HOW COLLABORATIVE THEORY FORMATION REDUCES EPISTEMIC UNCERTAINTY

DAVID KINNEY

1. Introduction

The goal of this document is to present a statistical model of collaborative theory formation. We also present simulation results that show the value of collaboration within this model. Special emphasis is placed on conditions wherein individuals are more adept at estimating unknown parameters when those parameters are not commonly estimated other members of the group. The simulation results show that under these conditions, *if* we add individuals to a group in a way that deliberately searches for people who measure uncommon parameters, then larger groups tend to see reduced epistemic uncertainty with respect to their predictions. We follow Hüllermeier and Waegeman (2021) and Mobiny et al. (2021) in defining epistemic uncertainty in terms of the expected Kullback-Leibler divergence between an observable outcome and the hypothesis used by a group aiming to predict that outcome.

2. The Model

Our toy model of a generic data-generating process is as follows. Let $\boldsymbol{\theta}$ be an unobserved parameter vector such that all entries $\theta_j \sim \mathcal{N}(0,1)$. Let $\mathbf{x}$ be an observed parameter vector such that all entries $x_j \sim \mathcal{N}(0,1)$. Both $\boldsymbol{\theta}$ and $\mathbf{x}$ are assumed to have M entries. The outcome y is the linear combination $y = \langle \boldsymbol{\theta}, \mathbf{x} \rangle$. Thus, the probability density function over possible values of y is given by:

$$p(y|\mathbf{x}) = f(y; 0, \sum_{j=1}^{M} \sqrt{x_j^2}), \tag{2.1}$$

where $f(y; \mu, \sigma)$ is the probability density function for the normal distribution with mean μ and standard deviation σ.

Next, we suppose that a set $\mathcal{I}$ of N individuals are trying to predict the value of y for a given observed vector of known parameters $\mathbf{x}$. They do this by *measuring* a subset of the unknown parameters. Precisely, for some $Q < M$, each individual samples Q integers in the interval $[1, M]$, to generate an **estimation set** E_i, and then measures the parameter θ_j for all $j \in E_i$. Each individual samples from $[1, M]$ without replacement, where the initial probability of of sampling each $j \in [1, M]$ is equal to $\frac{|\theta_j|}{\sum_{l=1}^{M} |\theta_l|}$. Thus, individuals are more likely to estimate more important parameters.

How do individuals measure the values of unknown parameters? An idea that we explore here is that, in keeping with the basic tenets of **standpoint epistemology**, individuals that measure parameters that are not measured by many individuals in the group make more accurate measurements than individuals who measure more commonly measured parameters. We formalize this using

a **standpoint function** γ, which is defined for an integer $j \in [1, M]$:

$$\gamma(j) := \frac{|\{i \in \mathcal{I} : j \in E_i\}|}{N} + \epsilon, \text{ where } 0 < \epsilon << 1. \tag{2.2}$$

The lower the value of $\gamma(j)$, the less common it is for any individual to estimate the value of the parameter θ_j. We then stipulate that each individual i measures the value of an unknown parameter θ_j, where $j \in E_i$, by generating an estimate $\hat{\theta}_{ij} \sim \mathcal{N}(\theta_j, \gamma(j))$. That is, the measured estimate $\hat{\theta}_{ij}$ is sampled from a normal distribution centered around the true parameter value θ_j, with a standard deviation equal to the value of the standpoint function for that parameter. Thus, each individual's estimate is more likely to be accurate when fewer other group members estimate that parameter. This represents the degree to which individuals with unique or uncommon perspectives are assumed to be capable of more accurate inference with respect to the objects of those uncommon perspectives. When an individual does not measure a parameter θ_j, they estimate its value as 0. This yields a matrix $\hat{\boldsymbol{\theta}}$ in which each entry $\hat{\theta}_{ij}$ is the individual i's estimate of the value of of the j'th parameter.

We then assume that the group uses their estimates $\hat{\boldsymbol{\theta}}$, along with the observed parameters $\mathbf{x}$, to make a prediction $\hat{y}$ of the value of y, using the formula:

$$\hat{y} = \sum_{j=1}^{M} \sum_{i=1}^{N} \frac{1}{N} \hat{\theta}_{ij} x_j. \tag{2.3}$$

That is, the prediction $\hat{y}$ is the linear combination of the known parameter vector $\mathbf{x}$ and the group's mean estimate of the unknown parameter vector $\boldsymbol{\theta}$. For our part, we are not directly interested in the accuracy of this prediction, but rather in the extent to which, in general, the actual outcome y of the data-generating process is informative as to the group parameter estimate $\hat{\boldsymbol{\theta}}$, given the known parameter vector $\mathbf{x}$. That is, letting Y be a random variable with range $\mathbb{R}$, letting $\hat{\Theta}$ be matrix-valued random variable with range $\mathbb{R}^{N \times M}$, and letting X be a vector-valued random variable with range $\mathbb{R}^M$, we follow Hüllermeier and Waegeman (2021) and Mobiny et al. (2021) in being interested in the value of the conditional mutual information

$$MI(Y; \hat{\Theta}|X) = \int_{\mathbb{R}^M} \Big(\int_{\mathbb{R}^{N \times M}} \int_{\mathbb{R}} p(y, \hat{\boldsymbol{\theta}}|\mathbf{x}) \log \frac{p(y, \hat{\boldsymbol{\theta}}|\mathbf{x})}{p(y|\mathbf{x}) p(\hat{\boldsymbol{\theta}}|\mathbf{x})} \mathrm{d}y \mathrm{d}\hat{\boldsymbol{\theta}} \Big) p(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{2.4}$$

or

$$MI(Y; \hat{\Theta}|X) = \int_{\mathbb{R}^M} \Big(\int_{\mathbb{R}^{N \times M}} \int_{\mathbb{R}} p(\hat{\boldsymbol{\theta}}|\mathbf{x}) p(y|\hat{\boldsymbol{\theta}}, \mathbf{x}) \log \frac{p(y|\hat{\boldsymbol{\theta}}, \mathbf{x})}{p(y|\mathbf{x})} \mathrm{d}y \mathrm{d}\hat{\boldsymbol{\theta}} \Big) p(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{2.5}$$

This is equivalent to the expected Kullback-Leibler divergence:

$$\mathbb{E}_{p(\hat{\Theta}, X)} D_{\mathrm{KL}}[p(Y|\hat{\boldsymbol{\theta}}, \mathbf{x}) || p(Y|\mathbf{x})]. \tag{2.6}$$

The greater this expectation, the more epistemic uncertainty there is with respect to the group's prediction.

We have already seen how the distribution $p(Y|\mathbf{x})$ can be written as the probability density function of a normal distribution. In our model, the implied definition of $p(Y|\hat{\boldsymbol{\theta}}, \mathbf{x})$ is

$$p(y|\hat{\boldsymbol{\theta}}, \mathbf{x}) = f(y; \hat{y}, \sum_{j=1}^{M} \sqrt{\gamma(j)^2 x_j^2}). \tag{2.7}$$

We are interested in how the expectation in Eq. 2.6 behaves as N increases (i.e., as we add more individuals to the group). However, this is difficult to evaluate for several reasons. Most saliently, the closed-form expression for $p(\hat{\Theta})$ is given by:

$$p(\hat{\theta}) = \prod_{i=1}^{N} g(\hat{\boldsymbol{\theta}}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \left[\frac{\theta_{E_i[1]}}{\sum_{j=1}^{M} \theta_j} \prod_{l=2}^{Q} \frac{\theta_{E_i[l]}}{\sum_{j=1}^{M} \theta_j - \sum_{t=1}^{l-1} \theta_{E_i[t]}} \right], \tag{2.8}$$

where:

- $g(\hat{\boldsymbol{\theta}}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution,
- $\boldsymbol{\mu}$ is a vector such that $\mu_j = \theta_j$ if $j \in E_i$ and 0 otherwise,
- $\boldsymbol{\Sigma}$ is a diagonal matrix such that $\sigma_{jj} = \gamma(j)$ if $j \in E_i$ and 1 otherwise, and
- each individual i's estimation set E_i is indexed $\{E_i[1], \ldots, E_i[Q]\}$ in the order that the elements are sampled from the integers in $[1, M]$.

It is far from obvious how to evaluate the behavior of the expected value of the Kullback-Leibler divergence $D_{\mathrm{KL}}[f(Y; \hat{y}, \sum_{j=1}^{M} \sqrt{\gamma(j)^2 x_j^2})) || f(Y; 0, \sum_{j=1}^{M} \sqrt{x_j^2})]$ under this distribution for a single vector of known parameters $\mathbf{x}$, let alone for all possible vectors $\mathbf{x} \in \mathbb{R}^M$.

3. Testing via Simulation

Thankfully, we can estimate the quantity in Eq. 2.6, and observe its behavior as we add individuals to the group, via simulation. First, we discretize the variable Y, replacing it with a variable Y^* defined as follows:

$$y^*(y) = \begin{cases} y_1^* & \text{if } y < -1000 \\ y_\alpha^* & \text{if } y \in [-1000 + 20(\alpha - 2), -1000 + 20(\alpha - 1)] \\ y_{101}^* & \text{if } y > 1000 \end{cases} \tag{3.1}$$

This effectively turns Y into a 101-valued variable representing whether y is less than -1000, greater than 1000, or in some length-20 band between these two extremes. The probability distributions $p(Y^*|\hat{\boldsymbol{\theta}}, \mathbf{x})$ and $p(Y^*|\mathbf{x})$ are then defined as follows:

$$p(y^*|\hat{\boldsymbol{\theta}}, \mathbf{x}) \begin{cases} \int_{-\infty}^{1000} p(y|\hat{\boldsymbol{\theta}}, \mathbf{x}) & \text{if } y < -1000 \\ \int_{-1000+20(\alpha-2)}^{-1000+20(\alpha-1)} p(y|\hat{\boldsymbol{\theta}}, \mathbf{x}) & \text{if } y \in [-1000 + 20(\alpha - 2), -1000 + 20(\alpha - 1)] \\ \int_{1000}^{\infty} p(y|\hat{\boldsymbol{\theta}}, \mathbf{x}) & \text{if } y > 1000 \end{cases} \tag{3.2}$$

$$p(y^*|\mathbf{x}) \begin{cases} \int_{-\infty}^{1000} p(y|\mathbf{x}) & \text{if } y < -1000 \\ \int_{-1000+20(\alpha-2)}^{-1000+20(\alpha-1)} p(y|\mathbf{x}) & \text{if } y \in [-1000 + 20(\alpha - 2), -1000 + 20(\alpha - 1)] \\ \int_{1000}^{\infty} p(y|\mathbf{x}) & \text{if } y > 1000 \end{cases} \tag{3.3}$$

We then run a simulation in the following steps:

1. Set a value $Q = .1 < M = 100$.
2. Generate a random vector of known parameters $\mathbf{x} \in \mathcal{N}(0, 1, M)$.
3. Generate a random vector of unknown parameters $\hat{\boldsymbol{\theta}} \in \mathcal{N}(0, 1, M)$.
4. Set a value $N = 10$ for the number of individuals in $\mathcal{I}$.

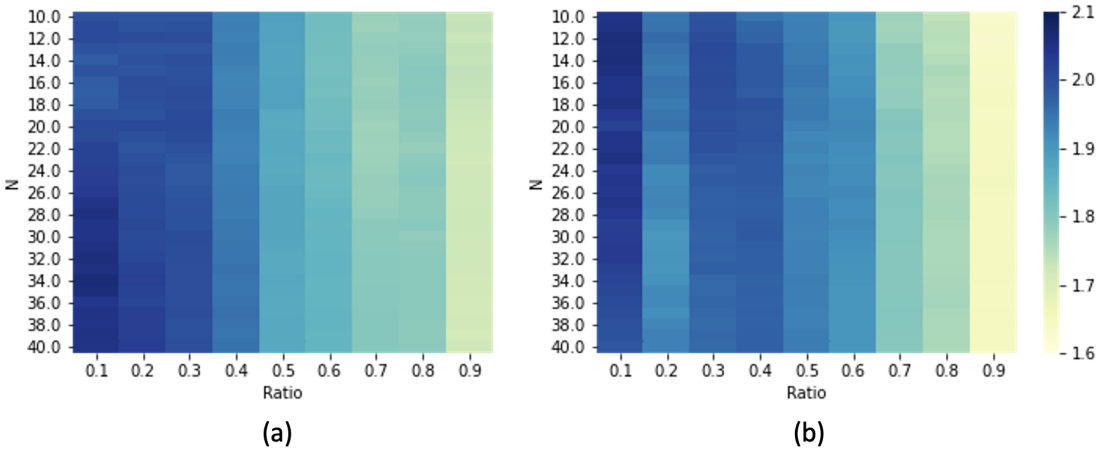

FIGURE 1. Heatmaps showing the mean Kullback-Leibler divergence between $p(Y^*|\hat{\boldsymbol{\theta}}, \mathbf{x})$ and $p(Y^*|\mathbf{x})$ as a function of the number of individuals in the group (N) and the ratio between the percentage of parameters each group member estimates (Q) and the total number of parameters determining the outcome (M). Figure (a) shows this heatmap when new members are not added to the group in a way that deliberately aims to recruit members to the group with standpoints not currently well-represented within the group. Figure (b) shows the heatmap when recruiting is designed so as to be more likely to attract individuals with uncommon standpoints.

(5) Generate a cardinality-Q estimation set E_i for each individual in $\mathcal{I}$ by sampling without replacement from the integers in $[1, M]$ with initial weights $\frac{|\hat{\theta}_j|}{\sum_{l=1}^{M}|\hat{\theta}_l|}$.
(6) Generate a matrix $\hat{\boldsymbol{\theta}}$ such that $\hat{\theta}_{ij} \sim \mathcal{N}(\theta_j, \gamma(j))$ whenever $j \in E_i$, and $\hat{\theta}_{ij} = 0$ otherwise.
(7) Calculate the Kullback-Leibler divergence between the distributions $p(Y^*|\hat{\boldsymbol{\theta}}, \mathbf{x})$ and $p(Y^*|\mathbf{x})$.
(8) Repeat steps (5)-(7) for an additional individual added to $\mathcal{I}$, so that the size of the entire group is $N + 1$.
(9) Repeat step (8) 30 times
(10) Repeat steps (1)-(9) 100 times.
(11) Repeat steps (1)-(10) for all $Q \in [.1, .9]$, in intervals of .1.

This generates $27,900$ estimates of the Kullback-Leibler divergence of interest, across nine different values of Q/M and all group sizes from 10 to 40.

Figure 1(a) shows the mean Kullback-Leibler divergence between $p(Y^*|\hat{\boldsymbol{\theta}}, \mathbf{x})$ and $p(Y^*|\mathbf{x})$ produced under this simulation for each combination of group size N and the ratio Q/M. As can be seen from this graph, as Q/M increases, epistemic uncertainty decreases, as would be expected. What is more difficult to see is there is no significant positive linear relationship between N and epistemic uncertainty ($\beta = .0003$, $p = .499$), such that adding more individuals to the group does nothing to affect epistemic uncertainty in the estimation task.

However, note that in this simulation, when adding a new member to the group we selected the parameters that they measured by sampling from a distribution weighted according to the true values of those parameters. What if we instead selected the parameters measured by new members of the group by sampling from a distribution more likely to return parameters that are not already

sampled by many members of the group? This would simulate a process wherein we aim to recruit new group members with uncommon perspectives. Specifically, we re-run the simulation but at step (8), instead of repeating step (5) and sampling without replacement from the integers $[1, M]$ with initial weights $\frac{|\theta_j|}{\sum_{l=1}^{M} |\theta_l|}$, we instead sample without replacement from the integers $[1, M]$ with initial weights given by a softmax transformation on an M-entry vector with entries $1 - \gamma(j)$ for each integer $j \in [1, M]$. That is, each new member of the group is more likely to measure a parameter when that parameter is *not* currently measured by many members of the group.

Re-running the simulation with this amendment yields the heatmap in Figure 1(b). Here, we find that N is negatively linearly correlated with epistemic uncertainty ($\beta = -.0021$, $p < .001$). This shows that when we actively seek individuals who estimate uncommon parameters, the epistemic uncertainty in an estimation task decreases as group size increases, under the assumptions made in this model. Moreover, comparison of the two heatmaps shows that for several values of Q/M, aiming to recruit new group members by seeking those with unusual perspectives tends to lead to lower epistemic uncertainty than aiming to recruit new members who measure important parameters.

References

Hüllermeier, Eyke and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods". In: *Machine Learning* 110.3 (2021), pp. 457–506.

Mobiny, Aryan et al. "Dropconnect is effective in modeling uncertainty of bayesian deep networks". In: *Scientific reports* 11.1 (2021), pp. 1–14.