

Estimating True Beliefs from Declared Opinions

Jennifer Tang, Aviv Adler, Amir Ajorlou, and Ali Jadbabaie

Abstract—A common feature of interactions and opinion exchanges on social networks, both real and digital, is the presence of social pressure, which may cause agents to alter their expressed opinions in order to fit in with those around them. In such systems, each agent has a true and unchanging *inherent belief* but broadcasts a *declared opinion* at each time step, influenced by both her inherent belief and the declared opinions of her neighbors. An important question in this setting is *parameter estimation*: how to disentangle the effects of social pressure and estimate the underlying true beliefs of the agents from their declared opinions. To address this question, Jadbabaie et al. [1] formulated the *interacting Pólya urn model* of opinion dynamics under social pressure and studied parameter estimation on complete-graph social networks using an aggregate estimator. They found that, under these settings, this estimator asymptotically estimates the true beliefs unless majority pressure causes the network to approach consensus over time.

In this work, we consider parameter estimation for the interacting Pólya urn model on arbitrary networks, and prove that the maximum likelihood estimator always asymptotically estimates the true beliefs – including the degree to which those beliefs are held – even when consensus is approached.

I. INTRODUCTION

Opinion dynamics is the study of how people’s opinions evolve over time as they interact with others on social networks. This can provide insights and predictions into how public opinion develops on a variety of political, social, commercial and cultural topics. For instance, Ancona et al. [2] used opinion dynamics models to model the spread of vaccine hesitancy and to develop marketing strategies to help combat it. In many opinion dynamics models, there is an assumption that people are truthful in the opinions they share. However, in reality this is not always the case, as people often alter their expressed views to better fit in with their social environment, which in turn feeds back into the social environment. The social pressure feedback loop can cause publicly-expressed opinions to become arbitrarily uniform over time [3], which can make parameter estimation difficult.

In this work, we study an *interacting Pólya urn model* for opinion dynamics under social pressure, originating from [1] and developed further in [4]. This model captures a system of agents with stochastic behaviors who additionally might be untruthful due to a desire to conform to their neighbors. This model consists of n agents on a fixed network communicating on an issue with two basic sides, 0 and 1. Each agent has an *inherent* (true and unchanging) belief, which is either 0 or 1, and also a *bias parameter* γ which indicates to

what degree they are willing to share their inherent belief as opposed to conforming. Then the agents communicate their *declared* opinions to their neighbors at discrete time steps: at each step $t = 1, 2, \dots$, all the agents simultaneously declare one of the two opinions (i.e. either ‘0’ or ‘1’), which is then observed by their neighbors; the declarations of all the agents at any given step are made at random and independently of each other but with probabilities determined by their inherent belief, bias parameter, and the ratio of the two opinions previously observed by the agent up to the current time. This can represent both scenarios where agents alter their statements (contrary to their actual beliefs) to better fit in with the opinions they have observed from others in the past and scenarios where the agents update their beliefs according to the declared opinions of others, but retain a bias towards their original beliefs.

A. Background Literature

We refer the reader to [1] and [4] for in-depth discussion of previous literature. Here, we discuss some relevant highlights.

A highly influential opinion dynamics model is the DeGroot model [5], where agents in a network average their neighbors’ opinions in an iterative manner. With this procedure, the entire group asymptotically approaches a state where they all share a single opinion, a phenomenon known as consensus. However in real social networks, consensus is not always reached. To deal with this, other opinion dynamics models were created. Among these is the Friedkin-Johnsen model [6]. Each agent in the Friedkin-Johnsen model updates her opinion at each step by averaging her neighbors’ opinions (as in the DeGroot model) and then averaging the result with her initial opinion.

Ye et al. [7] study a model in which each agent has both a private and expressed opinion, which evolve differently. Agents’ private opinions evolve using the same update as in the Friedkin-Johnsen model, while their public opinions are updated as the average of their own private opinion and the average public opinion of their neighbors. Both [3] and [7] are very similar to [1], since agents’ expressed opinions may not match their internal beliefs. However, unlike [1], [7] assumes opinions are precisely expressed on a continuous interval, which is unrealistic for certain applications. On the other hand [3] works with binary opinions like [1], though with a significantly more complex model that includes additional terms and parameters.

The analysis in [1] is primarily focused on studying whether inherent beliefs are recoverable using an aggregate estimator. This is carried out by establishing the convergence

The work was supported by ARO MURI W911 NF-19-1-0217 and a Vannevar Bush Fellowship from the Office of the Under Secretary of Defense.

of the dynamics in the network and analyzing the equilibrium state, though the analysis is limited to the complete graph and all agents having the same amount of resistance to social pressure. In [4], the authors generalize this result to allow for any network structure and for any bias parameters. They found that the proportion of declared opinions of each agent converges almost surely to an equilibrium point in any network configuration. They also determined necessary and sufficient conditions for a network to approach consensus. We note that the definition of consensus used for the interacting Pólya urn model is that all agents will declare a single opinion (all ‘0’ or all ‘1’) with probability tending to 1.

B. Contributions

In [1], the authors consider when it is possible to asymptotically determine the inherent beliefs of the agents based on their history of declared opinions and those of their neighbors. They study a simplified case in which the social network is an (unweighted) complete graph and all agents have the same, known, degree of bias towards their true beliefs, and consider a specific aggregate estimator which tries to first estimate the proportion of agents with true belief 1 and then determine which agents those are. In this setting, they show that the aggregate estimator estimates the proportion of agents with true belief 1 if and only if the agents do not asymptotically approach consensus (where a large majority causes all agents declare the same opinion with probability approaching 1).

In this work, we consider the general setting:

- 1) the social network is an arbitrary weighted (and connected) undirected graph, possibly with self-loops;
- 2) the agents can have heterogeneous bias parameters, indicating different levels of resistance to social pressure or certainty in their inherent opinions, which are not known.

This greatly increases the applicability of the model, as real-life social networks have a variety of different structures and people often react very differently to social pressure.

In this setting, we study the maximum likelihood estimator (MLE), which estimates bias parameters from the history of declared opinions, rather than the aggregate estimator from [1]. We also derive a simplified estimator for inherent beliefs from the MLE, which takes a clean form with a low-dimensional sufficient statistic, consisting of two values which are simple to update at each step. We show that if the history of the agents’ declared opinions is known, the MLE almost surely asymptotically converges to the correct inherent beliefs and bias parameters of all the agents in all such networks (even when the network approaches consensus). This resolves the fundamental question posed in [1] of whether such estimation is always possible.

II. MODEL DESCRIPTION

We use the model from [4] which is a slight generalization of the model from [1]. We refer the reader to [4, Sec II] for more details on the model and in particular [4, Sec II F] for intuition on the model.

A. Graph Notation

Let undirected graph $G = (V, E)$ (possibly including self-loops) be a network of n agents (corresponding to the vertices) labeled $i = 1, 2, \dots, n$. For each edge $(i, j) \in E$, there is a weight $a_{i,j} \geq 0$, where by convention $a_{i,j} = 0$ if $(i, j) \notin E$. We denote the weighted degree of vertex i as $\text{deg}(i) = \sum_j a_{i,j}$.

B. Inherent Beliefs and Declared Opinions

Each agent i has an *inherent belief* $\phi_i \in \{0, 1\}$, which does not change. At each time step t , each agent i (simultaneously) announces a *declared opinion* $\psi_{i,t} \in \{0, 1\}$. We denote by \mathcal{H}_t the *history* of the process, consisting of all $\psi_{i,\tau}$ for $\tau \leq t$. Each agent has a bias parameter $\gamma_i \in (0, \infty)$ where $\gamma_i \neq 1$.

If $\phi_i = 1$, then $\gamma_i > 1$. If $\phi_i = 0$, then $\gamma_i < 1$. If an agent has $\gamma_i = 1$, this signifies that the agent does not have an inherent belief, which is case we do not allow in this work on inferring inherent beliefs.

The declarations $\psi_{i,t}$ are based on a probabilistic rule which we define after the following definitions:

Define

$$\mu_i^0(t) \triangleq \frac{m_i^0 + \sum_{\tau=2}^t \sum_{j=1}^n a_{i,j} \mathbb{I}[\psi_{i,\tau} = 0]}{m_i^0 + m_i^1 + (t-1) \text{deg}(i)} \quad (1)$$

$$\mu_i^1(t) \triangleq \frac{m_i^1 + \sum_{\tau=2}^t \sum_{j=1}^n a_{i,j} \mathbb{I}[\psi_{i,\tau} = 1]}{m_i^0 + m_i^1 + (t-1) \text{deg}(i)}. \quad (2)$$

The parameter $\mu_i^1(t)$ is essentially the sufficient statistic that summarizes the proportion of declared opinions in the neighborhood of given agent i up to time t . Since $\mu_i^0(t) = 1 - \mu_i^1(t)$, we simplify the notation to $\mu_i(t) \triangleq \mu_i^1(t)$.

Define the function (note that μ, γ are scalars)

$$f(\mu, \gamma) \triangleq \frac{\gamma\mu}{1 + (\gamma-1)\mu} = \frac{1}{1 + \gamma\left(\frac{1}{\mu} - 1\right)}. \quad (3)$$

The probabilistic rule for declared opinions $\psi_{i,t+1}$ can be written as

$$\psi_{i,t+1} \triangleq \begin{cases} 1 & \text{with probability } f(\mu_i(t), \gamma_i) \\ 0 & \text{with probability } 1 - f(\mu_i(t), \gamma_i) \end{cases}. \quad (4)$$

Note that the bias parameter γ_i is always defined as agent i ’s bias towards opinion 1. However, the model is symmetric in the following way: a γ bias towards 1 is equivalent to a $1/\gamma$ bias towards 0,¹ which is captured by the equation

$$f(\mu_i^1(t), \gamma) = 1 - f(\mu_i^0(t), 1/\gamma). \quad (5)$$

We also define a sufficient statistic that summarizes agent i ’s declarations. Let $b_i^0, b_i^1 > 0$ (the initialization) be such

¹The honesty parameter in [1] is equivalent to the bias towards the agent’s true belief, i.e. a honesty parameter of γ with a true belief of 0 corresponds to a bias parameter of $1/\gamma$.

that $b_i^0 + b_i^1 = 1$ for each i . Let

$$\beta_i^0(t) = \frac{b_i^0}{t} + \frac{1}{t} \sum_{\tau=2}^t (1 - \psi_{i,\tau}) \quad (6)$$

$$\beta_i^1(t) = \frac{b_i^1}{t} + \frac{1}{t} \sum_{\tau=2}^t \psi_{i,\tau}. \quad (7)$$

These are counts and proportions of declarations of each opinion (or “time-averaged declarations”) for each agent (plus initial conditions). We similarly use $\beta_i(t) \triangleq \beta_i^1(t)$. So long as

$$m_i^0 = \sum_{j=1}^n a_{i,j} b_j^0 \text{ and } m_i^1 = \sum_{j=1}^n a_{i,j} b_j^1, \quad (8)$$

it follows from the definition that

$$\mu_i(t) = \frac{1}{\deg(i)} \sum_{j=1}^n a_{i,j} \beta_j(t). \quad (9)$$

We define the vectors

$$\boldsymbol{\mu}(t) \triangleq [\mu_1(t), \dots, \mu_n(t)]^\top \quad (10)$$

$$\boldsymbol{\beta}(t) \triangleq [\beta_1(t), \dots, \beta_n(t)]^\top. \quad (11)$$

C. Evolution of Declared Ratio

Given the history, we can write the expected value of the next $\boldsymbol{\beta}(t+1)$ as

$$\mathbb{E}[\boldsymbol{\beta}(t+1)|\mathcal{H}_t] = \frac{t}{t+1} \boldsymbol{\beta}(t) + \frac{1}{t+1} \begin{bmatrix} f(\mu_1(t), \gamma_1) \\ f(\mu_2(t), \gamma_2) \\ \vdots \\ f(\mu_n(t), \gamma_n) \end{bmatrix} \quad (12)$$

$$= \frac{t}{t+1} \boldsymbol{\beta}(t) + \frac{1}{t+1} f(\boldsymbol{\mu}(t), \boldsymbol{\gamma}), \quad (13)$$

or alternatively

$$\mathbb{E}[\boldsymbol{\beta}(t+1) - \boldsymbol{\beta}(t)|\mathcal{H}_t] = \frac{1}{t+1} (f(\boldsymbol{\mu}(t), \boldsymbol{\gamma}) - \boldsymbol{\beta}(t)). \quad (14)$$

The equilibrium points of the expected dynamics are given by the solutions to the equations

$$0 = (\gamma_i - 1)\beta_i\mu_i + \beta_i - \gamma_i\mu_i \quad (15)$$

where $i \in \{1, \dots, n\}$ and $\mu_i = \frac{1}{\deg(i)} \sum_{j=1}^n a_{i,j} \beta_j$. (See [4] for more discussion.)

D. Consensus

An important term for this work is *consensus*, which needs to be defined appropriately for our stochastic system.

Definition 1. Consensus is approached if

$$\boldsymbol{\beta}(t) \rightarrow \mathbf{1} \text{ or } \boldsymbol{\beta}(t) \rightarrow \mathbf{0} \text{ as } t \rightarrow \infty. \quad (16)$$

Since $\beta_i(t)$ represents the ratio of 1’s agent i has declared, consensus is approached when this ratio goes to 0 or 1.

III. ESTIMATORS FOR INFERRING INHERENT BELIEFS AND BIAS PARAMETERS

One of the key questions in [1] is whether it is possible to infer the inherent beliefs of agents from the history of declared opinions. The authors of [1] studied the interacting Pólya urn model on the complete graph using an aggregate estimator which keeps track of the overall fraction of declared opinion of all agents throughout time, and showed that this estimator may not correctly estimate the inherent belief of all agents (even in the limit) if they approach consensus. Consensus presents difficulties for estimators since asymptotically all agents approach the same behavior regardless of their inherent beliefs.

However, we show that estimators based on maximizing the log-likelihood (almost surely) infer the inherent belief of any agent i in the limit, even when consensus is approached. This fact is connected to [4, Lemma 2] – each agent declares both opinions infinitely often, yielding sufficient information to determine inherent beliefs over time.

Additionally, unlike [1], our formulation also allows agents to have different bias parameters. Thus, it is natural to ask is how to estimate the bias parameter of any agent. Intuitively, after enough time has passed, the values of $\mu_i(t)$ and $\beta_i(t)$ will converge to values close to the equilibrium point. In such a case, we can use (15) to estimate the bias parameter γ_i and inherent belief ϕ_i with

$$\widehat{\gamma}_i(t) = \frac{\beta_i(t)}{1 - \beta_i(t)} \frac{1 - \mu_i(t)}{\mu_i(t)} \quad (17)$$

$$\widehat{\phi}_i = \mathbb{I}\{\beta_i(t) < \mu_i(t)\} \quad (18)$$

These estimators are asymptotically consistent, i.e.

$$\lim_{t \rightarrow \infty} \widehat{\gamma}_i(t) = \gamma_i \quad (19)$$

$$\lim_{t \rightarrow \infty} \widehat{\phi}_i = \phi_i \quad (20)$$

when the dynamics converge to an interior equilibrium point (see [4] for details). However, plugging the equilibrium values into (17) is not well-defined if $\beta_i(t)$ and $\mu_i(t)$ both converge to either 0 or 1, i.e. when consensus is approached. This shows that more careful analysis needs to be done in order to estimate the bias parameters and inherent beliefs in all circumstances.

IV. DEFINITION OF ESTIMATORS

We assume at time t the estimator has at its disposal the history of agent i and agent i ’s neighbors’ declarations up to and including time t , which we denote as \mathcal{H}_t . Given \mathcal{H}_{t-1} , we can compute exactly the values of

$$\mathbb{P}[\psi_{i,t} = 1 | \mathcal{H}_{t-1}] = f(\mu_i(t-1), \gamma_i). \quad (21)$$

Note that in general $\mathbb{P}[\psi_{i,t} = 1]$ is a random variable dependent on \mathcal{H}_{t-1} , while $\mathbb{P}[\psi_{i,t} = 1 | \mathcal{H}_{t-1}]$ is constant. The sequence $\mathcal{H}_0 \subseteq \mathcal{H}_1 \subseteq \dots$ is also a filtration.

Our estimator to predict γ_i is based on the maximum log-likelihood estimator:

Definition 2. The single-step negative log-likelihood for a given agent i at time t and parameter γ is

$$\ell_i(\gamma, t) \triangleq - \left(\mathbb{I}\{\psi_{i,t} = 1\} \log(f(\mu_i(t-1), \gamma)) + \mathbb{I}\{\psi_{i,t} = 0\} \log(1 - f(\mu_i(t-1), \gamma)) \right) \quad (22)$$

The negative log-likelihood for a given agent i at time t and parameter γ is

$$L_i(\gamma, t) \triangleq \sum_{\tau=2}^t \ell_i(\gamma, \tau). \quad (23)$$

The MLE for bias parameter γ_i gives the value of γ that maximizes the likelihood of agent i 's declarations, which also minimizes the negative log-likelihood.

Definition 3 (Estimator for Bias Parameter). The maximum likelihood estimator (MLE) for the bias parameter γ at time t is given by

$$\hat{\gamma}(t) \triangleq \arg \min_{\gamma} L_i(\gamma, t) \quad (24)$$

Since the inherent belief of an agent is defined as whether the bias parameter is greater than or less than 1, given the MLE estimator, we can always predict the inherent belief of agent i by taking

$$\text{sign}(\log(\hat{\gamma}(t))). \quad (25)$$

However, if we assume that $\gamma_i \neq 1$, and are only interested estimating the inherent beliefs, this reduces to a simpler form.

Let $\bar{\beta}_i(t) = \frac{1}{t-1} \sum_{\tau=2}^t \mathbb{I}[\psi_{i,\tau} = 1]$, which is a similar quantity to $\beta_i(t)$ except that the arbitrary initial conditions are not included. (If t is large, then difference between $\beta_i(t)$ and $\bar{\beta}_i(t)$ is negligible.)

Definition 4 (Estimator for Inherent Belief). Let

$$\hat{\phi}_i(t) = \frac{1}{2} \text{sign} \left((t-1) \bar{\beta}_i(t) - \left(\sum_{\tau=1}^{t-1} \mu_i(\tau) \right) \right) + \frac{1}{2}. \quad (26)$$

We note that the transformation of multiplying by $1/2$ and adding $1/2$ is simply to map the output of the $\text{sign}(\cdot)$ operation to 0 and 1. Fundamentally, this estimator requires only comparing

$$\bar{\beta}_i(t) > \frac{1}{t-1} \sum_{\tau=1}^{t-1} \mu_i(\tau). \quad (27)$$

Note that $\hat{\phi}_i(t)$ does not depend on knowing the bias parameter, as it only assumes that $\gamma \neq 1$, and the estimator is simple to compute as it only requires the aggregate count of an agent's declarations and her neighborhood's declarations.

Intuitively, this compares agent i 's actual declarations against its expected declarations if $\gamma_i = 1$ (i.e. if the agent were unbiased); however, the consistency of this estimator is derived from that of the MLE for the bias parameter given in Definition 3. We show this derivation in Section VI.

Lastly, note that while both the estimator in Definition 4 and the estimator in (18) have the same asymptotic values

when the network does not approach consensus, only the estimator in Definition 4 is guaranteed to work when the network approaches consensus.

A. Preliminaries: Bounds on $\mu_i(t)$

The following property is essential to our the analysis:

- We have $M_i(t) = m_i^0 + m_i^1 + (t-1) \text{deg}(i)$. This is because each of agent i 's neighbors contributes a signal of weight $a_{i,j}$ to agent i at each time step.
- We know that

$$\mu_i(t) \in \left[\frac{1}{M_i(t)}, 1 - \frac{1}{M_i(t)} \right] \quad (28)$$

As an important consequence,

$$\mu_i(t) \geq \frac{\kappa}{t} \text{ and } 1 - \mu_i(t) \geq \frac{\kappa}{t} \quad (29)$$

for some constant κ and $t > 1$. We can set $\kappa = (\max_i \text{deg}(i))^{-1}$.

B. Negative Log-likelihood Properties

We analyze in depth the log-likelihood estimator which is key to our analysis. We start by introducing an alternative representation for $\ell_i(\gamma, t)$. Let $\tilde{\psi}_{i,t} = 2\psi_{i,t} - 1$, which takes values -1 and $+1$, instead of 0 and 1, which gives a more symmetric representation of the process.

Noting that $f(\mu_i(t), \gamma)$ then represents the probability of $\tilde{\psi}_{i,t} = 1$, we have

$$\ell_i(\gamma, t) = -\log \left(\frac{1}{1 + e^{-\tilde{\psi}_{i,t} \log(\gamma \frac{\mu_i(t-1)}{1-\mu_i(t-1)})}} \right) \quad (30)$$

$$= \log \left(1 + e^{-\tilde{\psi}_{i,t} \log(\gamma \frac{\mu_i(t-1)}{1-\mu_i(t-1)})} \right). \quad (31)$$

It is useful to change the parameterization:

$$\lambda \triangleq \log \gamma \text{ and } \nu_i(t) \triangleq \log \frac{\mu_i(t)}{1 - \mu_i(t)}. \quad (32)$$

Using λ is a nice alternative to the bias parameters. We thus define some quantities which take $\lambda = \log \gamma$ as the argument instead of γ and use them where convenient:

$$\tilde{\ell}_i(\lambda, t) \triangleq \ell_i(\gamma, t) \text{ and } \tilde{L}_i(\lambda, t) \triangleq L_i(\gamma, t). \quad (33)$$

For this section to Section V we will fix i and then use γ_1 and γ_2 to represent any two possible choices for γ_i . Define

$$Z(\gamma_1, \gamma_2, t) \triangleq L_i(\gamma_2, t) - L_i(\gamma_1, t). \quad (34)$$

If $Z(\gamma_1, \gamma_2, t)$ is positive, intuitively, γ_1 is more likely than γ_2 , so we expect γ_1 to be the true parameter. Indeed, if γ_1 is the true parameter, then

$$\mathbb{E}[Z(\gamma_1, \gamma_2, t)] = \sum_{\tau=2}^t \mathbb{E} \left[\mathbb{I}\{\psi_{i,s} = 1\} \log \frac{f(\mu_i(\tau-1), \gamma_1)}{f(\mu_i(\tau-1), \gamma_2)} + \mathbb{I}\{\psi_{i,\tau} = 0\} \log \frac{1 - f(\mu_i(\tau-1), \gamma_1)}{1 - f(\mu_i(\tau-1), \gamma_2)} \right] \quad (35)$$

$$= \sum_{\tau=2}^t D_{\text{kl}}(f(\mu_i(\tau-1), \gamma_1) \| f(\mu_i(\tau-1), \gamma_2)) \quad (36)$$

which is always a nonnegative quantity.

Proposition 1. $L_i(\gamma, t)$ is a stochastic process which satisfies the following properties:

- (a) For fixed γ , $L_i(\gamma, t)$ (and $\tilde{L}_i(\lambda, t)$) is an increasing function in t
- (b) For fixed t , $\tilde{L}_i(\lambda, t)$ is a strictly convex function in λ
- (c) $\ell_i(\gamma, t) \in [0, \infty)$, and for a fixed t ,
 - If $\tilde{\psi}(t) = -1$, then $\ell_i(\gamma, t)$ is a decreasing function in γ (and $\tilde{\ell}_i(\lambda, t)$ is decreasing in λ)
 - If $\tilde{\psi}(t) = 1$, then $\ell_i(\gamma, t)$ is an increasing function in γ (and $\tilde{\ell}_i(\lambda, t)$ is increasing in λ)
- (d) If there exists $t_1, t_2 \leq t$ where $\tilde{\psi}_{i,t_1} = 1$ and $\tilde{\psi}_{i,t_2} = -1$, then $\tilde{L}_i(\lambda, t)$ has unique finite minimum as a function in λ . Also $L_i(\gamma, t)$ has the same minimum at $\gamma = e^\lambda$.
- (e) If γ^* is the true (bias) parameter, then for any $\gamma \neq \gamma^*$,

$$\mathbb{E}[\ell_i(\gamma, t) | \mathcal{H}_{t-1}] > \mathbb{E}[\ell_i(\gamma^*, t) | \mathcal{H}_{t-1}] \quad (37)$$

We show the proof of property (b) below; proofs of the other properties are omitted.

Proof.

$$\frac{d^2}{d\lambda^2} \tilde{\ell}_i(\lambda, t) = \frac{d^2}{d\lambda^2} \log(1 + e^{-\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))}) \quad (38)$$

$$= \frac{d}{d\lambda} \frac{-\tilde{\psi}_i(t) e^{-\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))}}{1 + e^{-\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))}} \quad (39)$$

$$= -\tilde{\psi}_i(t) \frac{d}{d\lambda} \frac{1}{1 + e^{\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))}} \quad (40)$$

$$= \tilde{\psi}_i(t)^2 \frac{e^{\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))}}{(1 + e^{\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))})^2} \quad (41)$$

$$= \frac{e^{\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))}}{(1 + e^{\tilde{\psi}_i(t)(\lambda + \nu_i(t-1))})^2} \quad (42)$$

$$> 0. \quad (43)$$

Thus $\tilde{\ell}_i(\lambda, t)$ is convex for all t , and so $\tilde{L}_i(\lambda, t) = \sum_{\tau=2}^t \tilde{\ell}_i(\lambda, \tau)$ is also convex. \square

V. LOG-LIKELIHOOD RATIOS AND MARTINGALES

To properly analyze the quantity (34), we need the following definitions. Unless otherwise stated, γ_1 is the true parameter the random data is generated from.

The *loss difference* is

$$Z(t) \triangleq Z(\gamma_1, \gamma_2, t) \quad (44)$$

$$z(t) \triangleq z(\gamma_1, \gamma_2, t) \triangleq \ell_i(\gamma_2, t) - \ell_i(\gamma_1, t). \quad (45)$$

The *predictable expected value* is

$$X(t) \triangleq X(\gamma_1, \gamma_2, t) \triangleq \sum_{\tau=2}^t \mathbb{E}[z(\tau) | \mathcal{H}_{\tau-1}] \quad (46)$$

$$x(t) \triangleq x(\gamma_1, \gamma_2, t) \triangleq \mathbb{E}[z(t) | \mathcal{H}_{t-1}]. \quad (47)$$

The *loss martingale* is

$$Y(t) \triangleq Y(\gamma_1, \gamma_2, t) \triangleq X(t) - Z(t) \quad (48)$$

$$y(t) \triangleq y(\gamma_1, \gamma_2, t) \triangleq x(t) - z(t). \quad (49)$$

The *predictable quadratic variation* is

$$W(t) \triangleq W(\gamma_1, \gamma_2, t) \triangleq \sum_{\tau=2}^t \text{Var}[z(\tau) | \mathcal{H}_{\tau-1}] \quad (50)$$

$$= \sum_{\tau=2}^t \text{Var}[y(\tau) | \mathcal{H}_{\tau-1}] \quad (51)$$

$$w(t) \triangleq w(\gamma_1, \gamma_2, t) \triangleq \text{Var}[z(t) | \mathcal{H}_{t-1}] \quad (52)$$

$$= \text{Var}[y(t) | \mathcal{H}_{t-1}]. \quad (53)$$

The quantity $W(t)$ is important for martingale concentration results. We also let $\lambda_1 = \log \gamma_1$ and $\lambda_2 = \log \gamma_2$.

We give some preliminary results about these processes (proofs are omitted).

Proposition 2. We have the following properties:

- (a) $Z(t)$ is a submartingale and $X(t)$ is strictly increasing
- (b) $Y(t)$ is a martingale
- (c) $W(t)$ is strictly increasing

Next we determine bounds on our quantities.

Lemma 1. If $\gamma_1 \neq \gamma_2$, then there is some $t_0 = t_0(\gamma_1, \gamma_2)$ and $c_0 = c_0(\gamma_1, \gamma_2) > 0$ such for all $t > t_0$

$$x(t) \geq c_0(\kappa/t). \quad (54)$$

Additionally, there are some constants k, t_1 (which depend on t_0, γ_1, γ_2) such that for all $t > t_1$,

$$X(t) > kc_0 \kappa \log(t). \quad (55)$$

Similarly, there exists a constant $c_1 = c_1(\gamma_1, \gamma_2) > 0$ and t_2 such that for all $t > t_2$

$$w(t) \leq c_1 x(t). \quad (56)$$

This also implies

$$W(t) \leq c_1 X(t). \quad (57)$$

Combining Lemma 1 and (36) gives that for $\gamma_1 \neq \gamma_2$,

$$\lim_{t \rightarrow \infty} \mathbb{E}[Z(\gamma_1, \gamma_2, t)] = \lim_{t \rightarrow \infty} X(\gamma_1, \gamma_2, t) = \infty. \quad (58)$$

Remark 1. The fact that $\mathbb{E}[Z(t)] \rightarrow \infty$ relies on the fact that $\mu_i(t) \in [\kappa/t, 1 - \kappa/t]$, as discussed in Section IV-A.

If instead, $\mu_i(t)$ scales as $1/t^2$, then the limit of $\mathbb{E}[Z(t)]$ would be finite. In such a scenario, randomness might make $Z(t)$ unreliable for distinguishing between γ_1 and γ_2 .

Changes to the model which may cause the condition $\mu_i(t) \in [\kappa/t, 1 - \kappa/t]$ to fail include putting higher weight on previously declared opinions or having the network add more agents at each time step t .

We want to show that the test $Z(t) > 0$ works to distinguish whether γ_1 or γ_2 is the true parameter. We do this by showing that if γ_1 is the true parameter, then almost surely $Z(t) \leq 0$ (i.e. the test fails) for only finitely many t . We show this by applying Freedman's inequality ([8] and [9] (Thm 1.6)) and Lemma 1 (proof omitted):

Theorem 1. Let γ^* be the true parameter. The likelihood ratio test $Z(t) = L_i(\gamma_2, t) - L_i(\gamma_1, t)$ is such that

$$Z(t) \begin{cases} > 0 & \text{if } \gamma^* = \gamma_1 \\ < 0 & \text{if } \gamma^* = \gamma_2 \end{cases} \quad (59)$$

for all but finitely many t .

VI. CONSISTENCY OF ESTIMATORS

The above concentration results show that the MLE (24) converges asymptotically to the true γ .

Theorem 2. For any agent i , almost surely,

$$\lim_{t \rightarrow \infty} \hat{\gamma}(t) = \gamma_i. \quad (60)$$

Proof. We take advantage of the alternative parameterization of $\lambda = \log \gamma$. Let the MLE for λ be

$$\hat{\lambda}(t) = \arg \min_{\lambda} \tilde{L}_i(\lambda, t) \quad (61)$$

We will show that $\hat{\lambda}(t)$ converges to the true parameter, which we call λ^* , so $\hat{\gamma}(t)$ converges to the true γ_i .

For any fixed $\epsilon > 0$, let $a = \lambda^* - \epsilon$ and $b = \lambda^* + \epsilon$. From Theorem 1, there exists some time t_a so that $\tilde{L}(a, t) > \tilde{L}(\lambda^*, t)$ for all $t > t_a$ and there exists some time t_b so that $\tilde{L}(b, t) > \tilde{L}(\lambda^*, t)$ for all $t > t_b$.

At all times $t > \max\{t_a, t_b\} \triangleq t(\epsilon)$, the value of $\tilde{L}(\lambda^*, t)$ is less than both $\tilde{L}(a, t)$ and $\tilde{L}(b, t)$. By Proposition 1(b), the function $\tilde{L}(\lambda, t)$ is convex in λ , and thus the minimum of $\tilde{L}(\lambda, t)$ at any $t > t(\epsilon)$ must be in $[a, b] = [\lambda^* - \epsilon, \lambda^* + \epsilon]$.

Thus, for every $\epsilon > 0$, we can always find a $t(\epsilon)$ where for all $t > t(\epsilon)$ we have that $\hat{\lambda}(t)$ is within ϵ of λ^* , and thus

$$\lim_{t \rightarrow \infty} \hat{\lambda}(t) = \lambda^* \quad (62)$$

completing the proof. \square

This also shows that the inherent belief estimator from Definition 4 almost surely converges to the correct result.

Theorem 3. Almost surely, if $\gamma_i \neq 1$, then

$$\lim_{t \rightarrow \infty} \hat{\phi}_i(t) = \phi_i \quad (63)$$

Proof. This is equivalent to

$$\lim_{t \rightarrow \infty} \left(\sum_{\tau=1}^{t-1} \mu_i(\tau) \right) - (t-1)\bar{\beta}_i(t) \begin{cases} < 0 & \text{if } \phi_i = 1 \\ > 0 & \text{if } \phi_i = 0 \end{cases} \quad (64)$$

The result follows from three facts:

- (i) letting $\lambda_i^* = \log(\gamma_i^*)$ and $\hat{\lambda}_i(t) = \arg \min_{\lambda} \tilde{L}_i(\lambda, t) = \log(\hat{\gamma}_i(t))$ be the maximum likelihood estimator of λ_i^* , then $\lim_{t \rightarrow \infty} \hat{\lambda}_i(t) = \lambda_i^*$;
- (ii) for any t , $\tilde{L}_i(\lambda, t)$ is strictly convex in λ ;
- (iii) $\left(\sum_{\tau=1}^{t-1} \mu_i(\tau) \right) - (t-1)\bar{\beta}_i(t) = \frac{\partial}{\partial \lambda} \tilde{L}_i(\lambda, t) \Big|_{\lambda=0}$.

Fact (i) follows directly from Theorem 2 and (ii) is Proposition 1(b). Fact (ii) also shows that

$$\hat{\lambda}_i(t) > 0 \iff \frac{\partial}{\partial \lambda} \tilde{L}_i(\lambda, t) \Big|_{\lambda=0} < 0; \quad (65)$$

and (assuming $\lambda_i^* \neq 0$), $\phi_i = 1 \iff \lambda_i^* > 0$. Thus facts (i) and (ii) show that (almost surely)

$$\phi_i = 1 \implies \hat{\lambda}_i(t) > 0 \text{ for all sufficiently large } t \quad (66)$$

$$\implies \frac{\partial}{\partial \lambda} \tilde{L}_i(\lambda, t) \Big|_{\lambda=0} < 0 \text{ for all sufficiently large } t \quad (67)$$

Thus, only fact (iii) remains to be shown.

Using $\tilde{L}_i(\lambda, t) = \sum_t \tilde{\ell}_i(\lambda, t)$ and

$$\tilde{\ell}_i(\lambda, t) = \log \left(1 + e^{-\tilde{\psi}_{i,t}(\lambda + \nu_i(t-1))} \right) \quad (68)$$

we evaluate the derivative at $\lambda = 0$, we get

$$\frac{\partial}{\partial \lambda} \tilde{\ell}_i(\lambda, t) \Big|_{\lambda=0} = \frac{-\tilde{\psi}_{i,t}}{e^{\tilde{\psi}_{i,t}} + 1} \quad (69)$$

$$= \begin{cases} \frac{1}{\frac{1-\mu_i(t-1)}{\mu_i(t-1)} + 1} & \text{if } \tilde{\psi}_{i,t} = -1 \\ \frac{-1}{\frac{\mu_i(t-1)}{1-\mu_i(t-1)} + 1} & \text{if } \tilde{\psi}_{i,t} = 1 \end{cases} \quad (70)$$

$$= \begin{cases} \mu_i(t-1) & \text{if } \tilde{\psi}_{i,t} = -1 \\ \mu_i(t-1) - 1 & \text{if } \tilde{\psi}_{i,t} = 1 \end{cases} \quad (71)$$

$$= \mu_i(t-1) - \mathbb{I}\{\psi_{i,t} = 1\}. \quad (72)$$

And thus the derivative of the entire negative log-likelihood evaluated at 0 is given by

$$\frac{\partial}{\partial \lambda} \tilde{L}_i(\lambda, t) \Big|_{\lambda=0} = \sum_{\tau=2}^t \mu_i(\tau-1) - \mathbb{I}\{\psi_{i,t} = 1\} \quad (73)$$

$$= \left(\sum_{\tau=1}^{t-1} \mu_i(\tau) \right) - (t-1)\bar{\beta}_i(t). \quad (74)$$

This shows (iii) and completes the proof. \square

REFERENCES

- [1] Ali Jadbabaie, Anuran Makur, Elchanan Mossel, and Rabi Salhab, "Inference in opinion dynamics under social pressure," *IEEE Transactions on Automatic Control*, vol. 68, no. 6, pp. 3377–3392, 2023.
- [2] Camilla Ancona, Francesco Lo Iudice, Franco Garofalo, and Pietro De Lellis, "A model-based opinion dynamics approach to tackle vaccine hesitancy," *Scientific Reports*, vol. 12, no. 1, pp. 11835, 2022.
- [3] Damon Centola, Robb Willer, and Michael Macy, "The emperor's dilemma: A computational model of self-enforcing norms," *American Journal of Sociology*, vol. 110, no. 4, pp. 1009–1040, 2005.
- [4] Jennifer Tang, Aviv Adler, Amir Ajorlou, and Ali Jadbabaie, "Stochastic opinion dynamics under social pressure in arbitrary networks," 2023.
- [5] Morris H. DeGroot, "Reaching a consensus," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 118–121, 1974.
- [6] Noah E Friedkin and Eugene C Johnsen, "Social influence and opinions," *Journal of Mathematical Sociology*, vol. 15, no. 3-4, pp. 193–206, 1990.
- [7] Mengbin Ye, Yuzhen Qin, Alain Govaert, Brian D.O. Anderson, and Ming Cao, "An influence network model to study discrepancies in expressed and private opinions," *Automatica*, vol. 107, pp. 371–381, 2019.
- [8] Joel Tropp, "Freedman's inequality for matrix martingales," *Electronic Communications in Probability*, vol. 16, 01 2011.
- [9] David A. Freedman, "On Tail Probabilities for Martingales," *The Annals of Probability*, vol. 3, no. 1, pp. 100 – 118, 1975.