Optimal Cooperative Inference

Scott Cheng-Hsin Yang, Yue Yu, Arash Givchi, Pei Wang, Wai Keen Vong, and Patrick Shafto

Department of Mathematics & Computer Science, Rutgers University—Newark

Abstract

Cooperative transmission of data fosters rapid accumulation of knowledge by efficiently combining experience across learners. Although well studied in human learning, there has been less attention to cooperative transmission of data in machine learning, and we consequently lack strong formal frameworks through which we may reason about the benefits and limitations of cooperative inference. We present such a framework. We introduce a novel index for measuring the effectiveness of probabilistic information transmission, and cooperative information transmission specifically. We relate our cooperative index to previous measures of teaching in deterministic settings. We prove conditions under which optimal cooperative inference can be achieved, including a representation theorem which constrains the form of inductive biases for learners optimized for cooperative inference. We conclude by demonstrating how these principles may inform the design of machine learning algorithms and discuss implications for human learning, machine learning, and human-machine learning systems.

1 Introduction

Human learning is characterized by the cooperative transmission of data [21]. In addition to direct observations and taking actions in one’s own environment, humans also engage in purposeful selection of data whose goal is conveying knowledge about the world to less knowledgeable agents. Moreover, less knowledgeable agents assume purposeful, cooperative selection and leverage cooperation to augment learning. The cooperative selection of data, and learning from such data, plays a central role in theories of cognition [1], cognitive development [11], and cultural evolution [22]. Indeed, this cooperative inference is argued to be the feature that drives accumulation of knowledge over generations [23, 6].

Such communication through cooperative inference has received relatively limited attention in machine learning. The principles that appear to drive human cultural knowledge accumulation may also be leveraged in machines to achieve similar ends. A recent effort in this direction is machine teaching [25, 10], which formalizes how to translate and communicate model inferences using data. While this development begins to address how communication between learning models could occur, it is mute on the effectiveness of communication between models, and does not formalize cooperative inference by both the teacher and the learner.

In this paper, we address this lack by introducing a measure of communication effectiveness in the cooperative setting. The role that this measure plays in cooperative knowledge accumulation is analogous to the role that training and test errors play in traditional machine learning. We also use the measure to extend the teaching dimension [12, 26]—a classical measure of communication efficiency—from deterministic to probabilistic settings. We show how analyzing this measure reveals the conditions, in terms of constraints on learning model’s inductive biases, under which cooperation may produce optimal communication.

The paper is organized as follows: In Section 2 we first introduce a Transmission Index that quantifies communication effectiveness for any pair of probabilistic inference and data selection processes. In Section 3

*scott.cheng.hsin.yang@gmail.com
†patrick.shafto@gmail.com
1Effectiveness is a measure of the quality of communication; efficiency is the size of the data necessary to reach a particularly effectiveness.
we make connection between this index and the Average Teaching Dimension, thereby connecting our measure of effectiveness with previous measures of efficiency. In Sections 4 and 5 we introduce cooperative inference based on previous research in human social learning [15, 17], present a Cooperative Index by extend the Transmission index to the cooperative setting, and identify the condition that must be satisfied to achieve optimal communication. In Section 5 we conclude with connections to human, machine, and human-machine learning.

2 The Transmission Index

In this section, we define Transmission Index to quantify the communication effectiveness. Communication occurs between two agents, which we call a teacher and a learner. Here the teacher represents the process of selecting data to convey a particular concept, and the learner represents the inference process of interpreting the received data. In a probabilistic setting, the effectiveness of communication is related to the probability that the learner’s interpretation matches the teacher’s intended concept.

**Definition 2.1.** Let \( h \) be a concept in a finite concept space \( \mathcal{H} \). A data set space, \( \mathcal{D} \), is a collection of subsets of a given finite set of data points. \( D \in \mathcal{D} \) is called a data set. Further, let \( P_T(D|h) \) be the teacher’s probability of selecting a data set \( D \) for communicating a given concept \( h \) and \( P_L(h|D) \) be the learner’s posterior for \( h \) given data set \( D \). We denote the size of \( \mathcal{H} \) and \( \mathcal{D} \) by \( |\mathcal{H}| \) and \( |\mathcal{D}| \), respectively.

Because \( \mathcal{H} \) and \( \mathcal{D} \) are both discrete, in matrix notation, we can form the row-stochastic learner’s inference matrix, \( L \in [0, 1]|\mathcal{D}| \times |\mathcal{H}| \), having elements \( P_L(h|D) \), and the column-stochastic teacher’s selection matrix, \( T \in [0, 1]|\mathcal{D}| \times |\mathcal{H}| \), having elements \( P_T(D|h) \).

**Definition 2.2.** The Transmission Index (TI) is defined as

\[
TI(L, T) = \frac{1}{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{D}|} L_{i,j} T_{i,j}.
\]

In connection to information theory, the Transmission Index can be understood as a measure of the effectiveness of information through coding information, sending the code, and then decoding. In our context, the channel is not arbitrary but is restricted to be the data, i.e., information is transmitted by passing data. Similarly, the decoding and coding are not arbitrary but correspond to the inference process and data selection process, and does not need to be agreed upon beforehand.

Now we give a few examples to show that TI captures how well on average a concept in a given concept space can be communicated with a given data set space. Also, note that in the case where \( \mathcal{H} \) and \( \mathcal{D} \) are clear from the context, we represent \( TI(L, T) \) simply by TI.

**Example 2.3.** Let \( |\mathcal{D}| = |\mathcal{H}| = 2 \). Consider this teacher’s selection matrix, \( T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \), and these three learner’s inference matrices, \( L^{(a)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \), \( L^{(b)} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \), and \( L^{(c)} = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \).

In the first case (a), \( TI(L^{(a)}, T) = 1 \), because the concept that the teacher intends to teach through a certain data set matches perfectly what the learner would infer given that data set. In the second case (b), \( TI(L^{(b)}, T) = 0 \), because the concept that the teacher intends to teach through a certain data set leads the learner to infer the other concept with certainty. In the last case (c), \( TI(L^{(c)}, T) = \frac{1}{2} \). Here the learner’s inference is ambiguous, and TI captures that. In summary, TI captures the expected probability that the learner will interpret the teacher’s intention correctly.

**Proposition 2.4.** The range of the Transmission Index is \( 0 \leq TI \leq 1 \), and TI = 1 if and only if two conditions hold: (i) \( L_{i,j} = 1 \) if \( T_{i,j} > 0 \) for all \( i, j \), and (ii) there is no zero column in \( L \) and \( T \). Also, TI = 1 implies that \( |\mathcal{D}| \geq |\mathcal{H}| \), with equality achieved when \( L \) and \( T \) are the same permutation matrix.

**Proof.** TI \( \geq 0 \) because \( L \) and \( T \) are stochastic matrices, and TI = 0 if and only if for any \( i, j \), either \( L_{i,j} = 0 \) or \( T_{i,j} = 0 \).
We show $\text{TI} \leq 1$:

$$
\text{TI}(\mathbf{L}, \mathbf{T}) = \frac{1}{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{D}|} \mathbf{L}_{i,j} \mathbf{T}_{i,j} \leq \frac{1}{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} \left( \sum_{i=1}^{|\mathcal{D}|} \mathbf{T}_{i,j} \right) \leq \frac{1}{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} 1 = 1. \quad (1)
$$

Inequality (a) in (1) becomes an equality if and only if condition (i) is satisfied. This is because in order for $\mathbf{L}_{i,j} \mathbf{T}_{i,j} = \mathbf{T}_{i,j}$, we need $(\mathbf{L}_{i,j} - 1) \mathbf{T}_{i,j} = 0$, and this implies that $\mathbf{L}_{i,j} = 1$ or $\mathbf{T}_{i,j} = 0$, for any $i,j$. Inequality (b) in (1) follows from $\mathbf{T}$ being a column-stochastic matrix, and it becomes an equality if and only if condition (ii) is satisfied.

Given that $\mathbf{L}$ is a row-stochastic matrix, if $\mathbf{L}_{i,j} = 1$, then there is no other non-zero elements in row $i$. This means that there are at most $|\mathcal{D}|$ elements with value one in $\mathbf{L}$; hence, by condition (i) the number of non-zero elements in $\mathbf{T}$ is at most $|\mathcal{D}|$. Also, condition (ii) requires that the number of non-zero elements in $\mathbf{T}$ be at least $|\mathcal{H}|$. Therefore, $|\mathcal{D}| \geq |\mathcal{H}|$, with equality achieved if and only if $\mathbf{T}$ has only one positive element for each column. Together with condition (i), this implies that $\mathbf{L}$ has at least one element with value one in each column. Because $\mathbf{L}$ is row-stochastic, this implies $\mathbf{L}$ is a permutation matrix. Condition (i) also implies that if $\mathbf{L}_{i,j} < 1$, then $\mathbf{T}_{i,j} = 0$. Together with condition (ii), $\mathbf{T}$ is the same permutation matrix.

**Remark 2.5.** $\text{TI}$ is invariant under joint row and column permutations of $\mathbf{L}$ and $\mathbf{T}$. When $|\mathcal{H}| = |\mathcal{D}|$ and $\text{TI} = 1$, row and column exchangeability implies that $\mathbf{L}$ and $\mathbf{T}$ can always be arranged into an identity matrix of order $|\mathcal{H}|$.

**Remark 2.6.** $\text{TI}$ is well-defined even when one of $\mathcal{H}$ and $\mathcal{D}$ is countably infinite; that is, the definition of $\text{TI}$ can be modified to

$$
\text{TI}(\mathbf{L}, \mathbf{T}) := \lim_{n \to |\mathcal{H}|} \lim_{m \to |\mathcal{D}|} \frac{1}{n} \sum_{j=1}^{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{D}|} \mathbf{L}_{i,j} \mathbf{T}_{i,j},
$$

and the limit exists in those two cases. For the case where $|\mathcal{D}| \to \infty$ and $|\mathcal{H}|$ is finite, one can show that the above limit exists because $\text{TI}$ increases with $m$ and has an upper bound. For the case where $|\mathcal{H}| \to \infty$ and $|\mathcal{D}|$ is finite, $\text{TI}(\mathbf{L}, \mathbf{T}) = 0$ because there is always infinitely many untaught concepts given finitely many examples. However, when both $\mathcal{H}$ and $\mathcal{D}$ are countably infinite, $\text{TI}$ is not generally well-defined. In particular, it is possible to construct a pair of $\mathbf{L}$ and $\mathbf{T}$ that causes the limit of $\text{TI}$ to bounce in the interval $[0,1]$. See Supplementary Material for full detail.

### 3 Connection to Average Teaching Dimension

In this section, we make the connection between the Transmission Index and the Average Teaching Dimension. The Average Teaching Dimension is a variant of Teaching Dimension, a classic measure for quantifying the efficiency of teaching. The Teaching Dimension is well-studied; it has formal connections with the VC Dimension [12] and has been analyzed for certain models in continuous concept space [13] and in cooperative settings [26] [8]. However, Teaching Dimension and these analyses assume a deterministic the learning model and focus on efficiency rather than effectiveness. To make connection to the analysis of Teaching Dimension, we first extend the Transmission Index, a measure of effectiveness, to the Expected Teaching Dimension, a measure of efficiency. Then we show that the Expected Teaching Dimension, which is well-defined for probabilistic knowledge transmission, is the same as the Average Teaching Dimension as the knowledge transmission becomes deterministic.

The analyses of Teaching Dimension are typically couched in the concept learning framework. In this framework, a concept, $h$, is a function that maps an instance, $x$, to a label, $y$. By observing examples, pairs of $(x,y)$, the learner can rule out concepts that are not consistent with the examples. With this notation, we can define the Average Teaching Dimension:

**Definition 3.1 (Average Teaching Dimension).** A concept $h \in \mathcal{H}$ is consistent with a data set $\mathcal{D}$ if and only if for every data point $(x,y) \in \mathcal{D}$, $h(x) = y$. $D \subseteq \mathcal{D}$ is a teaching set for concept $h \in \mathcal{H}$ if $h$, but no other concept in $\mathcal{H}$, is consistent with $D$. Let $\mathcal{D}(h) \subseteq \mathcal{D}$ be the collection of teaching sets in $\mathcal{D}$ for concept $h$. The classical version of Average Teaching Dimension [8] is defined as follows: First, for any $h \in \mathcal{H}$, let
\[ T\!D(h) = \begin{cases} \infty & \text{if } D^*(h) \text{ is empty}, \\ \min_{D \in D^*(h)} |D| & \text{otherwise} \end{cases} \]

where \(|D|\) is the size of the data set \(D\). Then, the Average Teaching Dimension (ATD) for the concept space \(\mathcal{H}\) is

\[ ATD(\mathcal{H}) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} T\!D(h). \]

Expected Teaching Dimension extends the Transmission Index to incorporate data set size as follows:

**Definition 3.2.** The Expected Teaching Dimension (ETD) is defined as

\[ ETD(\mathcal{H}) = \frac{\sum_{h \in \mathcal{H}} \sum_{D \in D^*(h)} |D| \cdot P_L(h|D) \cdot P_T(D|h) \cdot P_{\alpha}(D|h)}{\sum_{h \in \mathcal{H}} \sum_{D \in D^*(h)} P_L(h|D) \cdot P_T(D|h)}. \]

**Definition 3.3.** Let \(M \in [0,1]^{|D| \times |\mathcal{H}|}\) be a matrix, where the element \(M_{i,j}\) represents the probability that \(h_i\) is consistent with \(D_j\). We define \(C \in \{0,1\}^{D \times |\mathcal{H}|}\) to be a consistency matrix, where \(C_{i,j} = 1\) if \(h_i\) is consistent with \(D_j\) and \(C_{i,j} = 0\) otherwise. \(C\) can be sampled from \(M\) by treating \(C_{i,j}\) as the outcome of a Bernoulli trial with parameter \(M_{i,j}\).

**Proposition 3.4.** Let \(|\mathcal{H}| = |D| = N\), and \(C\) be a consistency matrix of size \(N \times N\). Let \(L\) and \(T\) be the row-normalized and column-normalized matrices of \(C\), respectively. Then, \(ATD(\mathcal{H})\) is finite if and only if \(TI(L,T) = 1\).

**Proof.** \(ATD(\mathcal{H})\) is finite if and only if \(T\!D(h)\) is finite for all \(h \in \mathcal{H}\). Finite \(T\!D(h)\) means that there is at least one teaching set \(D \in D\) for \(h\). Let \(\alpha_i \in \{1,2,\ldots,N\}\) be the index set for the teaching sets of \(h_i\). Because every \(D\) can only belong to at most one \(D^*(h_i)\), so \(\alpha_i \subseteq \{1,\ldots,N\}\) \(\cup_{j \neq i} \alpha_j\) for every \(i \in \{1,2,\ldots,N\}\). Further because \(|D| = |\mathcal{H}|\), this construction of \(\alpha_i\) implies that if \(|\alpha_i| > 1\) for some \(i\), then there must exist at least one \(j \neq i\) with the property that \(|\alpha_j| = 0\). However, because \(T\!D(h_i)\) is finite, \(\alpha_i\) cannot be an empty set for any \(i\). Hence, \(|\alpha_i| = 1\) for all \(i\). In particular, this implies that \(C\) is a permutation matrix. Thus, \(ATD(\mathcal{H})\) is finite if and only if \(C\) is a permutation matrix. \(C\) being a permutation matrix implies that \(C = L = T\), which by Proposition [2.3] is equivalent to \(TI(L,T) = 1\). \(\Box\)

**Example 3.5.** If \(L = T\) and is a permutation matrix, \(C = T\). As we proved in Proposition [3.4] \(ETD\) is the same as \(ATD\).

**Example 3.6.** We give an example when \(ETD\) is finite but \(ATD\) is infinite in the probabilistic setting. Let \(|\mathcal{H}| = |D| = 2\), \(M = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}\). There are four possible consistency matrices that can be sampled from \(M\): \(C^{(a)} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\), \(C^{(b)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\), \(C^{(c)} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}\), \(C^{(d)} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\). For \(C^{(a)}, C^{(c)}\) and \(C^{(d)}\), the corresponding \(ATD(\mathcal{H})\) is \(\infty\), and for \(C^{(b)}\) it is \(\frac{|D_1| + |D_2|}{2}\). Let \(L^{(s)}\) and \(T^{(s)}\) be the row-normalized and column-normalized matrices of \(C^{(s)}\), respectively, for \(s \in \{a,b,c,d\}\). Then, \(TI(L^{(a)}, T^{(a)}) = \frac{1}{7}, TI(L^{(b)}, T^{(b)}) = 1, TI(L^{(c)}, T^{(c)}) = \frac{1}{2},\) and \(TI(L^{(d)}, T^{(d)}) = \frac{5}{8}\), with \(ETD(\mathcal{H}) = |D_1|, \frac{|D_1| + |D_2|}{2}, |D_1|, \frac{3|D_1| + 2|D_2|}{5}\), respectively. Thus, \(ETD\) can be seen as an generalization of \(ATD\) from scenarios of perfect transmission \((TI = 1)\) to those of imperfect transmission \((0 \leq TI \leq 1)\) as well.

In Definition [3.3] probabilistic transmission is expressed as the sampling of consistency matrices. This corresponds to the view that there is a distribution of learning scenarios and the learner can be uncertain which scenario is selected. Another way probabilistic transmission can enter is that \(M\) represents the degree of consistency between data and hypotheses. In this case, the learner would need to make a decision on what the underlying true consistency matrix is. Consider \(M = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}\) again. A simple decision rule is to round up \(M_{i,j}\) to 1 if it exceeds a threshold and down to 0 otherwise. This decision rule would result in either \(C^{(a)}\) or \(C^{(d)}\), both of which correspond to \(ATD = \infty\).
4 Optimal cooperative inference

The Transmission Index introduced in Section 2 assumes that the learner and teacher, or more abstractly, the inference process and the data selection process, are independent. However, communication for the transmission of knowledge is often cooperative (e.g., in pedagogy [9, 10] and conversations [13]). Here, cooperation implies that the teacher’s selection of data depends on what the learner is likely to infer and vice versa. In this section, we formalize cooperative inference, which captures this inter-dependency between the two processes of inference and selection [15, 17]. It can be seen as a way to map a common convention to another one that is more effective at transmitting knowledge without a priori agreement on the encoding of data-concept pairs [26]. We define Cooperative Index as a measure of communication effectiveness in the cooperative setting by applying the Transmission Index to cooperative inference. Finally, we provide proofs regarding the form of the shared likelihood matrix required to maximize the cooperative index and optimize cooperative inference.

Definition 4.1. [Cooperative inference] Let \( D \in \mathcal{D} \) and \( h \in \mathcal{H} \), we define cooperative inference as a system of two equations:

\[
P_L(h|D) = \frac{P_T(D|h) P_{L_0}(h)}{P_T(D)}, \tag{2a}
\]

\[
P_T(D|h) = \frac{P_L(h|D) P_{T_0}(D)}{P_T(h)}, \tag{2b}
\]

where \( P_L(h|D) \) and \( P_T(D|h) \) are defined in Definition 2.1. \( P_{L_0}(h) \) is the learner’s prior of \( h \), \( P_{T_0}(D) \) is the teacher’s prior of selecting \( D \), \( P_L(D) = \sum_{h \in \mathcal{H}} P_T(D|h) P_{L_0}(h) \) is the normalizing constant for \( P_L(h|D) \), and \( P_T(h) = \sum_{D \in \mathcal{D}} P_L(h|D) P_{T_0}(D) \) is normalizing constant for \( P_T(D|h) \).

The cooperative inference equations in (2) can be solved using fixed-point iteration [15, 17]: Define an initial likelihood, or common convention, \( P_T(D|h) = P_D(h|D) \), for the first evaluation of (2a). Then, given \( P_{L_0}(h) \) and \( P_{T_0}(D) \), one can evaluate (2a), use the resulting \( P_{L}(h|D) \) to evaluate (2b), and iterate until convergence. By symmetry, the iteration can also begin with (2b). This symmetry implies that the initial likelihood matrix, \( \mathbf{M} \in [0, 1]^{\mathcal{D} \times \mathcal{H}} \) with elements \( P_D(D|h) \), can be an arbitrary non-negative matrix because it always gets appropriately normalized in the first iteration.

For the remainder of the paper, we assume that \( P_{L_0} \) and \( P_{T_0} \) are uniform distributions over \( \mathcal{H} \) and \( \mathcal{D} \), respectively. In this case, the the fixed-point iteration of (2) depends only on \( \mathbf{M} \) and is simply the repetition of column and row normalization on \( \mathbf{M} \). Without loss of generality, we also assume that the iteration begins with (2a).

Definition 4.2. Let \( \mathbf{L}^{(k)} \) and \( \mathbf{T}^{(k)} \) be the matrices with elements \( P_L(h|D) \) and \( P_T(D|h) \), respectively, at the \( k \)th iteration of (2). If the iteration of (2) converges, we define \( L^{(\infty)} := \lim_{k \to \infty} \mathbf{L}^{(k)} \) and \( T^{(\infty)} := \lim_{k \to \infty} \mathbf{T}^{(k)} \).

Definition 4.3 (Cooperative Index, CI). Given \( \mathbf{M} \) and assuming that the fixed-point iteration of (2) converges, we define the cooperative index as

\[
\text{CI}(\mathbf{M}) = T_I(L^{(\infty)}, T^{(\infty)}) = \frac{1}{\mathcal{H}} \sum_{j=1}^{\mathcal{H}} \sum_{i=1}^{\mathcal{D}} \mathbf{L}_{i,j}^{(\infty)} \mathbf{T}_{i,j}^{(\infty)}.
\]

We further assume that \( \mathbf{M} \) is a square matrix unless otherwise stated. Then, the iteration of (2) becomes the well-known Sinkhorn-Knopp algorithm, which provably converges under certain conditions by Sinkhorn’s theorem [19]. Below, we state a simpler version of Sinkhorn’s theorem.

Definition 4.4 (Positive diagonal). If \( \mathbf{M} \) is a \( n \times n \) square matrix and \( \sigma \) is a permutation of \( \{1, \cdots, n\} \), then a sequence of positive elements \( \{\mathbf{M}_{i, \sigma(i)}\}_{i=1}^{n} \) is called a positive diagonal. If \( \sigma \) is the identity permutation, the diagonal is called the main diagonal.

Theorem 4.5 (A simpler version of Sinkhorn’s theorem [19]). Given any nonnegative square matrix \( \mathbf{M} \) with at least one positive diagonal, \( \mathbf{L}^{(k)} \) and \( \mathbf{T}^{(k)} \) in the fixed-point iteration of (2) converges to the same doubly stochastic matrix, \( \mathbf{M}^{(\infty)} \), as \( k \to \infty \).
Proof. Here we provide a sketch of the proof (see Supplementary Material for full detail). We pick one positive diagonal. First we show the product of all elements on that diagonal is positive and upper-bounded by $1$ throughout the fixed-point iteration of (2). Given uniform priors on both hypothesis and data set space, we then use inequality of arithmetic and geometric means to prove that the product either stay the same or increase throughout the iteration. Finally, monotone convergence theorem of real numbers guarantees that the product will converge to its supremum, at which point $L$ and $T$ must have converged to the same doubly stochastic matrix.

As is for $TI$, if $M$ is clear from the context, we denote $CI(M)$ simply by $CI$ for brevity. Now, we give two simple examples: The first demonstrates the fixed-point iteration of (2), the second compares cooperative inference with machine teaching.

**Example 4.6.** Let $M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, then $L^{(k)} = \begin{pmatrix} 1 - \frac{1}{2^k} & \frac{1}{2^k} \\ \frac{1}{2^k} & 1 - \frac{1}{2^k} \end{pmatrix}$ and $T^{(k)} = \begin{pmatrix} 1 & 2k+1 \\ 0 & 1 \end{pmatrix}$. Notice that zero elements remain zero throughout the iteration process, but non-zero elements may converge to zero. Since $L^{(\infty)}$ and $T^{(\infty)}$ are both the identity matrix, $CI = 1$. In contrast, after one iteration of (2), $L^{(1)} = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}$, $T^{(1)} = \begin{pmatrix} 1 & 1/3 \\ 0 & 2/3 \end{pmatrix}$, and $TI(L^{(1)}, T^{(1)})$ is only $\frac{2}{3}$. Similarly, for any $k$, $TI(L^{(k)}, T^{(k)}) < 1$. Thus, cooperative inference increases the effectiveness of communication.

**Example 4.7.** We apply cooperative index to examples previously used in human teaching and machine teaching, and compare the effectiveness of cooperative inference and other communication protocols. The first example tested children’s learning of hidden functions of a novel toy after observing data presented by an experimenter. Empirical results have shown that when the experimenter activated one function accidentally, children inferred the toy to afford at least one function, and they also explored possible additional functions. However, when the experimenter activated one function pedagogically, children inferred that the toy afford one and only one function, which restrained their exploration of additional functions.

These results can be explained as two different communication protocols having different effectiveness. In the case of accidental demonstration (A), the learner assigns equal likelihood to all hypotheses that are consistent with the observed data, and assumes the experimenter to randomly choose a random subset of hypotheses to demonstrate. The experiment considered three concepts (the novel toy has zero, one, or two functions) and three data sets (zero, one, or two functions demonstrated). Given this order for the concepts and data sets, $L^{(A)} = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$ and $T^{(A)} = \begin{pmatrix} 1 & 1/2 & 1/3 \\ 0 & 1/2 & 1/3 \\ 0 & 0 & 1/3 \end{pmatrix}$.

This communication protocol entails much uncertainty for the correspondence between data and hypothesis, which is captured by a low $TI$: $TI(L^{(A)}, T^{(A)}) = 0.453$. In contrary, under pedagogical demonstration (P), both parties apply cooperative inference on the selection and interpretation of data, which leads to a perfectly effective communication protocol, $TI(L^{(P)}, T^{(P)}) = CI(M = L^{(A)}) = 1$. This explains the empirical finding that children inferred the toy to afford exactly the number of functions demonstrated by the experimenter.

The second example considers a version-space learner who is trying to learn a threshold classifier $h_\theta$, $\theta \in \{1, 2, 3\}$. For $x \in \{0, 1\}$, $h_\theta$ returns $y = -$ if $x < \theta$ and $y = +$ if $x \geq \theta$. Assume a teacher provides training sets $D = \{(x_1, y_1), (x_2, y_2)\}$, and the learner assigns the same likelihood to all hypotheses that are consistent with the data. Following machine teaching chooses data that maximize the likelihood for the learner to infer the correct hypothesis:

$$L = \begin{pmatrix} \{x_1, y_1, x_2, y_2\} \backslash h_\theta & h_1 & h_2 & h_3 \\ \\ \{0, -, 1, +\} & 1 & 0 & 0 \\ \{0, -, 2, +\} & 1/2 & 1/2 & 0 \\ \{0, -, 3, +\} & 1/3 & 1/3 & 1/3 \\ \{1, -, 2, +\} & 0 & 1 & 0 \\ \{1, -, 3, +\} & 0 & 1/2 & 1/2 \\ \{2, -, 3, +\} & 0 & 0 & 1 \end{pmatrix}, T^{(mt)} = \begin{pmatrix} \{x_1, y_1, x_2, y_2\} \backslash h_\theta & h_1 & h_2 & h_3 \\ \\ \{0, -, 1, +\} & 1 & 0 & 0 \\ \{0, -, 2, +\} & 0 & 0 & 0 \\ \{0, -, 3, +\} & 0 & 0 & 0 \\ \{1, -, 2, +\} & 0 & 1 & 0 \\ \{1, -, 3, +\} & 0 & 0 & 0 \\ \{2, -, 3, +\} & 0 & 0 & 1 \end{pmatrix}.$$
TI of \( L \) and \( T^{(mt)} \) can be understood as the effectiveness of machine teaching as a communication protocol. When all data sets are considered (as shown above), the effectiveness of machine teaching is perfect, \( TI(L, T^{(mt)}) = 1 \), which is higher than the effectiveness of cooperative inference, \( CI(M = L) = 0.734 \). However, when data sets are restricted, such as when only considering the first three rows of the data above, the effectiveness of cooperative inference can be higher than that of machine teaching \( CI(L') = 1. TI(L, T^{(mt)'} = 0.611 \).

Given the cooperative index, which quantifies the effectiveness of transmission for cooperative inference, we can prove conditions under which \( M \) maximizes CI:

**Definition 4.8.** A square matrix is triangular if it has a positive main diagonal, and has only zeros below (upper-triangular) or above (lower-triangular) the main diagonal.

**Theorem 4.9** (Representation theorem for cooperative inference). Let \( M \) be a nonnegative square matrix with at least one positive diagonal, then the following are equivalent.

(a) The corresponding cooperative index is optimal, i.e. \( CI(M) = 1 \);
(b) \( M \) has exactly one positive diagonal;
(c) \( M \) is a permutation of an upper-triangular matrix.

**Proof.** From Proposition 2.4 we know that \( CI(M) = TI(M^{(\infty)}, M^{(\infty)}) = 1 \) if and only if \( M^{(\infty)} \) is a permutation matrix. Since elements of \( M \) that lie in a positive diagonal do not tend to zero during cooperative inference \([19]\) (i.e. if \( M_{i,j} \neq 0 \) lies in a positive diagonal, then \( M_{i,j}^{(\infty)} \neq 0 \), \( M^{(\infty)} \) is a permutation matrix if and only if \( M \) has exactly one positive diagonal. So we have \((a) \iff (b)\). \((b) \iff (c)\) is a fact of linear algebra which can be proved by induction on the dimension \( n \) of \( M \) (see Supplementary Material for full detail).

Theorem 4.9 shows that in order to achieve optimal cooperative inference and thereby effective knowledge accumulation, the shared inductive bias should be one that constraints the form of \( M \) to be upper triangular (or permutation thereof). This in turn constraints the learner’s likelihood function such that it applies zero probability to particular data-concept relationships. Below, we show an example of using CI to investigate the form of the likelihood that leads to optimal transmission effectiveness.

**Example 4.10.** Consider polynomial regression. In order to have a triangular \( M \), the likelihood must have finite support. We explore the behavior of CI under different likelihood functions, ranging from fat-tailed to compact. In particular, we explore the conditions under which the different distributions lead to optimal CI.

Let \( \{x_i\}_{i=1}^6 = \{-1, -1, 0, 0, 1, 1\} \) and \( \{y_i\}_{i=1}^6 = \{a, -a, \Delta + a, \Delta - a, a, -a\} \). \( \Delta / a \) can be viewed as a signal-to-noise ratio for a second-order polynomial. Let \( D = \{D_1, D_2\} \), where \( D_1 = \{x_1, \ldots, x_4, y_1, \ldots, y_4\} \) and \( D_2 = \{x_1, \ldots, x_6, y_1, \ldots, y_6\} \). Let \( H = \{h_1, h_2\} \), where \( h_i \) is a polynomial of order \( i \) with a likelihood function that defines the assumed noise distribution. The likelihood function is a q-Gaussian \( N_q(z; \mu) \) set to have unit variance. We construct the \( M \) via maximum likelihood as a function of \( \Delta \) and \( a \) for \( q = (0, 1, 1.5) \). Briefly, we find the maximum-likelihood estimate of \( h_i \) to \( D_j \), then assign \( M_{i,j} \) the likelihood produced by that estimate (see Supplementary Material for more details). Having obtained these \( M \) matrices, we iterate them according to (2) to explore the behavior of CI.

In Figure 1 we show the phase diagrams of CI for the three q-Gaussian distributions, which correspond to a compact (\( q = 0 \)), normal (\( q = 1 \)), and fat-tailed (\( q = 1.5 \)) distributions respectively. This result shows that when the error likelihood is a compact distribution, there exists at least one setting of \( a \) such that CI = 1 for all \( \Delta > 0 \). This is not the case when the error likelihood has infinite support, i.e. \( q = 1 \) or \( q = 1.5 \). As suggested by Theorem 4.9, modeling choices that yield \( M \)s that are closer to triangular, such as compact likelihood functions, can produce optimal cooperative inference.

5 Discussion

Cooperative inference is central to accounts of human learning, language, and cultural evolution, but has not been deeply investigated in machine learning (but see [2]). We have presented a Transmission Index to
Figure 1: Comparison of CI across three different error likelihood functions (based on a $q$-Gaussian distribution) fit to a polynomial regression example. Each of the plots illustrate how CI varies as a function of the parameters $a$ and $\Delta$ that specify each of the different data sets $D$. We find that only having a compact error distribution, i.e., when $q = 0$, results in optimal CI.

quantify the effectiveness of knowledge transmission via data, connected to previous approaches focusing on efficiency, presented a Cooperative Index to quantify effectiveness of cooperative transmission, proven bidirectional conditions that relate inductive biases to optimal cooperative inference, and provided a simulation that illustrates how to modify classic learning models to improve cooperative transmission of knowledge.

Our results inform theory across fields. For fields interested in human learning, our representation theorem for cooperative inference (Theorem 4.9) provides the first predictions about how cognition would be structured if it were optimized to facilitate transmission of knowledge in over generations. For fields interested in machine learning, there has been much interest in whether machine learning is interpretable or explainable (e.g., banking [20], medicine [18], self-driving cars [3]). We have provided a theoretical framework for evaluating the degree to which an algorithm is explainable, and presented one case where simple modifications may yield improved explanation by examples.

References

[1] Robert A Barton and Robin IM Dunbar. Evolution of the social brain. *Machiavellian intelligence II: Extensions and evaluations*, 2:240, 1997.

[2] Garrett Birkhoff. Three observations on linear algebra. *Univ. Nac. Tucumán. Revista A*, 5:147–151, 1946.

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv:1604.07316*, 2016.

[4] Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Isabel Chang, Sydney Katz, and Laura Schulz. The double-edged sword of pedagogy: Modeling the effect of pedagogical contexts on preschoolers' exploratory play. In *Proceedings of the 31th annual conference of the Cognitive Science Society*, 2009.

[5] Elizabeth Bonawitz, Patrick Shafto, Hyowon Gweon, Noah D Goodman, Elizabeth Spelke, and Laura Schulz. The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330, 2011.

[6] Robert Boyd, Peter J Richerson, and Joseph Henrich. The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10918–10925, 2011.
[7] Jacob W Crandall, Mayada Oudah, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A Goodrich, Iyad Rahwan, et al. Cooperating with machines. arXiv preprint arXiv:1703.06207, 2017.

[8] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. Journal of Machine Learning Research, 15(1):3107–3131, 2014.

[9] Kelley Durkin and Patrick Shafto. Epistemic trust and education: Effects of informant reliability on student learning of decimal concepts. Child Development, 87(1):154–164, 2016.

[10] Baxter S Eaves and Patrick Shafto. Parameterizing developmental changes in epistemic trust. Psychonomic Bulletin & Review, pages 1–30, 2016.

[11] Nancy Eisenberg, Richard A Fabes, and Tracy L Spinrad. Prosocial development. Wiley Online Library, 1998.

[12] Sally A Goldman and Michael J Kearns. On the complexity of teaching. Journal of Computer and System Sciences, 50(1):20–31, 1995.

[13] Justine T Kao, Jean Y Wu, Leon Bergen, and Noah D Goodman. Nonliteral understanding of number words. Proceedings of the National Academy of Sciences, 111(33):12002–12007, 2014.

[14] Ji Liu and Xiaojin Zhu. The teaching dimension of linear learners. Journal of Machine Learning Research, 17(162):1–25, 2016.

[15] Patrick Shafto and Noah Goodman. Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In Proceedings of the 30th annual conference of the Cognitive Science Society, pages 1632–1637. Cognitive Science Society Austin, TX, 2008.

[16] Patrick Shafto, Noah D Goodman, and Michael C Frank. Learning from others: The consequences of psychological reasoning for human learning. Perspectives on Psychological Science, 7(4):341–351, 2012.

[17] Patrick Shafto, Noah D Goodman, and Thomas L Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. Cognitive Psychology, 71:55–89, 2014.

[18] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 2017.

[19] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics, 21(2):343–348, 1967.

[20] Justin Sirignano, Apaar Sadhwani, and Kay Giesecke. Deep learning for mortgage risk. SSRN: https://ssrn.com/abstract=2799443, 2016.

[21] M. Tomasello. The cultural origins of human cognition. Harvard University Press, Cambridge, MA, 1999.

[22] Michael Tomasello. Why we cooperate. MIT Press, Cambridge, MA, 2009.

[23] Michael Tomasello. Origins of human communication. MIT Press, Cambridge, MA, 2010.

[24] John Von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. Contributions to the Theory of Games, 2:5–12, 1953.

[25] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In AAAI, pages 4083–4087, 2015.

[26] Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Teaching dimensions based on cooperative learning. In COLT, pages 135–146, 2008.
6 Supplementary Material

6.1 Details of Remark 2.6

Case 1. \(|\mathcal{D}| \to \infty\) and \(|\mathcal{H}|\) is finite. \(\text{TI}(L, T)\) is well-defined since

\[
\lim_{m \to |\mathcal{D}|} \sum_{i=1}^{m} L_{i,j} T_{i,j} \leq \sum_{i=1}^{|\mathcal{D}|} T_{i,j} = 1
\]

implies that \(\lim_{m \to |\mathcal{D}|} \sum_{i=1}^{m} L_{i,j} T_{i,j}\) exists (limits exist for increasing sequences with upper bounds).

Case 2. \(|\mathcal{H}| \to \infty\) and \(|\mathcal{D}|\) is finite. Similarly we have \(\lim_{n \to |\mathcal{H}|} \sum_{j=1}^{n} L_{i,j} T_{i,j} \leq \sum_{j=1}^{|\mathcal{H}|} T_{i,j} = 1\). Then

\[
0 \leq \text{TI}(L, T) = \lim_{n \to |\mathcal{H}|} \frac{1}{n} \sum_{i=1}^{|\mathcal{D}|} \sum_{j=1}^{n} L_{i,j} T_{i,j} \leq \lim_{n \to |\mathcal{H}|} \frac{1}{n} \sum_{i=1}^{|\mathcal{H}|} 1 \leq \lim_{n \to |\mathcal{H}|} \frac{|\mathcal{D}|}{n} \to 0.
\]

So \(\text{TI}(L, T) = 0\) when \(|\mathcal{H}| \to \infty\). This is reasonable because in this case there is always infinitely untaught concepts given finitely many examples.

Case 3. \(|\mathcal{H}|, |\mathcal{D}| \to \infty\). \(\text{TI}\) may not exist. For example, let \(L = I_\infty\) be the identity matrix of infinite dimension and \(T = \text{Diag}(I_2, J_2, J_4, J_4, \ldots, I_{n_k}, J_{n_k}, I_{n_{k+1}}, J_{n_{k+1}}, \ldots)\) be a block diagonal matrix where \(I_{n_k}\) is the identity matrix of dimension \(n_k\), \(J_{n_k}\) is the matrix with 1’s on its skew-diagonal and 0’s elsewhere and \(n_1 = 2, n_{k+1} = 2\sum_{i=1}^{k} n_i\). Denote \(S_n = \frac{1}{n} \sum_{j=1}^{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{H}|} L_{i,j} T_{i,j}\). Then if \(\text{TI}\) exists, \(\text{TI} = \lim_{n \to \infty} S_n\). However, \(S_{2n_k} = \frac{1}{2}, S_{2n_k+n_{k+1}} = \frac{3}{4}\) for any \(k\). Therefore, \(\text{TI}\) does not exist.

6.2 Proof of Theorem 4.5

For convenience, we first write the fixed-point iteration of (2) explicitly in vector form. We denote the matrix with elements \(P_L(h|D)\) by \(L \in [0, 1]|\mathcal{D}| \times |\mathcal{H}|\), the matrix with elements \(P_T(D|h)\) by \(T \in [0, 1]|\mathcal{D}| \times |\mathcal{H}|\), and the matrix with elements \(P_D(h|D)\) by \(M \in [0, 1]|\mathcal{D}| \times |\mathcal{H}|\). Further, denote the vectors consisting of \(P_{L_0}(h)\) and \(P_{T_0}(h)\) by \(a, d \in [0, 1]|\mathcal{H}| \times 1\), vectors consisting of \(P_{T_0}(D)\) and \(P_{L_0}(D)\) by \(b, c \in [0, 1]|\mathcal{D}| \times 1\), respectively. Given \(a, b,\) and \(M\), the fixed-point iteration of the cooperative inference equations can be expressed as:

\[
P_{L_k}(h|D) = \frac{P_{D}(h|D) P_{L_0}(h)}{P_{L_k}(D)} \iff L^{(k)} = \text{Diag}\left(\frac{1}{M a}\right) M \text{Diag}(a)
\]

\[
P_{T_{k+1}}(D|h) = \frac{P_{T_{k+1}}(h|D) P_{T_0}(D)}{P_{T_{k+1}}(h)} \iff T^{(k+1)} = \text{Diag}(b) L^{(k+1)} \text{Diag}\left(\frac{1}{d^{(k+1)}}\right)
\]

\[
P_{T_{k+1}}(h) = \sum_{D \in \mathcal{D}} P_{L_k}(h|D) P_{T_0}(D) \iff d^{(k+1)} = (L^{(k+1)})^T b
\]

\[
P_{L_{k+1}}(h|D) = \frac{P_{T_k}(D|h) P_{L_0}(h)}{P_{L_{k+1}}(D)} \iff L^{(k+1)} = \text{Diag}\left(\frac{1}{c^{(k+1)}}\right) T^{(k)} \text{Diag}(a)
\]

\[
P_{L_{k+1}}(D) = \sum_{h \in \mathcal{H}} P_{T_k}(D|h) P_{L_0}(h) \iff c^{(k+1)} = T^{(k)} a,
\]

where, \(k\) denotes the iteration step, and \(\text{Diag}(z)\) denotes the diagonal matrix with elements of the vector \(z\) on its diagonal, and \(\frac{1}{z}\) denotes element-wise inverse for any vector \(z\).

Note that (3b) and (3c) are the operations to column normalize \(\text{Diag}(b) L^{(k)}\), and (3d) and (3e) are the operations to row normalize \(T^{(k)} \text{Diag}(a)\). Zero rows in \(L^{(k)}\) and zero columns in \(T^{(k)}\) are fixed throughout the iteration of (3). This is equivalent to removing the zero rows and zero columns of \(M\) for (3) and inserting them back at convergence or when the iteration is stopped.
Now we provide a version of the proof using the notations introduced in the paper. The original proof can be found in [19]. Remember that \( a \) and \( b \) are assumed to be uniform.

Proof. Let \( \sigma \) be a permutation of \( \{1, \cdots, n\} \) that makes \( \{M_{i,\sigma(i)}\}_{i=1}^n \) a positive diagonal. Define
\[
e^{(k)} := \prod_{i=1}^n L_{i,\sigma(i)}^{(k)}; \quad f^{(k)} := \prod_{i=1}^n i,\sigma(i).
\]
Applying \( 3a \), \( L^{(1)} \) is a row-stochastic matrix, and \( \{L_{i,\sigma(i)}^{(1)}\}_{i=1}^n \) is a positive diagonal, hence \( e^{(1)} \) is positive. Also, by applying \( 3b \),
\[
f^{(1)} = \prod_{i=1}^n T_{i,\sigma(i)}^{(1)} = \prod_{i=1}^n b_i L_{i,\sigma(i)}^{(1)} = \frac{e^{(1)}}{n\prod_{i=1}^n d_{\sigma(i)}^{(1)}} = \frac{e^{(1)}}{n\prod_{i=1}^n d_{\sigma(i)}^{(1)}}.
\]
By the inequality of arithmetic and geometric means, \( \left( \prod_{i=1}^n d_{\sigma(i)}^{(1)} \right)^{\frac{1}{n}} \leq \frac{1}{n} \sum_{i=1}^n d_{\sigma(i)}^{(1)}. \) Also, \( L^{(1)} \) is a row-stochastic matrix and we assumed uniform prior on data set space, hence by \( 3c \),
\[
n^n \prod_{j=1}^n d_{j}^{(1)} \leq \left( \sum_{j=1}^n d_{j}^{(1)} \right)^n = \left( \sum_{i=1}^n \sum_{j=1}^n b_i L_{i,j}^{(1)} \right)^n = \left( \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n L_{i,j}^{(1)} \right)^n = 1
\]
The Equality in \( 3 \) is achieved if and only if \( d = \left( \frac{1}{n}, \ldots, \frac{1}{n} \right) \), or equivalently, \( L^{(1)} \) being a doubly stochastic matrix. Because \( f^{(1)} \) is the product of \( n \) values between 0 and 1,
\[
0 < e^{(1)} \leq f^{(1)} \leq 1,
\]
with equality in \( a \) if and only if \( L^{(1)} \) is a doubly stochastic matrix, and equality in \( b \) if and only if \( L^{(1)} \) is a permutation matrix. By the same logic applied to equations \( 3d \) and \( 3e \),
\[
0 < f^{(1)} \leq e^{(2)} \leq 1,
\]
with equality in \( c \) if and only if \( T^{(1)} \) is a doubly stochastic matrix, and equality in \( d \) if and only if \( T^{(1)} \) is a permutation matrix. Repeating this argument, we get the increasing sequence
\[
0 < e^{(1)} \leq f^{(1)} \leq e^{(2)} \leq f^{(2)} \leq \cdots \leq 1.
\]
Monotone convergence theorem of real numbers guarantees that this sequence converges to its supremum
\[
\lim_{k \to \infty} e^{(k)} = \lim_{k \to \infty} f^{(k)} = \sup \{e, f\}.
\]
e\( (k) = f^{(k)} = e^{(k+1)} \) asymptotically, therefore \( L^{(k)} \) and \( T^{(k)} \) are both doubly stochastic matrices. Because doubly stochastic matrices are stable under row and column normalization, \( L \) and \( T \) converge to the same doubly stochastic matrix
\[
M^{(\infty)} := \lim_{k \to \infty} L^{(k)} = \lim_{k \to \infty} T^{(k)}.
\]

6.3 Proof of Theorem 4.9

Proof. \( 1) (a) \iff (b) \): We first prove that \( (a) \) CI(\( M \)) = 1, and \( (b) \) \( M \) has exactly one positive diagonal, are equivalent. Since \( M \) is an \( n \times n \) nonnegative matrix with at least one positive diagonal, Theorem 4.5 guarantees that the iteration of equation set \( 3 \) converges to a doubly stochastic matrix, \( M^{(\infty)} \). According to Birkhoff–von Neumann theorem [2] [24], there exist \( \theta_1, \ldots, \theta_k \in (0, 1] \) with \( \sum_i \theta_i = 1 \) and distinct permutation
matrices $P_1, \ldots, P_k$ such that $M^{(\infty)} = \theta_1 P_1 + \cdots + \theta_k P_k$. To simplify, we adopt the inner product notation between matrices: $A \cdot B = \sum_{i,j} A_{i,j} B_{i,j}$, for any two $n \times n$ square matrices $A$ and $B$. Then the following holds.

$$CI = TI(M^{(\infty)}, M^{(\infty)}) = \frac{1}{n} M^{(\infty)} \cdot M^{(\infty)} = \frac{1}{n} \left( \sum_i \theta_i P_i \right) \cdot \left( \sum_j \theta_j P_j \right) = \frac{1}{n} \sum_{i,j} \theta_i \theta_j P_i \cdot P_j$$

Equality (I) comes from rewriting TI in the inner product notation. Equality (II) comes from substituting $M^{(\infty)}$ by its Birkhoff–von Neumann decomposition. Equality (III) comes from distribution.

Further as permutation matrices, $P_i \cdot P_j \leq n$, and the equality holds if and only if $P_i = P_j$. So we have that,

$$CI(M) = \frac{1}{n} \sum_{i,j} \theta_i \theta_j P_i \cdot P_j \leq \frac{1}{n} \sum_{i,j} \theta_i \theta_j n = \sum_{i,j} \theta_i \theta_j \leq (\sum_i \theta_i) \times (\sum_j \theta_j) = 1$$

The equality in (IV) holds if and only if $P_i = P_j$ for any $i, j$. Note that $P_1, \ldots, P_k$ are distinct, i.e. $P_i \neq P_j$ when $i \neq j$. So the equality in (IV) is achieved precisely when $k = 1$ and $M^{(\infty)} = P_1$. Hence, $CI(M)$ is maximized if and only if $M^{(\infty)}$ is a permutation matrix.

We then prove that $M^{(\infty)}$ is a permutation matrix if and only if $M$ has exactly one positive diagonal. This follows the claim (1): elements of $M$ that lie in a positive diagonal do not tend to zero during the cooperative inference iteration [19] (i.e. if $M_{i,j} \neq 0$ lies in a positive diagonal, then $M_{i,j}^{(\infty)} \neq 0$). Claim (1) implies that $M^{(\infty)}$ and $M$ have the same number of positive diagonals. Further note that a doubly stochastic matrix has exactly one diagonal if and only it is a permutation matrix. So as a doubly stochastic matrix, $M^{(\infty)}$ is a permutation matrix if and only if $M$ has exactly one positive diagonal. Thus, $CI(M)$ is maximized if and only if $M$ has exactly one positive diagonal.

To complete the proof for (a) $\iff$ (b), we only need to justify Claim (1). Note that the product of any positive diagonal converges to a positive number sup{e, f} (shown in the proof for Theorem 4.5) and all elements on the positive diagonal is upper-bounded by 1 and lower-bounded by sup{e, f}. Therefore elements on a diagonal of $M$ cannot converge to 0.

(2) (b) $\iff$ (c): This follows immediately from a slightly more general claim below.

Claim (2): Let $A$ be an $n \times n$-square matrix. Then $A$ has exactly one non-zero diagonal (i.e. a diagonal with no zero element) if and only if $A$ is a permutation of an upper-triangular matrix.

We now prove Claim (2). The if direction is clear since an upper-triangular matrix always has exactly one non-zero diagonal, which is its main diagonal. The only if direction is proved by induction on the dimension $n$ of $A$.

**Step 1 - Induction basis:** When $n = 2$, it is easy to check that any $2 \times 2$ matrix with exactly one diagonal is either of the form $\begin{pmatrix} a & b \\ 0 & c \end{pmatrix}$ or $\begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$, where $a, c \neq 0$. So it is a permutation of an upper-triangular matrix.

**Step 2 - Inductive step:** Suppose that the claim - an $n \times n$-square matrix $A$ has exactly one non-zero diagonal if and only if it is a permutation of an upper-triangular matrix - holds for any $n < N$. We need to show that the claim also holds when $n = N$.

The following notation will be used. Let $A$ be an $n \times n$-square matrix. $A_{i,j}$ denotes the element of $A$ at row $i$ and column $j$. $\tilde{A}_{i,j}$ denotes the $(n - 1) \times (n - 1)$ sub-matrix obtained from $A$ by crossing out row $i$ and column $j$.

First, we will prove three handy observations.

**Observation 1** If $A$ has exactly one non-zero diagonal and $A_{i,j} \neq 0$, then $\tilde{A}_{i,j}$ has at most one non-zero diagonal. In particular, if $A_{i,j}$ is on that non-zero diagonal, then $\tilde{A}_{i,j}$ has exactly one non-zero diagonal.

**Proof of Observation 1:** Suppose that $\tilde{A}_{i,j}$ has more than one diagonal. Then these diagonals for $\tilde{A}_{i,j}$ along with $A_{i,j}$ form different diagonals for $A$, contradiction.

**Observation 2** If $A$ has exactly one non-zero diagonal and $A$ has a row or a column with exactly one non-zero element, then $A$ is a permutation of an upper-triangular matrix.
Proof of Observation 2: Suppose that $A$ has a column with exactly one non-zero element. Then by permutation, we may assume that it is the first column of $A$ and the only non-zero element is $A_{1,1}$. $A_{1,1}$ must be on the non-zero diagonal of $A$. Hence according to observation 1, $\tilde{A}_{1,1}$ is a $(N - 1 \times N - 1)$-square matrix has exactly one non-zero diagonal. Then by the inductive assumption, we may permute $\tilde{A}_{1,1}$ into an upper-triangular matrix. Note that each permutation of $\tilde{A}_{1,1}$ induces a permutation of $A$. So there exist permutations that convert $A$ into $A'$ such that $A'_{i,j} = 0$ when $j > 1$ and $i > j$. Moreover, permutations that convert $A$ to $A'$ never switch column 1 (row 1) of $A$ with any other columns (rows). So $A'_{1,1} = 0$ for $i \neq 1$ as $A_{1,1}$ is the only non-zero element in the first column of $A$. Thus, we have $A'_{i,j} = 0$ when $i > j$ which implies that $A'$ is an upper-triangular matrix.

If $A$ has a row with exactly one non-zero element. Then up to permutation, we may assume it is the last row of $A$ and the only non-zero element is $A_{N,N}$. Following similar argument as above, we may show that $\tilde{A}_{N,N}$ can be arranged into an upper-triangular matrix by permutations. And corresponding permutations of $A$ will also convert $A$ into an upper-triangular matrix. So observation 2 holds.

Observation 3 If the main diagonal of A is the only non-zero diagonal of A, then $A_{t_1,t_2} A_{t_2,t_3} \cdots A_{t_{k-1},t_k} A_{t_k,t_1} = 0$ for any distinct $t_1,t_2, \ldots, t_k$.

Proof of Observation 3: Suppose that $A_{t_1,t_2} A_{t_2,t_3} \cdots A_{t_{k-1},t_k} A_{t_k,t_1} \neq 0$. Then a different non-zero diagonal for $A$ other than the main diagonal is form by $\{A_{i,i} | i \neq t_1, \ldots, t_k\}$ and $A_{t_1,t_2}, A_{t_2,t_3}, \ldots, A_{t_{k-1},t_k}, A_{t_k,t_1}$.

Now back to the inductive step. Suppose that $A$ is an $N \times N$-square matrix with exactly one non-zero diagonal. By permutation, we may assume that the main diagonal of $A$ is the only non-zero diagonal. In particular, $A_{1,1} \neq 0$. According to Observation 1, $\tilde{A}_{1,1}$ has exactly one non-zero diagonal and so can be arranged into an upper-triangular matrix by permutations. The corresponding permutations convert $A$ into a new form, denoted by $A^1$ with the property that $A^1_{i,j} = 0$ when $j > 1$ and $i > j$. In particular, $A^1_{1,j} = 0$ when $j \neq 1$ and $j \neq N$. $\tilde{A}_{1,1}$ is an upper-triangular matrix implies that $A^1_{N,N} \neq 0$. If $A^1_{1,1} = 0$, then the last row of $A^1$ contains only one non-zero element $A^1_{N,N}$. So we are done by Observation 2.

Otherwise, according to Observation 1, $\tilde{A}^1_{N,N}$ can be arranged into an upper-triangular matrix by permutation. Hence, after the corresponding permutations, we may convert $A^1$ into a new form, denoted by $A^2$ with the property that $A^2_{2,j} = 0$ when $i > 1$ and $i \neq N$. Moreover, permutations that convert $A^1$ to $A^2$ never switch row $N$ (column $N$) of $A^1$ with any other rows (columns). So only one of $\{A^2_{N,j} | j \neq N\}$ is not zero. If $A^2_{N,1} = 0$, along with $A^2_{1,1} = 0$ for $N > i > 1$, we have that first column of $A^2$ contains exactly one non-zero element $A^2_{1,1}$. So by Observation 2, we are done.

Otherwise, $A^2_{N,1} \neq 0$. According to Observation 3, $A^2_{N,1} A^2_{1,k} A^2_{k,N} = 0$, for $k = 2, \ldots, N - 1$. So we have that $A^2_{1,k} A^2_{k,N} = 0$, for $k = 2, \ldots, N - 1$. We will proceed by analyzing cases from $k = 2$ to $k = N - 1$.

When $k = 2$, if $A^2_{1,2} = 0$, then column 2 of $A^2$ contains only one non-zero element $A^2_{1,2}$ and we are done by Observation 2. Otherwise, we may assume that $A^2_{1,2} \neq 0$ and $A^2_{2,N} = 0$.

When $k = 3$, if $A^2_{3,N} \neq 0$, then $A^2_{1,3} = 0$. According to Observation 3, $A^2_{3,1} A^2_{1,2} A^2_{2,3} A^2_{3,N} = 0$ and this implies that $A^2_{2,3} = 0$. Hence column 3 of $A^2$ contains only one non-zero element $A^2_{3,3}$ and again we are done by Observation 2. Otherwise, we may assume that $A^2_{1,3} = 0$ and one of $\{A^2_{1,3}, A^2_{2,3}\}$ is not zero.

When $k = k$, if $A^2_{k,N} \neq 0$, then $A^2_{1,4} = 0$. Similarly as in the case where $k = 3$ (by Observation 3), $A^2_{k,1} A^2_{1,2} A^2_{2,3} A^2_{3,4} A^2_{4,N} = 0$ and this implies that $A^2_{3,4} = 0$. One of $\{A^2_{1,3}, A^2_{2,3}\}$ is not zero $\implies$ either $A^2_{k,1} A^2_{1,2} A^2_{2,3} A^2_{3,4} A^2_{4,N} = 0$ or $A^2_{k,1} A^2_{1,2} A^2_{2,3} A^2_{3,4} A^2_{3,N} = 0 \implies A^2_{3,4} = 0$. Hence column 4 of $A^2$ contains only one non-zero element $A^2_{4,4}$ and again we are done by Observation 2. Otherwise, we may assume that $A^2_{4,N} = 0$ and at least one of $\{A^2_{1,4}, A^2_{2,4}, A^2_{3,4}\}$ is not zero.

Inductively, either one of column $k$’s of $A^2$ contains only one non-zero element or $A^2_{k,N} = 0$ for all $k = 2, \ldots, N - 1$. Note the latter case implies that the column $N$ of $A^2$ contains only one non-zero element $A^2_{N,N}$ as $A^2_{N,1} \neq 0 \implies A^2_{1,N} = 0$. Either way, the proof is then completed by Observation 2.
To construct $M$, first notice that if maximum likelihood is achieved, $M_{1,1} = M_{1,2}$ under all settings of $\Delta$, $a$, and $q$. This is because a first- and second-order polynomial give the same fit to $D_1$.

For $M_{2,1}$, by symmetry arguments we know that the maximum-likelihood fit of a first-order polynomial to $D_2$ is a horizontal line ($f(x) = b$). We can find this value of $b$ through a grid search. Given this $b$,

$$M_{2,1} = N_q(a; b)^2 N_q(-a; b)^2 N_q(\Delta + a; b) N_q(\Delta - a; b),$$

where

$$N_q(z; b) = \frac{\sqrt{\beta}}{C_q} e_q(-\beta(x_i - \mu)^2).$$

Here, $\beta = \frac{1}{\sqrt{3-3q}}$ so that the variance is 1. $e_q(x)$ is the $q$-exponential function defined by $[1 + (1-q)x]^{\frac{1}{q-1}}$ when $q \neq 1$, and $\exp(x)$ when $q = 1$. The normalizing constant $C_q$ is given by:

$$C_q = \begin{cases} 
\frac{2\sqrt{\Gamma(\frac{3-q}{2})}}{\sqrt{\pi} \sqrt{1-q} \sqrt{\Gamma(\frac{3-3q}{2})}} & \text{for } -\infty < q < 1 \\
\sqrt{\pi} & \text{for } q = 1 \\
\frac{\sqrt{\pi} \Gamma(\frac{3-q}{1-q})}{\sqrt{q-1} \Gamma(\frac{3-3q}{2})} & \text{for } 1 < q < 3.
\end{cases}$$

For $M_{2,2}$, again by symmetry arguments we know that the maximum-likelihood fit of a second order polynomial to $D_2$ is a parabola that passes through the middle of each of the three pairs of data points. Thus, $M_{2,2} = N_q(a; 0)^6$. 

\[\text{Page 14}\]