Non-asymptotic Concentration Rates in Cooperative Learning Part II: Inference on Compact Hypotheses Sets

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 the exponential families. Additionally, 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. In Part II, we provide the first non-asymptotic belief concentration rate analysis for distributed non-Bayesian learning over networks on compact hypotheses sets. Additionally, we provide extensive numerical analysis for various distributed inference tasks on networks for observational models in the exponential family of distributions.**

***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 main contribution of this paper (Part II) are as follows:

- We provide the first non-asymptotic belief concentration analysis for non-Bayesian distributed learning over ***compact hypotheses sets***.
- We show the proposed update rule concentrates its beliefs on compact balls around the optimal set at geometric rate.

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.

- We provide extensive numerical results for various distributed inference tasks with observational models in the exponential family of distributions.

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 shows our main results about the exponential concentration of beliefs around the true parameter. Section III 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 $\mathbb{R}^d$. Section IV presents a set of numerical analysis and simulation results for the proposed algorithms for the distributed estimation of parameters of distributions from the exponential family for various networks topologies and number of agents. 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 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 shown in Part I of this paper series, the above problem

[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.

can be written as the optimization problem

$$\min_{\theta\in\Theta} F(\theta) \triangleq D_{KL}\left(\boldsymbol{P}\|\boldsymbol{P}_\theta\right) = \sum_{i=1}^{n} D_{KL}(P^i\|P^i_\theta). \tag{2}$$

We propose the following algorithm as a distributed version of the stochastic mirror descent for the solution of problem Eq. (2):

$$d\mu^i_{k+1} = \arg\min_{\pi\in\Delta_\Theta}\left\{\langle -\log p^i_\theta(x^i_{k+1}),\pi\rangle + \sum_{j=1}^{n} a_{ij} D_{KL}(\pi\|d\mu^j_k)\right\} \quad \text{where } \theta\sim\pi, \tag{3}$$

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. (3) 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

$$\mu^i_{k+1}(B) \propto \int_B p^i_\theta(x^i_{k+1}) \prod_{j=1}^{n} (d\mu^j_k(\theta))^{a_{ij}}. \tag{4}$$

The update in Eq. (4) 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].

III. Belief Concentration Rates

We now turn to the presentation of our main results about rate at which beliefs generated by the update rule in Eq. (4) concentrate around the true parameter θ^*. In Part I of this paper series, we will focus on the case when Θ is a finite set, and prove a concentration rate on the beliefs on a Hellinger ball around the optimal hypothesis. Contrary to Part I, in this section we focus on the case when Θ is a compact subset of $\mathbb{R}^d$. 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,

$$h^2(P,Q) = \frac{1}{2}\int\left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}}\right)^2 d\lambda, \tag{5}$$

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

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. 5 defines the squared Hellinger distance h^2, rather than h.

Definition 1. *Define an n-Hellinger ball of radius r centered at θ as*

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

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 $\mathcal{B}^c_r$. Specifically, we are going to express $\mathcal{B}^c_r$ 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 $\mathcal{B}^c_r$ as the union of annuli generated by the sequence $\{r_l\}$ as

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

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. (4) 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 1. *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. A Concentration Result for a Compact Set of Hypotheses

Next, we will study the non-asymptotic belief concentration process when the hypothesis set Θ is a compact subset of $\mathbb{R}^d$. We additionally require the map from Θ to $\prod_{i=1}^n P_\theta^i$ be continuous (where the topology on the space of distributions comes from the Hellinger metric). This will be useful in defining coverings, which will be made clear shortly.

Definition 2. *Let (M,d) be a metric space. A subset $S \subseteq M$ is called ε-separated with $\varepsilon > 0$ if $d(x,y) \geq \varepsilon$ for any $x, y \in S$. Moreover, for a set $B \subseteq M$, let $N_B(\varepsilon)$ be the smallest number of Hellinger balls with centers in S of radius ε needed to cover the set B, i.e., such that $B \subseteq \bigcup_{m\in S} \mathcal{B}_\varepsilon(m)$.*

As before, given a decreasing sequence $1 = r_1 \geq r_2 \geq \cdots \geq r_L = r$, we will define the annulus $\mathcal{F}_l$ to be $\mathcal{F}_l = \mathcal{B}_{r_l} \backslash \mathcal{B}_{r_{l+1}}$. Furthermore, S_{ε_l} will denote maximal ε_l-separated subset of $\mathcal{F}_l$. Finally, $K_l = |S_{\varepsilon_l}|$.

Remark 1. *Note that Definition 2 induces a covering of the sets $\mathcal{F}_l$ by K_l balls of radius ε_l, centered at points $m \in S_{\varepsilon_l}$, i.e., $\mathcal{B}_{\varepsilon_l}(m)$. From this covering we can deduce a partition $\mathcal{F}_l = \bigcup_{m\in S} \mathcal{F}_{m,l}$, where each $\mathcal{F}_{m,l} \in \mathcal{B}_{\varepsilon_l}(m)$.*

We note that, as a consequence of our assumption that the map from Θ to $\prod_{i=1}^n P_\theta^i$ is continuous, we have that each K_l is finite (since the image of a compact set under a continuous map is compact). Thus, we have the following covering of $\mathcal{B}_r^c$:

$$\mathcal{B}_r^c \subseteq \bigcup_{l=1}^{L-1} \bigcup_{m\in S_{\varepsilon_l}} \mathcal{F}_{l,m},$$

where each $\mathcal{F}_{l,m}$ is the intersection of a ball centered at an element in S_{ε_l} with $\mathcal{F}_l$. Figure 1 shows the elements of a covering for a set $\mathcal{B}_r^c$. The cluster of circles at the top right corner represents the balls $\mathcal{B}_{\varepsilon_l}$ and, for a specific case in the left of the image, we illustrate the set $\mathcal{F}_{l,m}$.

Without loss of generality we will make the following technical assumption that will be convenient for the analysis of the concentration of beliefs on compact sets.

We will require a continuity assumption of the likelihood modes with respect to the parameter space Θ for our non-asymptotic analysis.

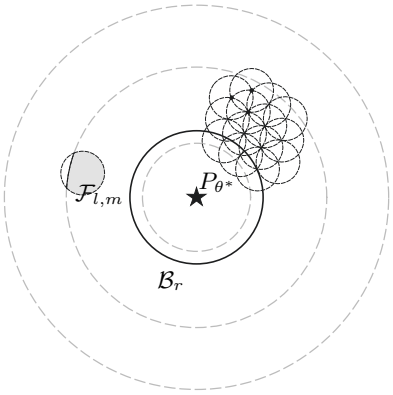

Fig. 1: Creating a covering for a set $\mathcal{B}_r$. ★ represents the correct hypothesis $\boldsymbol{P}_{\theta^*}$.

Assumption 2. *The likelihood function $p_\theta(x)$ is continuous on Θ with respect to θ for any $x \in \mathcal{X}$.*

Assumption 2 will hold for large classes of likelihood functions, in particular for the exponential family of distributions. For example, it trivially holds for Gaussian distributions with known variance and known mean. In general this assumption forbids arbitrarily large changes in the likelihood model for infinitesimal changes in the parameter. We will use this assumption later to guarantee the existence of a parameter inside a closed ball in the parameter space that minimizes the integral likelihood model defined on the ball for any measurable subset of the observation space. Assumption 2 is only a sufficient condition. The interested reader can see [70, Chapter 5] for an extensive account of weaker assumption to guarantee congruence of an estimator. Moreover, Assumption 2 will allows us to state the following auxiliary result.

Proposition 2. *Let B be a closed n-Hellinger ball (c.f. Definition 1) centered at $\theta_B \in \Theta$, and let Assumption 2 hold for the family of distributions $\mathscr{P}^i = \{P_\theta^i : \theta \in \Theta\}$ for $i \in V$. Then, there exists a θ such that for all measurable sets $\{X_t^i\}$ for $i \in V$ and $t = 1, \cdots k$. It holds that*

$$\int_B \prod_{t=1}^{k}\prod_{j=1}^{n} p_{\hat{\theta}}^j(X_t^j)^{[A^{k-t}]_{ij}} d\mu_0(\hat{\theta}) \geq \prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}.$$

Proof. The closedness of the n-Hellinger ball B, and the continuity with respect to θ in Assumption 2 are sufficient for extreme values to exists by the Weierstrass extreme value theorem. □

We next provide a concentration result for the logarithmic likelihood of a ratio of densities, which will serve the same technical function as Lemma **??** in the countable hypothesis case. We begin by defining two measures. For a hypothesis θ and a measurable set $B \subseteq \Theta$, let $\boldsymbol{P}_B^{\otimes k}$ be the probability distribution with density, (i.e., Radon-Nikodym derivative with respect to $\lambda^{\otimes nk}$),

$$g_B(\boldsymbol{x}^k) = \frac{1}{\mu_0(B)} \int_B \prod_{t=1}^{k}\prod_{j=1}^{n} p_\theta^j(x_t^j) d\mu_0(\theta). \tag{6}$$

Similarly, let $\bar{\boldsymbol{P}}_B^{\otimes k}$ be the measure with density

$$\bar{g}_B(\boldsymbol{x}^k) = \frac{1}{\mu_0(B)} \int_B \prod_{t=1}^{k} \prod_{j=1}^{n} (p_\theta^j(x_t^j))^{[A^{k-t}]_{ij}} d\mu_0(\theta). \quad (7)$$

Moreover, with some notation abuse define

$$g_\theta(\boldsymbol{x}^k) = \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(x_t^j), \quad (8)$$

$$\bar{g}_\theta(\boldsymbol{x}^k) = \prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(x_t^j)^{[A^{k-t}]_{ij}}. \quad (9)$$

Note that $\bar{\boldsymbol{P}}_B^{\otimes k}$'s are not probability distributions due to the exponential weights. Nonetheless, they are bounded and positive. The next lemma shows the concentration of the logarithmic ratio of a weighted density as defined in Eq. (7) for a sets B and a density at an arbitrary hypotheses $\hat{\theta} \in \Theta$, in terms of the probability distribution $\boldsymbol{P}_{\hat{\theta}}^{\otimes k}$.

Lemma 3. *Let Assumptions 1, and 2 hold. Consider a measurable sets $B \subset \Theta$ with positive measures, and assume that $B \subset \mathcal{B}_r(\theta_B)$ where $\mathcal{B}_r(\theta_B)$ and $\theta_B \in \Theta$. Moreover, let $\theta \in \Theta$ be an arbitrary element of the parameter space. Then, for all $y \in \mathbb{R}$*

$$\mathbb{P}_\theta \left[\log \frac{\bar{g}_{B_1}(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)} \geq y \right] \leq \exp\left(-\frac{y}{2} + \frac{4 \log n}{1-\delta} - \frac{k}{n} \sum_{j=1}^{n} \left(h(P_{\theta_B}^j, P_\theta^j) - r \right)^2 \right),$$

where $\mathbb{P}_\theta$ is the probability measure that gives $\boldsymbol{X}^k$ a distribution $\boldsymbol{P}_\theta^{\otimes k}$ with density g_θ as defined in Eq. (8).

Proof. By the Markov inequality, it follows that

$$\mathbb{P}_\theta \left[\log \frac{\bar{g}_B(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)} \geq y \right] \leq \exp(-y/2) \mathbb{E}_\theta \left[\sqrt{\frac{\bar{g}_B(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)}} \right] = \exp(-y/2) \int_{\boldsymbol{\mathcal{X}}^k} \sqrt{\frac{\bar{g}_B(\boldsymbol{x}^k)}{\bar{g}_\theta(\boldsymbol{x}^k)}} g_\theta(\boldsymbol{x}^k) d\lambda^{\otimes kn}(\boldsymbol{x}^k).$$

Initially, note that by Jensen's inequality[3], with $x^{[A^{k-t}]_{ij}}$ being a concave function and $1/\mu_0(B) \int_B d\mu_0 = 1$, we have that

$$\bar{g}_B(\boldsymbol{x}^k) = \frac{1}{\mu_0(B)} \int_B \prod_{t=1}^{k} \prod_{j=1}^{n} \left(p_\theta^j(x_t^j) \right)^{[A^{k-t}]_{ij}} d\mu_0(\theta) \leq \prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{1}{\mu_0(B)} \int_B p_{\hat{\theta}}^j(x_t^j) d\mu_0(\hat{\theta}) \right)^{[A^{k-t}]_{ij}}.$$

[3]For a concave function ϕ and $\int_\Omega f(x)dx = 1$, it holds that $\int_\Omega \phi(g(x)) f(x) dx \leq \phi\left(\int_\Omega g(x) f(x) dx \right)$.

Therefore,

$$\sqrt{\frac{\bar{g}_B(\boldsymbol{x}^k)}{\bar{g}_\theta(\boldsymbol{x}_t^{j\,k})}} \leq \sqrt{\frac{\prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{1}{\mu_0(B)} \int_B p_{\hat{\theta}}^j(x) d\mu_0(\hat{\theta}) \right)^{[A^{k-t}]_{ij}}}{\prod_{t=1}^{k} \prod_{j=1}^{n} p_\theta^j(x_t^j)^{[A^{k-t}]_{ij}}}} = \sqrt{\prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{\frac{1}{\mu_0(B)} \int_B p_{\hat{\theta}}^j(x_t^j) d\mu_0(\hat{\theta})}{p_\theta^j(x_t^j)} \right)^{[A^{k-t}]_{ij}}}$$

Next, applying the same argument with the Jensen's inequality and $x^{[A^{k-t}]_{ij}}$ but now with respect to the measure $g_\theta(\boldsymbol{x}^k) d\lambda^{\otimes kn}(\boldsymbol{x}^k)$ we obtain

$$\begin{aligned} \mathbb{P}_\theta \left[\log \frac{\bar{g}_B(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)} \geq y \right] &\leq \exp(-y/2) \times \\ &\times \int_{\boldsymbol{\mathcal{X}}^k} \sqrt{\prod_{t=1}^{k} \prod_{j=1}^{n} \left(\frac{\frac{1}{\mu_0(B)} \int_B p_{\hat{\theta}}^j(x_t^j) d\mu_0(\hat{\theta})}{p_\theta^j(x_t^j)} \right)^{[A^{k-t}]_{ij}}} \\ &\times g_\theta(\boldsymbol{x}^k) d\lambda^{\otimes kn}(\boldsymbol{x}^k) \\ &\leq \exp(-y/2) \times \\ &\times \prod_{t=1}^{k} \prod_{j=1}^{n} \left(\int_{\boldsymbol{\mathcal{X}}^k} \sqrt{\frac{\frac{1}{\mu_0(B)} \int_B p_\theta^j(x_t^j) d\mu_0(\theta)}{p_\theta^j(x_t^j)}} \right. \\ &\left. p_\theta^j(x_t^j) d\lambda^{\otimes kn}(\boldsymbol{x}^k) \right)^{[A^{k-t}]_{ij}}. \end{aligned}$$

Thus, we obtain

$$\mathbb{P}_\theta \left[\log \frac{\bar{g}_B(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)} \geq y \right] \leq \exp\left(-\frac{y}{2} \right) \prod_{t=1}^{k} \prod_{j=1}^{n} \rho\left(P_B^j, P_\theta^j \right)^{[A^{k-t}]_{ij}},$$

where P_B^j is the measure with Radon-Nikodym derivative $g_B^j(x) = \frac{1}{\mu_0(B)} \int_B p_\theta^j(x) d\mu_0(\theta)$ with respect to λ.

Now, recall from the call that the Hellinger distance is bounded by above by 1. Thus similarly as in Eq. (**??**), we have that

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

Moreover we can add and subtract $\sum_{t=1}^{k} \sum_{j=1}^{n} \frac{1}{n} h^2(P_B^j, P_\theta^j)$, thus,

$$\prod_{t=1}^{k} \prod_{j=1}^{n} \rho(P_B^j, P_\theta^j)^{[A^{k-t}]_{ij}} \leq \exp\left(-\sum_{t=1}^{k} \sum_{j=1}^{n} \left([A^{k-t}]_{ij} - \frac{1}{n} \right) h^2(P_B^j, P_\theta^j) \right.$$

$$-\frac{k}{n}\sum_{j=1}^{n} h^2(P_B^j, P_\theta^j)\Bigg).$$

Additionally, from Lemma 1,

$$\sum_{t=1}^{k}\sum_{j=1}^{n}\left([A^{k-t}]_{ij}-\frac{1}{n}\right)h^2(P_B^j,P_\theta^j)\leq\frac{4\log n}{1-\delta},$$

from which we can conclude that

$$\mathbb{P}_\theta\left[\log\frac{\bar{g}_B(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)}\geq y\right] \leq \exp\left(-\frac{y}{2}+\frac{4\log n}{1-\delta}-\frac{k}{n}\sum_{j=1}^{n}h^2(P_B^j,P_\theta^j)\right).$$

Finally, note that by [42, Corollary 1], and Definition 1, it follows that $h(P_B^j,P_\theta^j)\geq h\big(P_{\theta_B}^j,P_\theta^j\big)-r$, and the desired result follows.

□

Lemma 3 provides a concentration result for the logarithmic ratio between a weighted densities over a subsets B and a density on an arbitrary point θ. The terms involving the auxiliary variable y and the influence of the graph, via δ are the same as in Lemma 4 in Part I of this paper series. Moreover, the rate at which this bound decays exponentially is influenced now by the radius of the Hellinger balls B and θ.

We are ready now to state our main result regarding the concentration of beliefs around θ^* for compact sets of hypotheses.

Theorem 4. *Let Assumptions 1, and 2 hold, and let $\sigma\in(0,1)$ be a given probability tolerance level. Moreover, for any $r\in(0,1)$. Moreover, assume all agents start with equal initial beliefs. Then, the beliefs $\{\mu_k^i\}$, $i\in V$, generated by the update rule in Eq. (4) have the following property: with probability $1-\sigma$,*

$$\mu_{k+1}^i(\mathcal{B}_r)\geq 1-C\exp\left(-\frac{k}{8}r^2\right)\ \ \forall i \text{ and } k\geq N,$$

where

$$N=\inf\left\{t\geq 1\middle|2\exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}K_l\exp\left(-\frac{t}{8}r_{l+1}^2\right)<\sigma\right\}$$

$C=\sum_{l=1}^{L-1}\exp(-\frac{1}{8}r_{l+1}^2)$ and $\delta=1-\eta/n^2$, where η is the smallest positive element of the matrix A.

Proof. Lets start by analyzing the evolution of the beliefs on a measurable set B with $\theta^*\in B$. From Eq. (4) we have that

$$\mu_k^i(B)=\frac{\int_B\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}\prod_{j=1}^{n}d\mu_0^j(\theta)^{[A^k]_{ij}}}{\int_\Theta\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}\prod_{j=1}^{n}d\mu_0^j(\theta)^{[A^k]_{ij}}}$$

$$\geq 1-\frac{\int_{B^c}\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}d\mu_0(\theta)}{\int_B\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}d\mu_0(\theta)}$$

where $\prod_{j=1}^{n}d\mu_0^j(\theta)^{[A^k]_{ij}}=d\mu_0(\theta)$ follows from the assumption of equal initial beliefs for all agents.

Now lets focus specifically on the case where B is a n-Hellinger ball of radius $r\in(0,1)$ with center at θ^*, i.e., $\mathcal{B}_r$. For analysis purposes, we will let the radius r to be fixed, and we are actually going to analyze the concentration of beliefs on a smaller ball with radius R_k. The radius R_k needs to be small enough, so we will impose the corresponding upper bound when needed.

In addition, since $R_k<r$, we get

$$\mu_k^i(\mathcal{B}_r)\geq 1-\frac{\int_{\mathcal{B}_r^c}\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}d\mu_0(\theta)}{\int_{\mathcal{B}_{R_k}}\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}d\mu_0(\theta)}.$$

Following Proposition 2, it follows that there exists a $\theta\in\mathcal{B}_r$ such that

$$\mu_k^i(\mathcal{B}_r)\geq 1-\frac{\int_{\mathcal{B}_r^c}\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}d\mu_0(\theta)}{\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}}.$$

Furthermore, we can use the covering of the set $\mathcal{B}_r^c$ to obtain,

$$\mu_k^i(\mathcal{B}_r)\geq 1-\frac{\sum_{l=1}^{L-1}\sum_{m=1}^{K_l}\int_{\mathcal{F}_{l,m}}\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}d\mu_0(\theta)}{\prod_{t=1}^{k}\prod_{j=1}^{n}p_\theta^j(X_t^j)^{[A^{k-t}]_{ij}}}$$

$$\geq 1-\frac{\sum_{l=1}^{L-1}\sum_{m=1}^{K_l}\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)\mu_0(\mathcal{F}_{l,m})}{\bar{g}_\theta(\boldsymbol{X}^k)}, \tag{10}$$

where by definition, each $\mathcal{F}_{l,m}$ is contained in a n-Hellinger ball of radius ε_l centered at a point $m\in S_{\varepsilon_l}$

Equation 10 defines a ratio between two densities, i.e. $\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)/\bar{g}_\theta(\boldsymbol{X}^k)$, where the numerator is defined over the set $\mathcal{F}_{l,m}$ and the denominator with respect to $\theta\in\mathcal{B}_r\subset\Theta$.

Lemma 3 provides a way to bound term $\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)/\bar{g}_\theta(\boldsymbol{X}^k)$ with high probability. Thus

$$\mathbb{P}_\theta\left(\left\{\boldsymbol{X}^k\middle|\sup_{l,m}\log\frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)}\geq y\right\}\right) \leq\sum_{l=1}^{L-1}\sum_{m=1}^{K_l}\mathbb{P}_{\mathcal{B}_{R_k}}\left(\log\frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)}\geq y\right)$$

$$\leq\sum_{l=1}^{L-1}\sum_{m=1}^{K_l}\exp\left(\frac{y}{2}+\frac{4\log n}{1-\delta}-\frac{k}{n}\sum_{j=1}^{n}\left(h(P_m^j,P_\theta^j)-\varepsilon_l\right)^2\right), \tag{11}$$

where P_m^j is the distribution at a point $m \in S_{\varepsilon_l}$, where S_{ε_l} is the maximal ε_l separated set of $\mathcal{F}_l$ as in Definition 2. Now, lets analyze the result in (11) with respect to the used covering.

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}\Big(h(P_m^j,P_\theta^j)-\varepsilon_l\Big)^2\\
&=\frac{1}{n}\sum_{j=1}^{n}\Big(h^2(P_m^j,P_\theta^j)-2\varepsilon_l h(P_m^j,P_\theta^j)+\varepsilon_l^2\Big)\\
&\geq\frac{1}{n}\sum_{j=1}^{n}\Big(h^2(P_m^j,P_\theta^j)-2\varepsilon_l h(P_m^j,P_\theta^j)\Big)\\
&=\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P_\theta^j)-2\varepsilon_l\frac{1}{n}\sum_{j=1}^{n}h(P_m^j,P_\theta^j).
\end{aligned}$$

Moreover, note that $\Big(\frac{1}{n}\sum_{j=1}^{n}h(P_m^j,P^j)\Big)^2 \leq \frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P^j)$, thus

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P_\theta^j)-2\varepsilon_l\frac{1}{n}\sum_{j=1}^{n}h(P_m^j,P_\theta^j)\\
&\geq\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P_\theta^j)-2\varepsilon_l\sqrt{\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P_\theta^j)}.
\end{aligned}$$

Now, applying the triangle inequality, we know that

$$\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P_\theta^j)\geq\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P^j)-\frac{1}{n}\sum_{j=1}^{n}h^2(P^j,P_\theta^j),$$

and by definition we know that $\frac{1}{n}\sum_{j=1}^{n}h^2(P^j,P_\theta^j)\leq R_k^2$, and $\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P^j)\geq r_{l+1}^2$, thus

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P_\theta^j)&\geq r_{l+1}^2-R_k^2\\
\frac{1}{n}\sum_{j=1}^{n}h^2(P_m^j,P_\theta^j)&\geq\frac{1}{2}r_{l+1}^2,
\end{aligned}$$

where the last inequality follows by setting $R_k\leq r_{l+1}/\sqrt{2}$. Therefore, so far we have

$$\frac{1}{n}\sum_{j=1}^{n}\Big(h(P_m^j,P_\theta^j)-\varepsilon_l\Big)^2\geq\frac{1}{2}r_{l+1}^2-2\varepsilon_l r_l.$$

Next, we can set $\varepsilon_l=r_{l+1}/16$, and $r_l\leq 2r_{l+1}$, and obtain

$$\frac{1}{n}\sum_{j=1}^{n}\Big(h(P_m^j,P_\theta^j)-\varepsilon_l\Big)^2\geq\frac{1}{4}r_{l+1}^2.$$

We can conclude that

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left\{\boldsymbol{X}^k\,\middle|\,\sup_{l,m}\log\frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)}\geq y\right\}\right)\\
&\leq\sum_{l=1}^{L-1}\sum_{m=1}^{K_l}\exp\left(\frac{y}{2}+\frac{4\log n}{1-\delta}-k\frac{r_{l+1}^2}{4}\right).
\end{aligned}\tag{12}$$

Thus, using Eq. (12), and setting $y=-\frac{k}{8}r_{l+1}^2$ in Eq. (11), it follows that

$$\begin{aligned}
&\mathbb{P}_\theta\left(\left\{\boldsymbol{X}^k\,\middle|\,\sup_{l,m}\log\frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)}\geq y\right\}\right)\\
&\leq\sum_{l=1}^{L-1}K_l\exp\left(\frac{4\log n}{1-\delta}+\frac{k}{8}r_{l+1}^2-k\frac{1}{4}r_{l+1}^2\right)\\
&\leq\exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1}K_l\exp\left(-\frac{k}{8}r_{l+1}^2\right).
\end{aligned}\tag{13}$$

The probability measure in Eq. (13) is computed for $\boldsymbol{X}^k$ distributed according to $\boldsymbol{P}_\theta^{\otimes k}$. Nonetheless, $\boldsymbol{X}^k$ is distributed according to the (slightly different) $\boldsymbol{P}^{\otimes k}$. Our next step is to relate these two measures.

First, note that it holds that the total variation distance $D(P^n,Q^n)\geq h^2(P^n,Q^n)=1-\rho^n(P,Q)$, see for example, [69, Proof of Lemma 1];. Now, by definition of the Hellinger metric for any measurable set B it holds that

$$\sup_B|\mathbb{P}_\theta^{\otimes k}(B)-\mathbb{P}^{\otimes k}(B)|^2\leq 1-\rho^2(\boldsymbol{P}_\theta^{\otimes k},\boldsymbol{P}^{\otimes k}),$$

and by definition of the Hellinger affinity we have that

$$\begin{aligned}
\sup_B\big(\mathbb{P}_\theta^{\otimes k}(B)-\mathbb{P}^{\otimes k}(B)\big)^2&=1-(1-h^2(\boldsymbol{P}_\theta^{\otimes k},\boldsymbol{P}^{\otimes k}))^2\\
&\leq 2h^2(\boldsymbol{P}_\theta^{\otimes k},\boldsymbol{P}^{\otimes k}),
\end{aligned}$$

where first we have used the relation that for any $x\in\mathbb{R}$, it holds that $1-(1-x^2)^2<2x^2$. Moreover, we know that for Hellinger distances the following inequality holds

$$\begin{aligned}
h^2(\boldsymbol{P}_\theta^{\otimes k},\boldsymbol{P}^{\otimes k})&=2-2\prod_{t=1}^{k}\prod_{j=1}^{n}(1-\frac{1}{2}h^2(P_\theta^j,P^j))&(14)\\
&\leq 2-2\exp\left(-nk\frac{1}{2}\frac{1}{n}\sum_{j=1}^{n}h^2(P_\theta^j,P^j)\right)&(15)\\
&\leq 2-2\exp\left(-nk\frac{1}{2}R_k^2\right)&(16)
\end{aligned}$$

Then, from the fact that $\theta\in\mathcal{B}_{R_k}$, we have

$$\sup_B\big(\mathbb{P}_\theta(B)-\mathbb{P}^{\otimes k}(B)\big)^2\leq 4\big(1-\exp\big(-nk\frac{1}{2}R_k^2\big)\big).$$

Therefore it suffices to have

$$R_k\leq\sqrt{\frac{2}{nk}\log\left(\frac{16}{1-\sigma^2}\right)},\tag{17}$$

to obtain

$$\sup_B\big(\mathbb{P}_\theta^{\otimes k}(B)-\mathbb{P}^{\otimes k}(B)\big)^2\leq\frac{\sigma}{2}.$$

Additionally, as a result we have the following restrictions about how small the radius R_k needs to be:

$$R_k\leq\min\left\{\sqrt{\frac{2}{nk}\log\left(\frac{16}{1-\sigma^2}\right)},r/\sqrt{2}\right\}.$$

Therefore, by considering the measurable subset $\Gamma^k=\left\{\boldsymbol{X}^k\,\middle|\,\sup_{l,m}\log\frac{\bar{g}_{\mathcal{F}_{l,m}}(\boldsymbol{X}^k)}{\bar{g}_\theta(\boldsymbol{X}^k)}\geq-\frac{k}{16}r_{l+1}^2\right\}$, we have that

$$\mathbb{P}\left(\Gamma^k\right)<\mathbb{P}_\theta\left(\Gamma^k\right)+\sqrt{4\big(1-k\exp\big(-n\frac{1}{2}R_k^2\big)\big)}.$$

$$\leq \exp\left(\frac{4\log n}{1-\delta}\right)\sum_{l=1}^{L-1} K_l \exp\left(-\frac{k}{16}r_{l+1}^2\right)+\frac{\sigma}{2}.$$

Furthermore, we are interested in finding a large enough k such that the probability described in Eq. (13) is at most σ. Thus, we define

$$N \geq \inf\left\{ t\geq 1 \middle| \exp\left(\frac{4\log n}{1-\delta}\right)\times \right. \\ \left. \times \sum_{l=1}^{L-1} K_l \exp\left(-\frac{t}{8}r_{l+1}^2\right) < \frac{\sigma}{2}\right\}.$$

Moreover, from Eq. (10) we obtain that with probability $1-\sigma$ for all $k \geq N$,

$$\begin{aligned}\mu_k^i(\mathcal{B}_r) &\geq 1-\sum_{l=1}^{L-1}\sum_{m=1}^{K_l}\exp\left(-\frac{k}{8}r_{l+1}^2\right)\mu_0(\mathcal{F}_{l,m}) \\ &= 1-\sum_{l=1}^{L-1}\exp\left(-\frac{k}{8}r_{l+1}^2\right)\mu_0(\mathcal{F}_l) \\ &\geq 1-\sum_{l=1}^{L-1}\exp\left(-\frac{k}{8}r_{l+1}^2\right).\end{aligned}$$

Note that in the last upper bound follows from the fact that $\mu_0(\mathcal{F}_l) \leq 1$ given that $\mathcal{F}_l \subseteq \Theta$. Moreover, we have that $\sum_{m=1}^{K_l}\mu_0(\mathcal{F}_{l,m}) = \mu_0(\mathcal{F}_l)$, see Remark 1.

Now, lets define $\chi = \sum_{l=1}^{L-1}\exp\left(-\frac{1}{8}r_{l+1}^2\right)$, then it follows that

$$\begin{aligned}\mu_k^i(\mathcal{B}_r) &\geq 1-\sum_{l=1}^{L-1}\exp\left(-\frac{k}{8}r_{l+1}^2\right) \\ &= 1-\sum_{l=1}^{L-1}\exp\left(-\frac{1}{8}r_{l+1}^2\right)\exp\left(-\frac{k-1}{8}r_{l+1}^2\right) \\ &\geq 1-C\exp\left(-\frac{k-1}{8}r^2\right),\end{aligned}$$

or equivalently $\mu_{k+1}^i(\mathcal{B}_r) \geq 1-C\exp(-\frac{k}{8}r^2)$.

□

Analogous to Theorem 5 in Part I, Theorem 4 provides a probabilistic concentration result for the agents' beliefs around a Hellinger ball of radius r with center at θ^* for sufficiently large k. We provide an explicit number of iterations after which an exponential concentration occurs. Moreover, the rate at which this happens is proportional to the radius r of a ball around the optimal hypotheses.

IV. Experimental Results

In this section, we show a number of experimental results for the problem of distributed estimation of network-wide parameters for various network topologies and various observational models. In particular we focus on observational models form the exponential family of distributions.

Table I recall the results from Part I for a number of distributed estimation problems with likelihood models coming from exponential families. Particularly, we describe the relation between the distribution of the observations, the parameter space and the belief distributions. Moreover, we provide explicit relations between the parameters in the canonical form and the corresponding parameters of the beliefs.

We present the experimental results with the following format.

We explore six different estimation problems

- Figure 2: Distributed estimation of network-wide mean parameter with Gaussian observations with local knowledge of private variances.
- Figure 3: Distributed estimation of network-wide variance parameter with Gaussian observations with local knowledge of private means.
- Figure 4: Distributed estimation of network-wide mean and variance parameters with Gaussian observations without knowledge of local means or variances.
- Figure 5: Distributed estimation of network-wide parameter with heterogeneous Bernoulli observations.
- Figure 6: Distributed estimation of network-wide parameter with heterogeneous Poisson observations.
- Figure 7: Distributed estimation of network-wide parameter with heterogeneous Exponential observations.

For each of the figures described above, we measure the performance of the proposed algorithm using its normalized distance to optimality and the distance to consensus, defined as follows

$$\begin{aligned}\text{Distance to Optimality: } &\frac{|F(\theta_k)-F(\theta^*)|}{|F(\theta_0)-F(\theta^*)|}, \\ \text{Distance to Consensus: } &\|\mathcal{L}\theta_k\|_2^2,\end{aligned}$$

where $\theta_k = (\theta_k^1, \theta_k^2, \cdots, \theta_k^n)$ is the aggregation of all the current parameters estimation for each of the agents, the function $F(\theta_k)$ is defined as

$$F(\theta_k) = \sum_{i=1}^{n} D_{KL}(P^i \| P^i_{\theta_k^i}),$$

and L is the graph Laplacian of the communication graph. We have used the graph Laplacian as a measure to distance of consensus since by definition the set where $\theta_k^1 = \theta_k^2 = \cdots = \theta_k^n$, i.e. consensus, is null space of the matrix $\mathcal{L}$.

Finally, we present the results for five classes of networks, namely: complete graphs, cycle graphs, path graphs, star graphs, and Erdős-Rényi random graphs. For each of the network classes we show the performance for 10 agents, 100 agents, and 1000 agents.

In all experimental results, the predicted geometric convergence rate is observed. Moreover, as the number of agents in the network increases, the effects of the network topology become more evident. Particularly, for highly connected graphs such as the complete graph or the Erdős-Rényi, the distance to optimality and consensus decays faster. One interesting observation is that contrary to what was expected, the performance of the proposed algorithm on graphs with a star topology is worst is most of the cases. This can be explained by the fact, that given that the agents are in a well connected graph, they are oblivious to the topology of the network, and thus cannot

Observations X_k^i	Parameter Space Θ	Beliefs Distribution	$T(x)$	$M(\theta)$	Belief Parameters
Bern(θ^i)	$\{\theta \in [0,1]\}$	Beta(α_k^i, β_k^i)	x	$\log \frac{\theta}{1-\theta}$	$\begin{bmatrix} \alpha_k^i = \chi_k^i + 1 \\ \beta_k^i = \chi_k^i + \nu_k^i + 1 \end{bmatrix}$
Binomial(θ^i, m^i)	$\{\theta \in [0,1]\}$	Beta(α_k^i, β_k^i)	x	$\log \frac{\theta}{1-\theta}$	$\begin{bmatrix} \alpha_k^i = \chi_k^i + 1 \\ \beta_k^i = \chi_k^i + m^i \nu_k^i + 1 \end{bmatrix}$
Multinomial(θ^i, m^i)	$\{\theta \in [0,1]^d, \sum \theta_i = 1\}$	Dirichlet($\alpha_k^i \in \mathbb{R}_+^d$)	x	$\log \theta$	$\begin{bmatrix} \alpha_k^i = \chi_k^i + \mathbf{1} \end{bmatrix}$
Poisson(θ^i)	$\{\theta > 0\}$	Gamma(α_k^i, β_k^i)	x	$\log \theta$	$\begin{bmatrix} \alpha_k^i = \chi_k^i + 1 \\ \beta_k^i = \nu_k^i \end{bmatrix}$
Exp(θ^i)	$\{\theta > 0\}$	Gamma(α_k^i, β_k^i)	x	$-\theta$	$\begin{bmatrix} \alpha_k^i = \nu_k^i + 1 \\ \beta_k^i = \chi_k^i \end{bmatrix}$
$\mathcal{N}(\theta^i \vert (\sigma^i)^2)$	$\{\theta \in \mathbb{R}\}$	$\mathcal{N}(\bar{\theta}_k^i, (\bar{\sigma}_k^i)^2)$	x	$\frac{\theta}{(\sigma^i)^2}$	$\begin{bmatrix} \bar{\theta}_k^i = \frac{\chi_k^i}{\nu_k^i} \\ (\bar{\sigma}_k^i)^2 = \frac{(\sigma^i)^2}{\nu_k^i} \end{bmatrix}$
$\mathcal{N}(\tau^i \vert \theta^i)$	$\{\tau > 0\}$	Gamma(α_k^i, β_k^i)	$\frac{1}{2}(x-\theta^i)^2$	$-\tau$	$\begin{bmatrix} \alpha_k^i = \frac{\nu_k^i}{2} + 1 \\ \beta_k^i = \chi_k^i \end{bmatrix}$
$\mathcal{N}(\theta^i, \tau^i)$	$\{\theta \in \mathbb{R}, \tau > 0\}$	$\mathcal{N}$-Gamma($\bar{\theta}_k^i, \bar{\tau}_k^i, \alpha_k^i, \beta_k^i$)	$\begin{bmatrix} x^2 \\ x \\ \frac{1}{2} \end{bmatrix}$	$\begin{bmatrix} -\frac{1}{2}\tau \\ \tau\theta \\ \log \tau \end{bmatrix}$	$\begin{bmatrix} \alpha_k^i = [\chi_k^i]_1 - \frac{1}{2} \\ \beta_k^i = \frac{1}{2}[\chi_k^i]_2 - \frac{1}{2}\frac{([\chi_k^i]_3)^2}{\nu_k^i} \\ \theta_k^i = \frac{[\chi_k^i]_3}{\nu_k^i} \\ \lambda_k^i = [\chi_k^i]_4 \end{bmatrix}$

TABLE I: Parameter Descriptions for Distributed Learning on the Exponential Family.

exploit the network structure. That is, the central node does not know it is a central node, and similarly for the other agents.

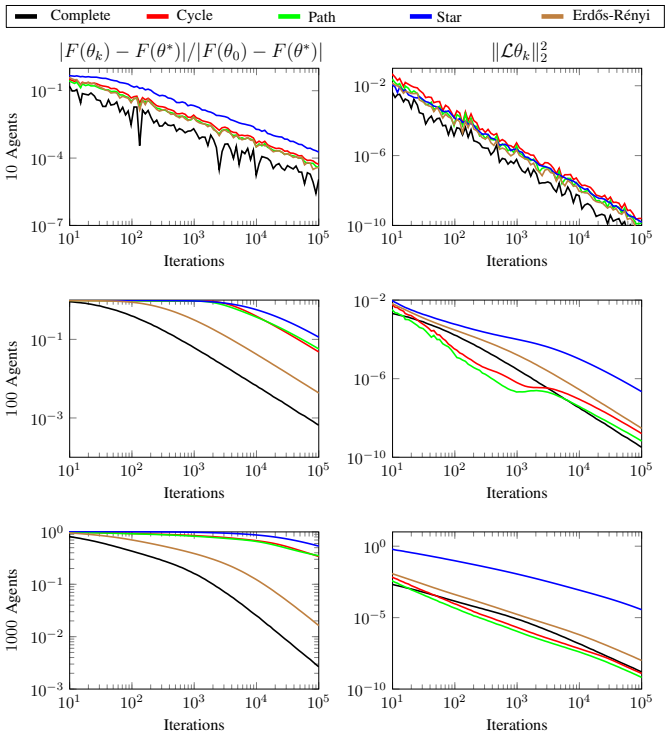

Fig. 2: Optimality and distance to consensus for the distributed estimation of a network-wide unknown **mean** parameter, from Gaussian observations, for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

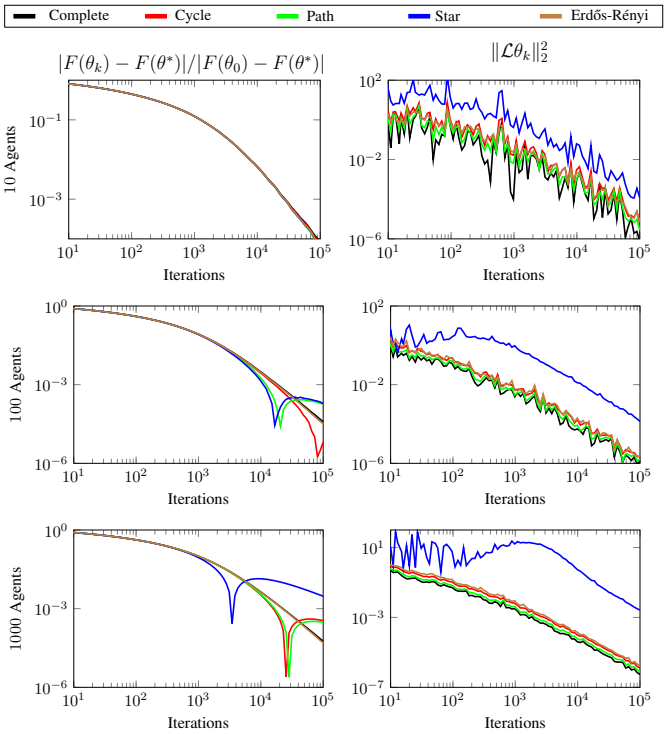

Fig. 3: Optimality and distance to consensus for the distributed estimation of a network-wide **variance** parameter, from Gaussian observations, for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

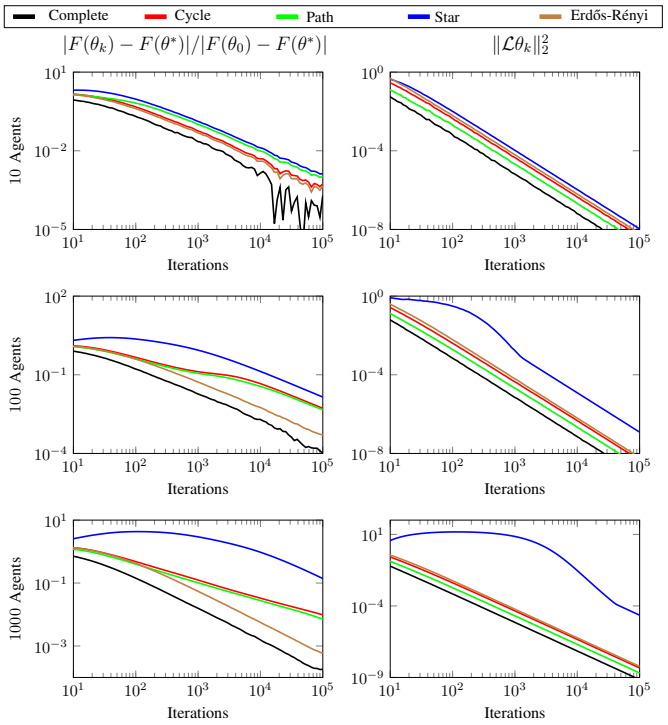

Fig. 4: Optimality and distance to consensus for the distributed estimation of a network-wide **mean and variance** parameters, from Gaussian observations, for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

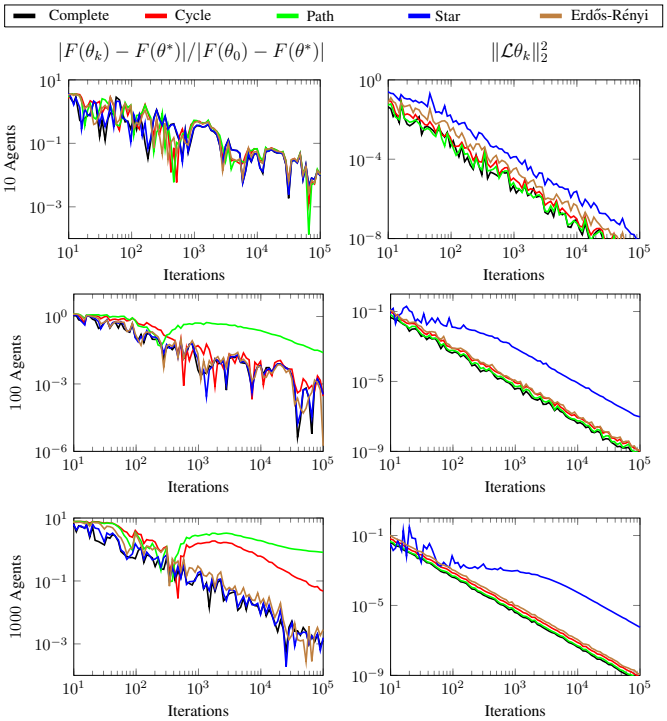

Fig. 6: Optimality and distance to consensus for the distributed estimation of a network-wide parameter of Poisson observations for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

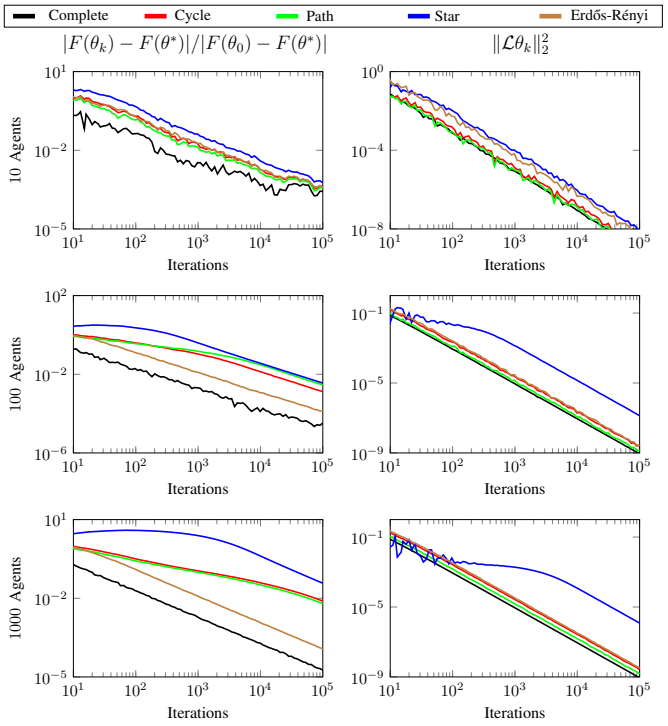

Fig. 5: Optimality and distance to consensus for the distributed estimation of a network-wide parameter of Bernoulli observations for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

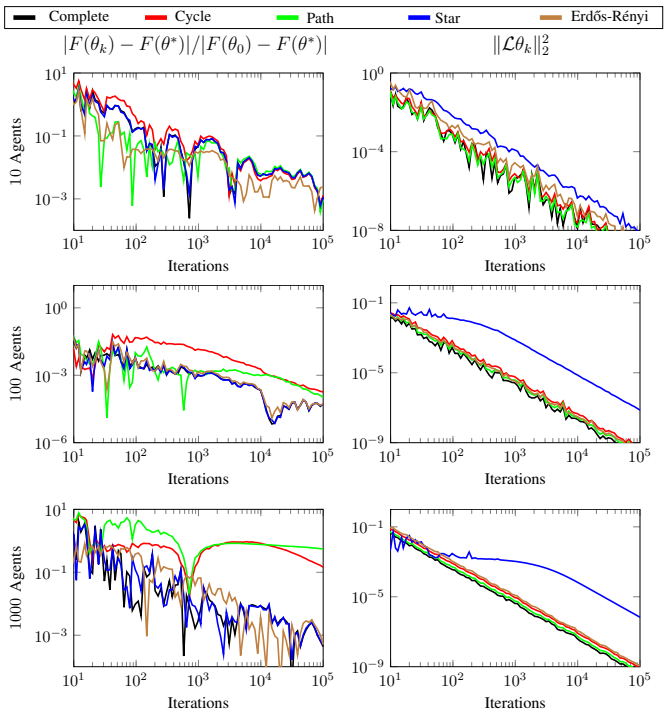

Fig. 7: Optimality and distance to consensus for the distributed estimation of a network-wide parameter of Exponential observations for various graph topologies (complete, cycle, path, star and Erdős-Rényi) of increasing size (10 agents, 100 agents and 100 agents).

V. 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.