Consensus and Distributed Inference Rates Using Network Divergence

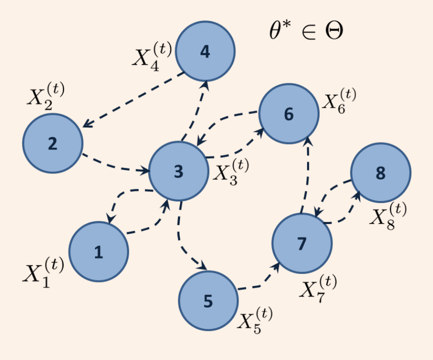
Anand D. Sarwate

Department of Electrical and Computer Engineering
Rutgers, The State University of New Jersey

August 23, 2017

(Joint work with Tara Javidi and Anusha Lalitha (UCSD))
Work sponsored by NSF under award CCF-1440033

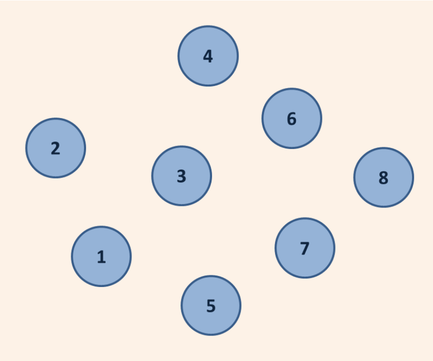
The model: finite hypothesis testing



Second simple model: estimate a global parameter θ^*.

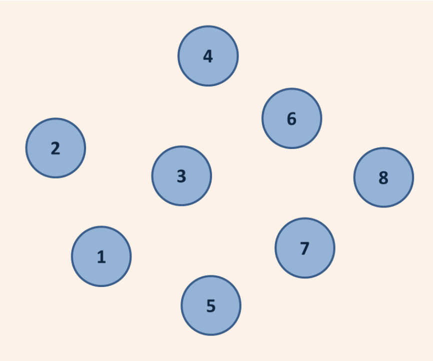
- Each agent takes observations over time conditioned on θ^*.
- Can do local updates followed by communication with neighbors.
- Main focus: simple rule and rate of convergence.

Model



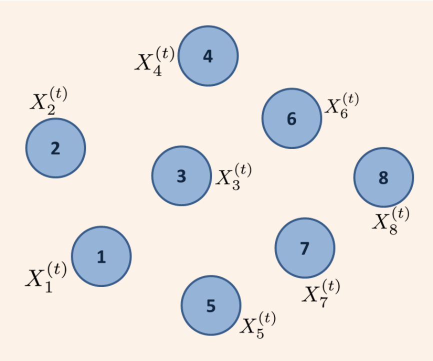
- Set of n nodes.

Model



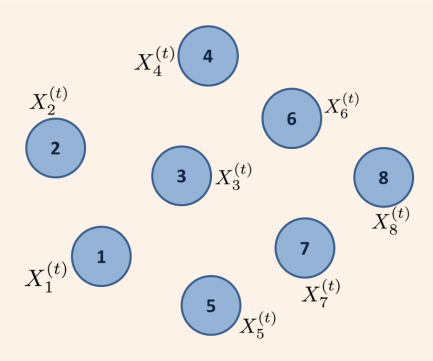
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.

Model



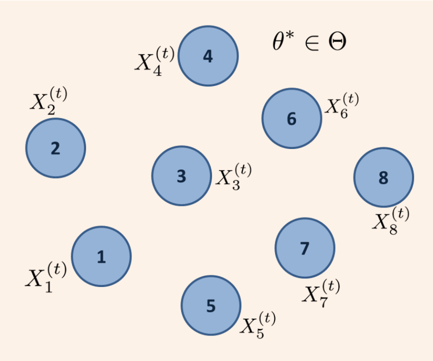
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.

Model



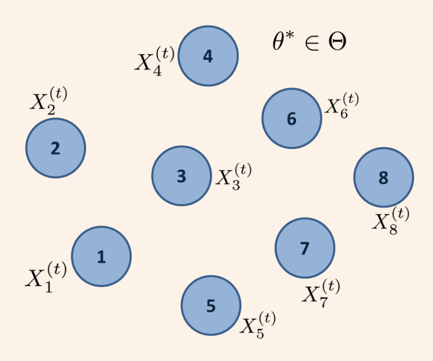
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.
- Fixed known distributions $\{f_i(\cdot;\theta_1), f_i(\cdot;\theta_2), \ldots, f_i(\cdot;\theta_M)\}$.

Model



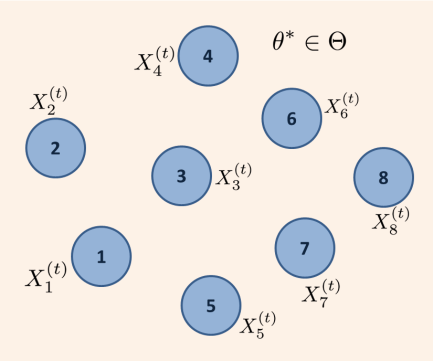
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.
- Fixed known distributions $\{f_i(\cdot;\theta_1), f_i(\cdot;\theta_2), \ldots, f_i(\cdot;\theta_M)\}$.
- $\theta^* \in \Theta$ is fixed global unknown parameter

Model



- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.
- Fixed known distributions $\{f_i(\cdot; \theta_1), f_i(\cdot; \theta_2), \ldots, f_i(\cdot; \theta_M)\}$.
- $\theta^* \in \Theta$ is fixed global unknown parameter
- $X_i^{(t)} \sim f_i(\cdot; \theta^*)$.

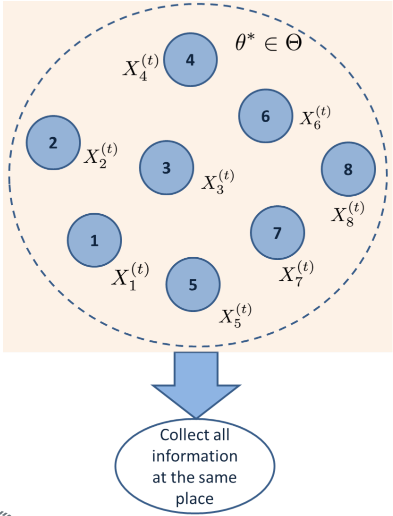
Model



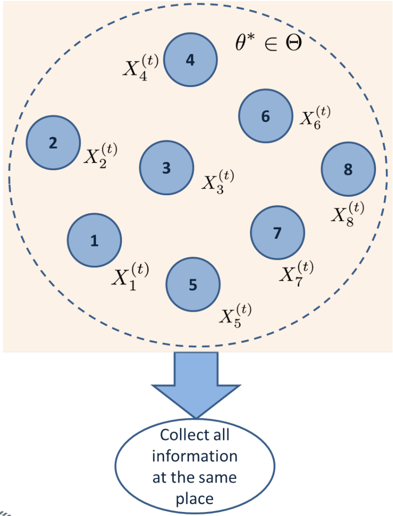
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.
- Fixed known distributions $\{f_i(\cdot;\theta_1), f_i(\cdot;\theta_2), \ldots, f_i(\cdot;\theta_M)\}$.
- $\theta^* \in \Theta$ is fixed global unknown parameter
- $X_i^{(t)} \sim f_i(\cdot;\theta^*)$.

GOAL Parametric inference of unknown θ^*

Hypothesis Testing



Hypothesis Testing

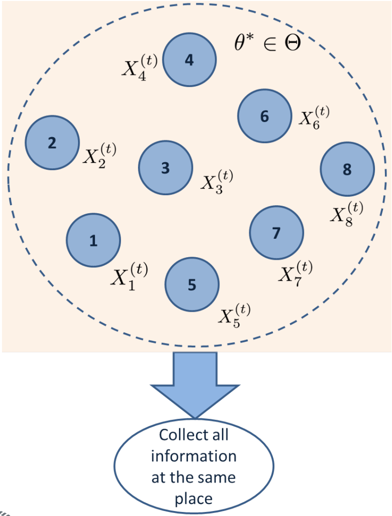


If θ^* is globally identifiable, then collecting all observations

$$\mathbf{X}^{(\mathbf{t})} = \{X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)}\}$$

at a central locations yields a *centralized hypothesis testing problem*.

Hypothesis Testing

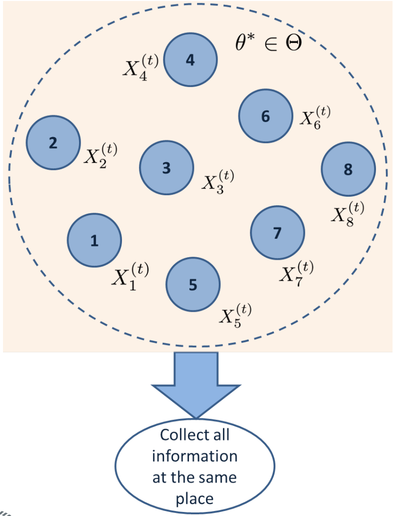


If θ^* is globally identifiable, then collecting all observations

$$\mathbf{X}^{(\mathbf{t})} = \{X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)}\}$$

at a central locations yields a *centralized hypothesis testing problem*. Exponentially fast convergence to the true hypothesis

Hypothesis Testing



If θ^* is globally identifiable, then collecting all observations

$$\mathbf{X}^{(\mathbf{t})} = \{X_1^{(t)}, X_2^{(t)}, \ldots, X_n^{(t)}\}$$

at a central locations yields a *centralized hypothesis testing problem*.
Exponentially fast convergence to the true hypothesis
Can this be achieved locally with low dimensional observations?

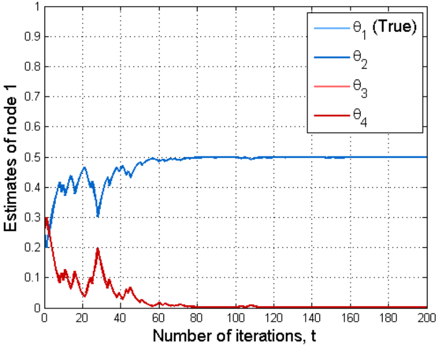
Example: Low-dimensional Observations

$$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$$

$$\theta^* = \theta_1$$

Color

Intensity



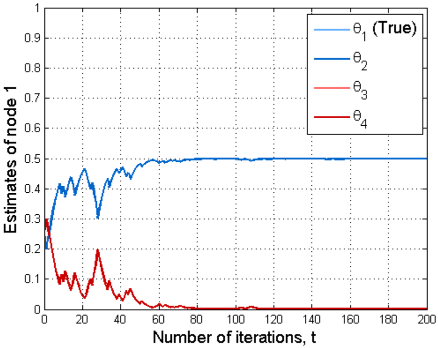
If all observations are not collected centrally, node 1 individually cannot learn θ^*.

Example: Low-dimensional Observations

$$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$$

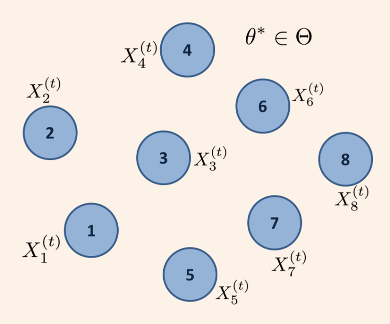
$$\theta^* = \theta_1$$

Color   Intensity



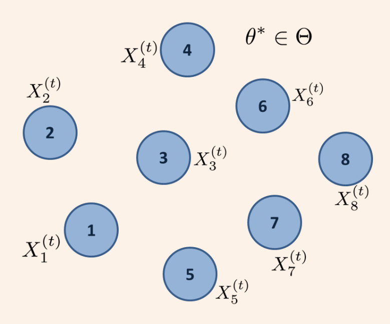
If all observations are not collected centrally, node 1 individually cannot learn θ^*. $\Longrightarrow$ nodes must communicate.

Distributed Hypothesis Testing



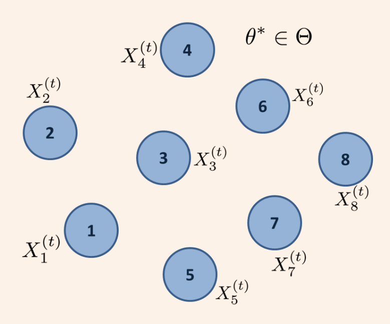
- Define $\bar{\Theta}_i = \{\theta \in \Theta : f_i(\cdot;\theta) = f_i(\cdot;\theta^*)\}$.

Distributed Hypothesis Testing



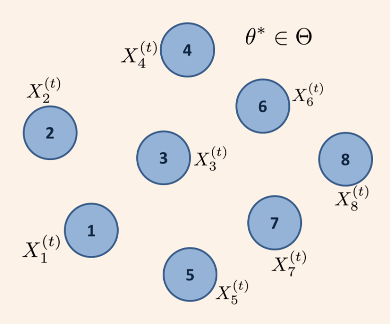
- Define $\bar{\Theta}_i = \{\theta \in \Theta : f_i(\cdot;\theta) = f_i(\cdot;\theta^*)\}$.
- $\theta \in \bar{\Theta}_i$ $\implies$ θ and θ^* are observationally equivalent for node i.

Distributed Hypothesis Testing



- Define $\bar{\Theta}_i = \{\theta \in \Theta : f_i(\cdot;\theta) = f_i(\cdot;\theta^*)\}$.
- $\theta \in \bar{\Theta}_i$ $\implies$ θ and θ^* are observationally equivalent for node i.
- Suppose $\{\theta^*\} = \bar{\Theta}_1 \cap \bar{\Theta}_2 \cap \ldots \cap \bar{\Theta}_n$.

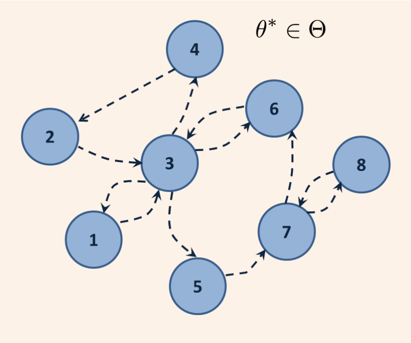
Distributed Hypothesis Testing



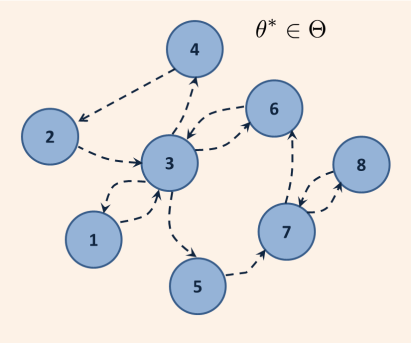
- Define $\bar{\Theta}_i = \{\theta \in \Theta : f_i(\cdot;\theta) = f_i(\cdot;\theta^*)\}$.
- $\theta \in \bar{\Theta}_i$ $\implies$ θ and θ^* are observationally equivalent for node i.
- Suppose $\{\theta^*\} = \bar{\Theta}_1 \cap \bar{\Theta}_2 \cap \ldots \cap \bar{\Theta}_n$.

GOAL Parametric inference of unknown θ^*

Learning Rule

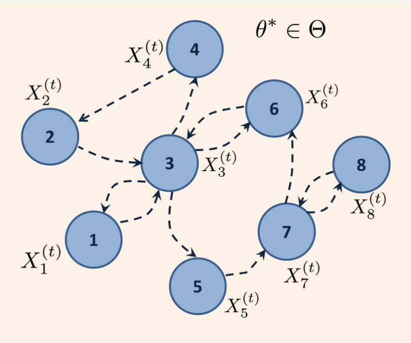


Learning Rule



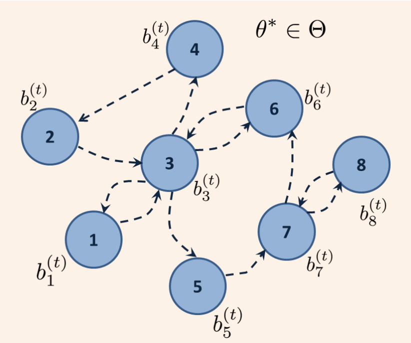
- At $t = 0$, node i begins with initial estimate vector $\mathbf{q_i^{(0)}} > 0$, where components of $\mathbf{q_i^{(t)}}$ form a probability distribution on Θ.

Learning Rule



- At $t = 0$, node i begins with initial estimate vector $\mathbf{q_i^{(0)}} > 0$, where components of $\mathbf{q_i^{(t)}}$ form a probability distribution on Θ.
- At $t > 0$, node i draws $X_i^{(t)}$.

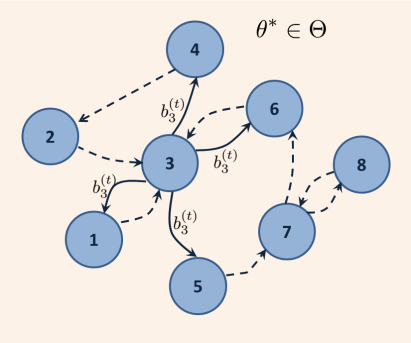
Learning Rule



- Node i computes belief vector, $\mathbf{b_i^{(t)}}$, via Bayesian update

$$b_i^{(t)}(\theta) = \frac{f_i\left(X_i^{(t)};\theta\right) q_i^{(t-1)}(\theta)}{\sum_{\theta' \in \Theta} f_i\left(X_i^{(t)};\theta'\right) q_i^{(t-1)}(\theta')}.$$

Learning Rule

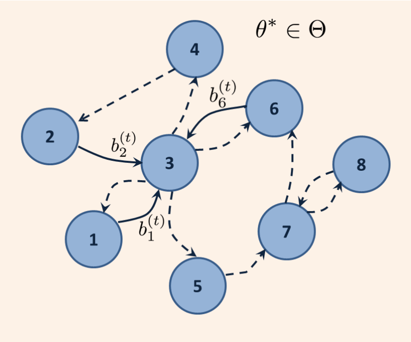


- Node i computes belief vector, $\mathbf{b_i^{(t)}}$, via Bayesian update

$$b_i^{(t)}(\theta) = \frac{f_i\left(X_i^{(t)};\theta\right) q_i^{(t-1)}(\theta)}{\sum_{\theta' \in \Theta} f_i\left(X_i^{(t)};\theta'\right) q_i^{(t-1)}(\theta')}.$$

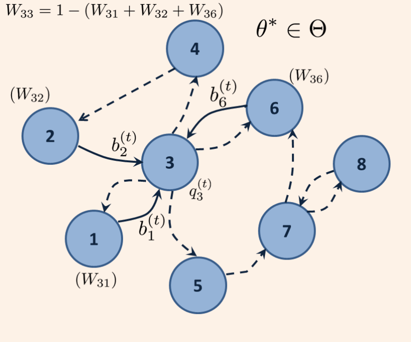
- Sends message $\mathbf{Y_i^{(t)}} = \mathbf{b_i^{(t)}}$.

Learning Rule



- Receives messages from its neighbors at the same time.

Learning Rule

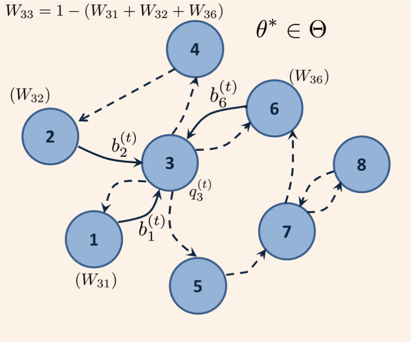


- Receives messages from its neighbors at the same time.
- Updates $\mathbf{q_i^{(t)}}$ via averaging of log beliefs,

$$q_i^{(t)}(\theta) = \frac{\exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta)\right)}{\sum_{\theta' \in \Theta} \exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta')\right)},$$

where weight W_{ij} denotes the influence of node j on estimate of node i.

Learning Rule



- Receives messages from its neighbors at the same time.
- Updates $\mathbf{q_i^{(t)}}$ via averaging of log beliefs,

$$q_i^{(t)}(\theta) = \frac{\exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta)\right)}{\sum_{\theta' \in \Theta} \exp\left(\sum_{j=1}^{n} W_{ij} \log b_j^{(t)}(\theta')\right)},$$

where weight W_{ij} denotes the influence of node j on estimate of node i.

- Put $t = t + 1$ and repeat.

In a picture

$Q_i(\theta, t)$ $X_i(t) \sim f_i(\cdot|\theta^*)$ **local observations**

Bayesian Update

$$b_i(\theta, t) = \frac{f_i(X_i(t)|\theta)}{\sum_{\theta'} f_i(X_i(t)|\theta) Q_i(\theta, t)}$$

Average log-beliefs

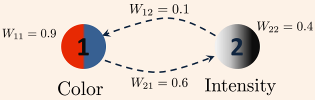
$\{b_j(\theta, t)\}$ **messages from neighbors**

$$Q_i(\theta, t+1) = \frac{\exp\left(\sum_{j=1}^n W_{ij} \log b_j(\theta, t)\right)}{\sum_{\theta' \in \Theta} \exp\left(\sum_{j=1}^n W_{ij} \log b_j(\theta', t)\right)}.$$

An example

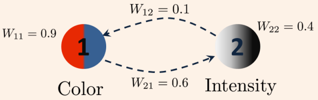
$$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$$

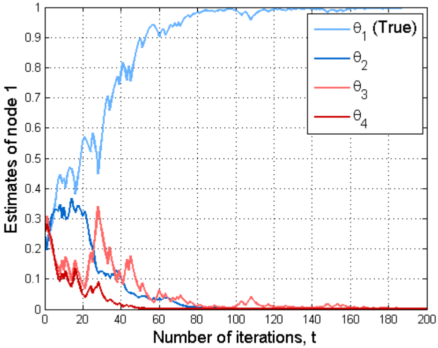
$$\theta^* = \theta_1$$



An example

$$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$$
$$\theta^* = \theta_1$$





When connected in a network, using the proposed learning rule node 1 learns θ^*.

Assumptions

Assumption 1

For every pair $\theta \neq \theta^*$, $f_i\left(\cdot;\theta^*\right) \neq f_i\left(\cdot;\theta\right)$ for at least one node, *i.e* the KL-divergence $D\left(f_i\left(\cdot;\theta^*\right) \| f_i\left(\cdot;\theta\right)\right) > 0$.

Assumptions

Assumption 1

For every pair $\theta \neq \theta^*$, $f_i(\cdot;\theta^*) \neq f_i(\cdot;\theta)$ for at least one node, *i.e* the KL-divergence $D\left(f_i(\cdot;\theta^*) \| f_i(\cdot;\theta)\right) > 0$.

Assumption 2

The stochastic matrix W is irreducible.

Assumptions

Assumption 1

For every pair $\theta \neq \theta^*$, $f_i(\cdot;\theta^*) \neq f_i(\cdot;\theta)$ for at least one node, *i.e* the KL-divergence $D(f_i(\cdot;\theta^*) \| f_i(\cdot;\theta)) > 0$.

Assumption 2

The stochastic matrix W is irreducible.

Assumption 3

For all $i \in [n]$, the initial estimate $q_i^{(0)}(\theta) > 0$ for every $\theta \in \Theta$.

Network Divergence

The eigenvector centrality $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is the left eigenvector of W corresponding to eigenvalue 1.

The central quantity of interest is what we call the *network divergence*

$$K(\theta^*, \theta) = \sum_{j=1}^{n} v_j D\left(f_j\left(\cdot; \theta^*\right) \| \, f_j\left(\cdot; \theta\right)\right)$$

Convergence Results

- Let θ^* be the unknown fixed parameter.
- Suppose assumptions $1 - 3$ hold.

Convergence Results

- Let θ^* be the unknown fixed parameter.
- Suppose assumptions $1-3$ hold.

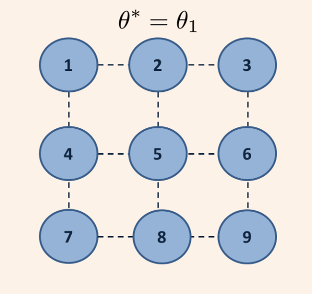
Theorem: Rate of rejecting $\theta \neq \theta^*$

Every node i's estimate of $\theta \neq \theta^*$ almost surely converges to 0 exponentially fast. Mathematically,

$$-\lim_{t\to\infty} \frac{1}{t} \log q_i^{(t)}(\theta) = K(\theta^*, \theta) \quad \mathbb{P}\text{-a.s.}$$

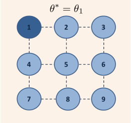
where $K(\theta^*, \theta) = \sum_{j=1}^{n} v_j D\left(f_j\left(\cdot;\theta^*\right) \| \, f_j\left(\cdot;\theta\right)\right)$.

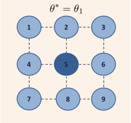
Example: Network-wide Learning



- $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and $\theta^* = \theta_1$.
- If i and j are connected, $W_{ij} = \frac{1}{\text{degree of node i}}$, otherwise 0.
- $\mathbf{v} = [\frac{1}{12}, \frac{1}{8}, \frac{1}{12}, \frac{1}{8}, \frac{1}{6}, \frac{1}{8}, \frac{1}{12}, \frac{1}{8}, \frac{1}{12}]$.

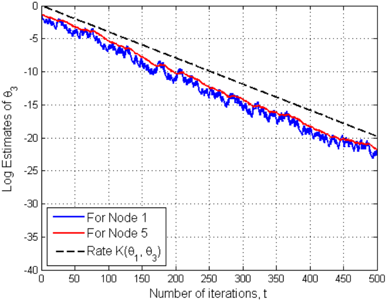
Example

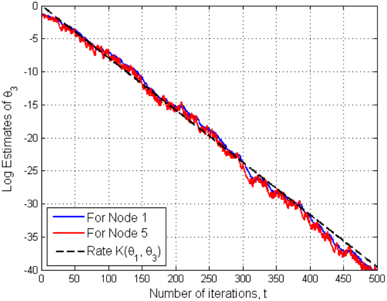




$\bar{\Theta}_1 = \{\theta^*\}$, $\bar{\Theta}_i = \Theta$ $i \neq 1$

$\bar{\Theta}_5 = \{\theta^*\}$, $\bar{\Theta}_i = \Theta$ $i \neq 5$





Corollaries

Theorem: Rate of rejecting $\theta \neq \theta^*$

Every node i's estimate of $\theta \neq \theta^*$ almost surely converges to 0 exponentially fast. Mathematically,

$$-\lim_{t \to \infty} \frac{1}{t} \log q_i^{(t)}(\theta) = K(\theta^*, \theta) \quad \mathbb{P}\text{-a.s.}$$

where $K(\theta^*, \theta) = \sum_{j=1}^{n} v_j D\left(f_j\left(\cdot; \theta^*\right) \| \, f_j\left(\cdot; \theta\right)\right)$.

Lower bound on rate of convergence to θ^*

For every node i, the rate at which error in the estimate of θ^* goes to zero can be lower bounded as

$$-\lim_{t \to \infty} \frac{1}{t} \log \left(1 - q_i^{(t)}(\theta^*) \right) = \min_{\theta \neq \theta^*} K(\theta^*, \theta) \quad \mathbb{P}\text{-a.s.}$$

Corollaries

Lower bound on rate of learning

The rate of learning λ across the network can be lower bounded as,

$$\lambda \geq \min_{\theta^* \in \Theta} \min_{\theta \neq \theta^*} K(\theta^*, \theta) \quad \mathbb{P}\text{-a.s.}$$

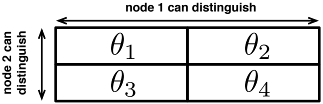
where,

$$\lambda = \liminf_{t \to \infty} \frac{1}{t} |\log e_t|,$$

and

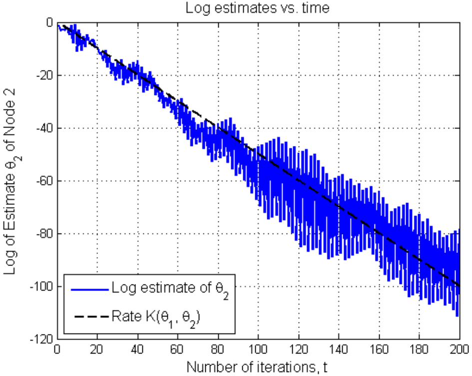
$$e_t = \frac{1}{2} \sum_{i=1}^{n} ||q_i^{(t)}(\cdot) - 1_{\theta^*}(.)||_1 = \sum_{i=1}^{n} \sum_{\theta \neq \theta^*} q_i^{(t)}(\theta).$$

Example: Periodicity



- $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and $\theta^* = \theta_1$.
- Underlying graph is periodic,

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$



Example: Networks with Large Mixing Times

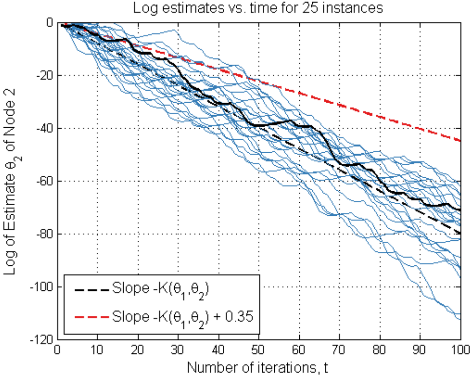
node 1 can distinguish

node 2 can distinguish

θ_1	θ_2
θ_3	θ_4

- $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and $\theta^* = \theta_1$.
- Underlying graph is aperiodic,

$$W = \begin{pmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{pmatrix}.$$



Concentration Result

Assumption 4

For $k \in [n]$, $X \in \mathcal{X}_k$, and for any given $\theta_i, \theta_j \in \Theta$ such that $\theta_i \neq \theta_j$, $\left| \log \frac{f_k(\cdot;\theta_i)}{f_k(\cdot;\theta_j)} \right|$ is bounded, denoted by L.

Theorem

Under Assumptions 1–4, for every $\epsilon > 0$ there exists a T such that for all $t \geq T$ and for every $\theta \neq \theta^*$ and $i \in [n]$ we have

$$\Pr\left(\log q_i^{(t)}(\theta) \geq -(K(\theta^*, \theta) - \epsilon)t\right) \leq \gamma(\epsilon, L, t),$$

and

$$\Pr\left(\log q_i^{(t)}(\theta) \leq -(K(\theta^*, \theta) + \epsilon)t\right) \leq \gamma(\frac{\epsilon}{2}, L, t),$$

where L is a finite constant and $\gamma(\epsilon, L, t) = 2\exp\left(-\frac{\epsilon^2 t}{2L^2 d}\right)$.

Related Work and Contribution

Jadbabaie *et al.* use local Bayesian update of beliefs followed by averaging the beliefs.

- Show exponential convergence with no closed form of convergence rate. ['12]
- Provide an upper bound on learning rate. ['13]

We average the log beliefs instead.

- Provide a lower bound on learning rate $\tilde{\lambda}$.
- *Lower bound* on learning rate is greater than the upper bound
  $\Longrightarrow$ Our learning rule *converges faster*.

Related Work and Contribution

Jadbabaie *et al.* use local Bayesian update of beliefs followed by averaging the beliefs.

- Show exponential convergence with no closed form of convergence rate. ['12]
- Provide an upper bound on learning rate. ['13]

We average the log beliefs instead.

- Provide a lower bound on learning rate $\tilde{\lambda}$.
- *Lower bound* on learning rate is greater than the upper bound
  $\Longrightarrow$ Our learning rule *converges faster*.

Shahrampour and Jadbabaie, '13 formulated a stochastic optimization learning problem; obtained a dual-based learning rule for doubly stochastic W,

- Provide closed-form lower bound on rate of identifying θ^*.
- Using our *rule* we achieve the *same lower bound* (from corollary 1)

$$\min_{\theta \neq \theta^*} \left(\frac{1}{n} \sum_{j=1}^{n} D(f_j(\cdot;\theta^*) || f_j(\cdot;\theta)) \right).$$

Related Work and Contribution

An update rule similar to ours was used in Rahnama Rad and Tahbaz-Salehi, 2010 to

- Show that node's belief converges in probability to the true parameter.
- However, under certain analytic assumptions.

For general model and discrete parameter spaces we show almost-sure exponentially fast convergence.

Related Work and Contribution

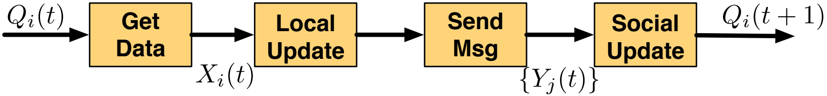
An update rule similar to ours was used in Rahnama Rad and Tahbaz-Salehi, 2010 to

- Show that node's belief converges in probability to the true parameter.
- However, under certain analytic assumptions.

For general model and discrete parameter spaces we show almost-sure exponentially fast convergence.

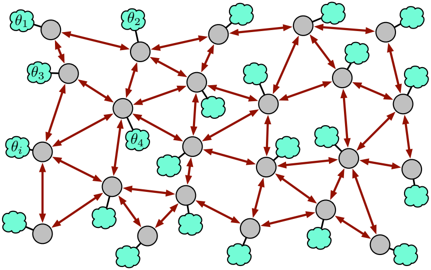
Shahrampour *et. al.* and Nedic *et. al.* (independently) showed that our learning rule coincides with distributed stochastic optimization based learning rule (W irreducible and aperiodic)

Hypothesis testing and "semi-Bayes"



- Combination of local Bayesian updates and averaging.
- Network divergence: an intuitive measure for the rate of convergence.
- "Posterior consistency" gives a Bayesio-frequentist analysis.

Looking forward



- Continuous distributions and parameters.
- Applications to distributed optimization.
- Further limiting messages via coordinate descent (Sarwate and Javidi '15).
- Time-varying parameters and distributed stochastic filtering.

Thank You!