Learning distributions and hypothesis testing via social learning

Anand D. Sarwate

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

September 29, 2015

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

Introduction

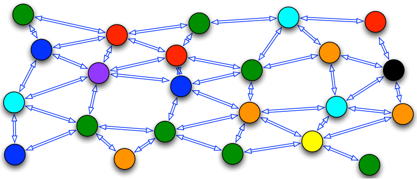
Some philosophical questions

- How we (as a network of social agents) make common choices or inferences about the world?
- If I want to help you learn, should I tell you my evidence or just my opinion?
- How much do we need to communicate with each other?

Which may have some applications (?)

- Distributed monitoring in networks (estimating a state).
- Hypothesis testing or detection using multi-modal sensors.
- Models for vocabulary evolution.
- Social learning in animals.

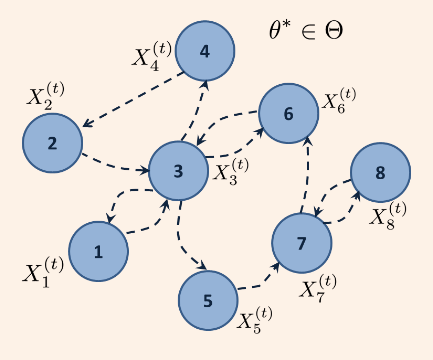
Estimation



First simple model: estimate a histogram of local data.

- Each agent starts with a single color.
- Pass message to learn the histogram of initial colors or sample from that histogram.
- Main focus: simple protocols with limited communication.

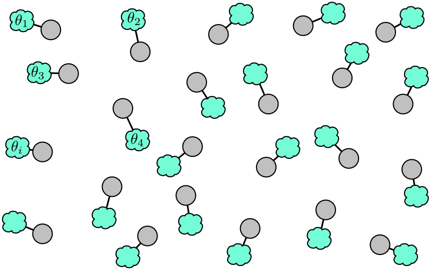
Hypothesis testing



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.

Social learning



Social learning focuses on simple models for how (human) networks can form consensus opinions:

- Consensus-based DeGroot model: gossip, average consensus etc.
- Bayesian social learning (Acemoglu et al., Bala and Goyal): agents make decisions and are observed by other agents.
- Opinion dynamics where agents change beliefs based on beliefs of nearby neighbors.

On limited messages

Both of our problems involve some sort of average consensus step. In the first part we are interested in exchanging approximate messages.

On limited messages

Both of our problems involve some sort of average consensus step. In the first part we are interested in exchanging approximate messages.

- Lots of work in quantized consensus (Aysal-Coates-Rabbat, Carli et al., Kashyap et al. Lavaei and Murray, Nedic et al, Srivastava and Nedic, Zhu and Martinez)

On limited messages

Both of our problems involve some sort of average consensus step. In the first part we are interested in exchanging approximate messages.

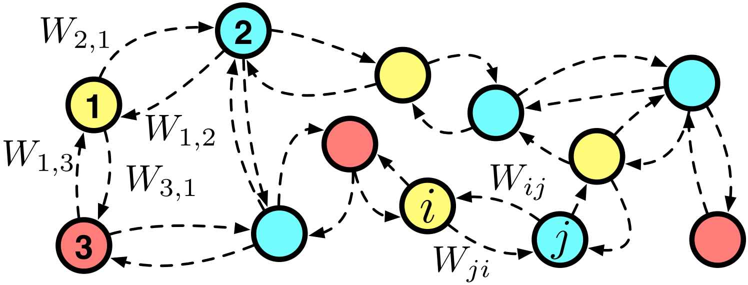
- Lots of work in quantized consensus (Aysal-Coates-Rabbat, Carli et al., Kashyap et al. Lavaei and Murray, Nedic et al, Srivastava and Nedic, Zhu and Martinez)
- Time-varying network topologies (even more references).

On limited messages

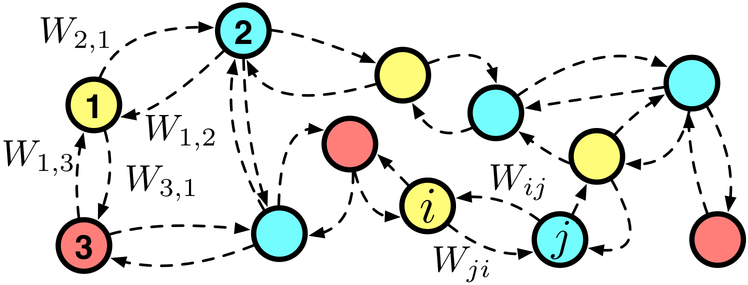
Both of our problems involve some sort of average consensus step. In the first part we are interested in exchanging approximate messages.

- Lots of work in quantized consensus (Aysal-Coates-Rabbat, Carli et al., Kashyap et al. Lavaei and Murray, Nedic et al, Srivastava and Nedic, Zhu and Martinez)
- Time-varying network topologies (even more references).
- Pretty mature area at this point.

A roadmap

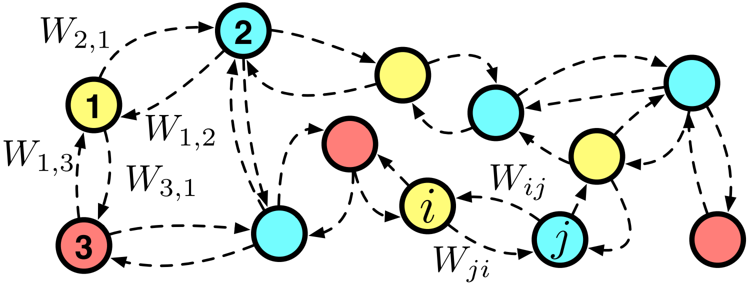


A roadmap



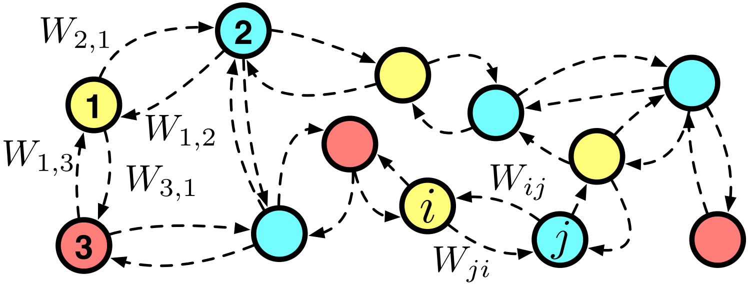
- "Social sampling" and estimating histograms

A roadmap



- "Social sampling" and estimating histograms
- Distributed hypothesis testing and network divergence

A roadmap

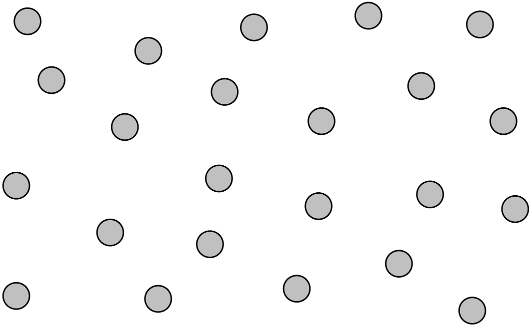


- "Social sampling" and estimating histograms
- Distributed hypothesis testing and network divergence
- Some ongoing work and future ideas.

Social sampling and merging opinions

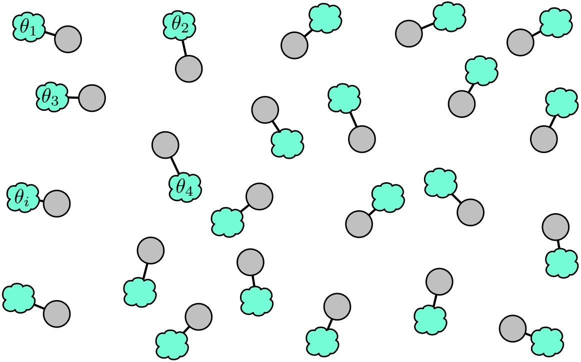
A.D. Sarwate, T. Javidi, Distributed Learning of Distributions via Social Sampling, *IEEE Transactions on Automatic Control* 60(1): pp. 34–45, January 2015.

Consensus and dynamics in networks



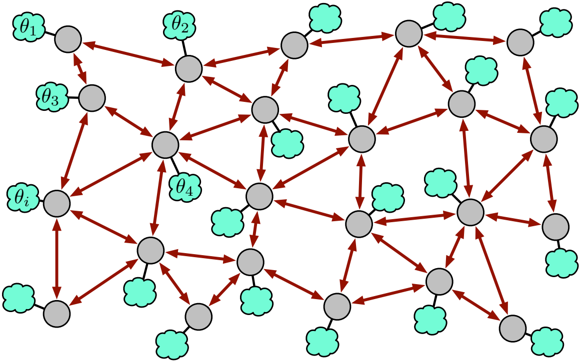
- Collection of individuals or agents

Consensus and dynamics in networks



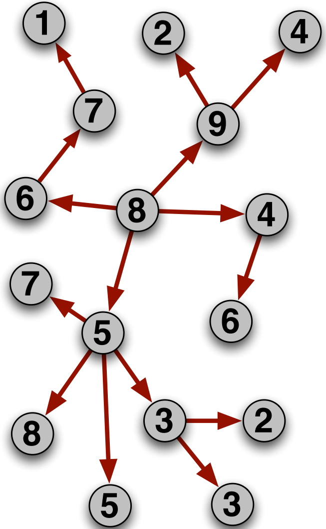
- Collection of individuals or agents
- Agents observe part of a global phenomenon

Consensus and dynamics in networks

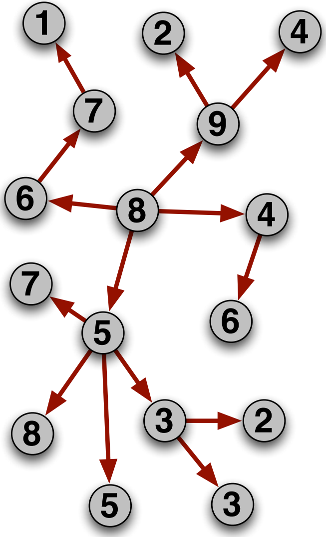


- Collection of individuals or agents
- Agents observe part of a global phenomenon
- Network of connections for communication

Phenomena vs. protocols



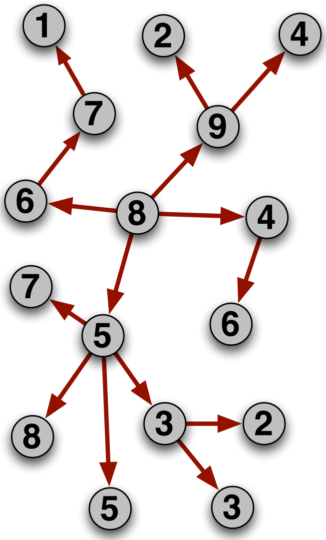
Phenomena vs. protocols



Engineering:

- Focus on algorithms
- Minimize communication cost
- How much do we lose vs. centralized?

Phenomena vs. protocols



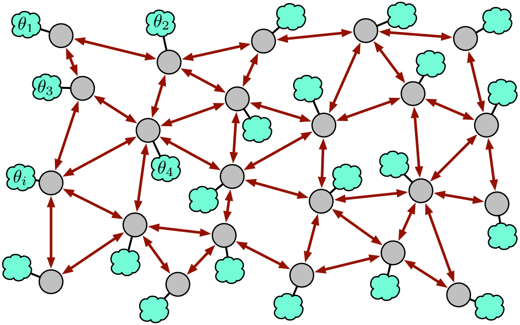
Engineering:

- Focus on algorithms
- Minimize communication cost
- How much do we lose vs. centralized?

Phenomenological:

- Focus on modeling
- Simple protocols
- What behaviors emerge?

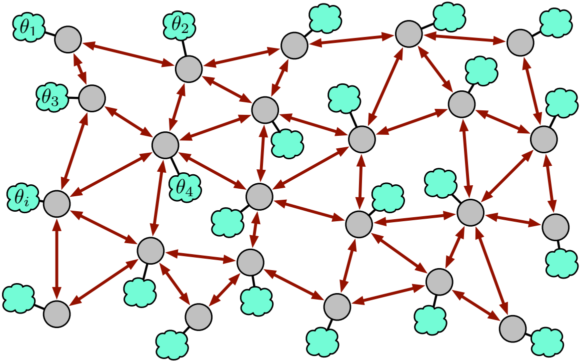
Why simple protocols?



We are more interested in developing simple models that can exhibit different phenomena.

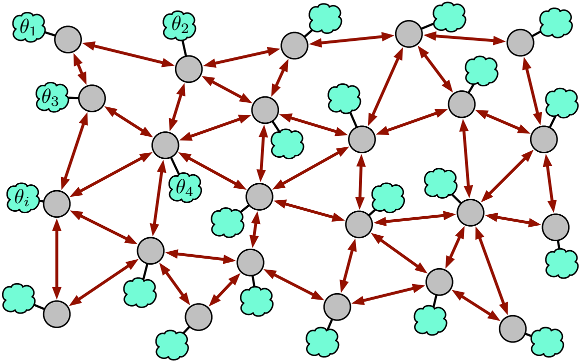
- Simple source models.
- Simple communication that uses fewer resources.
- Simple update rules that are easier to analyze.

Communication and graph



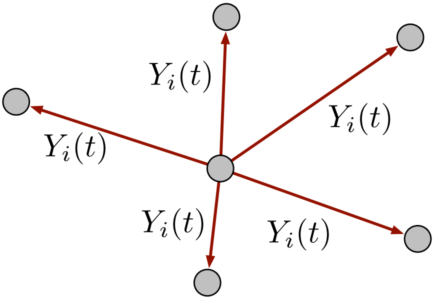
- The n agents are arranged in a connected graph G.

Communication and graph



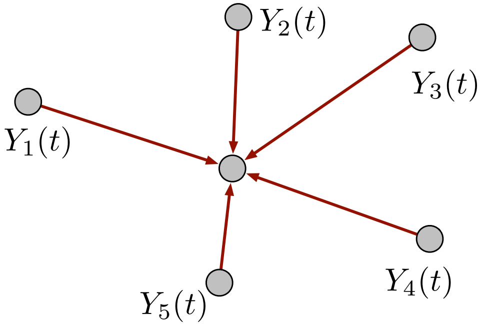
- The n agents are arranged in a connected graph G.

Communication and graph



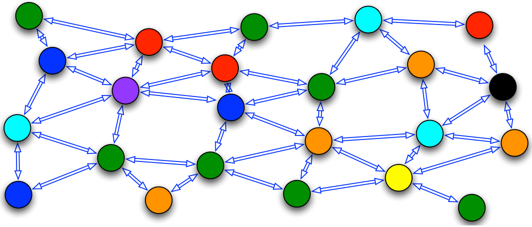
- The n agents are arranged in a connected graph G.
- Agent i broadcasts to neighbors $\mathcal{N}_i$ in the graph.

Communication and graph

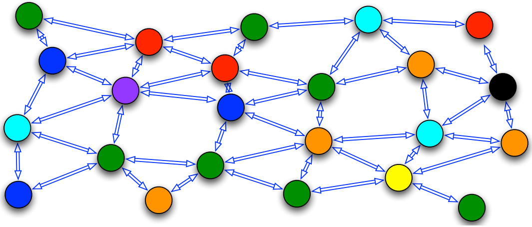


- The n agents are arranged in a connected graph G.
- Agent i broadcasts to neighbors $\mathcal{N}_i$ in the graph.
- Message $Y_i(t)$ lies in a discrete set.

The problem

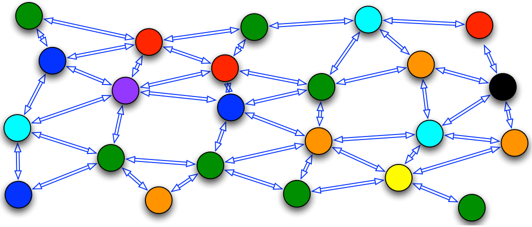


The problem



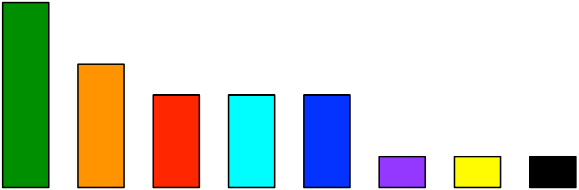
- Each agent starts with $\theta_i \in \{1, 2, \ldots, M\}$

The problem



- Each agent starts with $\theta_i \in \{1, 2, \ldots, M\}$
- Agent i knows θ_i (no noise)

The problem



- Each agent starts with $\theta_i \in \{1, 2, \ldots, M\}$
- Agent i knows θ_i (no noise)
- Maintain estimates $Q_i(t)$ of the empirical distribution Π of $\{\theta_i\}$

Social sampling

We model the messages as *random samples* from local estimates.

1. Update rule from $Q_i(t-1)$ to $Q_i(t)$:

$$Q_i(t) = W_i \left(Q_i(t-1), X_i(t), Y_i(t-1), \{Y_j(t-1) : j \in \mathcal{N}_i\}, t \right).$$

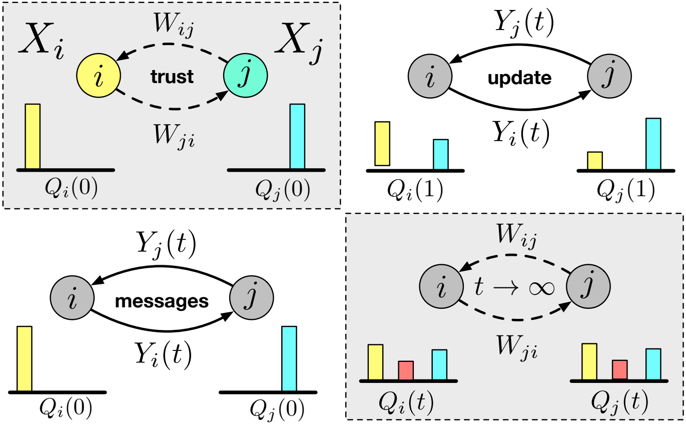
2. Build a sampling distribution on $\{0, 1, \ldots, M\}$:

$$P_i(t) = V_i(Q_i(t), t).$$

3. Sample message:

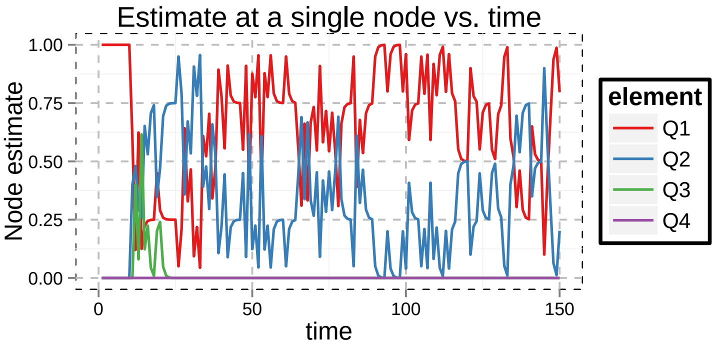
$$Y_i(t) \sim P_i(t).$$

Social sampling



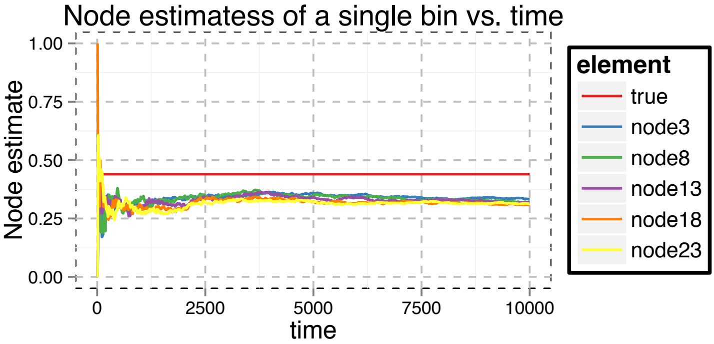
Possible phenomena

Possible phenomena



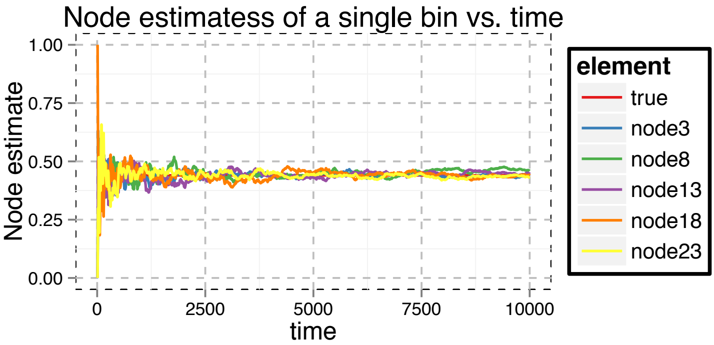
Coalescence: all agents converge to singletons

Possible phenomena



Consensus: agents converge to common $\hat{\Pi} \neq \Pi$

Possible phenomena



Convergence: agents converge to Π

Linear update rule

$$Q_i(t) = A_i(t)Q_i(t-1) + B_i(t)Y_i(t-1) + \sum_{j \in \mathcal{N}_i} W_{ij}(t)Y_j(t-1)$$

- Linear update rule combining $Y_i \sim P_i$ and Q_i.

Linear update rule

$$Q_i(t) = A_i(t)Q_i(t-1) + B_i(t)Y_i(t-1) + \sum_{j \in \mathcal{N}_i} W_{ij}(t)Y_j(t-1)$$

- Linear update rule combining $Y_i \sim P_i$ and Q_i.
- Exhibits different behavior depending on $A_i(t)$, $B_i(t)$, and $W(t)$.

Convergence

Main idea : massage the update rule into matrix form:

$$\mathbf{Q}(t+1) = \mathbf{Q}(t) + \delta(t) \left[\bar{H} \mathbf{Q}(t) + \mathbf{C}(t) + \mathbf{M}(t) \right].$$

with

1. Step size $\delta(t) = 1/t$
2. Perturbation $\mathbf{C}(t) = O(\delta(t))$
3. Martingale difference term $\mathbf{M}(t)$

This is a *stochastic approximation*: converges to a fixed point of $\bar{H}$.

Example: censored updates

Suppose we make distribution $P_i(t)$ a *censored* version of $Q_i(t)$:

$$P_{i,m}(t) = Q_{i,m}(t) \cdot \mathbf{1}\left(Q_{i,m}(t) > \delta(t)(1 - W_{ii})\right)$$

$$P_{i,0}(t) = \sum_{m=1}^{M} Q_{i,m}(t) \cdot \mathbf{1}\left(Q_{i,m}(t) \leq \delta(t)(1 - W_{ii})\right)$$

Example: censored updates

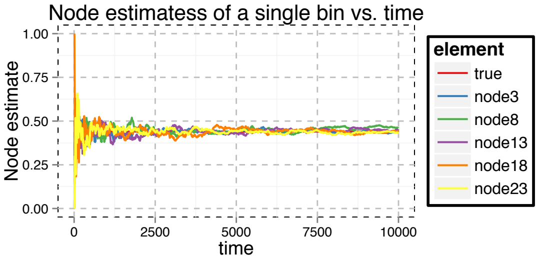
Suppose we make distribution $P_i(t)$ a *censored* version of $Q_i(t)$:

$$P_{i,m}(t) = Q_{i,m}(t) \cdot \mathbf{1}\left(Q_{i,m}(t) > \delta(t)(1 - W_{ii})\right)$$

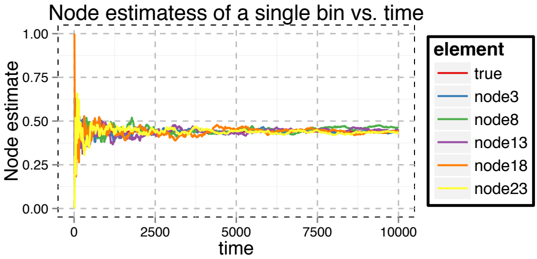
$$P_{i,0}(t) = \sum_{m=1}^{M} Q_{i,m}(t) \cdot \mathbf{1}\left(Q_{i,m}(t) \leq \delta(t)(1 - W_{ii})\right)$$

Agent sends $Y_i(t) = \mathbf{0}$ if it samples a “rare” element in Q_i.

Phenomena captured by censored model

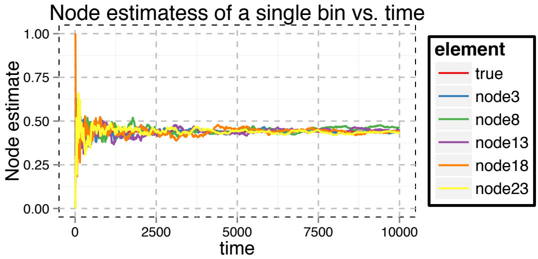


Phenomena captured by censored model



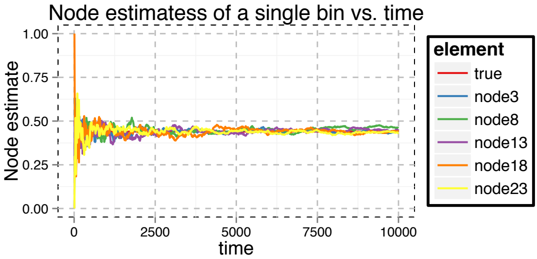
- Censored distribution $P_i(t)$ guards against "marginal opinions."

Phenomena captured by censored model



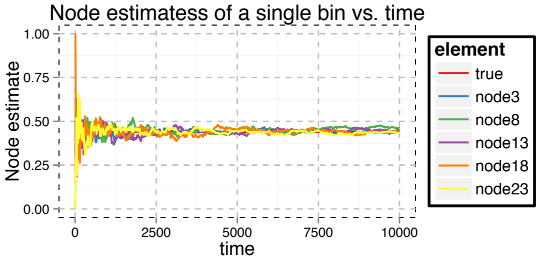
- Censored distribution $P_i(t)$ guards against "marginal opinions."
- Sampled messages $Y_i(t) \sim P_i(t)$ are simple messages.

Phenomena captured by censored model



- Censored distribution $P_i(t)$ guards against "marginal opinions."
- Sampled messages $Y_i(t) \sim P_i(t)$ are simple messages.
- Decaying weights $\delta(t)$ represent solidifying of opinions.

Phenomena captured by censored model



- Censored distribution $P_i(t)$ guards against "marginal opinions."
- Sampled messages $Y_i(t) \sim P_i(t)$ are simple messages.
- Decaying weights $\delta(t)$ represent solidifying of opinions.

Result : all estimates converge almost surely to Π.

Future directions

Future directions

- Find the rate of convergence and dependence of the rate on the parameters

Future directions

- Find the rate of convergence and dependence of the rate on the parameters
- Investigate the robustness of the update rule to noise and perturbations

Future directions

- Find the rate of convergence and dependence of the rate on the parameters
- Investigate the robustness of the update rule to noise and perturbations
- Continuous distributions?

Future directions

- Find the rate of convergence and dependence of the rate on the parameters
- Investigate the robustness of the update rule to noise and perturbations
- Continuous distributions?
- Other message passing algorithms?

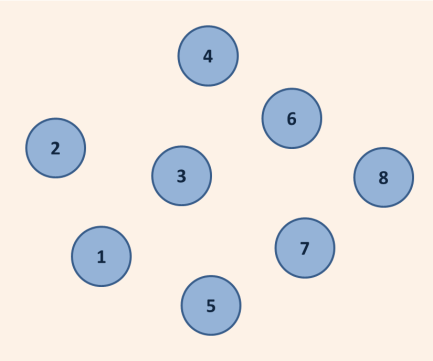
Future directions

- Find the rate of convergence and dependence of the rate on the parameters
- Investigate the robustness of the update rule to noise and perturbations
- Continuous distributions?
- Other message passing algorithms?
- Distributed optimization?

“Non-Bayesian” social learning

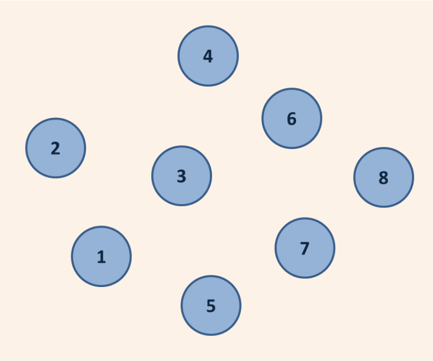
A. Lalitha, T. Javidi, A. Sarwate, Social Learning and Distributed Hypothesis Testing, ArXiV report number arXiv:1410.4307 [math.ST], October, 2014.

Model



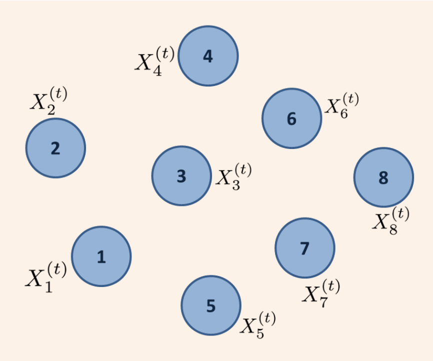
- Set of n nodes.

Model



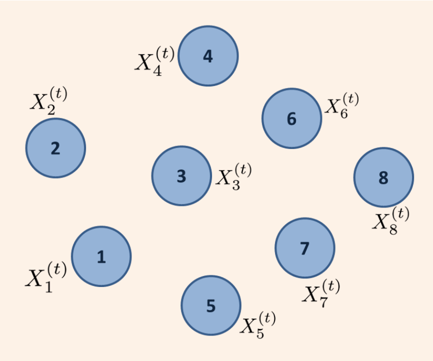
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.

Model



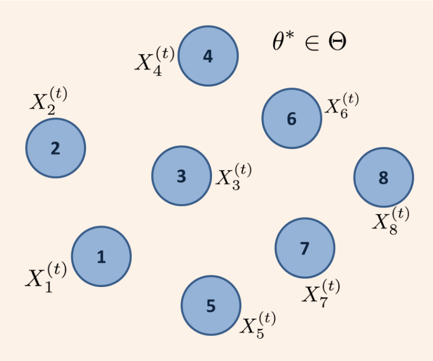
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.

Model



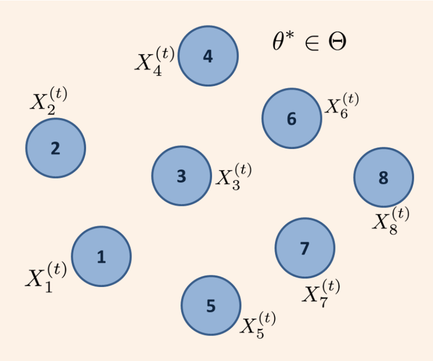
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.
- Fixed known distributions $\{f_i(\cdot;\theta_1), f_i(\cdot;\theta_2), \ldots, f_i(\cdot;\theta_M)\}$.

Model



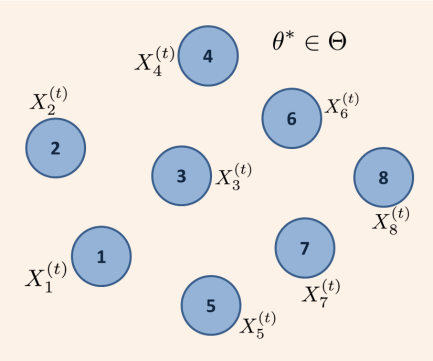
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.
- Fixed known distributions $\{f_i(\cdot;\theta_1), f_i(\cdot;\theta_2), \ldots, f_i(\cdot;\theta_M)\}$.
- $\theta^* \in \Theta$ is fixed global unknown parameter

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
- $X_i^{(t)} \sim f_i(\cdot;\theta^*)$.

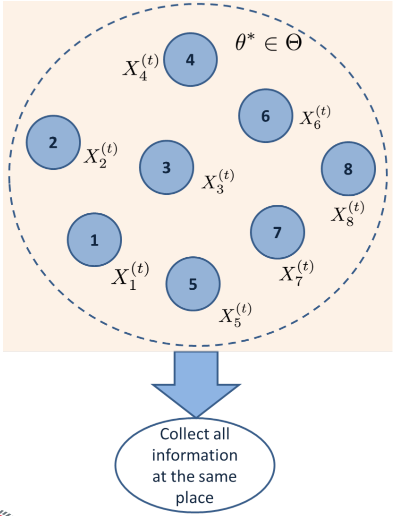
Model



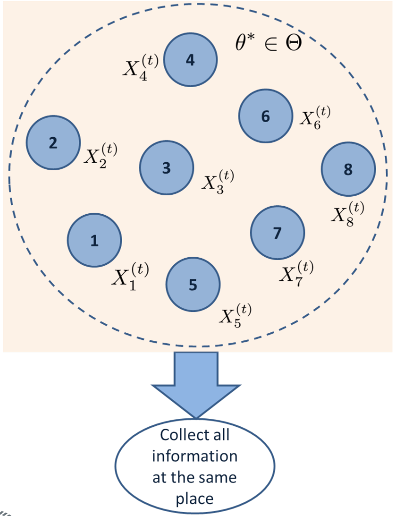
- Set of n nodes.
- Set of hypotheses $\Theta = \{\theta_1, \theta_2, \ldots, \theta_M\}$.
- Observations $X_i^{(t)}$ are i.i.d.
- Fixed known distributions $\{f_i(\cdot; \theta_1), f_i(\cdot; \theta_2), \ldots, f_i(\cdot; \theta_M)\}$.
- $\theta^* \in \Theta$ is fixed global unknown parameter
- $X_i^{(t)}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

GOAL Parametric inference of unknown θ^*

Hypothesis Testing



Hypothesis Testing

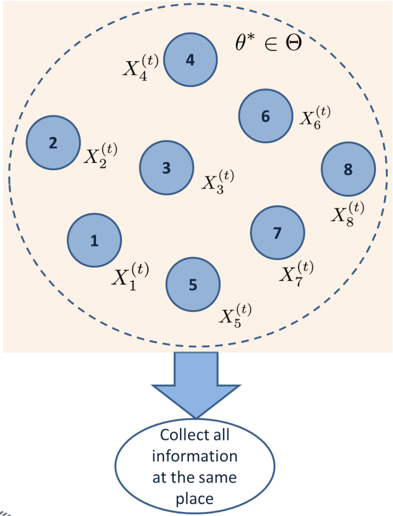


If θ^* is globally identifiable, then collecting all observations

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

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

Hypothesis Testing

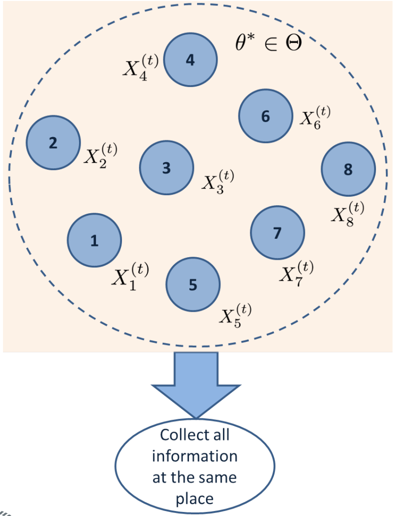


If θ^* is globally identifiable, then collecting all observations

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

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

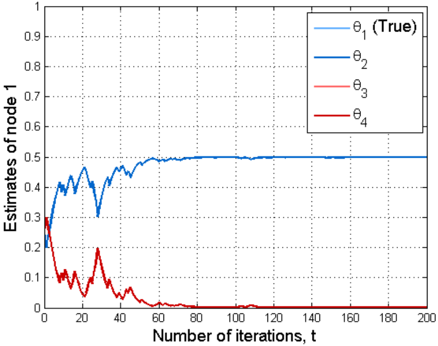
Can this be achieved locally with low dimensional observations?

Example: Low-dimensional Observations

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

$$\theta^* = \theta_1$$

Color Intensity



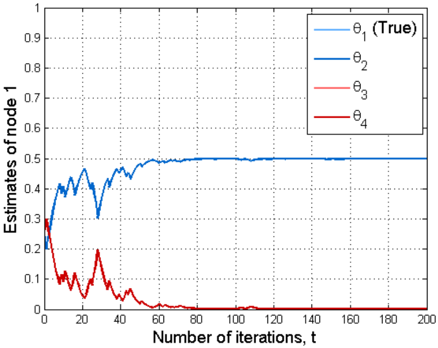
If all observations are not collected centrally, node 1 individually cannot learn θ^*.

Example: Low-dimensional Observations

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

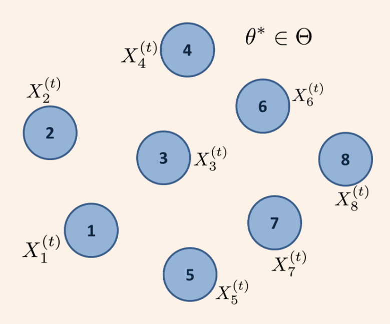
$$\theta^* = \theta_1$$

Color Intensity



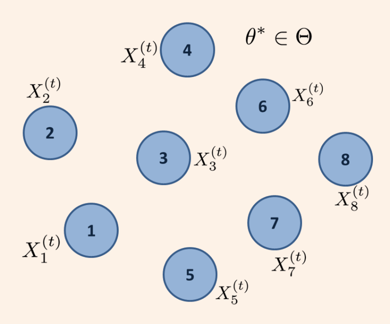
If all observations are not collected centrally, node 1 individually cannot learn θ^*. $\implies$ nodes must communicate.

Distributed Hypothesis Testing



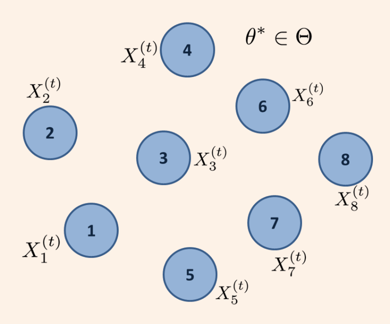
- Define $\bar{\Theta}_i = \{\theta \in \Theta : f_i(\cdot;\theta) = f_i(\cdot;\theta^*)\}$.

Distributed Hypothesis Testing



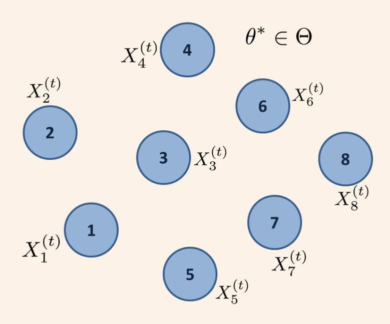
- Define $\bar{\Theta}_i = \{\theta \in \Theta : f_i(\cdot;\theta) = f_i(\cdot;\theta^*)\}$.
- $\theta \in \bar{\Theta}_i$ $\implies$ θ and θ^* are observationally equivalent for node i.

Distributed Hypothesis Testing



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

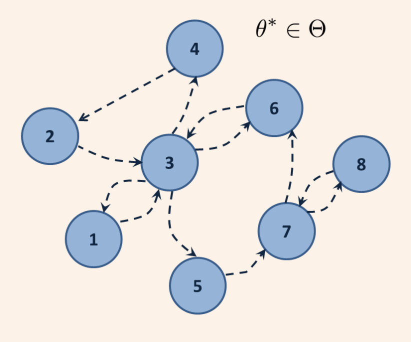
Distributed Hypothesis Testing



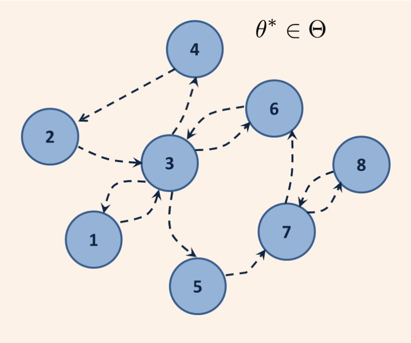
- Define $\bar{\Theta}_i = \{\theta \in \Theta : f_i(\cdot;\theta) = f_i(\cdot;\theta^*)\}$.
- $\theta \in \bar{\Theta}_i$ $\implies$ θ and θ^* are observationally equivalent for node i.
- Suppose $\{\theta^*\} = \bar{\Theta}_1 \cap \bar{\Theta}_2 \cap \ldots \cap \bar{\Theta}_n$.

GOAL Parametric inference of unknown θ^*

Learning Rule

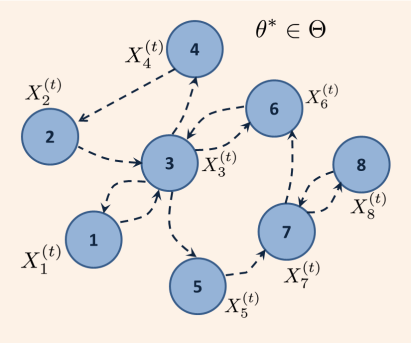


Learning Rule



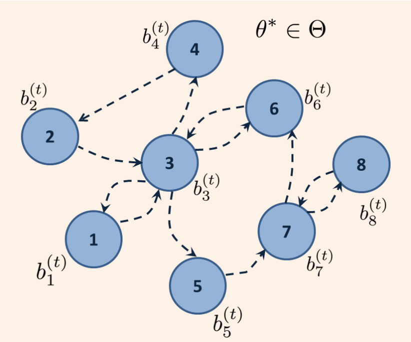
- At $t = 0$, node i begins with initial estimate vector $\mathbf{q_i^{(0)}} > 0$, where components of $\mathbf{q_i^{(t)}}$ form a probability distribution on Θ.

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 $X_i^{(t)}$.

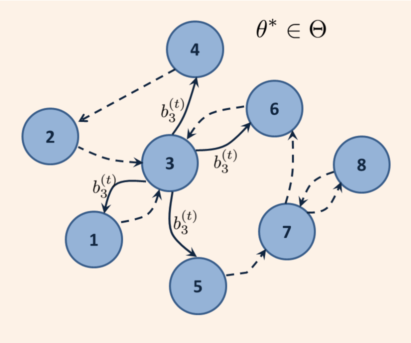
Learning Rule



- 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

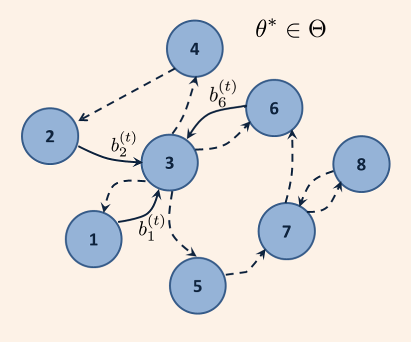


- Node i computes belief vector, $\mathbf{b_i^{(t)}}$, via Bayesian update

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

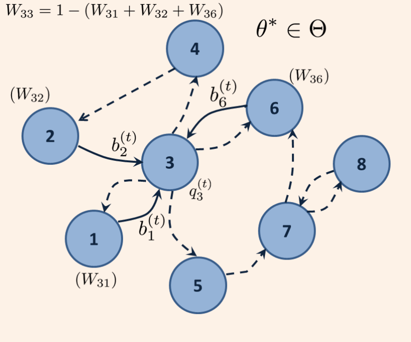
- Sends message $\mathbf{Y_i^{(t)}} = \mathbf{b_i^{(t)}}$.

Learning Rule



- Receives messages from its neighbors at the same time.

Learning Rule

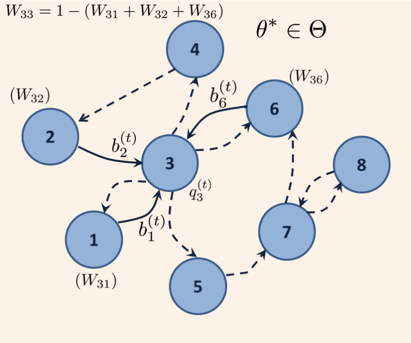


- Receives messages from its neighbors at the same time.
- Updates $\mathbf{q_i^{(t)}}$ via averaging of log beliefs,

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

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

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

Bayesian Update

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

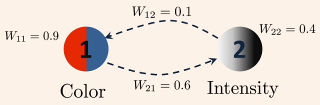
Average log-beliefs $\leftarrow$ $\{b_j(\theta, t)\}$ **messages from neighbors**

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

An example

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

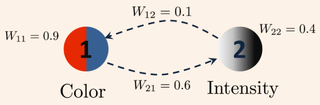
$$\theta^* = \theta_1$$

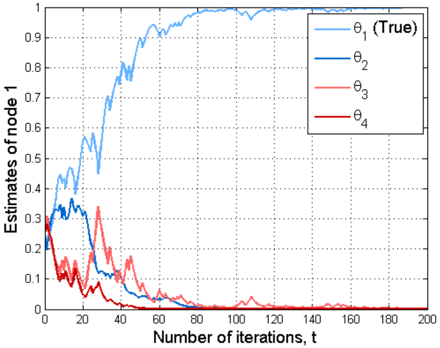


An example

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

$$\theta^* = \theta_1$$





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 $D(f_i(\cdot;\theta^*) \| f_i(\cdot;\theta)) > 0$.

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 $\theta \in \Theta$.

Convergence Results

- Let θ^* be the unknown fixed parameter.
- Suppose assumptions $1-3$ hold.
- The eigenvector centrality $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is the left eigenvector of W for eigenvalue 1.

Convergence Results

- Let θ^* be the unknown fixed parameter.
- Suppose assumptions $1 - 3$ hold.
- The eigenvector centrality $\mathbf{v} = [v_1, v_2, \ldots, v_n]$ is the left eigenvector of W for eigenvalue 1.

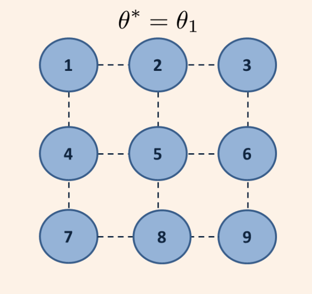
Theorem: Rate of rejecting $\theta \neq \theta^*$

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

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

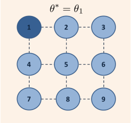
where $K(\theta^*, \theta) = \sum_{j=1}^{n} v_j D\left(f_j\left(\cdot;\theta^*\right) \| \, f_j\left(\cdot;\theta\right)\right)$.

Example: Network-wide Learning

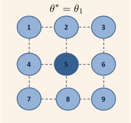


- $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and $\theta^* = \theta_1$.
- If i and j are connected, $W_{ij} = \frac{1}{\text{degree of node i}}$, otherwise 0.
- $\mathbf{v} = [\frac{1}{12}, \frac{1}{8}, \frac{1}{12}, \frac{1}{8}, \frac{1}{6}, \frac{1}{8}, \frac{1}{12}, \frac{1}{8}, \frac{1}{12}]$.

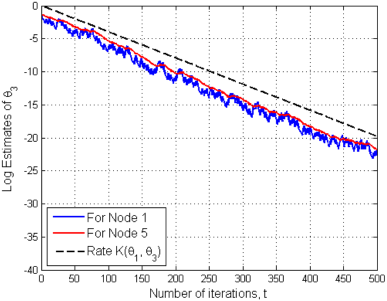
Example

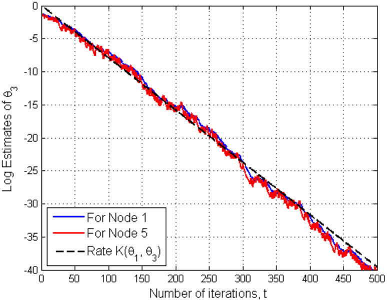


$$\bar{\Theta}_1 = \{\theta^*\},\ \bar{\Theta}_i = \Theta\ i \neq 1$$



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





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,

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

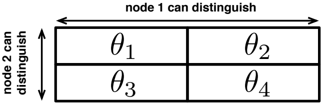
where,

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

and

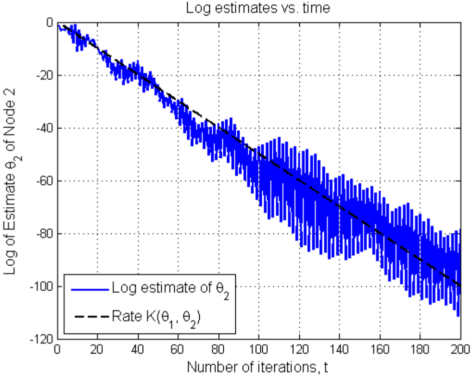
$$e_t = \frac{1}{2} \sum_{i=1}^{n} ||q_i^{(t)}(\cdot) - 1_{\theta^*}(.)||_1 = \sum_{i=1}^{n} \sum_{\theta \neq \theta^*} q_i^{(t)}(\theta).$$

Example: Periodicity

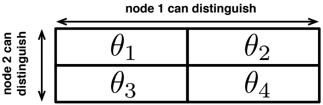


- $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ and $\theta^* = \theta_1$.
- Underlying graph is periodic,

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

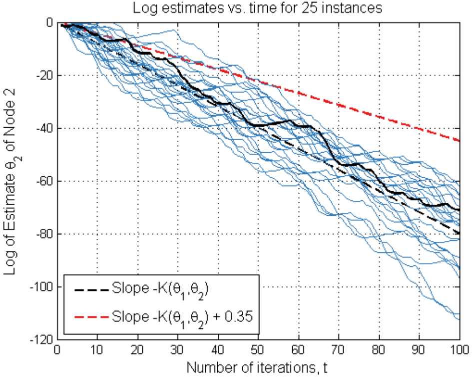


Example: Networks with Large Mixing Times



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

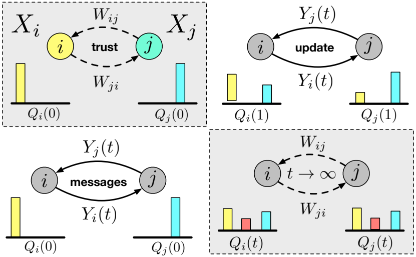
An update rule similar to ours was used in Rahnama Rad and Tahbaz-Salehi, 2010 to

- 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.

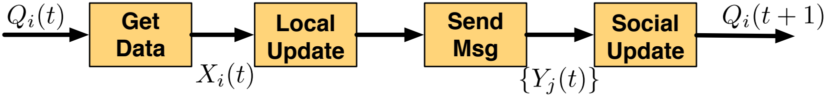
Shahrampour *et. al.* and Nedic *et. al.* (independently) showed that our learning rule coincides with distributed stochastic optimization based learning rule (W irreducible and aperiodic)

Social sampling to estimate histograms



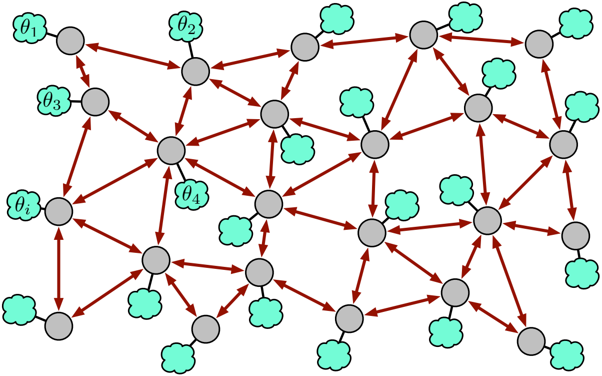
- Simple model of randomized message exchange.
- Unified analysis captures different qualitative behaviors.
- “Censoring rule” to achieve consensus to true histogram.

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.
- Time-varying case.

Thank You!