Paper Number: 3249

# Social Learning Under Inferential Attacks

Konstantinos Ntemos, Virginia Bordignon, Stefan Vlaski, and Ali H. Sayed

School of Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

## Problem Statement and our Contributions

**Inferential attacks in Social Learning**

Adversaries (which are unaware of the true hypothesis) aim at driving the network beliefs to a wrong hypothesis. In doing so, they construct *fake likelihood functions* to update their beliefs.

**In this paper we address the following questions**

- When the network is misled? Interplay among:
  - Agents' centrality.
  - Normal agents' observation models.
  - Adversaries' attack strategies.
- How adversaries can mislead the network and what information is required for them to do so? Adversaries can always construct *fake likelihood functions* given that they have access to:
  - Overall normal agents' KL divergences weighted by their centrality.
- What happens if adversaries do not have access to this information?
  - We formulate an optimization problem for adversaries' attack strategy and investigate its performance.

## System Model

- Set of agents: $\mathcal{N}=\mathcal{N}^n \bigcup \mathcal{N}^m$, where $\mathcal{N}^n$: set of normal agents, $\mathcal{N}^m$: set of adversaries.
- Agents interact over an undirected graph $\mathcal{G}=\langle \mathcal{N}, \mathcal{E}\rangle$, where $\mathcal{E}$ includes bidirectional links between agents.
- True hypothesis/state: $\theta^\star \in \Theta=\{\theta_1,\theta_2\}$.
- Normal agents aim at finding $\theta^\star$, while adversaries aim at forcing normal agents' beliefs towards the wrong state $(\Theta \setminus \theta^\star)$.
- Each agent $k \in \mathcal{N}$ has access to observations $\boldsymbol{\zeta}_{k,i} \in \mathcal{Z}_k$, time $i \geq 1$.
- Beliefs: $\boldsymbol{\mu}_{k,i}(\theta) \in (0,1)$, $k \in \mathcal{N}, \theta \in \Theta$.
- Normal agents follow the *log-linear social learning* protocol [Lalitha et al. '19]:
  1. Bayesian update step:

$$\boldsymbol{\psi}_{k,i}(\theta)=\frac{L_k(\boldsymbol{\zeta}_{k,i}|\theta)\boldsymbol{\mu}_{k,i-1}(\theta)}{\sum_{\theta'} L_k(\boldsymbol{\zeta}_{k,i}|\theta')\boldsymbol{\mu}_{k,i-1}(\theta')}, \quad k \in \mathcal{N}^n. \tag{1}$$

  2. Combination step:

$$\boldsymbol{\mu}_{k,i}(\theta)=\frac{\prod_{\ell \in \mathcal{N}_k} \boldsymbol{\psi}_{\ell,i}^{a_{\ell k}}(\theta)}{\sum_{\theta'} \prod_{\ell \in \mathcal{N}_k} \boldsymbol{\psi}_{\ell,i}^{a_{\ell k}}(\theta')}, \quad k \in \mathcal{N}^n \tag{2}$$

  where $a_{\ell k} \in [0,1]$ is the *combination weight* assigned by $k \in \mathcal{N}$ to its neighbor $\ell \in \mathcal{N}_k$ satisfying $0 < a_{\ell k} \leq 1$, for all $\ell \in \mathcal{N}_k$, $a_{\ell k}=0$ for all $\ell \notin \mathcal{N}_k$ and $\sum_{\ell \in \mathcal{N}_k} a_{\ell k}=1$.
- Adversaries' behavior: Instead of step 1 above they follow:

$$\boldsymbol{\psi}_{k,i}(\theta)=\frac{\widehat{L}_k(\boldsymbol{\zeta}_{k,i}|\theta)\boldsymbol{\mu}_{k,i-1}(\theta)}{\sum_{\theta'} \widehat{L}_k(\boldsymbol{\zeta}_{k,i}|\theta')\boldsymbol{\mu}_{k,i-1}(\theta')}, \quad k \in \mathcal{N}^m. \tag{3}$$

where $\widehat{L}_k(\cdot|\theta)$ are the *distorted likelihood functions* for adversary $k$.

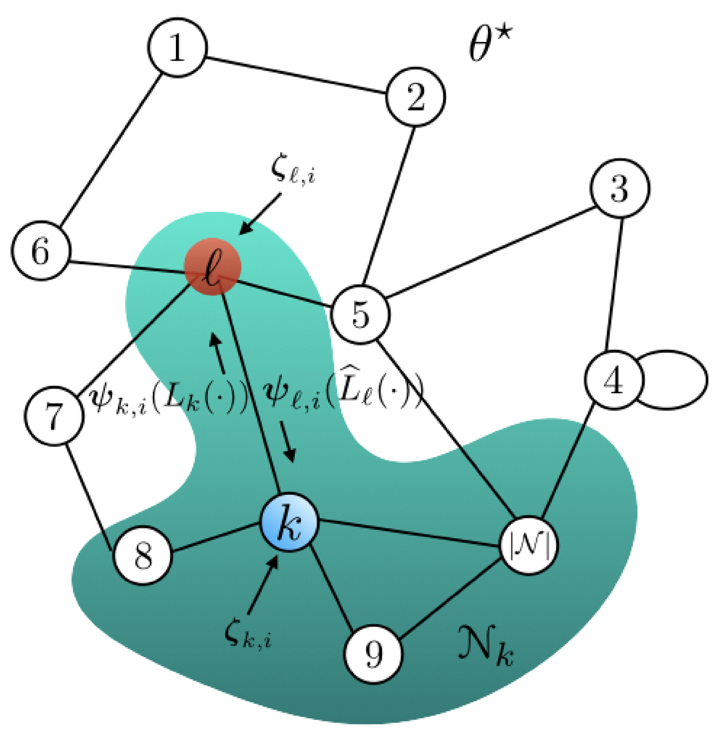


Figure 1. Illustration of the network model and the interactions between a normal agent ($k$) and an adversary ($\ell$).

This work was supported in part by the Swiss National Science Foundation grant 205121-184999.

## Modelling Assumptions

**(Finiteness of KL divergences)**
For any agent $k \in \mathcal{N}$ and for any $\theta \neq \theta^\star$, $D_{KL}\Big(L_k(\theta^\star)||L_k(\theta)\Big)$ is finite.

**(Positive initial beliefs)**
$\mu_{k,0}(\theta) > 0, \forall \theta \in \Theta, k \in \mathcal{N}$.

**(Strongly-connected network)**
The communication graph is *strongly connected* (i.e., there always exists a path with positive weights linking any two agents and at least one agent has a self-loop (there is at least one $k \in \mathcal{N}$ with $a_{kk} > 0$).

**(Distorted likelihood functions with full support)**
For every agent $k \in \mathcal{N}^m$, the distorted likelihood function satisfies $\epsilon \leq \widehat{L}_k(\zeta_{k,i}|\theta)$ for all $\zeta_{k,i} \in \mathcal{Z}_k$, $\theta \in \Theta$, where $0 < \epsilon \ll 1$ is a small positive real constant that satisfies $\epsilon < \min_k \frac{1}{|\mathcal{Z}_k|}$.

## When is the network deceived?

**Theorem 1 (Belief convergence with adversaries)**
The following are true:

1. The agents' beliefs converge a.s. to the wrong state if

$$\sum_{k \in \mathcal{N}^n} u_k \mathbb{E}\left\{\log \frac{L_k(\boldsymbol{\zeta}_k|\theta^\star)}{L_k(\boldsymbol{\zeta}_k|\theta)}\right\}=\sum_{k \in \mathcal{N}^n} u_k D_{KL}\Big(L_k(\theta^\star))||L_k(\theta)\Big) < \sum_{k \in \mathcal{N}^m} u_k \mathbb{E}\left\{\log \frac{\widehat{L}_k(\boldsymbol{\zeta}_k|\theta)}{\widehat{L}_k(\boldsymbol{\zeta}_k|\theta^\star)}\right\}, \quad \theta^\star,\theta \in \Theta, \theta^\star \neq \theta. \tag{4}$$

2. The agents' beliefs converge a.s. to the true state if

$$\sum_{k \in \mathcal{N}^n} u_k \mathbb{E}\left\{\log \frac{L_k(\boldsymbol{\zeta}_k|\theta^\star)}{L_k(\boldsymbol{\zeta}_k|\theta)}\right\}=\sum_{k \in \mathcal{N}^n} u_k D_{KL}\Big(L_k(\theta^\star))||L_k(\theta)\Big) > \sum_{k \in \mathcal{N}^m} u_k \mathbb{E}\left\{\log \frac{\widehat{L}_k(\boldsymbol{\zeta}_k|\theta)}{\widehat{L}_k(\boldsymbol{\zeta}_k|\theta^\star)}\right\}, \quad \theta^\star,\theta \in \Theta, \theta^\star \neq \theta. \tag{5}$$

$u$ is the Perron eigenvector associated with the eigenvalue at 1.

## Is it always possible to deceive the network?

- **Uninformative Probability Mass Functions (PMFs)**: The likelihood functions are *uninformative* if $L_k(\zeta_k|\theta_1)=L_k(\zeta_k|\theta_2)$ for all $\zeta_k \in \mathcal{Z}_k$, otherwise the likelihood functions are *informative*.
- Normal sub-network divergence:

$$S_j \triangleq \sum_{k \in \mathcal{N}^n} u_k \mathbb{E}\left\{\log \frac{L_k(\boldsymbol{\zeta}_k|\theta_j)}{L_k(\boldsymbol{\zeta}_k|\theta_{j'})}\right\}, \quad \theta_j=\theta^\star, j, j' \in \{1,2\}, j \neq j'. \tag{6}$$

- To characterize fake PMFs that mislead the network for any hypothesis $\theta^\star \in \Theta$ (since adversaries are unaware of the true hypothesis) the system of inequalities resulting from (4) needs to be solved.

**We consider the following construction of fake likelihood functions $\widehat{L}(\cdot|\theta_1), \widehat{L}(\cdot|\theta_2)$ for an adversary $k \in \mathcal{N}$:**

$$\widehat{L}_k(\zeta_k|\theta_j)=\begin{cases}\epsilon_{j'}, & \text{if } \zeta_k=\zeta_k^{j'},\\ \alpha-\epsilon_{j'}, & \text{if } \zeta_k=\zeta_k^{j},\\ \epsilon, & \text{otherwise}\end{cases} \tag{7}$$

where $\alpha=1-(|\mathcal{Z}_k|-2)\epsilon$, $j,j' \in \{1,2\}$, $j \neq j'$, $\zeta_k^1, \zeta_k^2 \in \mathcal{Z}_k$ are such that $L_k(\zeta_k^1|\theta_1)L_k(\zeta_k^2|\theta_2) \neq L_k(\zeta_k^1|\theta_2)L_k(\zeta_k^2|\theta_1)$.

**Theorem 2 (Distorted PMFs with known divergences)**
There is always a construction of fake likelihood functions of the form (7) that misleads the network for any $\theta^\star \in \Theta$, given that there exists at least one adversary with informative PMFs, for sufficiently small $\epsilon$. Adversaries need to know the normal sub-network divergences $S_1, S_2$.

## What if the normal sub-network divergences ($S_1$, $S_2$) are unknown?

- Rearranging (4), we can define the following cost function.

$$\mathcal{C}(\theta^\star)=\sum_{k \in \mathcal{N}^n} u_k D_{KL}(L_k(\theta^\star))||L_k(\theta))+\sum_{\ell \in \mathcal{N}^m} u_\ell \sum_{\zeta_\ell} L_\ell(\zeta_\ell|\theta^\star)\log \frac{\widehat{L}_\ell(\zeta_\ell|\theta^\star)}{\widehat{L}_\ell(\zeta_\ell|\theta)}, \quad \theta^\star,\theta \in \Theta, \theta_1 \neq \theta_2. \tag{8}$$

- Adversaries can minimize $C(\theta^\star)$ over $\widehat{L}_\ell(\theta_1), \widehat{L}_\ell(\theta_2)$, by assuming some prior distribution over the states $\pi=(\pi_{\theta_1},\pi_{\theta_2})$ (common prior among adversaries).
- Taking expectation over $\theta^\star$ in (8) leads to the following minimization problem:

$$\begin{aligned}&\min_{\widehat{L}_\ell(\theta_1),\widehat{L}_\ell(\theta_2)} \sum_{\theta \in \Theta} \pi_\theta C(\boldsymbol{\theta}^\star=\theta), \quad \ell \in \mathcal{N}^m\\ &\text{s.t. } \widehat{L}_\ell(\zeta_\ell|\theta) \geq \epsilon, \qquad \forall \zeta_\ell \in \mathcal{Z}_\ell, \theta \in \Theta,\\ &\sum_{\zeta_\ell \in \mathcal{Z}_\ell} \widehat{L}_\ell(\zeta_\ell|\theta)=1, \quad \forall \theta \in \Theta.\end{aligned} \tag{9}$$

## Attack strategies without any knowledge about the network model

- Define the coefficients $Z_\ell(\zeta_\ell)$ expressing the *relative confidence* that $\zeta_\ell$ resulted from state $\theta_1$ instead of $\theta_2$ as:

$$Z_\ell(\zeta_\ell) \triangleq \pi_{\theta_1} L_\ell(\zeta_\ell|\theta_1)-\pi_{\theta_2} L_\ell(\zeta_\ell|\theta_2), \quad \zeta_\ell \in \mathcal{Z}_\ell. \tag{10}$$

- Define the sets:

$$\mathcal{D}_\ell^1 \triangleq \{\zeta_\ell : Z(\zeta_\ell) \geq 0, \quad \ell \in \mathcal{N}^m\} \tag{11}$$

$$\mathcal{D}_\ell^2 \triangleq \{\zeta_\ell : Z(\zeta_\ell) < 0, \quad \ell \in \mathcal{N}^m\} \tag{12}$$

- The solution of opt. problem (9) depends on whether $\mathcal{D}_\ell^1, \mathcal{D}_\ell^2$ are both non-empty or not.

**Theorem 3 (Distorted PMFs with unknown divergences and mixed confidence)**

If both $\mathcal{D}_\ell^1, \mathcal{D}_\ell^2$ are non-empty sets, then the attack strategy optimizing (9) for every adversary $\ell \in \mathcal{N}^m$ is given by

$$\widehat{L}_\ell(\zeta_\ell|\theta_j)=\begin{cases}\epsilon, & \text{if } \zeta_\ell \in \mathcal{D}_\ell^j,\\ \dfrac{Z_\ell(\zeta_\ell)(1-|\mathcal{D}_\ell^j|\epsilon)}{\sum_{\zeta_\ell \notin \mathcal{D}_\ell^j} Z_\ell(\zeta_\ell)}, & \text{if } \zeta_\ell \notin \mathcal{D}_\ell^j\end{cases} \tag{13}$$

where $j \in \{1,2\}$.

**Theorem 4 (Distorted PMFs with unknown divergences and pure confidence**

Let $\mathcal{D}_\ell^1=\emptyset$ or $\mathcal{D}_\ell^2=\emptyset$. Then, the attack strategy optimizing (9) for an agent $\ell \in \mathcal{N}^m$ is given by

$$\widehat{L}_\ell(\zeta_\ell|\theta_j)=\begin{cases}1-(|\mathcal{Z}_\ell|-1)\epsilon, & \text{if } \mathcal{D}_\ell^j=\mathcal{Z}_\ell, \zeta_\ell=\zeta_{min},\\ \epsilon, & \text{if } \mathcal{D}_\ell^j=\mathcal{Z}_\ell \text{ and } \zeta_\ell \neq \zeta_{min},\\ \dfrac{Z_\ell(\zeta_\ell)}{\sum_{\zeta_\ell \in \mathcal{Z}_\ell} Z_\ell(\zeta_\ell)}, & \text{if } \mathcal{D}_\ell^j=\emptyset\end{cases} \tag{14}$$

where $j \in \{1,2\}$ and $\zeta_{min}=\arg\min_{\zeta_\ell}\{Z_\ell(\zeta_\ell)\}$.

## Intuition behind Theorems 3 and 4 - "Flip and inflate" strategy

Examples of the solutions of Theorems 3 and 4 are given with $|\mathcal{Z}_k|=5$. Red color depicts the higher value of $L_k(\zeta_k|\theta)$ for every observation $\zeta_k$ w.r.t. states (i.e., $L_k(\zeta_k|\theta)$ in red are such that $\pi(\theta)L_k(\zeta_k|\theta) > \pi(\theta')L_k(\zeta_k|\theta')$, $\theta \neq \theta'$). We set $\epsilon=10^{-3}$.

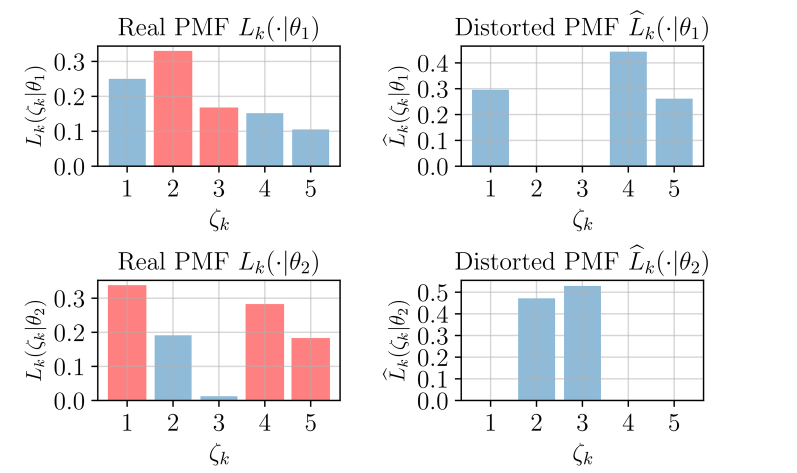


Figure 2. Theorem 3 ($\pi_{\theta_1}=\pi_{\theta_2}=0.5$).

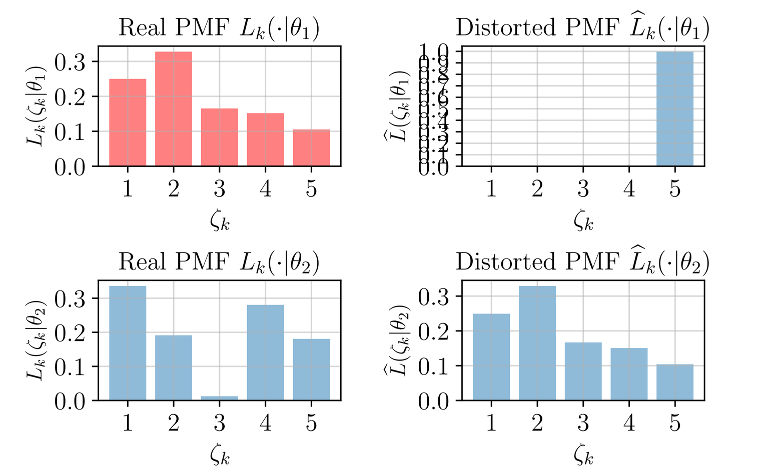


Figure 3. Theorem 4 ( $\pi_{\theta_1}=0.99, \pi_{\theta_2}=0.01$).

## Simulations

- 15 agents, with 11 normal agents and 4 adversaries interact over a random network topology (strongly connected network) and a star topology (the central agent is adversary).
- All agents assign uniform combination weights to their neighbors and we set $\epsilon=10^{-3}$.
- All agents observe the state through a *binary symmetric channel*, (i.e., $\mathcal{Z}_k=\{\zeta_1,\zeta_2\}$ for all $k \in \mathcal{N}$) with observation probabilities $L_k(\zeta_1|\theta_1)=L_k(\zeta_2|\theta_2)=p$ and $L_k(\zeta_2|\theta_1)=L_k(\zeta_1|\theta_2)=1-p$ for all $k \in \mathcal{N}$.

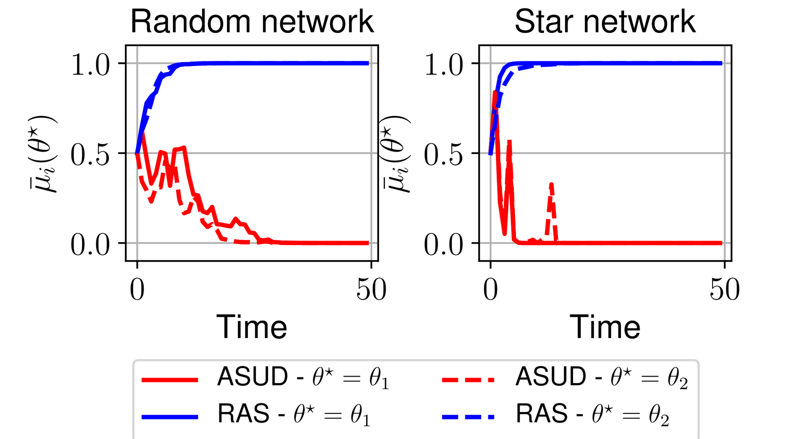


Figure 4. Evolution of agents' average belief on $\theta^\star$ (i.e., $\bar{\mu}_i(\theta^\star) \triangleq \frac{\sum_{k \in \mathcal{N}} \mu_{k,i}(\theta^\star)}{|\mathcal{N}|}$) for $p=0.8$. Left: random topology, Right: star topology. ASUD: Attack Strategy with Unknown Divergences (given by Theorem 3, $\pi_{\theta_1}=\pi_{\theta_2}=0.5$), RAS: Random Attack Strategy.

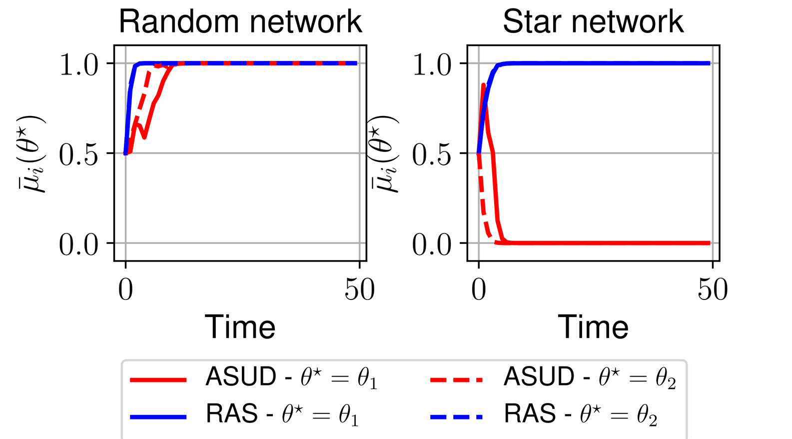


Figure 5. Evolution of agents' average belief on $\theta^\star$ with highly discriminating models ($p=0.95$). Left: random topology, Right: star topology. ASUD: Attack Strategy with Unknown Divergences (given by Theorem 3, $\pi_{\theta_1}=\pi_{\theta_2}=0.5$), RAS: Random Attack Strategy.

konstantinos.ntemos@epfl.ch, virginia.bordignon@epfl.ch, stefan.vlaski@epfl.ch, ali.sayed@epfl.ch

2021 IEEE International Conference on Acoustics, Speech and Signal Processing

6-11 June 2021 – Toronto, Ontario, Canada