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 to their neighbors. Both the agents' inherent opinions and bias parameters are hidden from their neighbors and outside observers. Then the agents communicate their *declared* opinions to their neighbors at discrete time steps: at each step $t = 1, 2, \ldots$, 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.

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.

A. Background Literature

We refer the reader to [1] and [4] for in-depth discussion of prior work. 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, on a connected aperiodic graph, 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 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 study the convergence properties of the interacting Pólya urn model introduced in [1] on arbitrary undirected networks, finding 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 problem of estimating the agents' parameters in the general setting presented in [4] (which analyzed the convergence properties of the interacting Pólya urn model); we also remove the restriction that the social network's graph needs to be undirected:

1) the social network is an arbitrary weighted (and connected) graph, possibly directed and with self-loops;
2) the agents can have heterogeneous bias parameters, indicating different levels of resistance to social pressure or certainty in their inherent beliefs.

Both the agents' inherent beliefs and bias parameters are unknown and must be inferred from observing the behavior of the network. This greatly increases the applicability of the model, as real-life social networks have a variety of different structures and people have varied reactions 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 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 graph $G = (V, E)$ (possibly directed and including self-loops) be a network of n agents (corresponding to the vertices) labeled $i = 1, 2, \ldots, n$. 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 matrix of these weights as $\boldsymbol{A} \in \mathbb{R}^{n \times n}$. We denote the weighted degree of vertex i as $\mathsf{deg}(i) = \sum_j a_{i,j}$. We assume that G is connected.

B. Inherent Beliefs and Declared Opinions

We define the interacting Pólya urn model of opinion dynamics under social pressure by defining the key parameters governing the behavior of the agents and their relationship to each other. The basic concept of the model is: each agent i declares at each step t an opinion $\psi_{i,t} \in \{0, 1\}$; in expectation, agent i imitates the (weighted) average opinion they have observed declared by their neighbors (including themselves via self-loops), but biased by an internal *bias parameter* $\gamma_i \geq 0$ towards their *inherent belief* $\phi_i \in \{0, 1\}$. The value of γ_i denotes how an observation of a neighbor declaring '1' is weighted compared to the same neighbor declaring '0', e.g. $\gamma_i = 2$ denotes that each observation of neighbor j declaring '1' counts twice as much as when they declare '0', while $\gamma_i = 1/2$ denotes the converse.[1]

Thus, the inherent belief of agent is i is the opinion they are biased toward:

$$\phi_i = \begin{cases} 1 & \text{if } \gamma_i > 1 \\ 0 & \text{if } \gamma_i < 1 \end{cases} \tag{1}$$

If $\gamma_i = 1$ then the agent is *unbiased* and is considered to not have an inherent belief; since the goal is to estimate the inherent beliefs of the agents, for the remainder of this work we assume that the agent under consideration is not unbiased.

To formally state the model, let $b_i^0, b_i^1 > 0$ be the *initialization* of agent i's declared opinions, where $b_i^0 + b_i^1 = 1$. Then we define the *declared proportion* of 0's (or 1's) declared by agent i up to time t as:

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

$$\beta_i^1(t) = \frac{b_i^1}{t} + \frac{1}{t} \sum_{\tau=2}^{t} \psi_{i,\tau} \,. \tag{3}$$

We note that by definition $\beta_i^0(t) + \beta_i^1(t) = 1$; thus to specify these values it is sufficient to specify just $\beta_i(t) \triangleq \beta_i^1(t)$ (i.e.

[1] The 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 $1/\gamma$.

the proportion of agent i's declared opinions up to time t that were 1's).

For any agent i we also denote the total (weighted) proportion of opinions 0 and 1 she has observed by time t from her neighbors (including herself via self loop) as

$$\mu_i^0(t) = \frac{1}{\mathsf{deg}(i)} \sum_{j=1}^{n} a_{i,j}\beta_j^0(t) \tag{4}$$

$$\text{and } \mu_i^1(t) = \frac{1}{\mathsf{deg}(i)} \sum_{j=1}^{n} a_{i,j}\beta_j^1(t)\,; \tag{5}$$

as before, $\mu_i^0(t) + \mu_i^1(t) = 1$ by definition so it suffices to specify $\mu_i(t) \triangleq \mu_i^1(t)$. This corresponds to the social environment that agent i finds herself in at time t.

Then, at time $t+1$, each agent i will (independently) declare an opinion $\psi_{i,t+1}$ where $\psi_{i,t+1} = 1$ with probability $p_i(t) = f(\mu_i(t), \gamma_i)$ (and $\psi_{i,t+1} = 0$ otherwise) where

$$f(\mu_i(t), \gamma_i) \triangleq \frac{\gamma_i \mu_i(t)}{\gamma_i \mu_i(t) + (1 - \mu_i(t))}\,. \tag{6}$$

The values of $\beta_i(t+1)$ and $\mu_i(t+1)$ for all i are updated according to the declared opinions at time t and the values of $\beta_i(t)$ and $\mu_i(t)$. Since $\mu_i(t) = \mu_i^1(t)$ and $1 - \mu_i(t) = \mu_i^0(t)$, this corresponds to weighting each observation of opinion '1' as γ_i times an equivalent observation of opinion '0'.

Finally, we denote the *history* of the network up to time t (which denotes all declared opinions, including initializations, and therefore can be used to compute all $\beta_i(\tau), \mu_i(\tau)$ for $\tau \le t$) as $\mathcal{H}_t$. We also denote the vectors of $\beta_i(t), \mu_i(t)$ over agents i as $\boldsymbol{\beta}(t), \boldsymbol{\mu}(t)$.

In [4], it was shown that these dynamics on undirected graphs must approach some equilibrium point satisfying

$$\beta_i = f(\mu_i, \gamma_i) \quad \text{for all } i \tag{7}$$

as $t \to \infty$ (with probability 1). In this work, we consider the following estimation problem (which was considered in [1] for a more restricted model on complete graphs): given the history $\mathcal{H}_t$ up to time t, can we estimate γ_i, ϕ_i for all agents i in the limit as $t \to \infty$?[2]

C. Consensus

An important term for this work is *consensus*, which needs to be defined appropriately for our stochastic system. This notion was defined and studied in [4].

Definition 1. Consensus *is approached if*

$$\boldsymbol{\beta}(t) \to \mathbf{1} \text{ or } \boldsymbol{\beta}(t) \to \mathbf{0} \quad as \quad t \to \infty\,. \tag{8}$$

Since $\beta_i(t)$ represents the fraction of agent i's declared opinions which are 1, consensus is approached when this ratio goes to 0 or 1. Let $\boldsymbol{J_1} = \boldsymbol{\Gamma}^{-1}\boldsymbol{W}$ and $\boldsymbol{J_0} = \boldsymbol{\Gamma}\boldsymbol{W}$, and let $\lambda_{\max}(\cdot)$ denote maximum eigenvalue; in [4] it was shown that consensus $\boldsymbol{\beta}(t) \to \mathbf{1}$ occurs when $\lambda_{\max}(\boldsymbol{J_1}) \le 1$ and $\boldsymbol{\beta}(t) \to \mathbf{0}$ occurs when $\lambda_{\max}(\boldsymbol{J_0}) \le 1$.

[2]While we assume for simplicity that b_i^0, b_i^1 are known to the estimator, this is not necessary as these terms become negligible in the limit as $t \to \infty$.

Consensus is important for the parameter estimation problem we consider in this work because it represents a major obstacle to solving the estimation problem, as it is an uninformative equilibrium (at consensus, each agent repeats the same opinion regardless of their internal parameters).

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 fraction of declared opinions of all agents throughout time, and showed that this estimator may not converge to the inherent belief of all agents 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 maximum likelihood estimation (MLE) 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 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 (7) to estimate the bias parameter γ_i and inherent belief ϕ_i with

$$\widehat{\gamma}_i^{eq}(t) = \frac{\beta_i(t)}{1 - \beta_i(t)} \frac{1 - \mu_i(t)}{\mu_i(t)} \tag{9}$$

$$\widehat{\phi}_i^{eq}(t) = \mathbb{I}\{\beta_i(t) < \mu_i(t)\} \tag{10}$$

These estimators are asymptotically consistent, i.e.

$$\lim_{t\to\infty} \widehat{\gamma}_i^{eq}(t) = \gamma_i \text{ and } \lim_{t\to\infty} \widehat{\phi}_i^{eq}(t) = \phi_i \tag{11}$$

when the dynamics converge to an interior equilibrium point. However, plugging the equilibrium values into (9) is not well-defined if $\beta_i(t)$ and $\mu_i(t)$ both converge to either 0 or 1 for all i, 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 value of

$$p_i(t) = \mathbb{P}\left[\psi_{i,t} = 1 | \mathcal{H}_{t-1}\right] = f(\mu_i(t-1), \gamma_i)\,. \tag{12}$$

Note that in general $\mathbb{P}\left[\psi_{i,t} = 1\right]$ is a random variable dependent on $\mathcal{H}_{t-1}$, while $\mathbb{P}\left[\psi_{i,t} = 1 | \mathcal{H}_{t-1}\right]$ is constant. The sequence $\mathcal{H}_0 \subseteq \mathcal{H}_1 \subseteq \ldots$ is also a filtration on which we can base stochastic processes.

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 $t > 1$ and parameter γ is*

$$\ell_i(\gamma, t) \triangleq -\Big(\mathbb{I}\{\psi_{i,t} = 1\} \log(f(\mu_i(t-1), \gamma)) + \mathbb{I}\{\psi_{i,t} = 0\} \log(1 - f(\mu_i(t-1), \gamma))\Big) \quad (13)$$

The negative log-likelihood for a given agent i at time t and parameter $\gamma \in (0, \infty)$ is

$$L_i(\gamma, t) \triangleq \sum_{\tau=2}^{t} \ell_i(\gamma, \tau)\,. \quad (14)$$

Note that γ_i is the actual bias parameter of agent i, whereas γ represents a proposed value whose loss we are measuring. 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*

$$\widehat{\gamma}_i(t) \triangleq \arg\min_{\gamma} L_i(\gamma, t) \quad (15)$$

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 $\text{sign}(\log(\widehat{\gamma}_i(t)))$.

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 a similar quantity to $\beta_i(t)$ except that the arbitrary initial conditions are not included. (If t is large, then the difference between $\beta_i(t)$ and $\bar{\beta}_i(t)$ is negligible.)

Definition 4 (Inherent Belief Estimator)**.** *Let*

$$\widehat{\phi}_i(t) = \frac{1}{2}\text{sign}\left((t-1)\bar{\beta}_i(t) - \left(\sum_{\tau=1}^{t-1} \mu_i(\tau)\right)\right) + \frac{1}{2}\,. \quad (16)$$

Multiplying by $1/2$ and adding $1/2$ maps the output of $\text{sign}(\cdot)$ to 0 and 1. Fundamentally, this estimator requires only comparing

$$\bar{\beta}_i(t) > \frac{1}{t-1}\sum_{\tau=1}^{t-1} \mu_i(\tau)\,. \quad (17)$$

Note that $\widehat{\phi}_i(t)$ 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 (10) 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 $\mu_i(t)$

Lemma 1. *Letting $\kappa \triangleq \min_i(\min(b_i^0, b_i^1)) > 0$, for any agent i and time t,*

$$\mu_i(t) \in \left[\frac{\kappa}{t}, 1 - \frac{\kappa}{t}\right]. \quad (18)$$

Proof. This follows since by definition $b_i^0, b_i^1 \geq \kappa$ for any i; thus by equations (2), (3) we know that $\beta_i^0(t), \beta_i^1(t) \geq \kappa/t$ so $\beta_i(t) = \beta_i^1(t) = 1 - \beta_i^0(t)$ satisfies $\beta_i(t) \in [\frac{\kappa}{t}, 1 - \frac{\kappa}{t}]$. But each $\mu_i(t)$ is a weighted average of $\beta_j(t)$, and hence $\mu_i(t) \in [\frac{\kappa}{t}, 1 - \frac{\kappa}{t}]$ for all i, t. □

Note that this means that any agent i will (almost surely) declare both '0' and '1' infinitely many times, even if the network approaches consensus, because either opinion has probability $\geq \Theta(1/t)$ at step t (and $\sum_t 1/t = \infty$).

B. Negative Log-Likelihood Properties

We analyze in depth the MLE 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.

Since $f(\mu_i(t), \gamma)$ is still the probability of $\tilde{\psi}_{i,t} = 1$,

$$\ell_i(\gamma, t) = -\log\left(\frac{1}{1 + e^{-\tilde{\psi}_{i,t}\log\left(\gamma\frac{\mu_i(t-1)}{1-\mu_i(t-1)}\right)}}\right) \quad (19)$$

$$= \log\left(1 + e^{-\tilde{\psi}_{i,t}\log\left(\gamma\frac{\mu_i(t-1)}{1-\mu_i(t-1)}\right)}\right). \quad (20)$$

We reparameterize γ and $\mu_i(t)$ as follows:

$$\chi \triangleq \log\gamma \quad \text{and} \quad \nu_i(t) \triangleq \log\frac{\mu_i(t)}{1-\mu_i(t)}\,. \quad (21)$$

Using χ symmetrizes the bias parameter across $\mathbb{R}$ (so $\chi = 0$ represents an unbiased agent).

We thus define some quantities which take $\chi = \log\gamma$ as the argument instead of γ and use them where convenient:

$$\tilde{\ell}_i(\chi, t) \triangleq \ell_i(\gamma, t) \text{ and } \tilde{L}_i(\chi, t) \triangleq L_i(\gamma, t)\,. \quad (22)$$

For this section to Section V we will fix an agent i and then use γ^1 and γ^2 to represent any two possible choices for γ_i. We then show that if we know that one of these is the true value of γ_i, in the limit it is almost surely possible to determine which one (Theorem 1); this result will then be used to show that $\lim_{t\to\infty}\widehat{\gamma}_i = \gamma_i$ almost surely (Theorem 2). Define

$$Z(t) = Z(\gamma^1, \gamma^2, t) \triangleq L_i(\gamma^2, t) - L_i(\gamma^1, t)\,. \quad (23)$$

If $Z(t)$ is positive, intuitively, γ^1 fits the observed behavior better than γ^2, so we expect γ^1 to be the true parameter. Indeed, if γ^1 is the true parameter, then

$$\mathbb{E}[Z(t)|\mathcal{H}_{t-1}] = \sum_{\tau=2}^{t} \mathbb{E}\Bigg[\mathbb{I}\{\psi_{i,\tau}=1\}\log\frac{f(\mu_i(\tau-1),\gamma^1)}{f(\mu_i(\tau-1),\gamma^2)} + \mathbb{I}\{\psi_{i,\tau}=0\}\log\frac{1-f(\mu_i(\tau-1),\gamma^1)}{1-f(\mu_i(\tau-1),\gamma^2)}\Big|\mathcal{H}_{\tau-1}\Bigg] \quad (24)$$

$$= \sum_{\tau=2}^{t} D_{\text{KL}}(f(\mu_i(\tau-1),\gamma^1)\|f(\mu_i(\tau-1),\gamma^2)) \quad (25)$$

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 $\tilde{L}_i(\chi,t)$) is an increasing function in t

(b) For fixed t, $\tilde{L}_i(\chi,t)$ is a strictly convex function in χ

(c) $\ell_i(\gamma,t)\in[0,\infty)$, and for a fixed t,
- *If $\tilde{\psi}_{i,t}=-1$, then $\ell_i(\gamma,t)$ is a decreasing function in γ (and $\tilde{\ell}_i(\chi,t)$ is decreasing in χ)*
- *If $\tilde{\psi}_{i,t}=1$, then $\ell_i(\gamma,t)$ is an increasing function in γ (and $\tilde{\ell}_i(\chi,t)$ is increasing in χ)*

(d) If there exists $t_1,t_2\le t$ where $\tilde{\psi}_{i,t_1}=1$ and $\tilde{\psi}_{i,t_2}=-1$, then $\tilde{L}_i(\chi,t)$ has unique finite minimum as a function in χ. Also $L_i(\gamma,t)$ has the same minimum at $\gamma=e^{\chi}$.

(e) For any $\gamma\neq\gamma_i$,

$$\mathbb{E}[\ell_i(\gamma,t)|\mathcal{H}_{t-1}] > \mathbb{E}[\ell_i(\gamma_i,t)|\mathcal{H}_{t-1}] \quad (26)$$

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

Proof.

$$\frac{d^2}{d\chi^2}\tilde{\ell}_i(\chi,t) = \frac{d^2}{d\chi^2}\log\left(1+e^{-\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))}\right) \quad (27)$$

$$= \frac{d}{d\chi}\frac{-\tilde{\psi}_{i,t}e^{-\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))}}{1+e^{-\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))}} \quad (28)$$

$$= -\tilde{\psi}_{i,t}\frac{d}{d\chi}\frac{1}{1+e^{\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))}} \quad (29)$$

$$= \tilde{\psi}_{i,t}^2\frac{e^{\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))}}{(1+e^{\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))})^2} \quad (30)$$

$$= \frac{e^{\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))}}{(1+e^{\tilde{\psi}_{i,t}(\chi+\nu_i(t-1))})^2} \quad (31)$$

$$> 0\,. \quad (32)$$

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

V. Log-Likelihood Ratios and Martingales

To properly analyze the quantity (23), we need the following definitions. Unless otherwise stated, γ^1 is the true parameter from which the random data is generated. The *loss difference* is

$$Z(t) \triangleq Z(\gamma^1,\gamma^2,t) \quad (33)$$

$$z(t) \triangleq z(\gamma^1,\gamma^2,t) \triangleq \ell_i(\gamma^2,t)-\ell_i(\gamma^1,t)\,. \quad (34)$$

The *predictable expected value* is

$$X(t) \triangleq X(\gamma^1,\gamma^2,t) \triangleq \sum_{\tau=2}^{t}\mathbb{E}[z(\tau)|\mathcal{H}_{\tau-1}] \quad (35)$$

$$x(t) \triangleq x(\gamma^1,\gamma^2,t) \triangleq \mathbb{E}[z(t)|\mathcal{H}_{t-1}]\,. \quad (36)$$

The *loss martingale* is

$$Y(t) \triangleq Y(\gamma^1,\gamma^2,t) \triangleq X(t)-Z(t) \quad (37)$$

$$y(t) \triangleq y(\gamma^1,\gamma^2,t) \triangleq x(t)-z(t)\,. \quad (38)$$

The *predictable quadratic variation* is

$$W(t) \triangleq W(\gamma^1,\gamma^2,t) \triangleq \sum_{\tau=2}^{t}\mathrm{Var}[z(\tau)|\mathcal{H}_{\tau-1}] \quad (39)$$

$$= \sum_{\tau=2}^{t}\mathrm{Var}[y(\tau)|\mathcal{H}_{\tau-1}] \quad (40)$$

$$w(t) \triangleq w(\gamma^1,\gamma^2,t) \triangleq \mathrm{Var}[z(t)|\mathcal{H}_{t-1}] \quad (41)$$

$$= \mathrm{Var}[y(t)|\mathcal{H}_{t-1}]\,. \quad (42)$$

We also let $\chi^1=\log\gamma^1$ and $\chi^2=\log\gamma^2$. We give some preliminary results about these processes.

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 2. *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 $t>t_0$*

$$x(t) \ge c_0(\kappa/t)\,. \quad (43)$$

Additionally, there are some constants k, t_1 (which depend on t_0,γ^1,γ^2) such that for all $t>t_1$,

$$X(t) > kc_0\kappa\log(t)\,. \quad (44)$$

Similarly, there exists a constant $c_1=c_1(\gamma^1,\gamma^2)>0$ and t_2 such that for all $t>t_2$

$$w(t) \le c_1 x(t)\,. \quad (45)$$

This also implies

$$W(t) \le c_1 X(t)\,. \quad (46)$$

Combining Lemma 2 and (25) gives that for $\gamma^1\neq\gamma^2$,

$$\lim_{t\to\infty}\mathbb{E}[Z(\gamma^1,\gamma^2,t)] = \lim_{t\to\infty}X(\gamma^1,\gamma^2,t)=\infty\,. \quad (47)$$

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

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 2 (proof omitted):

Theorem 1. *If $\gamma_i \in \{\gamma^1, \gamma^2\}$, then the likelihood ratio test $Z(t) = L_i(\gamma^2, t) - L_i(\gamma^1, t)$ is such that*

$$Z(t) \begin{cases} > 0 \text{ if } \gamma_i = \gamma^1 \\ < 0 \text{ if } \gamma_i = \gamma^2 \end{cases} \tag{48}$$

for all but finitely many t.

VI. Consistency of Estimators

The above concentration results show that the MLE (15) $\widehat{\gamma}_i(t)$ converges asymptotically to the true γ_i.

Theorem 2. *For any agent i, almost surely,*

$$\lim_{t\to\infty} \widehat{\gamma}_i(t) = \gamma_i \,. \tag{49}$$

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

$$\widehat{\chi}_i(t) = \arg\min_{\chi} \tilde{L}_i(\chi, t) \,. \tag{50}$$

We will show that $\widehat{\chi}_i(t)$ converges to the true parameter, which we call χ_i, so $\widehat{\gamma}_i(t)$ converges to the true γ_i.

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

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

Thus, for every $\epsilon > 0$, we can always find a $t(\epsilon)$ where for all $t > t(\epsilon)$ we have that $\widehat{\chi}_i(t)$ is within ϵ of χ_i, and thus $\lim_{t\to\infty} \widehat{\chi}_i(t) = \chi_i$ completing the proof. □

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*

$$\lim_{t\to\infty} \widehat{\phi}_i(t) = \phi_i \tag{51}$$

Proof. This is equivalent to

$$\lim_{t\to\infty} \left(\sum_{\tau=1}^{t-1} \mu_i(\tau) \right) - (t-1)\bar{\beta}_i(t) \begin{cases} < 0 & \text{if } \phi_i = 1 \\ > 0 & \text{if } \phi_i = 0 \end{cases} \tag{52}$$

The result follows from three facts:

(i) letting $\chi_i = \log(\gamma_i)$ and $\widehat{\chi}_i(t) = \arg\min_\chi \tilde{L}_i(\chi, t) = \log(\widehat{\gamma}_i(t))$ be the maximum likelihood estimator of χ_i, then $\lim_{t\to\infty} \widehat{\chi}_i(t) = \chi_i$;

(ii) for any t, $\tilde{L}_i(\chi, t)$ is strictly convex in χ;

(iii) $\left(\sum_{\tau=1}^{t-1} \mu_i(\tau) \right) - (t-1)\bar{\beta}_i(t) = \frac{\partial}{\partial \chi} \tilde{L}_i(\chi, t)\Big|_{\chi=0}$.

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

$$\widehat{\chi}_i(t) > 0 \iff \frac{\partial}{\partial \chi} \tilde{L}_i(\chi, t)\Big|_{\chi=0} < 0 \,; \tag{53}$$

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

$$\phi_i = 1 \implies \widehat{\chi}_i(t) > 0 \text{ for all sufficiently large } t \tag{54}$$

$$\implies \frac{\partial}{\partial \chi} \tilde{L}_i(\chi, t)\Big|_{\chi=0} < 0 \text{ for all sufficiently large } t \tag{55}$$

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

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

$$\tilde{\ell}_i(\chi, t) = \log\left(1 + e^{-\tilde{\psi}_{i,t}(\chi + \nu_i(t-1))}\right) \tag{56}$$

at $\chi = 0$, we get

$$\frac{\partial}{\partial \chi} \tilde{\ell}_i(\chi, t)\Big|_{\chi=0} = \frac{-\tilde{\psi}_{i,t}}{e^{\tilde{\psi}_{i,t}\nu_i(t-1)} + 1} \tag{57}$$

$$= \begin{cases} \frac{1}{\frac{1-\mu_i(t-1)}{\mu_i(t-1)} + 1} & \text{if } \tilde{\psi}_{i,t} = -1 \\ \frac{-1}{\frac{\mu_i(t-1)}{1-\mu_i(t-1)} + 1} & \text{if } \tilde{\psi}_{i,t} = 1 \end{cases} \tag{58}$$

$$= \begin{cases} \mu_i(t-1) & \text{if } \tilde{\psi}_{i,t} = -1 \\ \mu_i(t-1) - 1 & \text{if } \tilde{\psi}_{i,t} = 1 \end{cases} \tag{59}$$

$$= \mu_i(t-1) - \mathbb{I}\{\psi_{i,t} = 1\} \,. \tag{60}$$

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

$$\frac{\partial}{\partial \chi} \tilde{L}_i(\chi, t)\Big|_{\chi=0} = \sum_{\tau=2}^{t} \mu_i(\tau - 1) - \mathbb{I}\{\psi_{i,t} = 1\} \tag{61}$$

$$= \left(\sum_{\tau=1}^{t-1} \mu_i(\tau) \right) - (t-1)\bar{\beta}_i(t) \,. \tag{62}$$

This shows (iii) and completes the proof. □

References

[1] Ali Jadbabaie, Anuran Makur, Elchanan Mossel, and Rabih 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," in *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2023, pp. 1360–1366.

[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.