Synthesizing New Expertise via Collaboration

Bijan Mazaheri[†], Siddharth Jain[*], and Jehoshua Bruck[*]

[*]Electrical Engineering, California Institute of Technology, U.S.A. {sidjain,bruck}@caltech.edu

[†]Computing & Mathematical Sciences, California Institute of Technology, U.S.A. bmazaher@caltech.edu

***Abstract*—Consider a set of classes and an *uncertain* input. Suppose, we do not have access to data and only have knowledge of perfect experts between a few classes in the set. What constitutes a consistent set of opinions? How can we use this to predict the opinions of experts on missing sub-domains? In this paper, we define a framework to analyze this problem. In particular, we define an expert graph where vertices represent classes and edges represent binary experts on the topics of their vertices. We derive necessary conditions for an expert graph to be valid. Further, we show that these conditions are also sufficient if the graph is a cycle, which can yield unintuitive results. Using these conditions, we provide an algorithm to obtain upper and lower bounds on the weights of unknown edges in an expert graph.**

I. Introduction

2.6 million scientific papers are published every year on many overlapping topics. Within this massive body of human knowledge, should we be concerned about apparent contradictions, or is this to be expected? What does it mean for these studies to be consistent?

There are two types of uncertainty that may contribute to perceived inconsistency. The first is uncertainty in factors whose behavior is not known to us, which represents a fundamental limit in our understanding of a problem. The second is uncertainty in parameters that are *understood* by our experts, but *not known* at the time of decision. These are often referred to as aleatoric uncertainty and epistemic uncertainty [10].

To illustrate this, consider the following table of probabilities of classes $y \in \{A, B, C, D\}$ over four equally probable states of unknown input $x \in \{x_1, x_2, x_3, x_4\}$ given known input z.

x	$\Pr(y = A \mid x, z)$	$\Pr(y = B \mid x, z)$	$\Pr(y = C \mid x, z)$	$\Pr(y = D \mid x, z)$
x_1	.9	.09	.009	.001
x_2	.001	.9	.09	.009
x_3	.009	.001	.9	.09
x_4	.09	.009	.001	.9

The two types of uncertainty correspond to columns and rows in this table. A row of probabilities demonstrates aleatoric uncertainty given all understood variables, whereas the multiple possible unknown states in a column represent epistemic uncertainty.

Expertise is built on *limited domains*, so experts never have access to the probabilities given in this table. Instead, they only have access to the conditional probabilities for a *specific* x, for example: $\Pr(y = A \mid y \in \{A, B\}, x) = \frac{\Pr(y=A \mid x)}{\Pr(y=A \mid x)+\Pr(y=B \mid x)}$. This is equivalent to having access to ratios of the probabilities. An A/B expert would not be able to tell the difference between the $.9/.09$ in x_1 and the $.09/.009$ in x_4.

Instead of requiring our experts to break down the cases (x_1, x_2, x_3, x_4) considered in their decision, we instead only consider their ultimate opinion: the *expected* conditional probability, e.g. $\sum_x \Pr(x \mid z) \Pr(A \mid z, x)$. An expert who can accurately report this expectation is called a "perfect expert." We will study how networks of opinions of such experts can behave. As we will see, the results can be highly unintuitive.

Returning to our example, knowing that the four states are equiprobable, the first expert will notice that A is much more likely than B in three out of the four scenarios (x_1, x_3, x_4):

$$\mathbb{E}_x[\Pr(A \mid \{A, B\}, x)] = \frac{1}{4}\left(\frac{90}{99} + \frac{1}{901} + \frac{9}{10} + \frac{90}{99}\right) \geqslant .675$$

The probabilities have been cyclically shifted, so the three other experts will also have similar calculations, with the class to the left being preferred to the one on the right in three out of the four scenarios, i.e. B is preferred to C, C is preferred to D, and D is preferred to A.

Hence, no matter which class is chosen, one expert will vehemently assert that there is at least a 67.5% chance we are wrong! This is an example of what is known in voting theory as the Condorcet Paradox [15], which arises when voter preferences are arranged in a cyclic fashion as we have simulated in our scenarios x_1, x_2, x_3, x_4.

The unintuitive nature of combining expertise in the presence of epistemic uncertainty calls for a careful treatment of this problem. This paper will develop a framework for understanding networks of experts with and overlapping knowledge and work to completely describe the feasibility of these networks under no assumptions of the distribution on unknown state variable u. Understanding which expert networks are consistent with each other will also allow us to create "synthetic experts" from bounds on consistent opinions on unspecified sub-domains given no assumptions on the distribution of x. The paper is organized as follows.

- **Section III**: We establish notation and define *expert graphs*, a representation of the knowledge from experts on pairs of classes.
- **Section IV**: We provide a necessary *curl condition* that must be satisfied by every *cycle* in the graph. We show that the question of sufficiency is reducible to understanding the linear ordering polytope [2], [7], [12]. We show this condition allows for the emergence of *non-transitive*

behavior in the expert opinions which we quantify by defining the notion of *non-transitive strength.*

- **Section V** We give an algorithm to create *synthetic experts* between classes for which there is no expert. This corresponds to filling in bounds for the weights of unspecified edges in an expert graph.
- **Section VI**: We conclude the paper and propose directions for future work.

II. Related Work

In voting theory, there is significant work on "induced binary probabilities," which correspond to networks of hypothetical pairwise elections given a population of rankings [7]. This field has separately come across what we call the "curl condition," and shown that it is necessary and sufficient for $n \leqslant 5$ classes, and just necessary for $n \geqslant 6$ classes. In this setting, each voter can be thought of as a "state" of epistemic uncertainty with an absolute ranking. Our problem introduces aleatoric uncertainty to this setting, resulting in soft preferences instead of binary rankings.

The task of synthesizing our experts' knowledge is essentially that of a multiclassifier built from pairwise classifiers. As a result, there are multiple frameworks that exist to use the outcome of pairwise comparisons to obtain multiclassifiers (see [1], [6]). [1]

The focus of this literature is on the *design* of classifiers to aggregate. For example, we may desire the combination of a philosopher who can distinguish between two fields ($\{A, B\}$ vs $\{C, D\}$) and two experts in each field (A vs B and C vs D), but it is often difficult to assemble such a complete team. We will instead consider the setting where we are given a set of *unengineered* binary classifiers. In the rest of the paper unless otherwise stated, we will be using already trained experts instead of classifiers.

III. The Expert Graph Framework

A. Preliminaries

We here provide notation used throughout the paper. $[\ell]$ is used to denote the set $\{0, 1, \cdots, \ell-1\}$ for any $\ell \in \mathbb{N}$. $|A|$ denotes the size of set A. $\mathbb{1}[c]$ will be used for an indicator function which is 1 if condition c is met and 0 otherwise. Any bold symbol unless otherwise stated will be used to denote a vector. $\boldsymbol{u} = (u^{(0)}, \ldots, u^{(\ell-1)})^T$ denotes a ℓ-length vector. $\mathbf{1}_\ell$ denotes an all 1 vector of size ℓ.

$\triangle_\ell$ will be used to denote vectors of size ℓ which can represent probability distributions. That is, $\lambda \in \triangle_\ell$ iff $\lambda \in [0,1]^\ell$ iff $\mathbf{1}^\top \lambda = 1$.

We use $\prec$, $\succ$, $\preceq$, $\succeq$ to denote element-wise inequality. For example, we say $\mathbf{u} \succeq \mathbf{v}$ if $u^{(i)} \geqslant v^{(i)}\ \forall\ i \in [\ell]$. The L1 norm of a vector $\boldsymbol{x}$ is given by $\|\boldsymbol{x}\|_1 = \sum_{i=0}^{\ell-1} |x_i|$ and the L2 norm is given by $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=0}^{\ell-1} |x_i|^2}$.

We will use $\mathrm{Co}(S)$ to denote the open convex hull of S, $\overline{\mathrm{Co}}(S)$ to denote the closed convex hull, and $\mathrm{Bo}(\cdot)$ to denote the boundary.

[1]Related literature also exists in ensemble methods [4], [14], [16]. This work generally focuses on combining experts trained on the *same* task, whose knowledge is *imperfect* due to differing training datasets. In contrast, our experts are perfect, but trained on *different* tasks.

B. Expert Graphs

Consider a domain $\mathcal{X}$. Each input $x \in \mathcal{X}$ will have a probability of being labeled one of n classes in the set $\mathcal{C} = \{C_0, C_1, \ldots, C_{n-1}\}$:

$$p_x^{(i)} = \Pr(y = C_i \mid x) \tag{1}$$

Assumption 1. *We assume for any $x \in \mathcal{X}$, $p_x^{(i)} > 0$ for all C_i.*

Instead of having access to $p_x^{(i)}$ directly, our experts will instead only have access to pairwise conditional probabilities

$$\hat{f}_x(C_i, C_j) = \Pr(y = C_i \mid y \in \{C_i, C_j\}, x) = \frac{p_x^{(i)}}{p_x^{(i)} + p_x^{(j)}} \tag{2}$$

Note that $\hat{f}_x(C_i, C_j) = 1 - \hat{f}_x(C_j, C_i)$. Further from assumption 1, $\hat{f}_x(C_i, C_j) \in (0,1)$. We will also work with distributions on the input domain. We will denote $d(\mathcal{X})$ to denote the probability distribution on the input domain $\mathcal{X}$, based on which we define: $\hat{f}_d(C_i, C_j) = \mathbb{E}_d[\hat{f}_x(C_i, C_j)]$. Here $\mathbb{E}_d[.]$ denotes expectation on distribution d.

Definition 1. Given $\ell \geqslant 2$, we define a permutation $\mathcal{A} = (a_0, a_1, \cdots, a_{\ell-1})$ with $a_i \in [n]\ \forall\ i \in [\ell]$.

Definition 2. Given $\mathcal{A}$, we define vectors $\hat{\boldsymbol{f}}_x(\mathcal{C}, \mathcal{A}) \in (0,1)^\ell$ and $\hat{\boldsymbol{f}}_d(\mathcal{C}, \mathcal{A}) \in (0,1)^\ell$ such that

$$\begin{aligned} \hat{f}_x^{(i)}(\mathcal{C}, \mathcal{A}) &= \hat{f}_x(C_{a_i}, C_{a_{i+1}})\ \forall i \in [\ell] \\ \hat{f}_d^{(i)}(\mathcal{C}, \mathcal{A}) &= \hat{f}_d(C_{a_i}, C_{a_{i+1}})\ \forall i \in [\ell] \end{aligned} \tag{3}$$

We are now ready to define the expert graph.

Definition 3. An *expert graph* $\mathcal{G}_d = (\mathcal{C}, S_\mathcal{C}, f_d(\cdot,\cdot))$ encodes experts opinions about undirected class pairs $(\{C_i, C_j\} \in S_\mathcal{C}$ given by pairwise experts $\hat{f}_d : \mathcal{C} \times \mathcal{C} \mapsto (0,1)$ on input distribution $d : \mathcal{X} \mapsto [0,1]$. An expert graph where $(\mathcal{C}, S_\mathcal{C})$ is a cycle will be referred to as an *expert cycle*.

It is important to note here that $d(x)$ is used to determine the weights of the edges, but is not necessarily known to the observer of the graph. Figure 1 shows an example of an expert graph with 4 classes and 5 experts.

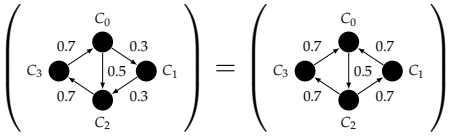

Fig. 1. Two equivalent expert graphs with $\mathcal{C} = (C_0, C_1, C_2, C_3)$ and $\hat{f}_d(C_0, C_1) = 0.3$, $\hat{f}_d(C_1, C_2) = 0.3$, $\hat{f}_d(C_2, C_3) = 0.7$, $\hat{f}_d(C_3, C_0) = 0.7$.

IV. Expert Cycles and Curl

A. Curl

Definition 4. Given class set $\mathcal{C}$, a cycle of indices $\mathcal{A} = (a_0, a_1, \ldots, a_{\ell-1})$, and an edge function $f(\cdot) : \mathcal{C} \times \mathcal{C} \mapsto [0,1]$, we define the **curl** to be

$$\mathrm{Curl}(\mathcal{C}, \mathcal{A}, f(\cdot,\cdot)) = \sum_{i=0}^{\ell-1} f(C_{a_i}, C_{a_{i+1}}) \tag{4}$$

Here, $a_{i+1} = a_{(i+1) \mod \ell}$. This definition is motivated by the notion of curl in vector calculus and physics.

In the context of this paper, we will consider the curl on expert graphs, so our edge function will be given by pairwise experts.

We denote $\text{Curl}_x(\mathcal{C}, \mathcal{A})$ for $f(\cdot,\cdot) = \hat{f}_x(\cdot,\cdot)$:

$$\text{Curl}_x(\mathcal{C}, \mathcal{A}) = \sum_{i=0}^{\ell-1} \hat{f}_x(C_{a_i}, C_{a_{i+1}}). \tag{5}$$

Similarly, we can define $\text{Curl}_d(\mathcal{C}, \mathcal{A})$ for $f(\cdot,\cdot) = \hat{f}_d(\cdot,\cdot)$, i.e.

$$\text{Curl}_d(\mathcal{C}, \mathcal{A}) = \sum_{i=0}^{\ell-1} \hat{f}_d(C_{a_i}, C_{a_{i+1}}) = \mathbb{E}_d[\text{Curl}_x(\mathcal{C}, \mathcal{A}) \tag{6}$$

Notice that if $\overline{\mathcal{A}} = (a_{\ell-1}, a_{\ell-2}, \cdots, a_1, a_0)$ is the reversed direction of the cycle $\mathcal{A}$, we have

$$\text{Curl}_x(\mathcal{C}, \overline{\mathcal{A}}) = \ell - \text{Curl}_x(\mathcal{C}, \mathcal{A}). \tag{7}$$

B. The Curl Condition

We can give upper and lower bounds on the curl.

Lemma 2. *Given $\ell \geqslant 2$ and a cycle $\mathcal{A}$ on ℓ classes in set $\mathcal{C}$, then the curl for any input x follows*

1) $\ell = 2$: $\quad Curl_x(\mathcal{C}, \mathcal{A}) = 1.$
2) $\ell \geqslant 3$: $\quad 1 < Curl_x(\mathcal{C}, \mathcal{A}) < \ell - 1.$

Proof:

1) $(\ell = 2)$: $\frac{p_x^{(a_0)}}{p_x^{(a_0)}+p_x^{(a_1)}} + \frac{p_x^{(a_1)}}{p_x^{(a_1)}+p_x^{(a_0)}} = 1.$
2) $(\ell \geqslant 3)$

Lower Bound: Recall that

$$\text{Curl}_x(\mathcal{C}, \mathcal{A}) = \sum_{i=0}^{\ell-1} \frac{p_x^{(a_i)}}{p_x^{(a_i)} + p_x^{(a_{i+1})}} \tag{8}$$

Note that the denominators of every term are strictly upper bounded by 1. This gives:

$$\text{Curl}_x(\mathcal{C}, \mathcal{A}) > \sum_{i=0}^{\ell-1} p_x^{(a_i)} = 1 \tag{9}$$

Upper Bound: By Equation 7 we have:

$$\text{Curl}_x(\mathcal{C}, \overline{\mathcal{A}}) = \ell - \text{Curl}_x(\mathcal{C}, \mathcal{A}) \overset{(a)}{<} \ell - 1$$

Where (a) is given by invoking the upper bound that was proved above.

■

Corollary 3. *Given $\mathcal{C}$ and a cycle $\mathcal{A}$ on ℓ classes, then* [2]

1) $\ell = 2$: $\quad Curl_d(\mathcal{C}, \mathcal{A}) = 1.$
2) $\ell \geqslant 3$: $\quad 1 < Curl_d(\mathcal{C}, \mathcal{A}) < \ell - 1.$

Proof: The proof follows from Lemma 2 and the definition of $\text{Curl}_d(\mathcal{C}, \mathcal{A})$. ■

[2]The notion of curl can be extended to one vs multiple class experts as well, in which case we can extend the upper bound of corollary 3 to $\ell - k + 1$, where k represents the total number of classes used by the expert. For details see Appendix Section A.

Corollary 3 provides a necessary condition for any $\hat{f}_d(\mathcal{C}, \mathcal{A})$.

Example 1. Below are some examples of $\hat{f}_d(\mathcal{C}, \mathcal{A})$ which are *not* possible,

- $\hat{f}_d(\mathcal{C}, \mathcal{A}) = (0.4, 0.8, 0.9)^T : \text{Curl}_d(\mathcal{C}, \mathcal{A}) = 2.1 \geqslant 2.$
- $\hat{f}_d(\mathcal{C}, \mathcal{A}) = (0.4, 0.8, 0.8)^T : \text{Curl}_d(\mathcal{C}, \mathcal{A}) = 2 \geqslant 2.$
- $\hat{f}_d(\mathcal{C}, \mathcal{A}) = (0.4, 0.3, 0.1)^T : \text{Curl}_d(\mathcal{C}, \mathcal{A}) = 0.8 \leqslant 1.$

C. Towards Sufficiency

A natural question is whether the curl condition is sufficient to describe all possible expert graphs. In this section we will reduce this question to a studied problem of decomposing graphs into convex combinations of acyclic tournaments, also referred to as the "linear ordering polytope" [7]. The proof of sufficiency for cycles will follow.

Definition 5. An *orientation* is an assignment of directions to the edges a graph G. An orientation $T = (\mathcal{C}, S_{\mathcal{C}}, f_T(\cdot))$ is specified using a binary edge-weight function $f_T(\cdot) : S_{\mathcal{C}} \mapsto \{0, 1\}$ with $f_T(C_i, C_j) = 1$ indicating $C_i \xrightarrow{T} C_j$ and $f_T(C_i, C_j) = 0$ indicating $C_i \xleftarrow{T} C_j$. An orientation on a complete graph is called a *tournament* [13].

We will first show how to give a set of class probabilities $p_x^{(C)}$ for which the pairwise expert output gets arbitrarily close to that of an acyclic tournament.

Lemma 4. *Let $T = (\mathcal{C}, S_{\mathcal{C}}, f_T(\cdot,\cdot))$ be a tournament on classes $\mathcal{C} = \{C_0, \ldots, C_{n-1}\}$. Consider the vectors of edge weights $\hat{\mathbf{f}}_\mathbf{x}$ and $\mathbf{f_T}$. For all ε, we can construct a state x with class probabilities $p_x^{(C_0)}, \ldots, p_x^{(C_{n-1})} \in (0, 1)$ such that $\left\|\hat{\mathbf{f}}_\mathbf{x} - \mathbf{f_T}\right\|_2 < \varepsilon$ if T is acyclic.*

Proof: Let k be an integer. Let $\alpha_t = \frac{1}{k^t}$ and $z = \sum_{t=1}^{n} \alpha_t$. By doing this, if we set $p_x^{(C_i)} = \frac{\alpha_{t_i}}{z}$ and $p_x^{(C_j)} = \frac{\alpha_{t_j}}{z}$, then

$$\hat{f}_x^{(k)}(C_i, C_j) \to \begin{cases} \leqslant \frac{1}{k} & \text{if } t_i > t_j \\ = \frac{1}{2} & \text{if } t_i = t_j \\ > 1 - \frac{1}{k} & \text{if } t_i < t_j \end{cases}$$

($\Rightarrow$) If T has a cycle, then there exists some set of indices $\mathcal{A} = (a_0, \ldots, a_{\ell-1})$ such that $\hat{f}_x^{(k)}(C_{a_i}, C_{a_{i+1}}) > 1 - \frac{1}{k}$ for $i = 0, \ldots, \ell - 1$. Recall that

$$\hat{f}_x^{(k)}(C_i, C_j) > 1 - \frac{1}{k} \Leftrightarrow p_x^{(C_i)} > p_x^{(C_j)}$$

Hence, $p_x^{(C_{a_0})} > p_x^{(C_{a_1})} > \ldots > p_x^{(C_{a_{\ell-1}})}$. But $\hat{f}_x(C_{a_{\ell-1}}, C_{a_0}) > 1 - \frac{1}{k}$ also implies $p_x^{(C_{a_{\ell-1}})} > p_x^{(C_{a_0})}$, a contradiction in transitivity.

($\Leftarrow$) We give a constructive assignment of probabilities according to the ranking implied by the tournament. All acyclic tournaments have one unique Hamiltonian (including all classes in $\mathcal{C}$) path $C_{a_0} \to C_{a_1} \to \ldots \to C_{a_{n-1}}$. Assign $p_x^{(C_{a_i})} = \frac{\alpha_i}{z}$.

Note that if $i < j$, we have a path from C_{a_i} to C_{a_j}, so $f_T(C_{a_j}, C_{a_i}) = 0$ in order to remain acyclic. Hence

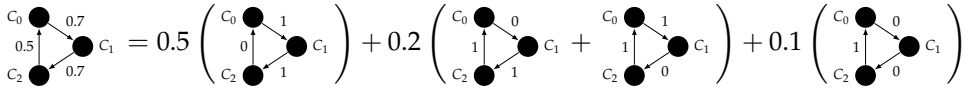

Fig. 2. Decomposing a knowledge cycle on 3 classes with edge weights $\hat{f}_d(\mathcal{C},\mathcal{A}) = (0.7, 0.7, 0.5)^\top$ into acyclic orientations.

$f_T(C_{a_i}, C_{a_j}) = \mathbb{1}[i < j]$. Thus, we have assigned successively larger probabilities to higher ranked classes such that the resulting experts have

$$|\hat{f}_x^{(k)}(C_{a_i}, C_{a_j}) - f_T(C_{a_i}, C_{a_j})| \leqslant \frac{1}{k}. \tag{10}$$

There are $\leqslant n^2$ edges, so set $\frac{1}{k} = \frac{\varepsilon}{n^2}$ to get $\left\|\hat{\mathbf{f}}_{\mathbf{x}}^{(\mathbf{k})} - \mathbf{f}_{\mathbf{T}}\right\|_2 \leqslant \varepsilon$ as desired. ∎

Lemma 4 allows us to harness results from voting theory literature to show our condition is sufficient for $n \leqslant 5$ classes [7]. While the curl condition is insufficient for the linear ordering polytope with $n \geqslant 6$, it is not yet clear whether this is also true for expert graphs.

Lemma 4 can also be used on acyclic orientations, since edges can be added to acyclic orientations to make acyclic tournaments. The task now reduces to finding a decomposition of the weighted directed graph into a convex composition of acyclic tournaments, for which we can find class probabilities via Lemma 4. Figure 2 above gives an example of such a decomposition of a cycle.

The tournaments described in the previous subsection can be achieved in the *limit*, but cannot be obtained directly via any legal assignment of class probabilities. Hence, it remains to show that any convex combination of tournaments can be expressed instead by a convex combination of *almost* tournaments from the converging series already given.

Convex hulls of finite sets in $\mathbb{R}^\ell$ are *convex* polytopes, which can be expressed as an intersection of h halfspaces indexed by f with $\{\mathbf{x} : \mathbf{a^{(f)}}^\top \mathbf{x} < b^{(f)}\}$ [9]. The perpendicular vectors $\mathbf{a^{(f)}}^\top$ can be combined as row-vectors of the matrix A so that any convex polytope can be expressed as

$$\{x : A\mathbf{x} \prec \mathbf{b}\} = \left\{\mathbf{x} : \begin{pmatrix} (\mathbf{a^{(1)}})^\top \\ \vdots \\ (\mathbf{a^{(h)}})^\top \end{pmatrix} \mathbf{x} \prec \begin{pmatrix} b^{(1)} \\ \vdots \\ b^{(h)} \end{pmatrix}\right\} \tag{11}$$

For convenience, the vectors $\mathbf{a^{(f)}}, \tilde{\mathbf{a}}^{(\mathbf{f})}$ are assumed to be unit vectors throughout.

Theorem 5. *Given $V = \{v_1, \ldots, v_m\}$ and "perturbed points" $\tilde{V} = \{\tilde{v}_1, \ldots \tilde{v}_m\}$ with $v_i \in \mathbb{R}^\ell$ and $\tilde{v}_j \in \mathbb{R}^\ell$ for all i, j. We have that if $\mathbf{x} \in Co(V)$ and is $\varepsilon > 0$ from the boundary $Bo(Co(V))$, then if we can find perturbed points $\tilde{V}$ such that they are within ε from the desired V, then $\mathbf{x} \in Co(\tilde{V})$.*

More precisely, let

$$Co(V) = \{\mathbf{x} : A\mathbf{x} \prec \mathbf{b}\}$$
$$Co(\tilde{V}) = \{\mathbf{x} : \tilde{A}\mathbf{x} \prec \tilde{\mathbf{b}}\}$$

as given by Equation 11. *If $A\mathbf{x} \prec \mathbf{b} - \varepsilon\mathbf{1}_\ell$ and $\|\mathbf{v_i} - \tilde{\mathbf{v}}_\mathbf{i}\|_2 < \varepsilon \; \forall i$, then $\tilde{A}\mathbf{x} \prec \tilde{\mathbf{b}}$.*

Proof: The proof is given in Appendix Section B. ∎

Theorem 6. *Say an expert graph $\mathcal{G}_d = (\mathcal{C}, S_\mathcal{C}, f_d(\cdot))$ can be decomposed into orientations $\mathcal{T} = \{T_1, \ldots, T_q\}$ with weights $(d(T_1), \ldots, d(T_q))^\top \in \triangle_q$*

$$f_d(e) = \sum_i d(T_i) f_{T_i}(e) \; \forall e \in S_\mathcal{C}. \tag{12}$$

Then $\mathcal{G}_d$ is also the convex combination of expert graphs achievable on states $x_1, \ldots, x_q$.

Proof: Use Lemma 4 to find the states $\mathbf{x}_\mathbf{i}^{(\mathbf{k})}$ with edge weights $\mathbf{f}_{\mathbf{x_i}}^{(\mathbf{k})}$ indexed by k which converge on T_i and apply Theorem 5. ∎

D. Sufficiency for Cycles

We can show sufficiency for expert cycles with a decomposition into acyclic orientations followed by the application of Theorem 6. The existence of such a decomposition is shown in Appendix Section C. A constructive decomposition is also given Appendix Section D, which is demonstrated in Figure 4.

E. Non-transitivity

In the previous subsections, we show that the curl condition derived in Corollary 3 is both necessary and sufficient for any cycle $\mathcal{A}$. This gives us tight upper and lower bounds on an unknown edge on cycle $\mathcal{A}$ given other edges as given in Theorem 7 below.

Theorem 7. *Given ℓ, $\mathcal{A}$, the expert provides $\hat{f}_d(C_{a_i}, C_{a_{i+1}})$ for $i \in [\ell - 1]$. Then, we have*

$$1 - \sum_{i=0}^{\ell-2} \hat{f}_d(C_{a_i}, C_{a_{i+1}}) < \hat{f}_d(C_{a_{\ell-1}}, C_{a_0}) < \ell - 1 - \sum_{i=0}^{\ell-2} \hat{f}_d(C_{a_i}, C_{a_{i+1}})$$

.

Proof: The proof follows from Corollary 3. ∎

The curl condition allows a degree of non-transitivity in pairwise experts. To explore this, we will first quantify non-transitivity in pairwise experts.

Definition 6. For an input distribution $d(\mathcal{X})$, classes $\mathcal{C}$, and cycle of classes $\mathcal{A}$, we say the **strength of non-transitivity**, $\sigma_d^{(\mathcal{C},\mathcal{A})}$ is

$$\sigma_d^{(\mathcal{C},\mathcal{A})} = \max\{\min_{i\in[\ell]} \hat{f}_d(C_{a_i}, C_{a_{i+1}}), \min_{i\in[\ell]} \hat{f}_d(C_{a_{i+1}}, C_{a_i})\} \tag{13}$$

Note that $\sigma_d^{(\mathcal{C},\mathcal{A})} = \sigma_d^{(\mathcal{C},\overline{\mathcal{A}})}$ by definition.

Definition 7. We say that a cycle $\mathcal{A}$ is **non-transitive** if $\sigma_d^{(\mathcal{C},\mathcal{A})} > .5$.

Intuitively, these definitions capture the notion of non-transitivity of preferences. If for classes C_0, C_1, C_2 we

have $\sigma_d^{(\mathcal{C},(0,1,2))} > .5$, this means that we have a "cycle of preference." That is, we prefer $(\leftarrow)$ $C_0 \leftarrow C_1 \leftarrow C_2 \leftarrow C_0$. No matter which class we decide on, we have a pairwise expert that suggests it is the incorrect choice. This notion is related to the Condorcet Paradox [15].

Example 2. Below we provide examples of *non-transitive* cycles.

- $\hat{f}_d(\mathcal{C},\mathcal{A}) = (0.6, 0.8, 0.55)^T$: $\sigma_d^{(\mathcal{C},\mathcal{A})} = 0.55$.
- $\hat{f}_d(\mathcal{C},\mathcal{A}) = (0.3, 0.4, 0.4)^T$: $\sigma_d^{(\mathcal{C},\mathcal{A})} = 0.6$.
- $\hat{f}_d(\mathcal{C},\mathcal{A}) = (0.8, 0.7, 0.7, 0.7)^T$: $\sigma_d^{(\mathcal{C},\mathcal{A})} = 0.7$.

Observation 8. *For a cycle of length ℓ, by definition we have*

$$\frac{\max\{Curl_d(\mathcal{C},\mathcal{A}), Curl_d(\mathcal{C},\overline{\mathcal{A}})\}}{\ell} \geqslant \sigma_d^{(\mathcal{C},\mathcal{A})}.$$

Corollary 9. *Given $\ell \geqslant 3$, $\mathcal{C}$, $\mathcal{A}$, the strength of non-transitivity $\sigma_d^{(\mathcal{C},\mathcal{A})}$ is upper bounded by*

$$\sigma_d^{(\mathcal{C},\mathcal{A})} < \frac{\ell - 1}{\ell}.$$

V. Synthetic Experts

One may wish to quantify knowledge about an input without direct access to the relevant expert. This question equates to filling in missing edges in the expert graph *without* knowledge of $d(x)$. An example of this is given in Figure 3. Motivated by this example, we provide an algorithm that uses the curl condition derived in Corollary 3 to obtain upper and lower bounds on missing edges in the expert graph.

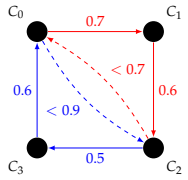

Fig. 3. The cycle $C_0 \to C_1 \to C_2 \to C_0$ gives $f_d(C_2, C_0) < 0.7$, which implies $f_d(C_0, C_2) > 0.3$. The cycle $C_0 \to C_2 \to C_3 \to C_0$ gives $f_d(C_0, C_2) < 0.9$. Together, we get $f_d(C_0, C_2) \in (0.3, 0.9)$.

Definition 8. Let $\Gamma = C_{b_0}, C_{b_1}, \cdots, C_{b_m}$ be a path on expert graph $\mathcal{G}$ where we assume $(C_{b_i}, C_{b_{i+1}}) \in S_{\mathcal{C}}$ $\forall$ $i \in [m]$, then the forward and backward weights of the path Γ are given by $W_F(\Gamma) = \sum_{i=0}^{m-1} \hat{f}_d(C_{b_i}, C_{b_{i+1}})$ and $W_B(\Gamma) = \sum_{i=0}^{m-1} \hat{f}_d(C_{b_{i+1}}, C_{b_i})$ respectively.

Lemma 10. *Let $C_i, C_j \in \mathcal{C}$, such that $(C_i, C_j) \notin S_{\mathcal{C}}$. Consider $\mathcal{D}_{ij}$ be a set of paths in $\mathcal{G}$ such that every $\Gamma \in \mathcal{D}_{ij}$ starts at vertex C_i and ends at vertex C_j. Then,*

$$1 - min_{\Gamma \in \mathcal{D}_{ij}} W_F(\Gamma) < \hat{f}_d(C_j, C_i) < min_{\Gamma \in \mathcal{D}_{ij}} W_B(\Gamma).$$

Proof: Using Theorem 7, for any $\Gamma \in \mathcal{D}_{uv}$, we have

$$1 - W_F(\Gamma) < \hat{f}_d(C_j, C_i) < W_B(\Gamma). \tag{14}$$

These bounds are true for all paths, which the lemma optimizes. ∎

Lemma 10 also provides a simple algorithm to obtain tight upper and lower bounds on $\hat{f}_d(C_v, C_u)$ for any $(C_v, C_u) \notin S_{\mathcal{C}}$. The Floyd-Warshall algorithm can be used to efficiently calculate the shortest path between any two unknown nodes in a graph [5]. Thus, for each $\{C_i, C_j\} \in S_{\mathcal{C}}$ we create a directed graph giving edge (C_i, C_j) weight $\hat{f}_d^{(C_i,C_j)}$ and edge (C_j, C_i) weight $\hat{f}_d^{(C_j,C_i)}$. This can be used as input for Floyd-Warshall and the output is applied to Lemma 10.

These bounds describe the range of possible edge-weights so as to not violate the curl condition. If the curl condition is sufficient, then any such edge weights are achievable and this bound is tight. Thus, from Section IV we have that these bounds are sufficient for expert cycles.

VI. Conclusion

We have defined the expert graph as a framework to understand the amalgamation of pairwise experts with overlapping domains. To analyze these graphs, we introduced the curl and derived *necessary* lower and upper bounds. From this notion of curl, we show the emergence of non-transitivity in the outputs given by experts and provide an algorithm to derive upper and lower bounds on missing edges of an expert graph.

We have shown that any weighed graph in the open convex hull of acyclic orientations can also be achieved as an expert graph. We used this to prove sufficiency of the curl condition for expert cycles. This allows for future work to apply results from literature on the linear ordering polytope [2], [7], [12].

In our analysis, we have assumed that epistemic uncertainty is the same for all experts. One natural extension would be to relax this assumption. Experts in different domains may understand epistemic uncertainty in terms of different and incomparable states. More concretely, each expert may not partition the uncertainty into the same states x_1, x_2, x_3, x_4. For example, a doctor may consider different unknown genetic factors, whereas a nutritionist may instead consider cases of different nutrition.

Future work may also expand the notion of the expert graph to an expert *hypergraph*, which includes experts with potentially *nonbinary* domains of expertise.

References

[1] E. L. Allwein, R. E. Schapire, and Y. Singer, *"Reducing multiclass to binary: a unifying approach for margin classifiers,"* J. Mach. Learn. Res. 1, pp. 113–141, September 2001.

[2] N Alon. *"Voting paradoxes and digraphs realizations,"* Advances in Applied Mathematics 29, no. 1, pp. 126-35. 2002.

[3] D. Bertsimas and J. Tsitsiklis, *"Introduction to Linear Optimization,"* (1st. ed.). Athena Scientific, 1997.

[4] L. Breiman. *"Bagging predictors,"* Machine learning 24, no. 2, pp. 123-140. 1996.

[5] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *"Introduction to Algorithms,"* Third Edition (3rd. ed.). The MIT Press, 2009.

[6] T. G. Dietterich and G. Bakiri, *"Solving multiclass learning problems via error-correcting output codes,"* J. Artif. Int. Res. 2 (1), pp. 263–286, August 1994.

[7] P. C. Fishburn, *"Induced binary probabilities and the linear ordering polytope: a status report,"* Mathematical Social Sciences 23, pp.67-80, 1992.

[8] E. Frazzoli, and M. Dahleh, *"6.241J Dynamic Systems and Control,"* Massachusetts Institute of Technology: MIT (Lecture 5), Spring 2011.

[9] B. Grünbaum. *Convex polytopes.* Vol. 221. Springer Science & Business Media, 2013.

[10] A. Indrayan, *Medical Biostatistics*, Second Edition, Chapman & Hall/CRC Press. 2008.

[11] B. Mazaheri, S. Jain, and J. Bruck, *"Synthesizing New Expertise via Collaboration,"* http://www.paradise.caltech.edu/papers/etr150.pdf

[12] D. C. McGarvey, *"A theorem on the construction of voting paradoxes,"* Econometrica 21, pp 608-610. 1953.
[13] J. Moon, *"Topics on tournaments in Graph Theory,"* Courier Dover Publications, 2015.
[14] R.E. Schapire. *"The strength of weak learnability,"* Machine learning 5, no. 2, pp. 197-227. 1990.
[15] J. S. Weber, *"An Elementary Proof of the Conditions for a Generalized Condorcet Paradox,"* Public Choice 77, no. 2, pp. 415-419. 1993.
[16] Z. H. Zhou, *"Ensemble methods: foundations and algorithms,"* CRC press. 2012.

APPENDIX

A. Extending the Curl Condition to Multi-classifiers

In previous sections, we considered pairwise experts, i.e. 1 vs 1 classifiers. In this section, we consider 1 vs $k-1$ classifiers, where $k \geqslant 3$. Given $\mathcal{C}$ such that $|\mathcal{C}| \geqslant k$, we define

$$\hat{f}_x(C_{i_0}, \cdots, C_{i_{k-1}})) = \Pr(y = C_{i_0} \mid y \in \{C_{i_0}, \cdots, C_{i_{k-1}}\}, x) = \frac{p_x^{(i_1)}}{\sum_{j=0}^{k-1} p_x^{(i_j)}}$$

Given $k \geqslant 3$, $\mathcal{C}$, $\mathcal{A}$, $\ell \geqslant k$, $\text{Curl}_x(\mathcal{C}, \mathcal{A}, k)$ is defined as

$$\text{Curl}_x(\mathcal{C}, \mathcal{A}, k) = \sum_{i=0}^{\ell-1} \hat{f}_x(C_{a_i}, (C_{a_{i+1}}, \cdots, C_{a_{i+k-1}})) \tag{15}$$

For a distribution $d(\mathcal{X})$, $\text{Curl}_d(\mathcal{C}, \mathcal{A}, k)$ is defined as

$$\text{Curl}_d(\mathcal{C}, \mathcal{A}, k) = \mathbb{E}_d[\text{Curl}_x(\mathcal{C}, \mathcal{A}, k)] \tag{16}$$

Lemma 11. *Given $x \in \mathcal{X}, \mathcal{C}, \mathcal{A}, k \geqslant 3$ and $\ell \geqslant k$,*

- $\ell = k$: $\quad Curl_x(\mathcal{C}, \mathcal{A}, k) = 1.$
- $\ell > k$: $\quad 1 < Curl_x(\mathcal{C}, \mathcal{A}, k) < \ell - k + 1.$

Proof: For convenience, let $p_x^{(a_0)}, p_x^{(a_1)}, \cdots, p_x^{(a_{\ell-1})}$ be denoted by a sequence $R = r_0, r_1, \cdots, r_{\ell-1}$ such that $r_i = p_x^{(a_i)}$ $\forall\ i \in [\ell]$, define

$$S_R(\ell, k) = \text{Curl}_x(\mathcal{C}, \mathcal{A}, k) = \sum_{i=0}^{\ell-1} \frac{r_i}{\sum_{j=i}^{i+k-1} r_j}$$

We will prove the statement of the theorem using induction on ℓ.

- Base case ($\ell = k$): $S_R(k, k)$ is given by

$$\sum_{i=0}^{k-1} \frac{r_i}{\sum_{j=i}^{i+k-1} r_j} = \frac{\sum_{j=0}^{k-1} r_j}{\sum_{j=0}^{k-1} r_j} = 1.$$

- Induction assumption
  - *Upper bound:*
  $$S_R(\ell, k) < \ell - k + 1 \quad \forall R, k < \ell \leqslant L.$$
  - *Lower bound:*
  $$S_R(\ell, k) > 1 \quad \forall R, k < \ell \leqslant L.$$
- To prove:
  - *Upper bound:*
  $$S_R(L+1, k) < L - k.$$
  - *Lower bound:*
  $$S_R(L+1, k) > 1.$$

1) *Upper Bound:* Let $r_m = max_{i=0}^{L} r_i$. Consider $U = u_0, u_1, \cdots, u_L$ such that $u_i = r_{i+m}$, i.e. U is a cyclic shift of R by m. Then, it is trivial to observe that

$$S_U(L+1, k) = S_R(L+1, k).$$

$$\begin{aligned} S_U(L+1,k) &= \frac{u_0}{u_0 + \sum_{j=1}^{k-1} u_j} + \sum_{i=1}^{L} \frac{u_i}{\sum_{j=i}^{i+k-1} u_j} \\ &\overset{(a)}{<} 1 + \sum_{i=1}^{L} \frac{u_i}{\sum_{j=i}^{i+k-1} u_j} \\ &= 1 + \sum_{i=1}^{L-k+1} \frac{u_i}{\sum_{j=i}^{i+k-1} u_j} + \sum_{i=L-k+2}^{L} \frac{u_i}{u_0 + \sum_{\substack{j=i \\ j \neq 0}}^{i+k-1} u_j} \\ &\overset{(b)}{\leqslant} 1 + \sum_{i=1}^{L-k+1} \frac{u_i}{\sum_{j=i}^{i+k-1} u_j} + \sum_{i=L-k+2}^{L} \frac{u_i}{u_{i+k} + \sum_{\substack{j=i \\ j \neq 0}}^{i+k-1} u_j} \\ &\overset{(c)}{=} 1 + S_{U'}(L, k) \\ &\overset{(d)}{<} 1 + L - k - 1 \\ &= L - k. \end{aligned}$$

Here (a) follows from $\frac{u_0}{u_0 + \sum_{j=1}^{k-1} u_j} < 1$, (b) follows from $u_0 = max_{i=0}^{L} u_i$, (c) follows by assuming $U' = u_1, u_2, \cdots, u_L$ and (d) follows from the induction assumption for the upper bound.

2) *Lower Bound:* Let $r_t = min_{i=0}^{L} r_i$. Consider $V = v_0, v_1, \cdots, v_L$ such that $v_i = r_{i+t}$, i.e. V is a cyclic shift of R by t. Then, it is trivial to observe that

$$S_V(L+1, k) = S_R(L+1, k).$$

$$\begin{aligned} S_V(L+1,k) &= \frac{v_0}{v_0 + \sum_{j=1}^{k-1} v_j} + \sum_{i=1}^{L} \frac{v_i}{\sum_{j=i}^{i+k-1} v_j} \\ &\overset{(a)}{>} \sum_{i=1}^{L} \frac{v_i}{\sum_{j=i}^{i+k-1} v_j} \\ &= \sum_{i=1}^{L-k+1} \frac{v_i}{\sum_{j=i}^{i+k-1} v_j} + \sum_{i=L-k+2}^{L} \frac{v_i}{v_0 + \sum_{\substack{j=i \\ j \neq 0}}^{i+k-1} v_j} \\ &\overset{(b)}{\geqslant} \sum_{i=1}^{L-k+1} \frac{v_i}{\sum_{j=i}^{i+k-1} v_j} + \sum_{i=L-k+2}^{L} \frac{v_i}{v_{i+k} + \sum_{\substack{j=i \\ j \neq 0}}^{i+k-1} v_j} \\ &\overset{(c)}{=} S_{V'}(L, k) \\ &\overset{(d)}{>} 1 \end{aligned}$$

Here (a) follows from $\frac{v_0}{v_0 + \sum_{j=1}^{k-1} v_j} > 0$, (b) follows from $v_0 = min_{i=0}^{L} v_i$, (c) follows by assuming $V' = v_1, v_2, \cdots, v_L$ and (d) follows from the induction assumption for the lower bound.

■

Lemma 12. *Given $\mathcal{C}, \mathcal{A}, k \geqslant 3$ and $\ell \geqslant k$, then for*

- $\ell = k$: $\quad Curl_d(\mathcal{C}, \mathcal{A}, k) = 1.$
- $\ell > k$: $\quad 1 < Curl_d(\mathcal{C}, \mathcal{A}, k) < \ell - k + 1.$

Proof: The statement follows from Lemma 12 and the definition of $\text{Curl}_d(\mathcal{C}, \mathcal{A}, k)$ given in Equation 16. ∎

B. Proof of Theorem 5

To prove this theorem, we will need to show that the boundaries of the polytopes do not move too far. We will do this using Lemma 13, which bounds how far $\text{Bo}(\text{Co}(V))$ can be from $\text{Bo}(\text{Co}(\tilde{V}))$ along a single "face."

Definition 9. Choose $f \in [h]$. Define:

$$W^{(f)} = \{\mathbf{w} : (\mathbf{a^{(f)}})^\top w = b^{(f)}, \mathbf{w} \in V\}$$
$$\tilde{W}^{(f)} = \{\tilde{\mathbf{v}}_\mathbf{i} : \mathbf{v_i} \in W^{(f)}\}$$

We restrict the size of $|W^{(f)}| = \ell$, which is the number of points needed to define a halfspace in $\mathbb{R}^\ell$. This can be done by allowing for multiple identical $\mathbf{a_f}, b_f$ combinations corresponding to all size l subsets of the v_i along the boundary.

Note that $\text{Co}(W^{(f)})$ describes a "face" of the polytope $\text{Co}(V)$ indexed by f which is perpendicular to $\mathbf{a^{(f)}}$. $\text{Co}(\tilde{W}^{(f)})$ describes the perturbed face.

Lemma 13. *Choose $f, g \in [h]$ arbitrarily and let $W^{(f)} = \{\mathbf{w_1^{(f)}}, \ldots, \mathbf{w_\ell^{(f)}}\}$ and $\tilde{W}^{(f)} = \{\tilde{\mathbf{w}}_\mathbf{1}^{(\mathbf{f})}, \ldots, \tilde{\mathbf{w}}_\ell^{(\mathbf{f})}\}$. For every $\mathbf{m^{(f)}} \in \overline{Co}(W^{(f)})$, we have $(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{m^{(f)}} < \tilde{b}^{(g)} + \varepsilon$.*

Proof: Because $m \in \overline{\text{Co}}(W^{(f)})$, there is some $\lambda \in \triangle_\ell$ with

$$\mathbf{m^{(f)}} = \sum_{i=1}^{\ell} \lambda_i \mathbf{w_i^{(f)}} \in \overline{\text{Co}}(W^{(f)}) \tag{17}$$

Consider also

$$\tilde{\mathbf{m}}^{(\mathbf{f})} = \sum_{i=1}^{\ell} \lambda_i \tilde{\mathbf{w}}_\mathbf{i}^{(\mathbf{f})} \in \overline{\text{Co}}(\tilde{W}^{(f)}) \tag{18}$$

Note that the norm of the difference between these two vectors is bounded:

$$\begin{aligned} \left\|\mathbf{m^{(f)}} - \tilde{\mathbf{m}}^{(\mathbf{f})}\right\|_2 &= \left\|\sum_{i=1}^{\ell} \lambda_i (\mathbf{w_i^{(f)}} - \tilde{\mathbf{w}}_\mathbf{i}^{(\mathbf{f})})\right\|_2 \\ &\leqslant \sum_{i=1}^{\ell} \lambda_i \underbrace{\left\|\mathbf{w_i^{(f)}} - \tilde{\mathbf{w}}_\mathbf{i}^{(\mathbf{f})}\right\|_2}_{<\varepsilon} < \varepsilon \end{aligned} \tag{19}$$

Also note that because $\tilde{\mathbf{m}}^{(\mathbf{f})} \in \overline{\text{Co}}(\tilde{W}^{(f)}) \subseteq \overline{\text{Co}}(\tilde{V})$, we have that $(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \tilde{\mathbf{m}}^{(\mathbf{f})} \leqslant \tilde{b}^{(g)}$. Now, a simple application of Cauchy-Schwartz gives:

$$\begin{aligned} (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{m^{(f)}} &= (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top (\tilde{\mathbf{m}}^{(\mathbf{f})} + (\mathbf{m^{(f)}} - \tilde{\mathbf{m}}^{(\mathbf{f})})) \\ &= \underbrace{(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \tilde{\mathbf{m}}^{(\mathbf{f})}}_{\leqslant \tilde{b}^{(g)}} + (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top (\mathbf{m^{(f)}} - \tilde{\mathbf{m}}^{(\mathbf{f})}) \\ &\leqslant \tilde{b}^{(g)} + \left\|\tilde{\mathbf{a}}^{(\mathbf{g})}\right\|_2 \left\|\mathbf{m^{(f)}} - \tilde{\mathbf{m}}^{(\mathbf{f})}\right\|_2 \\ &< \tilde{b}^{(g)} + \varepsilon \end{aligned} \tag{20}$$

∎

With this, we are now ready to prove Theorem 5.

Proof: Choose an arbitrary face $g \in [h]$. Recall we have $\mathbf{x} \in \text{Co}(V)$ with $(\mathbf{a^{(g)}})^\top \mathbf{x} < b - \varepsilon$ and we wish to show $(\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{x} < \tilde{b}^{(g)}$.

Let $\mathbf{m_x^{(f)}}$ be the result of extending $\tilde{\mathbf{a}}^{(\mathbf{g})}$ from $\mathbf{x}$ to $\text{Bo}(V)$. This must hit some face with $(\mathbf{a^{(f)}})^\top \mathbf{m_x^{(f)}} = b^{(f)}$, so $\mathbf{m_x^{(f)}} \in \text{Co}(W^{(f)})$. That is, find β such that

$$\mathbf{m_x^{(f)}} = \beta \tilde{\mathbf{a}}^{(\mathbf{g})} + \mathbf{x} \in \text{Co}(W^{(f)}) \tag{21}$$

First, lets bound β. Notice that because $\mathbf{m_x^{(f)}} \in \text{Co}(W^{(f)})$, we have

$$\begin{aligned} (\mathbf{a^{(f)}})^\top \mathbf{m_x^{(f)}} &= (\mathbf{a^{(f)}})^\top \left(\sum_{i=1}^{\ell} \lambda_i \mathbf{w_i^{(f)}}\right) \\ &= \sum_{i=1}^{\ell} \lambda_i (\mathbf{a^{(f)}})^\top \mathbf{w_i^{(f)}} = b^{(f)} \end{aligned} \tag{22}$$

So, we have

$$b^{(f)} = (\mathbf{a^{(f)}})^\top \mathbf{m_x^{(f)}} = \beta \underbrace{(\mathbf{a^{(f)}})^\top \tilde{\mathbf{a}}^{(\mathbf{g})}}_{\leqslant 1} + \underbrace{(\mathbf{a^{(f)}})^\top \mathbf{x}}_{<b^{(f)}-\varepsilon} \Rightarrow \varepsilon < \beta \tag{23}$$

Now, apply Lemma 13

$$\begin{aligned} (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{m_x^{(f)}} &< \tilde{b}^{(g)} + \varepsilon \\ (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{x} + (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \tilde{\mathbf{a}}^{(\mathbf{g})} \beta &< \tilde{b}^{(g)} + \varepsilon \\ (\tilde{\mathbf{a}}^{(\mathbf{g})})^\top \mathbf{x} &< \tilde{b}^{(g)} \end{aligned} \tag{24}$$

Recall we chose face $g \in [h]$ arbitrarily, so this holds for all halfspaces in the convex polytope. Hence, we have $A\mathbf{x} \prec \mathbf{b}$. ∎

C. Existence Proof for Decomposing Cycles into Acyclic Orientations

We can show that a decomposition of any curl consistent cycle into acyclic orientations must exist. Consider $\mathbf{v}$ the vector of edge-weights of an arbitrary cycle with $v^{(i)} = \hat{f}_d^{(C_i, C_{i+1})}$. Acyclic orientations in this framework correspond to the set $G = \{0,1\}^\ell \setminus \{\mathbf{0}_\ell, \mathbf{1}_\ell\}$.

Lemma 14. *Given $\mathbf{v} \in (0,1)^\ell$, such that $1 < \|\mathbf{v}\|_1 < \ell - 1$, then for some $d(\cdot) : G \mapsto (0,1)$ in $\triangle_{2^\ell - 2}$, we have*

$$\mathbf{v} = \sum_{\mathbf{t} \in G} d(\mathbf{t})\mathbf{t} \tag{25}$$

Proof: Consider $H = \{\mathbf{h} \in [0,1]^\ell : 1 \leqslant \|\mathbf{h}\|_1 \leqslant \ell - 1\}$. Note that $\mathbf{v} \in H$. Since G and H are closed and bounded sets, they are compact. Further, H is convex. Therefore, $\text{Co}(G)$ is also convex and compact using compactness and Carathéodory's theorem [3]. We can now prove the required statement by showing $H = \overline{\text{Co}}(G)$.

1) $\overline{Co}(G) \subset H$: Let $\mathbf{y} \in \overline{\text{Co}}(G)$, then $\mathbf{y} = \sum_{\mathbf{t} \in G} d(\mathbf{t})\mathbf{t}$, for some $d(\cdot) : G \mapsto (0,1)$ in $\triangle_{2^\ell - 2}$. We notice that $\mathbf{y} \in [0,1]^\ell$ and $\|\mathbf{y}\|_1 \in [1, \ell - 1]$. Therefore $\mathbf{y} \in H$ and hence $\overline{\text{Co}}(G) \subset H$.

2) $H \subset \overline{Co}(G)$: Let E be the set of extreme points of H. Since H is convex and compact, we can use the Krein-Milman Theorem [3] to get $H = \overline{\text{Co}}(E)$. Further, we notice that $E \subset G$ [3], therefore $H \subset \overline{\text{Co}}(G)$.

■

D. Decomposing Cycles into Acyclic Orientations

For this section, consider a knowledge cycle with edges $C_0 \to C_1 \to \ldots \to C_{\ell-1} \to C_0$. Let the edge weights be represented by a vector $\mathbf{f_0} \in (0,1)^\ell$ with $f_0^{(i)} = f_d^{(C_i, C_{i+1 \bmod \ell})}$. We will refer to *not fully decomposed* components as vectors $\mathbf{f_j} \in [0,1]^\ell$ and all *orientation* components as $\mathbf{t_j} \in \{0,1\}^\ell$ with weight $d(\mathbf{t_j})$. We will also use $\mathbf{1}_\ell[I] \in \{0,1\}^\ell$ to denote a vector for which $\mathbf{1}_\ell[I]^{(i)} = \mathbb{1}[i \in I]$.

We begin by observing that all $t_j \in \{0,1\}^\ell$ are allowed other than $\mathbf{1}_\ell$ and the origin.

Observation 15. *We can decompose a scaled version of $\mathbf{1}_\ell$ as follows:*

$$\left(1 - \frac{1}{\ell}\right)\mathbf{1}_\ell = \sum_{i=1}^{\ell-1} \frac{\mathbf{1}_\ell[\{j : j \neq i\}]}{\ell} = \sum_{i=1}^{\ell-1} \frac{\mathbf{1}_\ell[-\{j\}]}{\ell} \tag{26}$$

Observation 16. *If $\|\mathbf{f_j}\|_1 = 1$, we can decompose:*

$$\mathbf{f_j} = \sum_{i=1}^{\ell-1} f_j^{(i)} \mathbf{1}_\ell[\{i\}] \tag{27}$$

Definition 10. We define the support of vector $\mathbf{f_j} \in [0,1]^\ell$ to be the set of nonzero indices.

$$\text{Supp}(\mathbf{f_j}) = \{i : \mathbf{f_j}^{(i)} > 0]\} \tag{28}$$

Lemma 17. *Let γ be a factor for handling the case of full support:*

$$\gamma = 1 - \frac{\mathbb{1}[|S_j| = \ell]}{\ell} \tag{29}$$

Given $\mathbf{f_j} \in [0,1]^\ell$ with Supp$(\mathbf{f_j}) = S_j$ and $\|\mathbf{f_j}\|_1 \in (1, \gamma|S_j|)$, we can decompose

$$\mathbf{f_j} = z_j \gamma \mathbf{1}_\ell[S_j] + (1 - z_j)\mathbf{f_{j+1}} \tag{30}$$

where either $\|\mathbf{f_{j+1}}\|_1 = 1$ or
(i) $|Supp(\mathbf{f_{j+1}})| \leqslant |S_j| - 1$.
(ii) $\|\mathbf{f_{j+1}}\|_1 \in (1, |S_j| - 1)$.

Proof: We begin by first observing

$$f_{j+1}^{(i)} = \frac{f_j^{(i)} - z_j \gamma \mathbb{1}[i \in S_j]}{1 - z_i} \leqslant \frac{f_j^{(i)} - z_j \mathbb{1}[i \in S_j]}{1 - z_i} \tag{31}$$

Recall that if $f_j^{(i)} > 0$, then $i \in S_j$. Hence $\mathbf{f_{j+1}} \in [0,1]^\ell$. We notice that the upper bound of (ii) $\|\mathbf{f_{j+1}}\|_1 \leqslant |S_j| - 1$ now follows from (i). It remains to find z_j such that $\|\mathbf{f_{j+1}}\|_1 \geqslant 1$ and if $\|\mathbf{f_{j+1}}\|_1 > 1$ then (i) holds.

$$\begin{aligned} \zeta_a &= \frac{1}{\gamma} \min_{i \in S_j} f_j^{(i)} \\ \zeta_b &= \frac{\|\mathbf{f_j}\|_1 - 1}{\gamma |S_j| - 1} \\ z_j &= \min(\zeta_a, \zeta_b) \end{aligned} \tag{32}$$

We can apply the Pigeonhole principle to $\sum_{i \in S_j} f_j^{(i)} \leqslant \ell - 1$ to get $z_j \leqslant \zeta_a \leqslant 1$. If $z_j = 1$ then the second term of the decomposition is irrelevant. We can treat the L1 norm as a linear function when the vectors are all the same sign giving:

$$\|\mathbf{f_{j+1}}\|_1 = \frac{\|\mathbf{f_j}\|_1 - z_j \gamma |S_j|}{1 - z_j} \tag{33}$$

Now consider two cases:

1) ($z_j = \zeta_b$) Notice ζ_b has been carefully chosen so that,

$$\|\mathbf{f_{j+1}}\|_1 = \frac{\|\mathbf{f_j}\|_1 - \zeta_b \gamma |S_j|}{1 - \zeta_b} = 1. \tag{34}$$

2) ($z_j = \zeta_a$) Notice that if $f_j^{(i^*)} = \min_{i \in S_j} f_j^{(i)}$, then $f_{j+1}^{(i^*)} = 0$, which satisfies condition (i).
To show the lower bound of (ii), rewrite Equation 33:

$$\|\mathbf{f_{j+1}}\|_1 = \|\mathbf{f_j}\|_1 \frac{1 - z_j \left(\frac{\gamma |S_j|}{\|\mathbf{f_j}\|_1}\right)}{1 - z_j} \tag{35}$$

Recall that $\|\mathbf{f_j}\|_1 < \gamma |S_k|$, so $\frac{\gamma |S_j|}{\|\mathbf{f_j}\|_1} > 1$. So, if Equation 35 gives $\|\mathbf{f_{j+1}}\|_1 = 1$ with $z_j = \zeta_b$, then $z_j \leqslant \zeta_b$ gives $\|\mathbf{f_{j+1}}\|_1 \geqslant 1$.

■

Theorem 18. *The curl condition is sufficient for knowledge cycles.*

Proof: There are two methods for proving the required statement which are described below.

- *Method 1:* The proof follows directly from Lemma 14 and Theorem 6.
- *Method 2:* There is also an alternative proof using Lemma 17 which is as follows. Begin with arbitrary vector $\mathbf{f_0}$ with $\|\mathbf{f_0}\|_1 \in (1, \ell - 1)$. Use Lemma 17 to decompose

$$\mathbf{f_0} = z_0 \gamma \mathbf{1}_\ell[S_0] + (1 - z_0)\mathbf{f_1} \tag{36}$$

If $\mathbf{f_0}$ had full support, then $\gamma = 1 - \frac{1}{\ell - 1}$ and $\mathbf{1}_\ell[S_0] = \mathbf{1}_\ell$. Thus, we can use Observation 15 to decompose $\gamma \mathbf{1}_\ell$ into acyclic orientations. If Supp$(\mathbf{f_0}) < \ell$, then $\gamma = 1$ and $\mathbf{1}_\ell[S_0]$ already represents an acyclic orientation.
We repeatedly apply Lemma 17. In this process if we reach some $\mathbf{f_j}$ with $\|\mathbf{f_j}\|_1 = 1$, we can apply Observation 16 to decompose $\mathbf{f_j}$ into acyclic orientations. If we do not, we will terminate at $\mathbf{f}_\ell$ with $|S_\ell| = 0$. See Figure 4 for an example of this decomposition.
This gives a decomposition into acyclic orientations, which allows the application of Theorem 6.

■

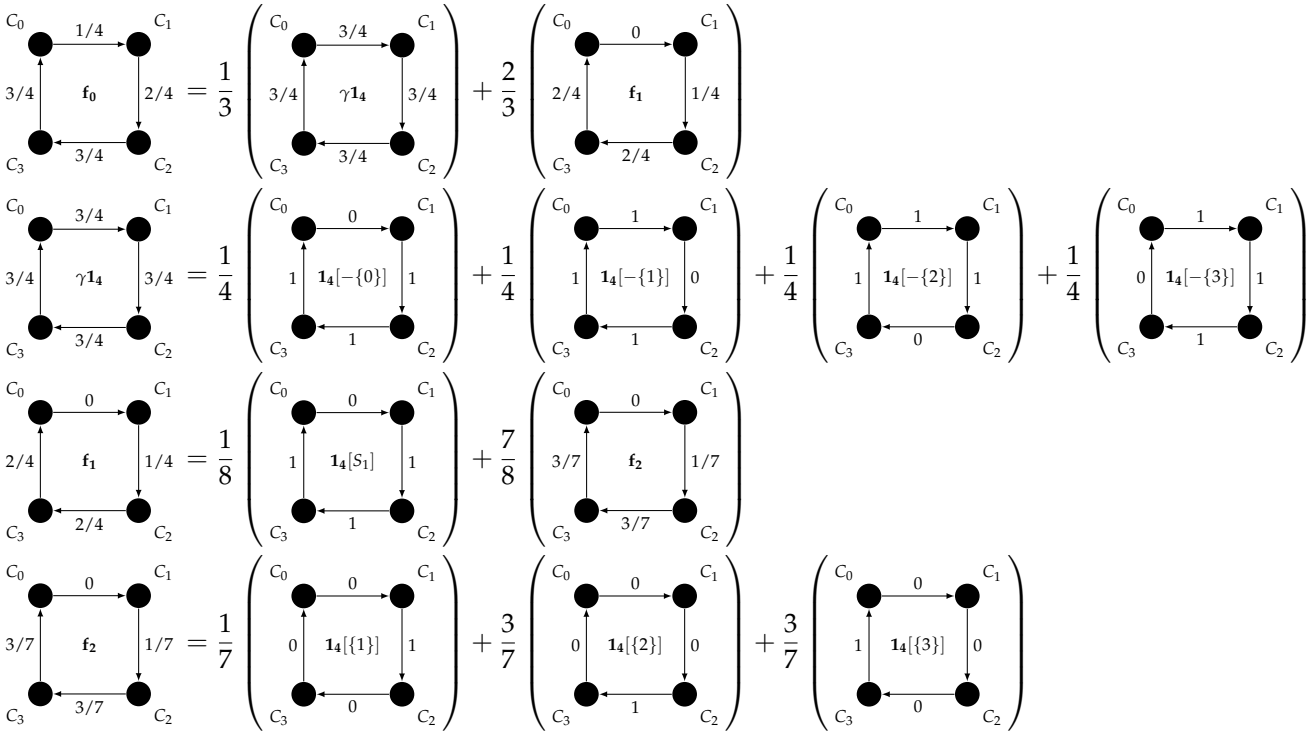

Fig. 4. Decomposing a knowledge $\mathbf{f_0}$ following the procedure given in Appendix Section D. The first line pulls off a scaled $\mathbf{1_4}$. The second line decomposes this by Observation 15. The third line decomposes $\mathbf{f_1}$ into $\mathbf{1_4}[S_1]$ and $\mathbf{f_2}$ with $\|f_2\|_1 = 1$. The fourth line decomposes $\mathbf{f_2}$ according to Observation 16.