1-\theta^j}\right) \end{aligned}$$

where $\theta^1 = 0.2$, $\theta^2 = 0.4$, $\theta^3 = 0.6$ and $\theta^4 = 0.8$. The optimal solution is $\theta^* = 0.5$ by the first-order optimality conditions or by exploiting symmetries in the objective function.

To summarize, we have given an interpretation of Bayes' rule as an instance of Stochastic Mirror Descent. We have shown how this interpretation motivates a distributed update rule. In the next section, we discuss explicit forms of this update rule for parametric models coming from exponential families.

IV. Cooperative Inference for Exponential Families

We begin with the observation that, for a general class of models $\{\mathscr{P}^i\}$, the direct computation of the posterior beliefs μ_{k+1}^i is intractable. Indeed, computing μ_{k+1}^i requires the solution of an integral of the form

$$\int_\Theta p_\theta^i(x_{k+1}^i) \prod_{j=1}^{n} (d\mu_k^j(\theta))^{a_{ij}}. \tag{9}$$

There is an entire area of research called *variational Bayes' approximations* dedicated to efficiently approximating integrals that appear in such context [61]–[63].

The purpose of this section is to show that for exponential family [64], [65] there are closed-form expressions for the posterior beliefs generated by the proposed distributed inference algorithm.

Definition 1. *The exponential family, for a parameter $\theta = [\theta^1, \theta^2, \dots, \theta^s]'$, is the set of probability distributions whose density can be represented as*

$$p_\theta(x) = H(x)\exp(M(\theta)'T(x)),$$

for specific functions $H(\cdot)$, $M(\cdot)$ and $T(\cdot)$ where $M(\theta) = [M(\theta^1), M(\theta^2), \dots, M(\theta^s)]'$ depends on the density parameters and $T(\cdot)$ depends on the observations.

For example, consider a Normal distribution parametrized by its mean θ with known variance σ^2. Then, it holds that

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2} + \frac{x\theta}{\sigma^2} - \frac{\theta^2}{2\sigma^2}\right) \\ &= \underbrace{\frac{\exp\left(-\frac{x^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}}_{H(x)} \exp\left(\underbrace{\begin{bmatrix}\theta & \theta^2\end{bmatrix}}_{M(\theta)} \underbrace{\begin{bmatrix}\frac{x}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{bmatrix}}_{T(x)}\right). \end{aligned} \tag{10}$$

Among the exponential family members, one can find distributions such as Normal, Poisson, Exponential, Gamma, Bernoulli, and Beta, among others [66]. In our case, we will take advantage of the existence of *conjugate priors* for all members of the exponential family. The definition of the conjugate prior is given below.

Definition 2. *Assume that the prior distribution p on a parameter space Θ belongs to the exponential family. Then, the distribution p is referred to as the conjugate prior for a likelihood function $p_\theta(x)$ if the posterior distribution $p(\theta|x) \propto p_\theta(x)p(\theta)$ is in the same family as the prior.*

Definition 2 implies that if the belief density at some time k is a conjugate prior for our likelihood model, then our belief at time $k+1$ will be of the same class as our prior. For example, if a likelihood function follows a Gaussian form, then having a Gaussian prior will produce a Gaussian posterior. This property simplifies the structure of the belief update procedure since we can express the evolution of the beliefs generated by the proposed algorithm in Eq. (7) by the evolution of the natural parameters of the member of the exponential family it belongs to. Naturally, by induction, if the prior belief at time $k = 0$ is a conjugate prior of the likelihood function, the beliefs for all $k > 0$ will belong to the same exponential family.

In the same way that a Gaussian likelihood function can be represented in its canonical form as in Eq. (10), we can find such representation as well for the belief density. Note, however, that in this case, the sample space is not $\mathcal{X}$ as in the likelihood function, but Θ because the belief is a distribution over Θ. Moreover, we will require some parametric characterization. Particularly we can write a belief density as

$$p_\chi(\theta) = f(\chi)\exp(M(\theta)'\chi).$$

where M is a function of the parameter space for $\theta \in \Theta$, and χ which is a parametric characterization of the belief density.

Going back to the example in Eq. (10), assume that our prior is a Normal distribution on θ with mean $\hat{\theta}$ and variance $\hat{\sigma}^2$, then $\chi = [\hat{\theta} \;\; \hat{\sigma}^2]'$ and

$$\begin{aligned} p_\chi(\theta) &= \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{(\theta-\hat{\theta})^2}{2\hat{\sigma}^2}\right) \\ &= \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{\theta^2}{2\hat{\sigma}^2} + \frac{\theta\hat{\theta}}{\hat{\sigma}^2} - \frac{\hat{\theta}^2}{2\hat{\sigma}^2}\right) \\ &= \underbrace{\frac{\exp\left(-\frac{\hat{\theta}^2}{2\hat{\sigma}^2}\right)}{\sqrt{2\pi\hat{\sigma}^2}}}_{f(\chi)} \exp\left(\underbrace{\begin{bmatrix}\theta & \theta^2\end{bmatrix}}_{M(\theta)} \underbrace{\begin{bmatrix}\frac{\hat{\theta}}{\hat{\sigma}^2} \\ -\frac{1}{2\hat{\sigma}^2}\end{bmatrix}}_{\chi}\right). \end{aligned} \tag{11}$$

Then, it can be shown that the posterior distribution, given some observation x, has the same exponential form as the prior, with updated parameter $\bar{\chi} = \chi + T(x)$ as follows:

$$p_{\bar{\chi}}(\theta|x) \propto p_\theta(x) p_\chi(\theta|x). \tag{12}$$

Particularly, for the example in Eq. (10) and Eq. (11), the posterior distribution is still Gaussian.

We will now exploit the structure of the exponential family of distributions to reformulate the distributed inference algorithm in Eq. (7) into an easy to implement algorithm in terms of the parametric representation of the beliefs for each agent.

Initially, consider that the set of agents have a belief at time k in the form of a distribution over the parameter space that is a member of the exponential family. That is, assume that each agent i has a belief over the parameters θ such that

$$d\mu_k^i(\theta) \propto \exp\left(M(\theta)'\chi_k^i\right),$$

then, according to the first step in Eq. (7), an agent i needs to compute the weighted geometric average of the beliefs of its neighbors including its own. Given the parametrization in the exponential family, it holds that,

$$\begin{aligned} \prod_{j=1}^n \left(d\mu_k^j(\theta)\right)^{a_{ij}} &\propto \prod_{j=1}^n \left(\exp\left(M(\theta)'\chi_k^j\right)\right)^{a_{ij}} \\ &= \exp\left(M(\theta)' \sum_{j=1}^n a_{ij}\chi_k^j\right). \end{aligned}$$

Now, if all agents have beliefs in the same exponential family and they are conjugate priors to their corresponding likelihood functions, then we can write the posterior of agent i as

$$\begin{aligned} d\mu_{k+1}^i(\theta) &\propto \exp\left(M(\theta)' \sum_{j=1}^n a_{ij}\chi_k^j\right) p_M^i(x_{k+1}^i) \\ &= \exp\left(M(\theta)' \sum_{j=1}^n a_{ij}\chi_k^j -\right) \exp(M(\theta)'T^i(x_{k+1}^i)) \\ &= \exp\left(M(\theta)'\left(\sum_{j=1}^n a_{ij}\chi_k^j + T^i(x_{k+1}^i)\right)\right) \\ &= \exp\left(M(\theta)'\chi_{k+1}^i\right). \end{aligned}$$

As an immediate conclusion, it follows that for distributed inference problems when the observation models are members of the exponential family, one can always construct a set of beliefs using prior conjugates, and the algorithm in Eq. (7) simplifies to updates in the parameters of the exponential family, as shown by the following proposition.

Proposition 2. *Assume the belief density $d\mu_k^i$ at time k has an exponential form with natural parameters χ_k^i and ν_k^i for all $1 \leq i \leq n$, and that these densities are conjugate priors of the likelihood models p_θ^i. Then, the belief density of agent i at time $k+1$, as computed in the update rule in Eq. (7), has the same form as the beliefs at time k with the natural parameters given by*

$$\chi_{k+1}^i = \sum_{j=1}^n a_{ij}\chi_k^j + T^i(x_{k+1}^i). \tag{13}$$

Proposition 2 simplifies the algorithm in Eq. (7) and facilitates its use in traditional estimation problems where members of the exponential family are used.

A. Examples

In this subsection, we explicitly state the general distributed algorithm in Eq. (13) presented in Proposition 2 for several distributed parameter estimation problems. Mainly, we explicitly write the definition of the vector $T^i(x_k^i)$ and χ_k^i, from which the parameters of the current beliefs for each agent can be computed. Later in Section **??** we will provide simulation results for several distributed inference problems over various graph topologies.

1) Distributed Gaussian Filter with unknown mean and known variance: Assume each agent in the network observes a signal of the form $X_k^i = \theta^i + \epsilon_k^i$, where θ^i is finite and unknown scalar quantity, while $\epsilon^i \sim \mathcal{N}(0, 1/\tau^i)$ is a zero mean Gaussian noise with precision $\tau^i = 1/(\sigma^i)^2$ known only by agent i. The objective of the network is to agree on a single θ^* that solves the optimization problem in Eq. (5).

In this case, the likelihood models, the prior and the posterior are Normal distributions. Thus, if the beliefs of the agents at time k are Gaussian, i.e., $\mu_k^i = \mathcal{N}(\theta_k^i, 1/\tau_k^i)$ for all $i = 1\ldots, n$, then their beliefs at time $k+1$ are also Gaussian. In particular, they are given by $\mu_k^i = \mathcal{N}(\theta_k^i, 1/\tau_k^i)$ for all $i = 1\ldots, n$, with

$$M(\theta) = \begin{bmatrix}\theta \\ \theta^2\end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix}x_k^i\tau^i \\ -\frac{1}{2}\tau^i\end{bmatrix}, \qquad \chi_k^i = \begin{bmatrix}\theta_k^i\tau_k^i \\ -\frac{1}{2}\tau_k^i\end{bmatrix}.$$

We note that this specific setup is known as Gaussian Learning and has been studied in [67], [68], where the expected parameter estimator is shown to converge at an $O(1/k)$ rate.

2) Distributed Gaussian Filter with unknown variance and known mean: In this case, the agents want to cooperatively estimate the value of a variance which is the parameter for Eq. (5). Specifically, each agent i observes a realization of the random variable $X_k^i = \theta^i + \epsilon_k^i$, with $\epsilon_k^i \sim \mathcal{N}(0, 1/\tau^i)$, where θ^i is known and τ^i is unknown. The beliefs of all agents are chosen to be a Gamma distribution $\mu_k^i = \text{Gamma}(\alpha_k^i, \beta_k^i)$ and it follows that

$$M(\tau) = \begin{bmatrix}\tau \\ \log\tau\end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix}-\frac{1}{2}(x_k^i - \theta^i)^2 \\ -\frac{1}{2}\end{bmatrix},$$

$$\chi_k^i = \begin{bmatrix} -\beta_k^i \\ -(\alpha_k^i - 1) \end{bmatrix}.$$

3) Distributed Gaussian Filter with unknown mean and variance: In the preceding examples, we have considered the cases when either the mean or the variance is known. Here, we will assume that both the mean and the variance are unknown and need to be estimated. Explicitly, we still have noise observations $X_k^i = \theta^i + \epsilon_k^i$, with $\epsilon_k^i \sim \mathcal{N}(0, 1/\tau^i)$. We are going to assume all agents have beliefs that follow the Normal-Gamma distribution, i.e. $\mu_k^i = \text{NormalGamma}(\theta_k^i, \lambda_k^i, \alpha_k^i, \beta_k^i)$ for $i = 1, \cdots, n$. Moreover, the it holds that

$$M(\theta, \tau) = \begin{bmatrix} \log \tau \\ \tau \\ \tau\theta \\ \tau\theta^2 \end{bmatrix}, \qquad T^i(x_k^i) = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2}(x_k^i)^2 \\ x_k^i \\ -\frac{1}{2} \end{bmatrix},$$

$$\chi_k^i = \begin{bmatrix} \alpha_k^i - \frac{1}{2} \\ -\frac{1}{2}\lambda_k^i(\theta_k^i)^2 - \beta_k^i \\ \lambda_k^i \theta_k^i \\ -\frac{1}{2}\lambda_k^i \end{bmatrix}.$$

4) Distributed Bernoulli Filter: Here, each of the agents receives private observations of the form $X_k^i \sim \text{Bernoulli}(p^i)$, with p^i unknown. In order to estimate the network-wide parameter, each agent constructs a sequence of beliefs following a Beta distribution, i.e. $\mu_k^i = \text{Beta}(\alpha_k^i, \beta_k^i)$. Then, the proposed algorithm in Eq. (13) updates its parameters. Moreover, it holds that

$$M(p) = \begin{bmatrix} \log p \\ \log(1-p) \end{bmatrix}, \; T^i(x_k^i) = \begin{bmatrix} x_k^i \\ 1 - x_k^i \end{bmatrix}, \; \chi_k^i = \begin{bmatrix} \alpha_k^i \\ \beta_k^i \end{bmatrix}.$$

5) Distributed Poisson Filter: Similarly as before, we consider an observation model where each agent i receives realization of a Poisson random variable with unknown parameter λ^i, i.e., $X_k^i \sim \text{Poisson}(\lambda^i)$ for all i. The conjugate prior of a Poisson likelihood model is the Gamma distribution. Thus, if at time k the beliefs of each agent i are given by $\mu_k^i = \text{Gamma}(\alpha_k^i, \beta_k^i)$. Moreover, it holds that

$$M(\lambda) = \begin{bmatrix} \log \lambda \\ \lambda \end{bmatrix}, \; T^i(x_k^i) = \begin{bmatrix} x_k^i \\ -1 \end{bmatrix}, \; \chi_k^i = \begin{bmatrix} \alpha_k^i - 1 \\ -\beta_k^i \end{bmatrix}.$$

6) Distributed Exponential Filter: As a final example, we consider an observation model where each agent i receives realization of an Exponential random variable with unknown rate λ^i, i.e., $X_k^i \sim \text{Exponential}(\lambda^i)$ for all i. The conjugate prior of an Exponential likelihood model is the Gamma distribution. Thus, if at time k the beliefs of each agent i are given by $\mu_k^i = \text{Gamma}(\alpha_k^i, \beta_k^i)$. Moreover, it holds that

$$M(\lambda) = \begin{bmatrix} \lambda \\ \log \lambda \end{bmatrix}, \; T^i(x_k^i) = \begin{bmatrix} -1 \\ x_k^i \end{bmatrix}, \; \chi_k^i = \begin{bmatrix} \alpha_k^i - 1 \\ -\beta_k^i \end{bmatrix}.$$

V. Belief Concentration Rates

We now turn to the presentation of our main results about the rate at which beliefs generated by the update rule in Eq. (7) concentrate around the true parameter θ^*. We will break up our analysis into two cases. Initially, Part I of this paper series will focus on when Θ is a finite set and will prove a concentration rate on the beliefs on a Hellinger ball around the optimal hypothesis. The case when Θ is a finite set has been previously studied in [27], [30], [33] with similar geometric concentration results for distributed learning has been shown. However, we take a fundamentally different proof approach that will allow us to gently introduce the techniques we will use later when we turn to our main scenario of interest, namely when Θ is a compact subset of $\mathbb{R}^d$. We analyze the case of compact hypotheses sets in Part II of this paper series. Our proof techniques use concentration arguments for beliefs on Hellinger balls from the recent work in [42] which, in turn, builds on the classic paper of [69].

We begin with two subsections focusing on background information, definitions, and assumptions.

A. Background: Hellinger Distance and Coverings

The *squared* Hellinger distance between two probability distributions P and Q is given by,

$$h^2(P, Q) = \frac{1}{2} \int \left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}} \right)^2 d\lambda, \tag{14}$$

where P and Q are dominated by λ. Moreover, the Hellinger distance satisfies the property that $0 \leq h(P, Q) \leq 1$.

We equip the set of all probability distributions $\mathscr{P}$ over the parameter set with the Hellinger distance to obtain the *metric* space $(\mathscr{P}, h)$. The metric space induces a topology, where we can define an open ball $\mathcal{B}_r(\theta)$ with a radius $r \in (0, 1)$ centered at a point $\theta \in \Theta$, which we use to construct a special covering of subsets $B \subset \mathscr{P}$. Recall that Eq. 14 defines the squared Hellinger distance h^2, rather than h.

Definition 3. *Define an n-Hellinger ball of radius r centered at θ as*

$$\mathcal{B}_r(\theta) = \left\{ \hat{\theta} \in \Theta \,\middle|\, \frac{1}{n} \sum_{i=1}^{n} h^2 \left(P_\theta^i, P_{\hat{\theta}}^i \right) \leq r^2 \right\}.$$

Additionally, when no center is specified, it should be assumed that it refers to θ^*, i.e. $\mathcal{B}_r = \mathcal{B}_r(\theta^*)$.

Given an n-Hellinger ball of radius r, we will use the following notation for a covering of its complement $\mathcal{B}_r^c$. Specifically, we are going to express $\mathcal{B}_r^c$ as the union of finite disjoint and concentric annuli. Let $r \in (0, 1)$ and $\{r_l\}$ be a finite strictly decreasing sequence such that $r_1 = 1$ and $r_L = r$ and express the set $\mathcal{B}_r^c$ as the union of annuli generated by the sequence $\{r_l\}$ as

$$\mathcal{B}_r^c = \bigcup_{l=1}^{L-1} \mathcal{F}_l,$$

where $\mathcal{F}_l = \mathcal{B}_{r_l} \setminus \mathcal{B}_{r_{l+1}}$.

B. Background: Assumptions on the Network and Mixing Weights

Naturally, we need some assumptions on the matrix A. For one thing, the matrix A has to be "compatible" with the underlying graph, in that information from node i should not affect node j if there is no edge from i to j in $\mathcal{G}$. At the other extreme, we want to rule out the possibility that A is

the identity matrix, which in terms of Eq. (7) means nodes do not talk to their neighbors. Formally, we make the following assumption.

Assumption 1. *The graph $\mathcal{G}$ and matrix A are such that:*
(a) A is doubly-stochastic with $[A]_{ij} = a_{ij} > 0$ for $i \neq j$ if and only if $(i,j) \in E$.
(b) A has positive diagonal entries, $a_{ii} > 0$ for all $i \in V$.
(c) The graph $\mathcal{G}$ is connected.

Assumption 1 is common in the distributed optimization literature. The construction of a set of weights satisfying Assumption 1 can be done in a distributed way, for example, by choosing the so-called "lazy Metropolis" matrix, which is a stochastic matrix given by

$$a_{ij} = \begin{cases} \frac{1}{2\max\{d^i+1, d^j+1\}} & \text{if } (i,j) \in E, \\ 0 & \text{if } (i,j) \notin E, \end{cases}$$

where d^i is the degree (the number of neighbors) of node i. Note that although the above formula only gives the off-diagonal entries of A, it uniquely defines the entire matrix (the diagonal elements are uniquely defined via the stochasticity of A). To choose the weights corresponding to a lazy Metropolis matrix, agents will need to spend an additional round at the beginning of the algorithm broadcasting their degrees to their neighbors.

Assumption 1 can be seen to guarantee that $A^k \to (1/n)\mathbf{1}\mathbf{1}^T$ where $\mathbf{1}$ is the vector of all ones. We will use the following result based on [30] and [33], that provides convergence rate for the difference $|A^k - (1/n)\mathbf{1}\mathbf{1}^T|$,:

Lemma 3. *Let Assumption 1 hold, then the matrix A satisfies the following relation:*

$$\sum_{t=1}^{k}\sum_{j=1}^{n}\left|\left[A^{k-t}\right]_{ij} - \frac{1}{n}\right| \leq \frac{4\log n}{1-\delta} \qquad \text{for } i = 1,\dots,n,$$

where $\delta = 1 - \eta/4n^2$ with η being the smallest positive entry of the matrix A. Furthermore, if A is a lazy Metropolis matrix associated with the graph $\mathcal{G}$, then $\delta = 1 - 1/\mathcal{O}(n^2)$.

C. Concentration Analysis for Finite Hypotheses Sets

We now turn to prove a concentration result when the set Θ of hypotheses is finite. We will show exponential convergence of beliefs on a Hellinger Ball around the true hypothesis θ^*. The purpose is to introduce the techniques gently we will use later in a compact set of hypotheses.

When the number of hypotheses is finite, the density update in Eq. (7) can be written in a simpler form for discrete beliefs over the parameter space Θ as

$$\mu^i_{k+1}(\theta) \propto p^i_\theta(x^i_{k+1}) \prod_{j=1}^{n} (\mu^j_k(\theta))^{a_{ij}}. \tag{15}$$

We will fix the radius r, and our goal will be to prove a concentration result for a Hellinger ball of radius r around the optimal hypothesis θ^*. We start by partitioning the complement of this ball, i.e., $\mathcal{B}^c_r$, as described above into the annuli $\mathcal{F}_l$. We introduce the notation $\mathcal{N}_l$ to denote the number of hypotheses within the annulus $\mathcal{F}_l$. We refer the reader to Figure 3, which shows a set of probability distributions, represented as black dots, where a star represents the true distribution P.

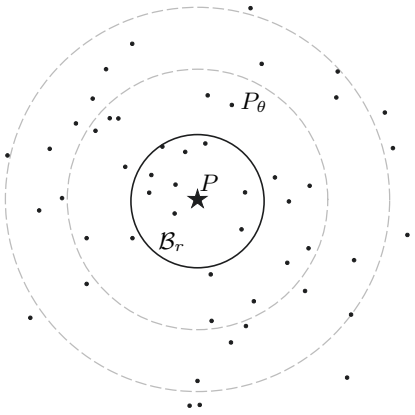

Fig. 3: Creating a covering for a ball $\mathcal{B}_r$. ★ represents the correct hypothesis $\boldsymbol{P}_{\theta^*}$, • indicates the location of other hypotheses and the dash lines indicate the boundary of the balls $\mathcal{B}_{r_l}$.

The distance between hypotheses is defined in terms of the Hellinger affinity between two distributions Q and P, given by

$$\rho(Q,P) = 1 - h^2(Q,P). \tag{16}$$

We are now ready to state our first result as a lemma that bounds the concentration of aggregated log-likelihood ratios.

Lemma 4. *Let Assumptions 1 hold. Given a set of independent random variables $\{X^i_t\}$ such that $X^i_t \sim P^i$ for $i = 1,\dots,n$ and $t = 1,\dots,k$, a set of distributions $\{Q^i\}$ where P^i dominates Q^i, then for all $y \in \mathbb{R}$,*

$$\mathbb{P}\left[\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{dQ^j}{dP^j}(X^j_t) \geq y\right] \leq \exp\left(-\frac{y}{2} + \frac{4\log n}{1-\delta} - k\frac{1}{n}\sum_{j=1}^{n} h^2(Q^j,P^j)\right).$$

Proof. By the Markov inequality we have

$$\begin{aligned}
&\mathbb{P}\left[\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{dQ^j}{dP^j}(X^j_t) \geq y\right] \\
&\leq \exp\left(-\frac{y}{2}\right)\mathbb{E}\left[\prod_{t=1}^{k}\prod_{j=1}^{n}\sqrt{\left(\frac{dQ^j}{dP^j}(X^j_t)\right)^{[A^{k-t}]_{ij}}}\right] \\
&\leq \exp\left(-\frac{y}{2}\right)\prod_{t=1}^{k}\prod_{j=1}^{n}\mathbb{E}\left[\sqrt{\left(\frac{dQ^j}{dP^j}(X^j_t)\right)}\right]^{[A^{k-t}]_{ij}} \\
&= \exp\left(-\frac{y}{2}\right)\prod_{t=1}^{k}\prod_{j=1}^{n}\rho(Q^j,P^j)^{[A^{k-t}]_{ij}},
\end{aligned}$$

where the last inequality follows from the definition of the Hellinger affinity function $\rho(Q,P)$ and Jensen's inequality.

Moreover, it follow from $\rho(Q^j, P^j) = 1 - h^2(Q^j, P^j)$ and $1 - x \leq \exp(-x)$ for $x \in [0, 1]$ that

$$\prod_{t=1}^{k}\prod_{j=1}^{n} \rho(Q^j, P^j)^{[A^{k-t}]_{ij}} \leq \exp\left(-\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij} h^2(Q^j, P^j)\right). \tag{17}$$

Now, by adding and subtracting $\sum_{t=1}^{k}\frac{1}{n}\sum_{j=1}^{n} h^2(Q^j, P^j)$ we have

$$\begin{aligned}
&\mathbb{P}\left[\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{dQ^j}{dP^j}(X_t^j) \geq y\right] \\
&\leq \exp\left(-\frac{y}{2} - \sum_{t=1}^{k}\sum_{j=1}^{n}\left([A^{k-t}]_{ij} - \frac{1}{n}\right) h^2(Q^j, P^j) \right. \\
&\quad \left. -\frac{k}{n}\sum_{j=1}^{n} h^2(Q^j, P^j)\right) \\
&\leq \exp\left(-\frac{y}{2} + \frac{4\log n}{1-\delta} - \frac{k}{n}\sum_{j=1}^{n} h^2(Q^j, P^j)\right).
\end{aligned}$$

Finally, the last line above follows from Lemma 3 applied to the second term inside the exponential. □

We are now ready to state our first main result, which bounds the concentration of Eq. (15) around the optimal hypothesis for a finite hypothesis set Θ. The following theorem shows that all agents' beliefs will concentrate around the Hellinger ball $\mathcal{B}_r$ at an exponential rate.

Theorem 5. *Let Assumption 1 hold, and let $\sigma \in (0, 1)$ be a desired probability tolerance. Then, the belief sequences $\{\mu_k^i\}$, $i \in V$ that are generated by the update rule in Eq. (15), with initial beliefs such that $\mu_0^i(\theta^*) > \epsilon$ for all i, have the following property: for any radius $r \in (0, 1)$ with probability $1 - \sigma$,*

$$\mu_{k+1}^i(\mathcal{B}_r) \geq 1 - \frac{1}{\epsilon}\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-kr_{l+1}^2\right) \quad \forall i \text{ and } k \geq N,$$

where

$$N = \inf\left\{t \geq 1 \,\middle|\, \exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-tr_{l+1}^2\right) < \sigma\right\},$$

and δ as defined in Lemma 3.

In Theorem 5, note that N indicates the time required for the beliefs on the ball $\mathcal{B}_r$ around the true hypothesis θ^* to start to concentrate at a geometric rate. Moreover, N is a function of δ and n, showing the impact of the network topology. The time N can also be interpreted as a transient time required for mixing of beliefs among the agents before the impact of the network disappears.

Proof. We are going to focus on bounding the beliefs of a measurable set B, such that $\theta^* \in B$. For such a set, it follows by induction from Eq. (15) that

$$\begin{aligned}
\mu_k^i(B) &= \frac{1}{Z_k^i}\sum_{\theta\in B}\prod_{j=1}^{n}\mu_0^j(\theta)^{[A^k]_{ij}}\prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}} \\
&= \left(1 + \frac{\sum_{\theta\in B^c}\prod_{j=1}^{n}\mu_0^j(\theta)^{[A^k]_{ij}}\prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}}{\sum_{\theta\in B}\prod_{j=1}^{n}\mu_0^j(\theta)^{[A^k]_{ij}}\prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}}\right)^{-1} \\
&\geq 1 - \frac{\sum_{\theta\in B^c}\prod_{j=1}^{n}\mu_0^j(\theta)^{[A^k]_{ij}}\prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}}{\sum_{\theta\in B}\prod_{j=1}^{n}\mu_0^j(\theta)^{[A^k]_{ij}}\prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}},
\end{aligned}$$

where Z_k^i is the appropriate normalization constant. Moreover, for $\theta^* \in B$ it follows that

$$\mu_k^i(B) \geq 1 - \sum_{\theta\in B^c}\prod_{j=1}^{n}\left(\frac{\mu_0^j(\theta)}{\mu_0^j(\theta^*)}\right)^{[A^k]_{ij}}\prod_{t=1}^{k}\prod_{j=1}^{n}\left(\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)}\right)^{[A^{k-t}]_{ij}},$$

Moreover, from the assumption that $\mu_0^i(\theta^*) > \epsilon$ for all $i = 1, \ldots, n$, it follows that

$$\mu_k^i(B) \geq 1 - \frac{1}{\epsilon}\sum_{\theta\in B^c}\prod_{t=1}^{k}\prod_{j=1}^{n}\left(\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)}\right)^{[A^{k-t}]_{ij}}. \tag{18}$$

The relation in Eq. (18) describes the iterative averaging of products of density functions, for which we can use Lemma 4 with $Q = P_\theta$ and $P = P_{\theta^*}$. Then,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\theta\in B^c}\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \geq y\right] \\
&\quad \leq \sum_{\theta\in B^c}\exp\left(-\frac{y}{2} + \frac{4\log n}{1-\delta} - \frac{k}{n}\sum_{j=1}^{n} h^2(P_\theta^j, P^j)\right),
\end{aligned}$$

and by setting $y = -\frac{k}{n}\sum_{j=1}^{n} h^2(P_\theta^j, P^j)$ we obtain

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\theta\in B^c}\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \geq -\frac{k}{n}\sum_{j=1}^{n} h^2(P_\theta^j, P^j)\right] \\
&\quad \leq \exp\left(\frac{4\log n}{1-\delta}\right)\sum_{\theta\in B^c}\exp\left(-\frac{k}{2n}\sum_{j=1}^{n} h^2(P_\theta^j, P^j)\right).
\end{aligned}$$

Now, we let the set B be the Hellinger ball of a radius r centered at θ^* and define a cover (as described above) to exploit the representation of $\mathcal{B}_r^c$ as the union of concentric Hellinger annuli, for which we have

$$\begin{aligned}
&\mathbb{P}\left[\sup_{\theta\in B^c}\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)} \geq -\frac{k}{n}\sum_{j=1}^{n} h^2(P_\theta^j, P^j)\right] \\
&\leq \exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\sum_{\theta\in\mathcal{F}_l}\exp\left(-\frac{k}{2}\frac{1}{n}\sum_{j=1}^{n} h^2(P_\theta^j, P^j)\right)
\end{aligned}$$

$$\leq \exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-\frac{k}{2}r_{l+1}^2\right).$$

We are interested in finding a value of k large enough such that the above probability is below σ. Thus, lets define the value of N as

$$N=\inf\left\{t\geq 1\middle|\exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-tr_{l+1}^2\right)<\sigma\right\}.$$

It follows that for all $k\geq N$ with probability $1-\sigma$, for all $\theta\in\mathcal{B}_r^c$

$$\sum_{t=1}^{k}\sum_{j=1}^{n}[A^{k-t}]_{ij}\log\frac{p_\theta^j(X_t^j)}{p^j(X_t^j)}\leq -\frac{k}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j).$$

Thus, from Eq. (18) with probability $1-\sigma$ we have

$$\begin{aligned}\mu_k^i(\mathcal{B}_r)&\geq 1-\frac{1}{\epsilon}\sum_{\theta\in\mathcal{B}_r^c}\exp\left(-\frac{k}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j)\right)\\&=1-\frac{1}{\epsilon}\sum_{l=1}^{L-1}\sum_{\theta\in\mathcal{F}_l}\exp\left(-\frac{k}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j)\right)\\&\geq 1-\frac{1}{\epsilon}\sum_{l=1}^{L-1}\mathcal{N}_{r_l}\exp\left(-kr_{l+1}^2\right).\end{aligned}$$

□

Note that in general, the belief concentration rate described in Theorem 5 depends on the geometry of the hypotheses set and how are they distributed on the parameter space. The next Corollary describes the scenario where the sequence $\{r_l\}$ is such that $L=2$, so $r_1=1$ and $r_2=r$.

Corollary 6. *Let Assumption 1 hold, and let $\sigma\in(0,1)$ be a desired probability tolerance. Then, the belief sequences $\{\mu_k^i\}$, $i\in V$ that are generated by the update rule in Eq. (15), with initial beliefs such that $\mu_0^i(\theta^*)>\epsilon$ for all i, have the following property: for any radius $r\in(0,1)$ with probability $1-\sigma$,*

$$\mu_{k+1}^i(\mathcal{B}_r)\geq 1-\frac{2}{\epsilon}\exp\left(\log\frac{\mathcal{N}}{\sigma}+\frac{4\log n}{1-\delta}-kr^2\right),$$

where $\mathcal{N}$ is the number of hypotheses outside $\mathcal{B}_r$ and δ as defined in Lemma 3.

D. Discussion and Comparison with Previous Approaches

Non-asymptotic belief concentration rates for non-Bayesian learning has been previously studied inin [27], [30], [33]. In this subsection, we provide some discussion and comparison with the result from Theorem 5, and Corollary 6 respectively.

We start by recalling a general form of the main result from [27], [30], [33].

Theorem 7 (Theorem 2 from [33])**.** *Let Assumptions 1 hold and let $\sigma\in(0,1)$. The update rule of Eq. (15), with positive and uniform initial belief on all hypotheses, has the following property: there is an integer $\boldsymbol{N}(\rho)$ such that, with probability $1-\rho$, for all $k\geq\boldsymbol{N}(\rho)$ and for all $\theta_v\notin\Theta^*$, we have*

$$\mu_k^i(\theta_v)\leq\exp\left(-\frac{k}{2}\gamma_2+\gamma_1^i\right)\quad\text{for all } i=1,\ldots,n,$$

where

$$\boldsymbol{N}(\rho)\triangleq\left\lceil\frac{1}{\gamma_2^2}8\left(\log\alpha\right)^2\log\frac{1}{\sigma}\right\rceil,$$

$$\gamma_1^i\triangleq\frac{12\log n}{1-\delta}\log\frac{1}{\alpha},\quad\gamma_2\triangleq\frac{1}{n}\min_{\theta_v\notin\Theta^*}\sum_{i=1}^{n}D_{KL}(P^i\|P_{\theta_v}^i),$$

where α is a positive lower bound on the likelihood functions.

For simplicity of presentation, we will focus our comparison with Corollary 6. Initially, note that Corollary 6 indicates the concentration of beliefs on a n-Hellinger ball of radius r around the optimal hypotheses, whereas Theorem 7 shows that the beliefs on the non-optimal hypotheses will decay to zero. These two statements are equivalent if the optimal hypotheses are unique, and the n-Hellinger ball of radius r contains only one hypothesis. Moreover, the rate at which such concentrations occur is exponential in the number of iterations for both cases. However, the rate in Corollary 6 is given by the radius r, whereas in Theorem 7 is given by the distance between the optimal and second-best hypotheses. These two statements seem equivalent. However, in Corollary 6 the distance is measured in terms of Hellinger distances, which are naturally upper bounded by 1. In Theorem 7, the Kullback-Leibler divergence is not upper-bounded. Thus, a larger value is expected. This weakness of the proposed method might be explained as the original problem in Eq. 5 involved KL divergences. However, this is a trade-off for a more general analysis that will allow us to work on compact hypotheses spaces. We believe this is a construction of the proof. Removing such construction is out of the scope of this paper and left for future work. Finally, the belief concentrations for both results happen after a time proportional to a term that depends on the network topology. They are equal up to a constant factor of 3. Finally, Corollary 6 removes the lower bounded likelihood assumption in Theorem 7. In both cases, the dependency on the high probability bound is only logarithmic.

VI. Conclusions

We have proposed an algorithm for distributed learning with both countable and compact sets of hypotheses. Our algorithm may be viewed as a distributed version of Stochastic Mirror Descent applied to the problem of minimizing the sum of Kullback-Leibler divergences. Our results show non-asymptotic geometric convergence rates for the beliefs concentration around the true hypothesis. Particularly in Part I, we provide an extensive application case of study for observational models in the exponential family of probability distributions. Moreover, we have developed a new belief concentration analysis for the case of finite hypotheses. Part II of this paper series extends this analysis to the compact hypotheses set case.

Future work should explore how variations on stochastic approximation algorithms will produce new non-Bayesian update rules for more general problems. Promising directions include acceleration results for proximal methods, other Bregman distances, or constraints within the space of probability distributions.

Furthermore, we have modeled interactions between agents as exchanges of local probability distributions (i.e., beliefs) between neighboring nodes in a graph. It remains open to understand to what extent this can be reduced when agents transmit only an approximate summary of their beliefs. We anticipate that future work will additionally consider the effect of parametric approximations allowing nodes to communicate only a finite number of parameters coming from, say, Gaussian Mixture Models or Particle Filters.

ACKNOWLEDGMENT

We would like to acknowledge support for this project from the National Science Foundation under grant no. CPS 15-44953 and by the Office of Naval Research under grant no. N00014-17-1-2195.

REFERENCES

[1] A. Jadbabaie, P. Molavi, A. Sandroni, and A. Tahbaz-Salehi, "Non-bayesian social learning," *Games and Economic Behavior*, vol. 76, no. 1, pp. 210–225, 2012.

[2] K. Rahnama Rad and A. Tahbaz-Salehi, "Distributed parameter estimation in networks," in *Proceedings of the IEEE Conference on Decision and Control*, pp. 5050–5055, 2010.

[3] M. Alanyali, S. Venkatesh, O. Savas, and S. Aeron, "Distributed bayesian hypothesis testing in sensor networks," in *Proceedings of the American Control Conference*, pp. 5369–5374, 2004.

[4] R. Olfati-Saber, E. Franco, E. Frazzoli, and J. S. Shamma, "Belief consensus and distributed hypothesis testing in sensor networks," in *Networked Embedded Sensing and Control*, pp. 169–182, Springer, 2006.

[5] R. J. Aumann, "Agreeing to disagree," *The Annals of Statistics*, vol. 4, no. 6, pp. 1236–1239, 1976.

[6] V. Borkar and P. P. Varaiya, "Asymptotic agreement in distributed estimation," *IEEE Transactions on Automatic Control*, vol. 27, no. 3, pp. 650–655, 1982.

[7] J. N. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Transactions on Automatic Control*, vol. 29, no. 1, pp. 42–50, 1984.

[8] C. Genest, J. V. Zidek, *et al.*, "Combining probability distributions: A critique and an annotated bibliography," *Statistical Science*, vol. 1, no. 1, pp. 114–135, 1986.

[9] R. Cooke, "Statistics in expert resolution: A theory of weights for combining expert opinion," in *Statistics in Science* (R. Cooke and D. Costantini, eds.), vol. 122 of *Boston Studies in the Philosophy of Science*, pp. 41–72, Springer Netherlands, 1990.

[10] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.

[11] G. L. Gilardoni and M. K. Clayton, "On reaching a consensus using degroot's iterative pooling," *The Annals of Statistics*, vol. 21, no. 1, pp. 391–401, 1993.

[12] J. A. Gubner, "Distributed estimation and quantization," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1456–1459, 1993.

[13] Y. Zhu, E. Song, J. Zhou, and Z. You, "Optimal dimensionality reduction of sensor data in multisensor estimation fusion," *IEEE Transactions on Signal Processing*, vol. 53, no. 5, pp. 1631–1639, 2005.

[14] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors i. fundamentals," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.

[15] S.-L. Sun and Z.-L. Deng, "Multi-sensor optimal information fusion kalman filter," *Automatica*, vol. 40, no. 6, pp. 1017–1023, 2004.

[16] D. Gale and S. Kariv, "Bayesian learning in social networks," *Games and Economic Behavior*, vol. 45, no. 2, pp. 329–346, 2003.

[17] E. Mossel and O. Tamuz, "Efficient bayesian learning in social networks with gaussian estimators," *arXiv preprint arXiv:1002.0747*, 2010.

[18] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," *The Review of Economic Studies*, vol. 78, no. 4, pp. 1201–1236, 2011.

[19] A. Jadbabaie, P. Molavi, and A. Tahbaz-Salehi, "Information heterogeneity and the speed of learning in social networks," *Columbia Business School Research Paper*, no. 13-28, 2013.

[20] S. Shahrampour and A. Jadbabaie, "Exponentially fast parameter estimation in networks using distributed dual averaging," in *Proceedings of the IEEE Conference on Decision and Control*, pp. 6196–6201, 2013.

[21] B. Golub and M. O. Jackson, "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, pp. 112–149, 2010.

[22] D. Acemoglu, A. Nedić, and A. Ozdaglar, "Convergence of rule-of-thumb learning rules in social networks," in *Proceedings of the IEEE Conference on Decision and Control*, pp. 1714–1720, 2008.

[23] A. Jadbabaie, J. Lin, and A. S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Transactions on Automatic Control*, vol. 48, no. 6, pp. 988–1001, 2003.

[24] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2015.

[25] A. Olshevsky, "Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control," *preprint arXiv:1411.4186*, 2014.

[26] E. Mossel, A. Sly, and O. Tamuz, "Asymptotic learning on bayesian social networks," *Probability Theory and Related Fields*, vol. 158, no. 1-2, pp. 127–157, 2014.

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

[28] L. Qipeng, F. Aili, W. Lin, and W. Xiaofan, "Non-bayesian learning in social networks with time-varying weights," in *30th Chinese Control Conference (CCC)*, pp. 4768–4771, 2011.

[29] L. Qipeng, Z. Jiuhua, and W. Xiaofan, "Distributed detection via bayesian updates and consensus," in *34th Chinese Control Conference (CCC)*, pp. 6992–6997, 2015.

[30] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Distributed detection: Finite-time analysis and impact of network topology," *IEEE Transactions on Automatic Control*, vol. 61, pp. 3256–3268, Nov 2016.

[31] S. Shahrampour, M. Rahimian, and A. Jadbabaie, "Switching to learn," in *Proceedings of the American Control Conference*, pp. 2918–2923, 2015.

[32] M. A. Rahimian, S. Shahrampour, and A. Jadbabaie, "Learning without recall by random walks on directed graphs," *preprint arXiv:1509.04332*, 2015.

[33] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-bayesian learning," *preprint arXiv:1508.05161*, 2015.

[34] A. Nedić, A. Olshevsky, and C. A. Uribe, "Nonasymptotic convergence rates for cooperative learning over time-varying directed graphs," in *Proceedings of the American Control Conference*, pp. 5884–5889, 2015.

[35] A. Nedić, A. Olshevsky, and C. A. Uribe, "Network independent rates in distributed learning," in *Proceedings of the American Control Conference*, pp. 1072–1077, 2016.

[36] L. Su and N. H. Vaidya, "Asynchronous distributed hypothesis testing in the presence of crash failures," *University of Illinois at Urbana-Champaign, Tech. Rep*, 2016.

[37] P. Molavi, A. Tahbaz-Salehi, and A. Jadbabaie, "A theory of non-Bayesian social learning," *Econometrica*, vol. 86, no. 2, pp. 445–490, 2018.

[38] A. Mitra, J. A. Richards, and S. Sundaram, "A new approach for distributed hypothesis testing with extensions to Byzantine-resilience," in *American Control Conference (ACC)*, pp. 261–266, IEEE, 2019.

[39] A. Mitra, J. A. Richards, and S. Sundaram, "A communication-efficient algorithm for exponentially fast non-Bayesian learning in networks," in *IEEE 58th Conference on Decision and Control (CDC)*, pp. 8347–8352, IEEE, 2019.

[40] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Distributed detection and estimation in wireless sensor networks," *preprint arXiv:1307.1448*, 2013.

[41] A. Nedić, A. Olshevsky, and C. A. Uribe, "A tutorial on distributed (non-bayesian) learning: Problem, algorithms and results," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 6795–6801, Dec 2016.

[42] L. Birgé, "About the non-asymptotic behaviour of bayes estimators," *Journal of Statistical Planning and Inference*, vol. 166, pp. 67–77, 2015.

[43] A. Nedić, A. Olshevsky, and C. A. Uribe, "Distributed learning with infinitely many hypotheses," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 6321–6326, Dec 2016.

[44] S. Ghosal, "A review of consistency and convergence of posterior distribution," in *Varanashi Symposium in Bayesian Inference, Banaras Hindu University*, 1997.

[45] L. Schwartz, "On bayes procedures," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 4, no. 1, pp. 10–26, 1965.

[46] S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart, "Convergence rates of posterior distributions," *Annals of Statistics*, pp. 500–531, 2000.
[47] S. Ghosal, A. Van Der Vaart, *et al.*, "Convergence rates of posterior distributions for noniid observations," *The Annals of Statistics*, vol. 35, no. 1, pp. 192–223, 2007.
[48] V. Rivoirard, J. Rousseau, *et al.*, "Posterior concentration rates for infinite dimensional exponential families," *Bayesian Analysis*, vol. 7, no. 2, pp. 311–334, 2012.
[49] M. Rabbat and R. Nowak, "Decentralized source localization and tracking wireless sensor networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 921–924, 2004.
[50] M. Rabbat, R. Nowak, and J. Bucklew, "Robust decentralized source localization via averaging," in *IEEE International Conference on Acoustics, Speech, and Signal Processing.*, vol. 5, pp. 1057–1060, 2005.
[51] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.
[52] A. Nedić and S. Lee, "On stochastic subgradient mirror-descent algorithm with weighted averaging," *SIAM Journal on Optimization*, vol. 24, no. 1, pp. 84–107, 2014.
[53] B. Dai, N. He, H. Dai, and L. Song, "Scalable bayesian inference via particle mirror descent," *preprint arXiv:1506.03101*, 2015.
[54] M. Rabbat, "Multi-agent mirror descent for decentralized stochastic optimization," in *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2015 IEEE 6th International Workshop on*, pp. 517–520, IEEE, 2015.
[55] A. Zellner, "Optimal information processing and bayes's theorem," *The American Statistician*, vol. 42, no. 4, pp. 278–280, 1988.
[56] S. G. Walker, "Bayesian inference via a minimization rule," *Sankhyā: The Indian Journal of Statistics (2003-2007)*, vol. 68, no. 4, pp. 542–553, 2006.
[57] T. P. Hill and M. Dall'Aglio, "Bayesian posteriors without bayes' theorem," *preprint arXiv:1203.0251*, 2012.
[58] A. Juditsky, P. Rigollet, A. B. Tsybakov, *et al.*, "Learning by mirror averaging," *The Annals of Statistics*, vol. 36, no. 5, pp. 2183–2206, 2008.
[59] G. Lan, A. Nemirovski, and A. Shapiro, "Validation analysis of mirror descent stochastic approximation method," *Mathematical programming*, vol. 134, no. 2, pp. 425–458, 2012.
[60] J. Li, G. Li, Z. Wu, and C. Wu, "Stochastic mirror descent method for distributed multi-agent optimization," *Optimization Letters*, pp. 1–19, 2016.
[61] C. W. Fox and S. J. Roberts, "A tutorial on variational bayesian inference," *Artificial intelligence review*, vol. 38, no. 2, pp. 85–95, 2012.
[62] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University of London United Kingdom, 2003.
[63] B. Dai, N. He, H. Dai, and L. Song, "Provable bayesian inference via particle mirror descent," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 985–994, 2016.
[64] B. O. Koopman, "On distributions admitting a sufficient statistic," *Transactions of the American Mathematical society*, vol. 39, no. 3, pp. 399–409, 1936.
[65] G. Darmois, "Sur les lois de probabilitéa estimation exhaustive," *CR Acad. Sci. Paris*, vol. 260, no. 1265, p. 85, 1935.
[66] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA, 2014.
[67] A. Nedić, A. Olshevsky, and C. A. Uribe, "Distributed gaussian learning over time-varying directed graphs," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 1710–1714, Nov 2016.
[68] C. Wang and B. Chazelle, "Gaussian learning-without-recall in a dynamic social network," *arXiv preprint arXiv:1609.05990*, 2016.
[69] L. LeCam, "Convergence of estimates under dimensionality restrictions," *The Annals of Statistics*, pp. 38–53, 1973.
[70] A. W. Van der Vaart, *Asymptotic statistics*, vol. 3. Cambridge university press, 2000.