Optimal Cooperative Inference

Scott Cheng-Hsin Yang **Yue Yu** **Arash Givchi**
Pei Wang **Wai Keen Vong** **Patrick Shafto**
Department of Mathematics & Computer Science, Rutgers University—Newark

Abstract

Cooperative transmission of data fosters rapid accumulation of knowledge by efficiently combining experiences across learners. Although well studied in human learning and increasingly in machine learning, we lack formal frameworks through which we may reason about the benefits and limitations of cooperative inference. We present such a framework. We introduce novel indices for measuring the effectiveness of probabilistic and cooperative information transmission. We relate our indices to the well-known Teaching Dimension in deterministic settings. We prove conditions under which optimal cooperative inference can be achieved, including a representation theorem that 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 and machine learning.

1 INTRODUCTION

Learning through cooperation is a foundational principle underlying human-human, human-machine, and (potentially) machine-machine interaction. In human-human interaction, cooperative information sharing has long been viewed as a foundation to human language (Grice, 1975; Goodman and Stuhlmüller, 2013; Kao et al., 2014), cognitive development (Csibra and Gergely, 2009), and cultural evolution (Tomasello, 1999; Tomasello et al., 2005). Cooperative learning has appeared in human-machine interaction (Crandall et al., 2017), social robotics (Thomaz and Breazeal, 2008; Knox et al., 2013; Chernova and Thomaz, 2014; Thomaz et al., 2016; Laskey et al., 2017; Bestick et al., 2016), machine teaching (Zhu, 2013, 2015; Patil et al., 2014; Simard et al., 2017), cooperative reinforcement learning (Hadfield-Menell et al., 2016; Ho et al., 2016), and deep neural networks (Lowe et al., 2017). Despite the importance of cooperative selection of, and learning from, data, we are unaware of any theory of when or why cooperation may be effective for increasing learning and the transmission of knowledge.

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. As training and test errors provide a framework for measuring how effectively a model selects the best model and generalizes, our new measure, Cooperative Index, provides a framework for measuring how effectively a model can be explained by way of examples from the data and for selecting models with inductive biases that are interpretable with respect to the data. We also use the measure to extend the Teaching Dimension (Goldman and Kearns, 1995; Zilles et al., 2008)—a classical measure of communication efficiency[1]—from deterministic to probabilistic settings. We show how analyzing this measure reveals the conditions, in terms of constraints on the 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, 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, we introduce cooperative inference based on

Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

[1] Effectiveness is a measure of the quality of communication; efficiency is the size of the data necessary to reach a particularly effectiveness.

previous research in human social learning (Shafto and Goodman, 2008; Shafto et al., 2014), present a Cooperative Index by extending the Transmission index to the cooperative setting, and identify the condition that must be satisfied to achieve optimal communication. In Section 6, we conclude with implications for human, machine, and human-machine learning.

2 THE TRANSMISSION INDEX

In this section we define Transmission Index to quantify 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 *concept space* $\mathcal{H}$. A *data set space*, $\mathcal{D}$, is a collection of subsets of a given 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.

When $\mathcal{H}$ and $\mathcal{D}$ are both discrete, in matrix notation, we can form the row-stochastic *learner's inference matrix*, $\mathbf{L} \in [0,1]^{|\mathcal{D}|\times|\mathcal{H}|}$, having elements $P_L(h|D)$, and the column-stochastic *teacher's selection matrix*, $\mathbf{T} \in [0,1]^{|\mathcal{D}|\times|\mathcal{H}|}$, having elements $P_T(D|h)$. As it is possible that there exist data sets (or concepts) whose probability of being selected is zero, here we allow a row (or column) stochastic matrix to have zero rows (or zero columns).

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

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

Note that in the above definition, both $|\mathcal{H}|$ and $|\mathcal{D}|$ can either be finite or countably infinite. TI is well-defined when $|\mathcal{D}|$ is countably infinite because $C_j := \sum_{i=1}^{|\mathcal{D}|}\mathbf{L}_{i,j}\mathbf{T}_{i,j}$ still converges in this case. ($C_j \le \sum_{i=1}^{|\mathcal{D}|}\mathbf{T}_{i,j} = 1$ and is thus bounded above, and each $\mathbf{L}_{i,j}\mathbf{T}_{i,j}$ is non-negative.) When $|\mathcal{H}|$ is countably infinite, TI should be interpreted as a limit. See remark 2.6 for more detail.

In connection to channel coding in information theory, the learner's inference process is analogous to the decoding process, and the teacher's data selection process can be thought of as the combination of choosing the code words and passing them through a noisy channel, which makes the transmitted signals stochastic. Therefore, the Transmission Index can be related to channel capacity and the mutual information between the code words and the observations. These relationships deserve a full treatment that is outside the scope of this paper.

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 $\mathrm{TI}(\mathbf{L},\mathbf{T})$ simply by TI.

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

In the first case (a), $\mathrm{TI}(\mathbf{L}^{(a)},\mathbf{T}) = 1$, 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), $\mathrm{TI}(\mathbf{L}^{(b)},\mathbf{T}) = 0$, 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), $\mathrm{TI}(\mathbf{L}^{(c)},\mathbf{T}) = \frac{1}{2}$. 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. Suppose that $|\mathcal{H}|$ is finite and $|\mathcal{D}|$ is finite or countably infinite [2], then the range of the Transmission Index is $0 \le \mathrm{TI} \le 1$, and $\mathrm{TI} = 1$ if and only if two conditions hold: (i) $\mathbf{L}_{i,j} = 1$ if $\mathbf{T}_{i,j} > 0$ for all i, j, and (ii) there is no zero column in $\mathbf{L}$ and $\mathbf{T}$. Also, $\mathrm{TI} = 1$ implies that $|\mathcal{D}| \ge |\mathcal{H}|$, with equality achieved when $\mathbf{L}$ and $\mathbf{T}$ are the same permutation matrix.

Proof. $\mathrm{TI} \ge 0$ because $\mathbf{T}$ and $\mathbf{L}$ are stochastic matrices, and $\mathrm{TI} = 0$ if and only if for any i, j, either $\mathbf{L}_{i,j} = 0$ or $\mathbf{T}_{i,j} = 0$.

We show $\mathrm{TI} \le 1$:

[2] Similar conclusion also holds when $|\mathcal{H}|$ is countable infinite. See remark 2.6 for more detail.

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

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. It is clear that when $|\mathcal{H}|$ is finite, TI is invariant under joint row and column permutations of $\mathbf{L}$ and $\mathbf{T}$. When $|\mathcal{H}| = |\mathcal{D}|$ and $\mathrm{TI} = 1$, 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. When $|\mathcal{H}|$ is countably infinite and $|\mathcal{D}|$ is either finite or countably infinite, the Transmission Index is generalized to:

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

This can be interpreted as the following. Let $S_n = \sum_{j=1}^{n}\sum_{i=1}^{|\mathcal{D}|}\mathbf{L}_{i,j}\mathbf{T}_{i,j}$, then $\mathrm{TI}(\mathbf{L},\mathbf{T}) = \lim_{n\to\infty}\frac{S_n}{n}$. Intuitively, columns of $\mathbf{T}$ provide an enumeration of concepts in $\mathcal{H}$ and $\frac{S_n}{n}$ measures how well on average the first n concepts can be communicated. Further, as all terms are non-negative, if the limit of $\{\frac{S_n}{n}\}$ exists, it does not depend on this particular enumeration. Therefore, naturally $\mathrm{TI}(\mathbf{L},\mathbf{T})$ is defined to be the limit of $\{\frac{S_n}{n}\}$.

Regrading the existence of TI, there are two cases. The proof of Proposition 2.4 implies that $0 \leq S_n \leq n$. One case is that the growth rate of S_n is strictly slower than any linear function, then $\mathrm{TI} = 0$. Otherwise, TI exists if and only if the sequence $\{C_j\}$ converges as $j \to \infty$, where $C_j = \sum_{i=1}^{|\mathcal{D}|}\mathbf{L}_{i,j}\mathbf{T}_{i,j}$. These results provide a guideline on constructing $\mathbf{L}$ and $\mathbf{T}$ to guarantee the existence of TI when $|\mathcal{H}|$ is countably infinite. See Supplementary Material for full detail.

In the rest of this paper, we assume that both $|\mathcal{H}|$ and $|\mathcal{D}|$ are finite. Adopting the limit notations, similar analysis can be made when $|\mathcal{H}|$ and $|\mathcal{D}|$ are countably infinite.

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 (Goldman and Kearns, 1995) and has been analyzed for certain models in continuous concept space (Liu and Zhu, 2016) and in cooperative settings (Zilles et al., 2008; Doliwa et al., 2014). However, Teaching Dimension and these analyses assume a deterministic 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 when 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 D if and only if for every data point $(x, y) \in D$, $h(x) = y$. $D \in \mathcal{D}$ is a *teaching set* for concept $h \in \mathcal{H}$ if h, but no other concept in $\mathcal{H}$, is consistent with D. Let $\mathcal{D}^*(h) \subset \mathcal{D}$ be the collection of teaching sets in $\mathcal{D}$ for concept h. The classical version of *Average Teaching Dimension* (Doliwa et al., 2014) is defined as follows: First, for any $h \in \mathcal{H}$, let

$$TD(h) = \begin{cases} \infty & \text{if } \mathcal{D}^*(h) \text{ is empty} \\ \min_{D\in\mathcal{D}^*(h)}|D| & \text{otherwise} \end{cases},$$

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

$$ATD(\mathcal{H}) = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} TD(h).$$

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

$$ETD(\mathcal{H}) = \frac{\sum_{h\in\mathcal{H}}\sum_{D\in\mathcal{D}} |D|\, P_L(h|D)\, P_T(D|h)}{\sum_{h\in\mathcal{H}}\sum_{D\in\mathcal{D}} P_L(h|D)\, P_T(D|h)}.$$

Definition 3.3. Let $\mathbf{M} \in [0,1]^{|\mathcal{D}|\times|\mathcal{H}|}$ be a matrix, where the element $\mathbf{M}_{i,j}$ represents the probability that h_i is consistent with D_j. We define $C \in \{0,1\}^{|\mathcal{D}|\times|\mathcal{H}|}$ 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 $\mathbf{M}$ by treating $C_{i,j}$ as the outcome of a Bernoulli trial with parameter $\mathbf{M}_{i,j}$.

Probabilistic consistency is an extension of deterministic consistency in the face of uncertainty. There are at least two cases where uncertainty can arise. The first case is when there are multiple possible learning scenarios but the learner is uncertain about which scenario is active. In this context, the probability of being consistent is the proportion of scenarios in which the concept is consistent with the data. The second case is when there is measurement noise. In this context, the learner has uncertainty about the true value of the data and therefore is also uncertain about whether the data is consistent with the concept.

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

Proof. $ATD(\mathcal{H})$ is finite if and only if $TD(h)$ is finite for all $h \in \mathcal{H}$. Finite $TD(h)$ means that there is at least one teaching set $D \in \mathcal{D}$ for h. Let $\alpha_i \subset \{1, 2, \ldots, N\}$ be the index set for the teaching sets of h_i. Because every D can only belong to at most one $\mathcal{D}^*(h_i)$, so $\alpha_i \subset \{1, \cdots, N\} \setminus \cup_{j\neq i}\, \alpha_j$ for every $i \in \{1, 2, \ldots, N\}$. Further, because $|\mathcal{D}| = |\mathcal{H}|$, this construction of α_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 $TD(h_i)$ is finite, α_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 $C = \mathbf{L} = \mathbf{T}$, which by Proposition 2.4 is equivalent to $\mathrm{TI}(\mathbf{L},\mathbf{T}) = 1$. □

Example 3.5. If $\mathbf{L} = \mathbf{T}$ and is a permutation matrix, $C = \mathbf{T}$. 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 $|\mathcal{H}| = |\mathcal{D}| = 2$, $\mathbf{M} = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}$. There are four possible consistency matrices that can be sampled from $\mathbf{M}$: $C^{(a)} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$, $C^{(b)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $C^{(c)} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$, $C^{(d)} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$. For $C^{(a)}$, $C^{(c)}$ and $C^{(d)}$, the corresponding $ATD(\mathcal{H})$ is ∞, and for $C^{(b)}$ it is $\frac{|D_1|+|D_2|}{2}$. Let $\mathbf{L}^{(*)}$ and $\mathbf{T}^{(*)}$ be the row-normalized and column-normalized matrices of $C^{(*)}$, respectively, for $* \in \{a, b, c, d\}$. Then, $\mathrm{TI}(\mathbf{L}^{(a)},\mathbf{T}^{(a)}) = \frac{1}{2}$, $\mathrm{TI}(\mathbf{L}^{(b)},\mathbf{T}^{(b)}) = 1$, $\mathrm{TI}(\mathbf{L}^{(c)},\mathbf{T}^{(c)}) = \frac{1}{2}$, and $\mathrm{TI}(\mathbf{L}^{(d)},\mathbf{T}^{(d)}) = \frac{5}{8}$, 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 ($\mathrm{TI} = 1$) to those of imperfect transmission ($0 \leq \mathrm{TI} \leq 1$) as well.

In addition to uncertain learning scenarios and measurement noise, another way probabilistic transmission can enter is that $\mathbf{M}$ represents the degree of consistency between data and hypotheses. In this case, a deterministic learner would need to make a decision on what the underlying true consistency matrix is. Consider $\mathbf{M} = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}$ again. A simple decision rule is to round $\mathbf{M}_{i,j}$ up 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 (Eaves and Shafto, 2016) and conversations (Kao et al., 2014)). 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 and has been proposed as a model of human language and teaching (Kao et al., 2014; Shafto and Goodman, 2008; Shafto et al., 2014). It can be seen as a way to map one common convention to another one that is more effective at transmitting knowledge without *a priori* agreement on the encoding of data-concept pairs (Zilles et al., 2008). We define *Cooperative Index* as a measure of communication effectiveness in the cooperative setting by applying the Transmission Index to cooperative inference. Then, we provide proofs regarding the form of the shared likelihood matrix required to maximize the cooperative index and hence 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:

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

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

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 (Shafto and Goodman, 2008; Shafto et al., 2014): First, define an initial likelihood[3], $P_T(D|h) = P_0(D|h)$, 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), use the resulting $P_T(D|h)$ to evaluate (2a), and iterate this process 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 $P_0(D|h)$, 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 of $\mathbf{M}$. 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 $\mathbf{L}^{(\infty)} := \lim_{k\to\infty} \mathbf{L}^{(k)}$ and $\mathbf{T}^{(\infty)} := \lim_{k\to\infty} \mathbf{T}^{(k)}$.

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

$$\mathrm{CI}(\mathbf{M}) = \mathrm{TI}(\mathbf{L}^{(\infty)}, \mathbf{T}^{(\infty)}) = \frac{1}{|\mathcal{H}|} \sum_{j=1}^{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{D}|} \mathbf{L}^{(\infty)}_{i,j} \mathbf{T}^{(\infty)}_{i,j}.$$

Remark 4.4. Similarly to TI, CI is also well-defined as a limit when both $|\mathcal{H}|$ and $|\mathcal{D}|$ are countably infinite, provided that the fixed-point iteration of (2) converges.

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 (Sinkhorn and Knopp, 1967). With this connection, we provide conditions under which optimal cooperative inference is achievable.

Definition 4.5 (Positive diagonal)**.** If $\mathbf{M}$ is an $n \times n$ square matrix and σ 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 σ is the identity permutation, the diagonal is called the *main diagonal*.

Theorem 4.6 (A simpler version of Sinkhorn's theorem (Sinkhorn and Knopp, 1967))**.** *Given any non-negative 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,* $\mathbf{M}^{(\infty)}$*, which contains neither zero columns nor zero rows, 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 the inequality of arithmetic and geometric means to prove that the product either stays 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 $\mathbf{L}$ and $\mathbf{T}$ must have converged to the same doubly stochastic matrix. □

As is for TI, if $\mathbf{M}$ is clear from the context, we denote $\mathrm{CI}(\mathbf{M})$ simply by CI for brevity. Now, we give two simple examples: The first demonstrates the fixed-point iteration of (2); the second compares full cooperative inference with a special case known as machine teaching (Zhu, 2015).

[3]This is the *shared likelihood* and *common convention* mentioned in the beginning of this section.

Example 4.7. Let $\mathbf{M} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$, then $\mathbf{L}^{(k)} = \begin{pmatrix} 1-\frac{1}{2k} & \frac{1}{2k} \\ 0 & 1 \end{pmatrix}$ and $\mathbf{T}^{(k)} = \begin{pmatrix} 1 & \frac{1}{2k+1} \\ 0 & 1-\frac{1}{2k+1} \end{pmatrix}$. Notice that zero elements remain zero throughout the iteration process, but non-zero elements may converge to zero. Since $\mathbf{L}^{(\infty)}$ and $\mathbf{T}^{(\infty)}$ are both the identity matrix, CI = 1. In contrast, after one iteration of (2), $\mathbf{L}^{(1)} = \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix}$, $\mathbf{T}^{(1)} = \begin{pmatrix} 1 & 1/3 \\ 0 & 2/3 \end{pmatrix}$, and $\mathrm{TI}(\mathbf{L}^{(1)}, \mathbf{T}^{(1)})$ is only $\frac{2}{3}$. Similarly, for any k, $\mathrm{TI}(\mathbf{L}^{(k)}, \mathbf{T}^{(k)}) < 1$. Thus, cooperative inference increases the effectiveness of communication.

Example 4.8. In this example we apply TI and CI to machine teaching in a simple setting. Following Liu and Zhu (2016), consider a version-space learner who is trying to learn a threshold classifier h_θ, $\theta \in \{1, 2, 3\}$. For $x \in \{0, 1, 2, 3\}$, h_θ returns $y = -$ if $x < \theta$ and $y = +$ if $x \geq \theta$. Assume a teacher provides training set $D = \{(x_1, y_1), (x_2, y_2)\}$ and the learner assigns the same likelihood to all concepts that are consistent with the data; then, the learner's inference matrix is:

$\mathbf{L} =$

$\{x_1, y_1, x_2, y_2\}\backslash h_\theta$	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

Following Liu and Zhu (2016), machine teaching chooses data that maximize the likelihood for the learner to infer the correct hypothesis. Note that this way of teaching can be considered as a special case of cooperative inference: The teacher selects data by maximizing $P_L(h|D)$ rather than sampling in proportion to the probability, and the learner does not reason about the teacher's selection and thus only the first step of the recursive cooperative inference is executed (see 2a and 2b). Machine teaching will choose data sets $\{0, -, 1, +\}$, $\{1, -, 2, +\}$, and $\{2, -, 3, +\}$ for h_1, h_2, and h_3, respectively, with probability 1. Let machine teaching's data selection matrix be $\mathbf{T}^{(mt)}$. The effectiveness of machine teaching can then be quantified by $\mathrm{TI}(\mathbf{L}, \mathbf{T}^{(mt)})$, which is 1 in this case. Depending on the concept space and data set space, machine teaching's effectiveness is not always perfect. For example, if the learner's inference matrix consists of only the first three rows of $\mathbf{L}$, TI for machine teaching becomes 0.611.

Given the cooperative index, which quantifies the effectiveness of transmission for cooperative inference, we can prove conditions under which $\mathbf{M}$ maximizes CI:

Definition 4.9. 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 4.10 (Representation theorem for cooperative inference)**.** *Let* $\mathbf{M}$ *be a nonnegative square matrix with at least one positive diagonal, then the following statements are equivalent:*

(a) The cooperative index is optimal, i.e., $\mathrm{CI}(\mathbf{M}) = 1$*;*

(b) $\mathbf{M}$ *has exactly one positive diagonal;*

(c) $\mathbf{M}$ *is a permutation of an upper-triangular matrix.*

Proof. From Proposition 2.4 we know that $\mathrm{CI}(\mathbf{M}) = \mathrm{TI}(\mathbf{M}^{(\infty)}, \mathbf{M}^{(\infty)}) = 1$ if and only if $\mathbf{M}^{(\infty)}$ is a permutation matrix. Since elements of $\mathbf{M}$ that lie in a positive diagonal do not tend to zero during cooperative inference (Sinkhorn and Knopp, 1967) (i.e., if $\mathbf{M}_{i,j} \neq 0$ lies in a positive diagonal, then $\mathbf{M}_{i,j}^{(\infty)} \neq 0$), $\mathbf{M}^{(\infty)}$ is a permutation matrix if and only if $\mathbf{M}$ 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 $\mathbf{M}$ (see Supplementary Material for full detail). ☐

Remark 4.11. Let C be a consistency matrix of size $N \times N$ as in Definition 3.3. Suppose that C is a permeation of an upper-triangular matrix, then $\mathrm{CI}(C) = 1$. Together with Proposition 3.4, we have that the Average Teaching Dimension of the corresponding concept space $\mathcal{H}$ is finite at the convergence of the cooperative inference iteration, but is infinite before that (unless C is a permutation matrix).

Theorem 4.10 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 $\mathbf{M}$ to be upper triangular (or a 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 4.12. Consider polynomial regression. In order to have a triangular $\mathbf{M}$, 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 $\{y_i\}_{i=1}^6 = \{a, -a, \Delta + a, \Delta - a, a, -a\}$. The quantity Δ/a can be viewed as the signal-to-noise ratio for a second-order polynomial. Let $\mathcal{D} = \{D_1, D_2\}$,

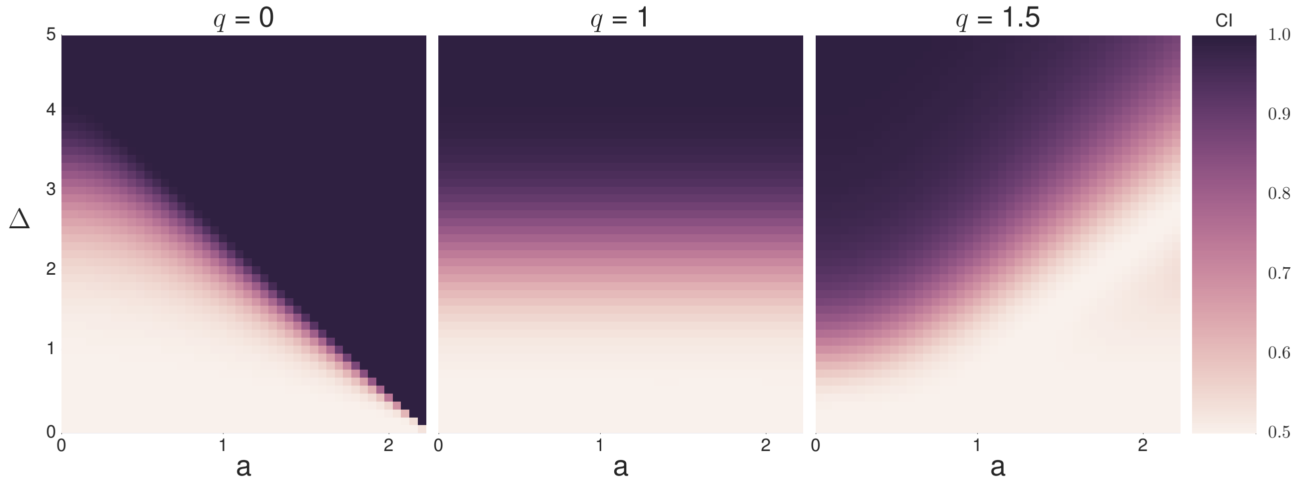

Figure 1: Comparison of CI across three different error likelihood functions (based on q-Gaussian distributions with different values of q) in polynomial regression. Each of the plots illustrates how CI varies as a function of the parameters a and Δ that specify different data set spaces (see main text and Supplementary Material for detail). We find that only having a compact error distribution, i.e., $q = 0$, results in optimal CI for all settings of Δ, which corresponds to the signal strength in the data.

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 $\mathcal{H} = \{h_1, h_2\}$, 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 $N_q(z; \mu)$ with unit variance (Tsallis et al., 2009). We construct the $\mathbf{M}$ via maximum likelihood as a function of Δ and a for $q = \{0, 1, 1.5\}$. For each value of q, we first find the maximum-likelihood estimate of h_i to D_j, then assign $\mathbf{M}_{i,j}$ the likelihood produced by that estimate (see Supplementary Material for more details). Having obtained these $\mathbf{M}$ 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 ($q = 1.5$) distribution. This result shows that when the error likelihood is a compact distribution, there exists at least one setting of a such that $\mathrm{CI} = 1$ 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 4.10, modeling choices that yield $\mathbf{M}$'s that are closer to triangular, such as compact likelihood functions, can produce optimal cooperative inference. This illustrates a simple modeling choice that allows a small set of examples to uniquely identify different parameterizations of the model. It is in this sense that optimization of the Cooperative Index may foster explainability and interpretability—by allowing small sets of examples to uniquely map to underlying parameterizations of the model, without requiring that the maps between hypotheses and data be bijective.

5 RELATED WORK

As briefly discussed in Example 4.8, machine teaching is a close cousin of cooperative inference in that both aim to choose good data to convey a target concept. Machine teaching can be thought of as performing only one step of the cooperative inference iteration and choosing deterministically the best choice available. In this setting, Liu and Zhu (2016) has derived the Teaching Dimension for linear learning models and discussed the connections to VC Dimension. For simpler version-space learner models, Doliwa et al. (2014) has made formal connections between the Teaching Dimension, VC Dimension, and sample compression in the iterative setting, and Searcy and Shafto (2016) investigated the representational implications of deterministic cooperation. These differ from CI in that they assume deterministic, rather than probabilistic, inference.

Furthermore, since cooperative inference is implemented via the Sinkhorn-Knopp algorithm, many connections stem from the body of work relating to Sinknorn's scaling (see Idel (2016) for review). To give a few examples, on the theoretical front, Sinkhorn's theorem has been analyzed with geometric interpretation (Dykstra, 1985), in a convex programming formulation (Macgill, 1977; Krupp, 1979), and as an entropy minimization problem with linear constraint (Brown et al., 1993). On the application side, Sinkhorn's theorem has been applied to modelling transportation (de Dios Ortuzar et al., 1994), designing condition numbers (Benzi, 2002), and ranking webpages (Knight, 2008).

6 DISCUSSION

Cooperative inference is central to human and machine learning. Previous work has introduced numerous accounts of the role of cooperation in learning and applied these across a host of problems in human and machine learning; however, to date, there has been no account of when or why we should expect cooperative inference to outperform simple learning. Building on prior models of cooperation from cognitive science of language and learning and demonstrating connections to models of machine teaching, we investigate this question. We introduced the Transmission and Cooperative Indices, which are metrics for the effectiveness of inference in standard learning and cooperative learning settings, respectively. We connect the Transmission Index with prior measures of efficiency in deterministic settings, namely, Teaching Dimension, and prove a representation theorem stating the conditions under which cooperation can yield optimally effective inference. We demonstrate how this model informs modification of a standard model of learning to ensure optimal cooperative transmission of the model class via a small subset of the data.

Beyond human learning, where this work provides foundational theory to inform accounts of human cognitive development, language, and cultural evolution, this work has strong implications for development of machine learning models that are designed for explainability and interpretability. Implicit in these is the existence of a shared goal, and cooperation is the natural formalization of this. Whereas models necessarily encode inferences about data in an internal language, and those internal languages may take many different forms depending on the task or domain, data provide a general purpose language in which inferences can be encoded to and decoded from. The promise of this work is that it provides an overarching framework for thinking about how to engineer models that are not only predictively accurate, but also understood well enough to be deployed correctly.

There are a number of practical and theoretical reasons to be concerned with the explanability of machine learning and AI algorithms. Practical reasons are related to algorithms' use in industry, for example, to decide who will get loans or determine prison sentences. Human intelligibility to ensure the algorithms are not simply propagating race, gender or other biases as well as to satisfy recent legal standards is necessary (see recent EU laws related to a right to an explanation; Goodman and Flaxman (2016)). Theoretical reasons are highlighted by the adversarial images that illustrate how little we understand the workings of deep learning (and probably other classes of) models. Our paper presents theoretical results upon which we may develop systems that are designed to be explainable by building models that adopt the structural constraints necessary to ensure optimal cooperative inference.

7 Acknowledgements

This material is based on research sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146 to P.S. and S.Y. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. This research was also supported by NSF SMA-1640816 to PS.

References

Michele Benzi. Preconditioning techniques for large linear systems: a survey. *Journal of Computational Physics*, 182(2):418–477, 2002.

Aaron Bestick, Ruzena Bajcsy, and Anca D Dragan. Implicitly assisting humans to choose good grasps in robot to human handovers. In *International Symposium on Experimental Robotics*, pages 341–354. Springer, 2016.

Jack B Brown, Phillip J Chase, and Arthur O Pittenger. Order independence and factor convergence in iterative scaling. *Linear Algebra and Its Applications*, 190:1–38, 1993.

Sonia Chernova and Andrea L Thomaz. Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(3):1–121, 2014.

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:1703.06207*, 2017.

Gergely Csibra and György Gergely. Natural pedagogy. *Trends in Cognitive Sciences*, 13(4):148–153, 2009.

Juan de Dios Ortuzar, Luis G Willumsen, et al. *Modelling transport.* Wiley New Jersey, 1994.

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.

Richard L Dykstra. An iterative procedure for obtaining i-projections onto the intersection of convex sets. *The Annals of Probability*, pages 975–984, 1985.

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

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

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv:1606.08813*, 2016.

Noah D Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184, 2013.

H Paul Grice. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and semantics*, pages 41–58. New York: Academic Press, 1975.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 3909–3917, 2016.

Mark K Ho, Michael Littman, James MacGlashan, Fiery Cushman, and Joseph L Austerweil. Showing versus doing: Teaching by demonstration. In *Advances in Neural Information Processing Systems*, pages 3027–3035, 2016.

Martin Idel. A review of matrix scaling and sinkhorn's normal form for matrices and positive maps. *arXiv:1609.06349*, 2016.

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.

Philip A Knight. The Sinkhorn-Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

W Bradley Knox, Peter Stone, and Cynthia Breazeal. Training a robot via human feedback: A case study. In *International Conference on Social Robotics*, pages 460–470. Springer, 2013.

RS Krupp. Properties of kruithof's projection method. *Bell Labs Technical Journal*, 58(2):517–538, 1979.

Michael Laskey, Caleb Chuck, Jonathan Lee, Jeffrey Mahler, Sanjay Krishnan, Kevin Jamieson, Anca Dragan, and Ken Goldberg. Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 358–365. IEEE, 2017.

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

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv:1706.02275*, 2017.

Sally M Macgill. Theoretical properties of biproportional matrix adjustments. *Environment and Planning A*, 9(6):687–701, 1977.

Kaustubh R Patil, Xiaojin Zhu, Łukasz Kopeć, and Bradley C Love. Optimal teaching for limited-capacity human learners. In *Advances in Neural Information Processing Systems*, pages 2465–2473, 2014.

Sophia Ray Searcy and Patrick Shafto. Cooperative inference: Features, objects, and collections. *Psychological Review*, 123(5):510–533, 2016.

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.

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.

Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. Machine teaching: A new paradigm for building machine learning systems. *arXiv:1707.06742*, 2017.

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

Andrea Thomaz, Guy Hoffman, Maya Cakmak, et al. Computational human-robot interaction. *Foundations and Trends in Robotics*, 4(2-3):105–223, 2016.

Andrea L Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, 172(6-7):716–737, 2008.

Michael Tomasello. *The cultural origins of human cognition.* Harvard University Press, Cambridge, MA, 1999.

Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne, and Henrike Moll. In search of the uniquely human. *Behavioral and Brain Sciences*, 28 (5):721–727, 2005.

Constantino Tsallis et al. *Introduction to nonextensive statistical mechanics*, volume 34. Springer, 2009.

Xiaojin Zhu. Machine teaching for Bayesian learners in the exponential family. In *Advances in Neural Information Processing Systems*, pages 1905–1913, 2013.

Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4083–4087, 2015.

Sandra Zilles, Steffen Lange, Robert Holte, and Martin Zinkevich. Teaching dimensions based on cooperative learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 135–146, 2008.