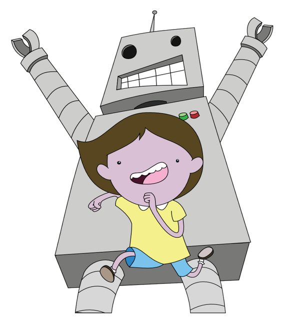
Optimal Cooperative Inference

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

RUTGERS
UNIVERSITY | NEWARK



shaftolab.com

Introduction

Learning through cooperation is a foundational principle underlying human-human (e.g., language, cultural evolution, education), human-machine (e.g., cooperative RL, social robotics, Bayesian teaching), and machine-machine (e.g., machine teaching) interaction.

Just as training error provides a framework for selecting models that generalize well, our *Cooperative Index* provides a framework for selecting models that can be explained well through data.

What are the implications of cooperation on representation?

Set up

Definitions

$h \in \mathcal{H}$ a concept in concept space

$D \in \mathcal{D}$ a data set in data space

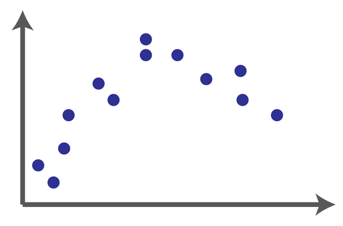
$P_L(h|D)$ the learner's posterior for a concept given a data set

$P_T(D|h)$ the teacher's probability of selecting a data set for communicating a given concept

Examples

Concept: boundary location
Data set: o's and x's.

Concept: order of polynomial
Data set: x, y pairs



Communication effectiveness

For discrete concept and data space,

$$P_L(\Theta|D) \to \mathbf{L} \in [0,1]^{|\mathcal{D}|\times|\mathcal{H}|}$$

(learner's inference matrix)

$$P_T(D|\Theta) \to \mathbf{T} \in [0,1]^{|\mathcal{D}|\times|\mathcal{H}|}$$

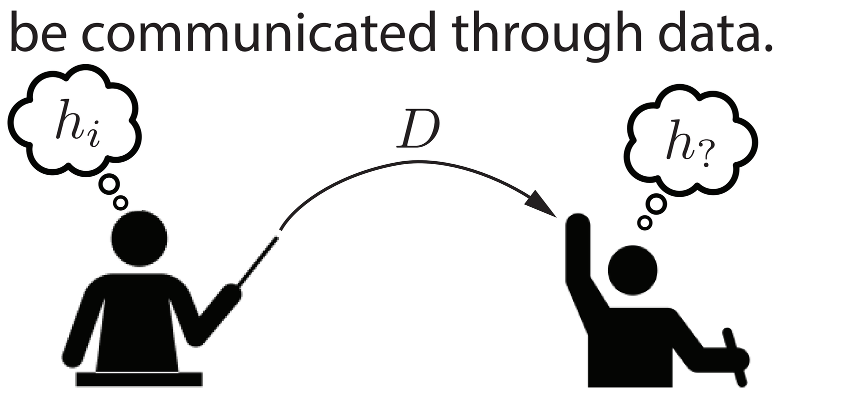
(teacher's selection matrix)

Define the *Transmission Index*:

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

$$0 \le \mathrm{TI}(\mathbf{L},\mathbf{T}) \le 1$$

Transmission index measures the effectiveness of communication—on average, how well a concept can be communicated through data.



For square matrices, Transmission Index = 1 iff the learner's inference and the teacher's selection matrices are the same permutation matrix.

Teaching Dimension (TD)

Concept: maps x to y.
h is *consistent* with D iff: $h(x) = y \;\; \forall (x,y) \in D$
D is a *teaching set* for h if h, but no other concept, is consistent with D.
Example: given consistency probability matrix (M)

	h_1	h_2
D_1	1	1
D_2	0	1

For h_1, no teaching set —> $TD(h_1) = \infty$
For h_2, teaching set is D_2 : $TD(h_2) = |D_2|$

Average Teaching Dimension [1]: $ATD(\mathcal{H}) = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} TD(h)$

ATD is finite iff the Transmission Index of M = 1, i.e., M is a permutation matrix.

Expected Teaching Dimension: $ETD(\mathcal{H}) = \frac{\sum_{h\in\mathcal{H}}\sum_{D\in\mathcal{D}} |D|\, P_L(h|D)\, P_T(D|h)}{\sum_{h\in\mathcal{H}}\sum_{D\in\mathcal{D}} P_L(h|D)\, P_T(D|h)}$

ETD is a generalization of ATD from deterministic to probabilistic setting.

Optimal cooperative inference

Cooperative inference: teacher's selection of data depends on what the learner is likely to infer and vice versa.

$$P_L(h|D) = \frac{P_T(D|h)\, P_{L_0}(h)}{P_L(D)} \quad (1a)$$

$$P_T(D|h) = \frac{P_L(h|D)\, P_{T_0}(D)}{P_T(h)} \quad (1b)$$

These coupled equations can be solved by fixed point iteration [2]. Machine teaching and Bayesian teaching are special cases (single iteration) of cooperative inference.

If the spaces are discrete and the priors are uniform, this iteration is the same as *Sinkhorn's algorithm* [3].

Sinkhorn's algorithm: starting with an *initial likelihood matrix*,

$$\mathbf{M} \in [0,1]^{|\mathcal{D}|\times|\mathcal{H}|}$$

repeat column normalization (1a) followed by row normalization (1b).

If the iteration converges, define the *Cooperative Index*

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

where the arguments to TI are the learner's inference and teacher's selection matrices at convergence.

Representation theorem for optimal cooperative inference:

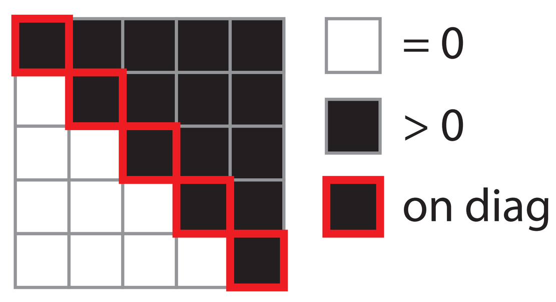
Let M be a nonnegative square matrix with at least one positive diagonal, then the following statements are equivalent:

(a) The cooperative index is optimal, i.e., CI(M) = 1;

(b) M has exactly one positive diagonal (an application of Sinkhorn's theorem [3]);

(c) M is a permutation of an upper-triangular matrix.

Examples:



TI = 0.29; CI = 1

$$\mathbf{M} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

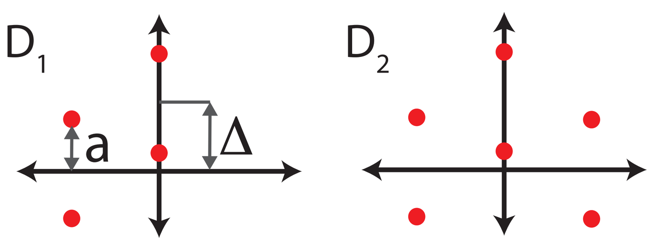
$$\mathbf{L}^{(k)} = \begin{pmatrix} 1-\frac{1}{2k} & \frac{1}{2k} \\ 0 & 1 \end{pmatrix}$$

$$\mathbf{T}^{(k)} = \begin{pmatrix} 1 & \frac{1}{2k+1} \\ 0 & 1-\frac{1}{2k+1} \end{pmatrix}$$

An application to likelihood choice

Concept: h_1 = linear fit h_2 = quadratic fit

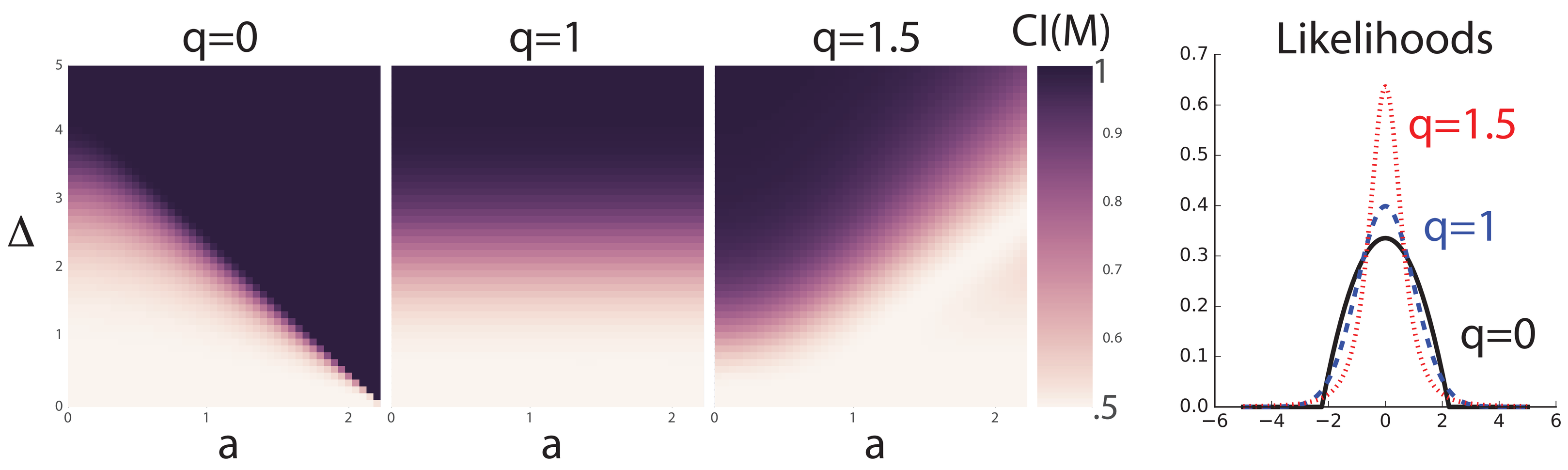
Data set: D_1 D_2



Given (a, Δ), construct M: MAP with q-Gaussian likelihoods

$$\mathbf{M} = \begin{array}{c|c|c} & h_1 & h_2 \\ \hline D_1 & \cdot & \cdot \\ D_2 & \cdot & \cdot \end{array}$$

(1a)
(1b)



Different likelihoods good at transmitting information in different regimes of signal to noise ratio.

Conclusions

- Introduced the Transmission and Cooperative Indices as metrics for the effectiveness of inference in standard and cooperative learning settings.
- Connected the Transmission Index with Teaching Dimension.
- Proved a representation theorem stating the conditions under which cooperation can yield optimally effective inference.

References:

[1] T. Doliwa, G. Fan, H. U. Simon & S. Zilles. Recursive teaching dimension, VC Dimension and sample compression. *Journal of Machine Learning Research*, 15(1):3107–3131, 2014.

[2] P. Shafto, N. D Goodman & T. L. Griffiths. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, 71:55–89, 2014.

[3] R. Sinkhorn & P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Acknowledgments: This research is sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146 to P.S. and S.Y. and also supported by NSF SMA-1640816 to P.S..