Generalizing the theory of cooperative inference

Pei Wang **Pushpi Paranamana** **Patrick Shafto**
Department of Mathematics & Computer Science, Rutgers University—Newark

Abstract

Cooperative information sharing is important to theories of human learning and has potential implications for machine learning. Prior work derived conditions for achieving optimal Cooperative Inference given relatively restrictive assumptions. We demonstrate convergence for any discrete joint distribution, robustness through equivalence classes and stability under perturbation, and effectiveness by deriving bounds from structural properties of the original joint distribution. We provide geometric interpretations, connections to and implications for optimal transport and to importance sampling, and conclude by outlining open questions and challenges to realizing the promise of Cooperative Inference.

1 Introduction

Cooperative information sharing is fundamental to human learning and finds applications in machine learning. The core idea of cooperation in human inference stems from work by Grice (Gricean Maxims; Grice, 1975) in linguistic pragmatics. Recent work in the linguistics literature has formalized the Rational Speech Act model (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Kao et al., 2014; Lassiter and Goodman, 2017), which builds on earlier models of cooperative information sharing (Shafto and Goodman, 2008; Shafto et al., 2014, 2012b). Indeed, these phenomena are not limited to language. Related models have been proposed to explain infants' learning from parents (Tomasello, 2009; Csibra and Gergely, 2009; Bonawitz et al., 2011; Shafto et al., 2012b; Buchsbaum et al., 2011; Gweon et al., 2014; Shneidman et al., 2016; Eaves Jr et al., 2016), learning from cooperative teachers (Shafto and Goodman, 2008; Shafto et al., 2014), and how people decide who to trust (Shafto et al., 2012a; Eaves Jr et al., 2016). This model is of interest for explaining human learning and communication, but it lacks overarching theory regarding when and why cooperation might facilitate learning and communication. Our paper is a step toward a general mathematical theory for this work.

Cooperative information sharing is of recent interest in machine learning. Explainability of machine learning models has been formalized in the Cooperative Inference (CI) framework (Yang and Shafto, 2017; Vong et al., 2018). There have also been several recent papers on learning from demonstrations that leverage cooperation in similar models (Ho et al., 2016, 2018). Finally, cooperative inverse reinforcement learning is explicitly centered around CI (Hadfield-Menell et al., 2016; Fisac et al., 2017).

We make four contributions toward strengthening the overarching theory of Cooperative Inference. (a) Section 3 proves convergence of CI for any rectangular matrix, which ensures that CI applies to any discrete model. (b) Cross ratio equivalence analysis in Section 4.1 shows that the space of possible joint distributions is reducible to only working with distributions that differ in cross ratio(s). This provides a natural geometric structure over possible machine learning models, which is a highly general and very interesting direction for future work detailed in Section 6. (c) Section 4.2 proves stability under perturbation which ensures robustness of inference where agents' beliefs differ. This shows that CI has the possibility of being viable in practice. (d) Section 5 provides general bounds on effectiveness that are derived from structural properties of the initial matrix, $\mathbf{M}$. Section 6 provides a geometric interpretation, and connects to optimal transport and importance sampling, and Section 7 conlcudes.

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

2 Overview and Background

All matrices in this paper are understood to be real, non-negative and have no zero rows or zero columns. Matrices are in uppercase and their elements are in the corresponding lowercase.

These matrices can be thought as joint distributions of models throughout. In more detail, let $\mathcal{H}$ be a concept space and $\mathcal{D}$ be a data space. For a given matrix $\mathbf{M}$, each column can be viewed as a concept in $\mathcal{H}$ and each row can be viewed as a data in $\mathcal{D}$. Normalizing by dividing the sum of its entries, $\mathbf{M}$ can be turned into a conditional distribution over $\mathcal{H}$ or $\mathcal{D}$.

In this paper, we study the cooperative communication between a teacher and a learner. Here, cooperation means that the teacher's selection of data depends on what the learner is likely to infer and vice versa. The idea of *cooperative inference* was introduced in (Yang et al., 2018). We now briefly review their work.

Definition 1. For a fixed *concept space* $\mathcal{H}$ and a *data space* $\mathcal{D}$, let $P_{L_0}(h)$ be the learner's prior of a *concept* h among $\mathcal{H}$ and $P_{T_0}(d)$ be the teacher's prior of selecting a *data* d from $\mathcal{D}$. The teacher's posterior of selecting d to convey h is denoted by $P_T(d|h)$ and the learner's posterior for h given d is denoted by $P_L(h|d)$. **Cooperative inference** is a system shown below:

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

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

where $P_L(d)$ and $P_T(h)$ are the normalizing constants.

Assuming uniform prior, (Yang et al., 2018) showed that Equation (1) can be solved using **Sinkhorn iteration** (**SK** for short; (Sinkhorn and Knopp, 1967a)). The solution (if it exists) depends only on the initial joint distribution matrix, $\mathbf{M}_{|\mathcal{D}|\times|\mathcal{H}|}$, which defines the consistency between data and concepts.

Sinkhorn iteration is simply the repetition of row and column normalization of $\mathbf{M}$. Denote the matrices obtained at the k^{th} row and column iteration of (1) by $\mathbf{L}^k$ and $\mathbf{T}^k$, respectively. Let their limits (if they exist) be $\mathbf{L} := \lim_{k\to\infty} \mathbf{L}^k$ and $\mathbf{T} := \lim_{k\to\infty} \mathbf{T}^k$.

Example 2. Consider a joint distribution matrix $\mathbf{M} = \begin{array}{cc} & \begin{array}{cc} h_1 & h_2 \end{array} \\ \begin{array}{c} d_1 \\ d_2 \end{array} & \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$, where $m_{ij} = 1$ if d_i is consistent with h_j and $m_{ij} = 0$ otherwise, for $i, j = 1, 2$. The SK iteration proceeds as the following: row normalization of $\mathbf{M}$ outputs: $\mathbf{L}^1 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$, column normalization of $\mathbf{L}^1$ outputs: $\mathbf{T}^1 = \begin{pmatrix} 1 & \frac{1}{3} \\ 0 & \frac{2}{3} \end{pmatrix}$. Iteratively, $\mathbf{L}^k = \begin{pmatrix} 1-\frac{1}{2k} & \frac{1}{2k} \\ 0 & 1 \end{pmatrix}$, $\mathbf{T}^k = \begin{pmatrix} 1 & \frac{1}{2k} \\ 0 & 1-\frac{1}{2k} \end{pmatrix}$, and the limits exist as $k \to \infty$: $\mathbf{L} = \mathbf{T} = \mathbf{M}^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

For this $\mathbf{M}$, a teacher and a leaner who reason *independently* can not reliably convey h_1 using $\mathcal{D}$; d_1, the only data that is consistent with h_1 is also consistent with h_2. However, a teacher and learner that assume *cooperation* can perfectly convey h_1 using d_1; in the converged joint distribution $\mathbf{M}^*$, d_1 is consistent only with h_1. Intuitively, a cooperative teacher will pick d_2 to teach h_2, because picking d_1 would cause confusion for the learner. Correspondingly, when receiving d_1, the cooperative leaner will reason that the teacher must intend to teach h_1, because otherwise he would pick d_2. In fact, the teaching between the cooperative pair is optimal, $\mathrm{CI}(\mathbf{M}) = 1$ (Definition 3).

The *Cooperative index* quantifies the effectiveness of the cooperative communication. It is the average probability that a concept in $\mathcal{H}$ can be correctly inferred by a learner given the teacher's selection of data.

Definition 3. Given $\mathbf{M}$ and assuming that SK iteration of (1) converges to a pair of matrices $\mathbf{L} = (l_{ij})$ and $\mathbf{T} = (t_{ij})$, we define the **cooperative index** as

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

Here, $\mathbf{L}\odot\mathbf{T}$ means the inner product between $\mathbf{L}$ and $\mathbf{T}$. The definition implies that $\mathrm{CI}(\mathbf{M})$ is invariant under row and column permutations of $\mathbf{M}$.

Next we define a few useful technical terms.

Definition 4. Let $A = (a_{ij})$ be an $n \times n$ matrix and S_n be the set of all permutations of $\{1, 2, \ldots, n\}$. For any $\sigma \in S_n$, the set of n-elements $\{a_{1\sigma(1)}, \ldots, a_{n,\sigma(n)}\}$ is called a **diagonal** of A. If every $a_{k\sigma(k)} > 0$, we say that the diagonal is **positive**. An element $a_{i_0 j_0}$ of A is called **on-diagonal** if it is contained in a positive diagonal, otherwise $a_{i_0 j_0}$ is called **off-diagonal**. In particular, A may have a positive off-diagonal element. We use $\overline{A}$ to denote the matrix obtained from A by setting all its off-diagonal elements into zeros. If A contains no positive off-diagonal element, i.e $A = \overline{A}$, A is said to have **total support**.

(Yang et al., 2018) focused on the case when the data set and the hypotheses set have the same size. They showed that $0 \leq \mathrm{CI}(\mathbf{M}) \leq 1$ for any $\mathbf{M}$ (if $\mathrm{CI}(\mathbf{M})$ exists). In particular, when $\mathbf{M}$ is a square matrix, they showed that Equation (1) has a solution if and only if $\mathbf{M}$ has at least one diagonal and $\mathrm{CI}(\mathbf{M})$ is optimal if and only if $\mathbf{M}$ has exactly one positive diagonal.

3 Convergence of rectangular matrices

It is typical that the sizes of a data set and a concept set are different. Therefore, considering only square models is too restrictive. We show that the solution of Equation (1) can be obtained using SK iteration for any rectangular joint distribution $\mathbf{M}$. This implies that cooperative inference can be performed on any discrete model.

First, we study the format of the limit of SK iteration on rectangular matrices. It is proven in (Sinkhorn and Knopp, 1967b) that the limit (if exists) of SK iteration on a **square** $\mathbf{M}$ is a single doubly stochastic matrix $\mathbf{M}^*$, i.e. $\mathbf{L} = \mathbf{T} = \mathbf{M}^*$. As the numbers of rows and columns are different in a **rectangular** $\mathbf{M}$, the limit of the SK iteration on $\mathbf{M}$ is a pair of distinct matrices $(\mathbf{L}, \mathbf{T})$, where, $\mathbf{L}$ is row normalized and $\mathbf{T}$ is column normalized. Such a pair is called *stable* defined below.

Definition 5. The **pattern** of a matrix A is the set of entries where $a_{ij} > 0$. Matrix B is said to have a **partial pattern** of A, denoted by $B \prec A$, if $a_{ij} = 0 \implies b_{ij} = 0$.

Definition 6. A pair of $u \times v$-matrices (P, Q) is called **stable** if column normalization of P equals Q and row normalization of Q equals P. A matrix is *stable* if it is contained in a *stable* pair.

Remark 7. If (P, Q) is *stable*, then P and Q are row and column normalized, respectively. *SK iteration* of P (or Q) results a sequence alternating between P and Q. Moreover, P and Q must have the same pattern.

As mentioned above, the limit of SK iteration is doubly stochastic for a square $\mathbf{M}$. The following proposition provides a similar analogy for the characteristics of the limit pair for rectangular $\mathbf{M}$.

Proposition 8. [1] *Suppose that (P, Q) is a **stable** pair of $u \times v$-matrices. Then up to permutations, P is a block-wise diagonal matrix of the form $P = \text{diag}(B_1, \ldots, B_k)^2$, where each B_i is row normalized and has a constant column sum denoted by c_i. In particular, $c_i = u_i/v_i$, where $u_i \times v_i$ is the dimension of B_i, for $i \in \{1, \ldots, k\}$.*

In addition to providing a convergence format for more general discrete joint distributions, the block diagonal form implies relations between subset of data and concepts that can be leveraged for developing structured models and joint distributions.

Let $(\mathbf{L}, \mathbf{T})$ be the limit pair of SK iteration on $\mathbf{M}$. $\mathbf{L}$ and $\mathbf{T}$ must have the same partial pattern of $\mathbf{M}$ as the SK iteration preserves zeros. Hence, the existence of a pair of *stable* matrices with partial pattern of $\mathbf{M}$ is necessary for the convergence of SK. In Proposition 9, we show that this condition is also sufficient.

Stable matrices with partial pattern of $\mathbf{M}$ can be partially ordered with respect to their patterns. We use $\overline{\mathbf{M}}$ to denote the matrix obtained from $\mathbf{M}$ by setting elements outside the maximum partial pattern to zeros. Note that elements outside the maximum partial pattern of a rectangular matrix shall be treated as off-diagonal elements in a square matrix.

Proposition 9. *A non-negative rectangular matrix $\mathbf{M}$ converges to a pair of stable matrices under SK iteration if and only if there exists a stable pair of matrices with partial pattern of $\mathbf{M}$.*

Proof. The 'only if' direction is clear from the above discussion. We now show the 'if' direction. Suppose there exists a *stable* pair (P, Q) such that $P \prec \mathbf{M}$. Let $\{\mathbf{L}^1, \mathbf{T}^1, \mathbf{L}^2, \mathbf{T}^2, \ldots\}$ be the sequence of matrices generated by SK iteration on $\mathbf{M}$, where $\mathbf{L}^k$ and $\mathbf{T}^k$ are row and column normalized respectively. This sequence is bounded since each element of $\mathbf{L}^k$ or $\mathbf{T}^k$ is bounded above by 1. Hence, according to Bolzano–Weierstrass theorem, the sequence must have as a limit a pair of matrices (may not be unique). Let $(\mathbf{L}, \mathbf{T})$ and $(\mathbf{L}', \mathbf{T}')$ be two pairs of such limits. To show that they are the same, we only need to prove that $\mathbf{L} = \mathbf{L}'$. Lemma A.1 and Remark A.2 indicate that $\mathbf{L}$ and $\mathbf{L}'$ must have the maximum partial pattern of $\mathbf{M}$, hence, they have the same pattern. Moreover, because they are limits, $\mathbf{L}$ and $\mathbf{L}'$ are *stable* as well. Therefore, it follows from Proposition 8 that up to permutations, $\mathbf{L}$ and $\mathbf{L}'$ have the same column sums. Further, Lemma A.3 implies that there exists X, Y and X', Y' such that $\mathbf{L} = X\overline{\mathbf{M}}Y$ and $\mathbf{L}' = X'\overline{\mathbf{M}}Y'$. Therefore, $\mathbf{L}$ and $\mathbf{L}'$ not only have the same row and column sums, but also are diagonally equivalent. Thus, Lemma A.4 implies that $\mathbf{L} = \mathbf{L}'$. □

In fact, the existence of a stable pair of matrices with partial pattern of $\mathbf{M}$ is naturally satisfied for all $\mathbf{M}$ under considerations(non-negative matrices without zero rows or zero columns), thus:

Proposition 10. *For any matrix $\mathbf{M}$, there exists a stable pair of matrices (P, Q) such that P and Q have a partial pattern of $\mathbf{M}$.*

Construction of a such (P, Q) is illustrated below.

Example 11. Let $\mathbf{M} = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{pmatrix}$ be a matrix without zero row or zero column. The first two columns are both non-zero implies that up to permutation, either $m_{11} \neq 0, m_{12} \neq 0$ or $m_{11} \neq 0, m_{22} \neq 0$. (1) If $m_{11} \neq 0, m_{22} \neq 0$, we may assume that $m_{23} \neq 0$ (up to permutation). In this case, let $A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$. (2) Otherwise $m_{11} \neq 0, m_{12} \neq 0$. (2-A) If further

[1] All proofs are included in the supplemental materials.
[2] The corresponding statement holds for Q too.

$m_{23} \neq 0$, let $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. (2-B) If $m_{23} = 0$, then $m_{13} \neq 0$. There must exist a non-zero element in the second row of $\mathbf{M}$. Up to permutation, we may assume that $m_{21} \neq 0$, let $A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$. In all cases, $A \prec \mathbf{M}$ is block-wise diagonal with each block in the form of a row or column vector. Let P, Q be row and column normalization of A respectively. It is straightforward to check that (P, Q) is *stable.*

Propositions 9 and 10 together imply our main result:

Theorem 12. *Every rectangular matrix converges to a pair of stable matrices under SK iteration.*

Remark 13. Theorem 12 is different from the classical convergence result for *scalar Sinkhorn iteration* (Menon and Schneider, 1969). Let $\mathbf{M}$ be a $u \times v$-matrix, $\mathbf{r} = (r_1, \ldots, r_u)^T$ be column vector and $\mathbf{c} = (c_1, \ldots, c_v)$ row vector. Similarly to the (regular) SK iteration, scalar SK iteration also alternates between row and column normalizing steps. However in each step of scalar SK, row-i (column-j) is normalized to have sum r_i (sum c_j) instead of 1. The convergence [3] of scalar SK on a given tuple $(\mathbf{M}, \mathbf{r}, \mathbf{c})$ has been intensively studied. A complete summary of equivalent convergence criteria are described in (Idel, 2016). Unfortunately, we can not simply apply these existing results: (1) Normalizing with respect to $\mathbf{r}$ and $\mathbf{c}$ has no statistical basis for our setting. (2) The convergence criteria are hard to verify. (3) For a given model, the teacher's data selection matrix needs not to be the same as the learner's concepts inferring matrix.

Corollary 14. *SK iteration of* $\mathbf{M}$ *and* $\overline{\mathbf{M}}$ *converge to the same limit. Therefore,* $CI(\mathbf{M}) = CI(\overline{\mathbf{M}})$.

Remark 15. Corollary 14 indicates that the elements outside the maximum partial pattern of $\mathbf{M}$ have no effect on the limit, and thus on *Cooperative Index.* For instance, in square matrices, such elements are precisely positive off diagonal entries. They are easy to detect using ideas from graph theory (Dulmage and Mendelsohn, 1958). Being able to pass to the maximal partial pattern makes the cooperative inference much more feasible. The convergence of $\overline{\mathbf{M}}$ is linear, where as the convergence of $\mathbf{M}$ slower (Soules, 1991).

In the rest of this paper, we assume $\mathbf{M}$ is square. With machinery developed in this section, similar analysis can be made for rectangular matrices.

4 Equivalence and sensitivity

We first introduce cross ratio equivalence and show that models whose joint distribution matrices are cross ratio equivalent, are the same under cooperative inference. Further, we will show that cooperative inference on models is robust to small perturbations on the joint distribution matrix $\mathbf{M}$. These features are essential because in most realistic situations we only have access to noisy data points, and because they provide flexibility in model choice by allowing selection of any joint distribution in a cross ratio equivalent class.

4.1 Cross Ratio Equivalence

Intuitively, given a model, SK iteration is a process that selects a representation for two cooperative agents. We develop a method to characterize the models that yield to the same representation.

SK iteration can be interpreted as a map between the initial and the limit matrices. Let $\mathcal{A}$ be the set of $n \times n$ matrices that has at least one positive diagonal, $\overline{\mathcal{A}} \subset \mathcal{A}$ be the set of $n \times n$ matrices with *total support* (Definition 4) and $\mathcal{B}$ be the set of $n \times n$ doubly stochastic matrices. According to (Sinkhorn and Knopp, 1967b), SK iteration of any $\mathbf{M} \in \mathcal{A}$ converges to a unique matrix $\mathbf{M}^* \in \mathcal{B}$. Hence SK iteration can be viewed as a map Φ from $\mathcal{A}$ to $\mathcal{B}$ where $\Phi(\mathbf{M}) = \mathbf{M}^*$.

It is important to note that Φ is not injective. For instance, in Example 22 below, with any choices of m_{12} and m_{32}, $\mathbf{M}$ maps to the same image under Φ. For a matrix $\mathbf{L} \in \mathcal{B}$, $\Phi^{-1}(\mathbf{L})$ is used to denote the set of all matrices in $\mathcal{A}$ that map to $\mathbf{L}$.

We will now introduce the notion-*cross ratio equivalence* between square matrices and show that the preimage set of a matrix $\mathbf{L} \in \mathcal{B}$ can be completely characterized by its *cross ratios.*

Definition 16. Let A, B be two $n \times n$ matrices and $D_1^A = \{a_{1,\sigma(1)}, \ldots, a_{n,\sigma(n)}\}$ and $D_2^A = \{a_{1,\sigma'(1)}, \ldots, a_{n,\sigma'(n)}\}$ be two positive diagonals of A determined by permutations $\sigma, \sigma' \in S_n$ (Definition 4). Denote the products of elements on D_1^A and D_2^A by $d_1^A = \Pi_{i=1}^n a_{i,\sigma(i)}, d_2^A = \Pi_{i=1}^n a_{i,\sigma'(i)}$ respectively. Then $\mathrm{CR}(D_1^A, D_2^A) = d_1^A / d_2^A$ is called the **cross ratio** between D_1^A and D_2^A of A. Further, let the diagonals in B determined by the same σ and σ' be $D_1^B = \{b_{1,\sigma(1)}, \ldots, b_{n,\sigma(n)}\}$ and $D_2^B = \{b_{1,\sigma'(1)}, \ldots, b_{n,\sigma'(n)}\}$. We say A is **cross ratio equivalent** to B, denoted by $A \overset{cr}{\sim} B$, if $d_i^A \neq 0 \iff d_i^B \neq 0$ and $\mathrm{CR}(D_1^A, D_2^A) = \mathrm{CR}(D_1^B, D_2^B)$ holds for any D_1^A and D_2^A.

Example 17. Let $A = \begin{pmatrix} 3 & 2 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}, B = \begin{pmatrix} 9 & 20 & 6 \\ 0 & 5 & 3 \\ 2 & 0 & 4 \end{pmatrix}$. A has three positive diagonals $D_1^A = \{a_{11}, a_{22}, a_{33}\}$, $D_2^A = \{a_{12}, a_{23}, a_{31}\}$ and $D_3^A = \{a_{13}, a_{22}, a_{31}\}$ with $d_1^A = 3, d_2^A = 2, d_3^A = 1$. B has three corresponding positive diagonals D_1^B, D_2^B and D_3^B with $d_1^B = 180, d_2^B = 120, d_3^B = 60$. It is easy to check that $\mathrm{CR}(D_i^A, D_j^A) = \mathrm{CR}(D_i^B, D_j^B)$ for any $i, j \in \{1, 2, 3\}$.

[3] Here convergence means the sequence generated by the iterative process converges to a single matrix.

Hence A is *cross ratio equivalent* to B.

Remark 18. (1) Definition 16 implies that if $A \overset{cr}{\sim} B$, then $\overline{A}$ and $\overline{B}$ (Definition 4) must have the same pattern. Otherwise there exists a positive diagonal D_1^A of A (or B) whose corresponding diagonal D_1^B in B (or A) contains zero ($d_1^A \neq 0$ whereas $d_1^B = 0$).
(2) Let A and B be matrices with the same pattern. Assume they both have N_d positive diagonals. To determine whether A is *cross ratio equivalent* to B, instead of examining $\binom{N_d}{2}$ pairs of cross ratios, it is sufficient to check whether $CR(D_1^A, D_i^A) = CR(D_1^B, D_i^B)$, $i \in \{1, \ldots, N_d\}$ holds for a fixed positive diagonal D_1^A.

Proposition 19. *Let* $\mathbf{M} \in A$ *be a consistency matrix and* $\mathbf{L} \in \mathcal{B}$ *be a doubly stochastic matrix. Then* $\mathbf{M} \in \Phi^{-1}(\mathbf{L})$ *if and only if* $\mathbf{M}$ *is cross ratio equivalent to* $\mathbf{L}$.

Sketch of proof. Let $\mathbf{M} \in \Phi^{-1}(\mathbf{L})$, we now show they have the same cross ratios. Since $\mathbf{M}$ and $\overline{\mathbf{M}}$ have exactly the same positive diagonals, we may assume that $\mathbf{M}$ has total support. Hence, (Sinkhorn and Knopp, 1967b) implies that there exist diagonal matrices $X = \text{diag}(x_1, \ldots, x_n)$ and $Y = \text{diag}(y_1, \ldots, y_n)$ such that $\mathbf{M} = X\mathbf{L}Y$. In particular, $m_{ij} = x_i \times l_{ij} \times y_j$ holds, for any element m_{ij}. Let $D_1^{\mathbf{M}} = \{m_{i,\sigma(i)}\}$, $D_2^{\mathbf{M}} = \{m_{i,\sigma'(i)}\}$ be two positive diagonals of $\mathbf{M}$ and $D_1^{\mathbf{L}} = \{l_{i,\sigma(i)}\}$, $D_2^{\mathbf{L}} = \{l_{i,\sigma'(i)}\}$ be the corresponding positive diagonals in $\mathbf{L}$. Then:

$$\begin{aligned} CR(D_1^{\mathbf{M}}, D_2^{\mathbf{M}}) &= \frac{\Pi_{i=1}^n m_{i,\sigma(i)}}{\Pi_{i=1}^n m_{i,\sigma'(i)}} = \frac{\Pi_{i=1}^n x_i \times l_{i,\sigma(i)} \times y_{\sigma(i)}}{\Pi_{i=1}^n x_i \times l_{i,\sigma'(i)} \times y_{\sigma'(i)}} \\ &= \frac{\Pi_{i=1}^n x_i \times \Pi_{i=1}^n l_{i,\sigma(i)} \times \Pi_{i=1}^n y_{\sigma(i)}}{\Pi_{i=1}^n x_i \times \Pi_{i=1}^n l_{i,\sigma'(i)} \times \Pi_{i=1}^n y_{\sigma'(i)}} \\ &= \frac{\Pi_{i=1}^n l_{i,\sigma(i)}}{\Pi_{i=1}^n l_{i,\sigma'(i)}} = CR(D_1^{\mathbf{L}}, D_2^{\mathbf{L}}) \qquad \square \end{aligned}$$

Corollary 20. *For* $\mathbf{M}_1, \mathbf{M}_2 \in \mathcal{A}$, *if* $\mathbf{M}_1 \overset{cr}{\sim} \mathbf{M}_2$ *then* $\text{CI}(\mathbf{M}_1) = \text{CI}(\mathbf{M}_2)$.

Proposition 19 captures the key ingredient, *cross ratios*, of a model. It indicates that cross ratio equivalent models can be treated the same for cooperative agents. Corollary 20 implies that their cooperative indices are the same and hence they have the same communication effectiveness. This can be very useful in practice: (1) Models with same representation can be effectively categorized, which avoids unnecessary implementation of similar models; (2) Models can be freely modified as long as the cross ratios are preserved which may increase computational efficiency.

4.2 Sensitivity Analysis

We now investigate sensitivity of Φ to perturbation of $\mathbf{M}$. Without loss of generality, we will assume that only one element in $\mathbf{M}$ is perturbed at a time as other perturbations may be treated as compositions of such. Let $\mathbf{M}^\epsilon = (m_{ij}^\epsilon)$ be a matrix obtained by varying the element m_{st} of $\mathbf{M} = (m_{ij})$ by ϵ, i.e. $m_{st}^\epsilon = m_{st} + \epsilon$ and $m_{ij} = m_{ij}^\epsilon$ for $(i,j) \neq (s,t)$. We may also assume that $\epsilon > 0$. Otherwise we may view $\mathbf{M}$ as a matrix obtained from a positive perturbation on $\mathbf{M}^\epsilon$.

Proposition 19 indicates that Φ is robust to any amount of perturbation on *off diagonal* elements. In more detail, suppose that both m_{st}^ϵ and m_{st} are off diagonal elements of $\mathbf{M}^\epsilon$ and $\mathbf{M}$ respectively. Then $\overline{\mathbf{M}} = \overline{\mathbf{M}}^\epsilon \implies \Phi(\mathbf{M}) = \Phi(\overline{\mathbf{M}}) = \Phi(\overline{\mathbf{M}}^\epsilon) = \Phi(\mathbf{M}^\epsilon) \implies \text{CI}(\mathbf{M}) = \text{CI}(\mathbf{M}^\epsilon)$. Thus we have:

Proposition 21. *Cooperative Inference is robust to any amount of off diagonal perturbations on* $\mathbf{M}$.

Example 22. Let $\mathbf{M} = \begin{array}{c c} & \begin{array}{ccc} h_1 & h_2 & h_3 \end{array} \\ \begin{array}{c} d_1 \\ d_2 \\ d_3 \end{array} & \begin{pmatrix} 1 & * & 1 \\ 0 & 1 & 0 \\ 1 & * & 1 \end{pmatrix} \end{array}$ be a consistency matrix. Suppose that the consistency between d_1, d_3 and h_2 can not be properly measured. With Proposition 21, CI($\mathbf{M}$) can still be easily obtained: $\overline{\mathbf{M}} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ converges to $\overline{\mathbf{M}}^* = \begin{pmatrix} 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}$ in one step of SK iteration. So we have that $\text{CI}(\mathbf{M}) = \text{CI}(\overline{\mathbf{M}}) = (4 \times 0.5^2 + 1^2)/3 = 2/3$.

Proposition 21 is not only important for sensitivity analysis, but also practical to efficiently perform cooperative inference as mentioned in Remark 15. For instance, if one $*$ in Example 22 is positive, it takes infinite many steps of SK iteration for $\mathbf{M}$ to reach its limit, whereas it takes only one step for $\overline{\mathbf{M}}$.

Proposition 21 also implies the main theorem in (Yang et al., 2018) stating CI($\mathbf{M}$) is optimal if $\mathbf{M}$ is a permutation of a triangular matrix. For an $n \times n$ *triangular* matrix $\mathbf{M} = (m_{ij})$, all the elements except $m_{i,i}$ are off diagonal. To efficiently apply cooperative inference, one only needs to consider $\overline{\mathbf{M}} = \text{diag}(m_{11}, \ldots, m_{nn})$. SK iteration on $\overline{\mathbf{M}}$ converges to $I_n = \text{diag}(1, \ldots, 1)$ in one step. Therefore, we have $\text{CI}(\mathbf{M}) = 1$.

By analogy, Corollary 14 implies that CI is robust to any perturbation on elements that are off maximal partial pattern for *rectangular* matrices as well.

Perturbations for *on-diagonal* elements are more complicated and interesting. To obtain $\mathbf{M}^\epsilon$, one may either perturb an on-diagonal element of $\mathbf{M}$ or perturb a zero element of $\mathbf{M}$ introducing a new diagonal(s) for $\mathbf{M}^\epsilon$.

A celebrated result in (Sinkhorn, 1972) shows that $\Phi : \mathcal{A} \to \mathcal{B}$ is a continuous function:

Theorem 23 (Continuity of SK iteration)**.** $\Phi(\mathbf{M}^\epsilon)$ *converges to* $\Phi(\mathbf{M})$ *as* $\mathbf{M}^\epsilon \to \mathbf{M}$.

Here, distance between matrices are measured by the maximum element-wise difference, e.g. $d(\mathbf{M}, \mathbf{M}^\epsilon) = \epsilon$.

This implies that small on-diagonal perturbations on a model with joint distribution $\mathbf{M}$, yield close solutions for cooperative inference.

Example 24. Let $\mathbf{M} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, $\mathbf{M}^{\epsilon_1} = \begin{pmatrix} 1.5 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, $\mathbf{M}^{\epsilon_2} = \begin{pmatrix} 1.1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, $\mathbf{M}^{\epsilon_3} = \begin{pmatrix} 1 & 1 & 0.1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ and $\mathbf{M}^{\epsilon_4} = \begin{pmatrix} 1 & 1 & 0.5 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$. Apply SK iterations on $\mathbf{M}$ and $\mathbf{M}^{\epsilon_i}$, we have: $\Phi(\mathbf{M}) = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0 & 0.5 \end{pmatrix}$, $\Phi(\mathbf{M}^{\epsilon_1}) = \begin{pmatrix} 0.534 & 0.466 & 0 \\ 0 & 0.534 & 0.466 \\ 0.466 & 0 & 0.534 \end{pmatrix}$, $\Phi(\mathbf{M}^{\epsilon_2}) = \begin{pmatrix} 0.508 & 0.492 & 0 \\ 0 & 0.508 & 0.492 \\ 0.492 & 0 & 0.508 \end{pmatrix}$, $\Phi(\mathbf{M}^{\epsilon_3}) = \begin{pmatrix} 0.478 & 0.478 & 0.044 \\ 0 & 0.5228 & 0.478 \\ 0.522, & 0 & 0.478 \end{pmatrix}$, $\Phi(\mathbf{M}^{\epsilon_4}) = \begin{pmatrix} 0.423 & 0.423 & 0.155 \\ 0 & 0.577 & 0.423 \\ 0.577 & 0 & 0.423 \end{pmatrix}$. It is clear that for perturbations on the same location, the variation on the limit matrix decreases as the size of the perturbation gets smaller. Moreover, perturbations of the same size cause different variations on the limits depending on whether a new diagonal is introduced. For instance, $\mathbf{M}^{\epsilon_4}$ introduces a new diagonal to $\mathbf{M}$ whereas $\mathbf{M}^{\epsilon_2}$ does not. Although both are 0.5 away from $\mathbf{M}$, after SK iteration $d(\Phi(\mathbf{M}), \Phi(\mathbf{M}^{\epsilon_2})) = 0.034$ and $d(\Phi(\mathbf{M}), \Phi(\mathbf{M}^{\epsilon_4})) = 0.155$.

In the following example, we illustrate how one may effectively bound the variation in the limit in terms of ϵ, even for perturbations that introduce new diagonals. However, in general, Φ is not Lipschitz [4].

Example 25. Let $\mathbf{M} = \begin{pmatrix} a_{11} & a_{12} & 0 \\ 0 & a_{22} & a_{23} \\ a_{31} & 0 & a_{33} \end{pmatrix}$ and $\mathbf{M}^{\epsilon} = \begin{pmatrix} a_{11} & a_{12} & \epsilon \\ 0 & a_{22} & a_{23} \\ a_{31} & 0 & a_{33} \end{pmatrix}$. While $\mathbf{M}$ has only two diagonals D_1 and D_2 with products of elements $d_1 = a_{11}a_{22}a_{33}$ and $d_2 = a_{12}a_{23}a_{31}$, the perturbation introduces one more diagonal D_3 with $d_3 = \epsilon \cdot a_{22}a_{31}$ to $\mathbf{M}^{\epsilon}$. The Birkhoff-von Neumann theorem (Theorem A.5) guarantees that doubly stochastic matrices $\Phi(\mathbf{M})$ and $\Phi(\mathbf{M}^{\epsilon})$ can be written as *convex combinations* of permutation matrices as shown below:

$$\Phi(\mathbf{M}) = \theta_1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \theta_2 \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \theta_1 & \theta_2 & 0 \\ 0 & \theta_1 & \theta_2 \\ \theta_2 & 0 & \theta_1 \end{pmatrix}$$

$$\begin{aligned}\Phi(\mathbf{M}^{\epsilon}) &= \alpha_1 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} + \alpha_3 \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ 0 & \alpha_1 + \alpha_3 & \alpha_2 \\ \alpha_2 + \alpha_3 & 0 & \alpha_1 \end{pmatrix},\end{aligned}$$

where $\theta_1 + \theta_2 = 1, \alpha_1 + \alpha_2 + \alpha_3 = 1$ and $\theta_i, \alpha_j > 0$. Notice that the variation between $\Phi(\mathbf{M})$ and $\Phi(\mathbf{M}^{\epsilon})$ is caused by α_3, we will now derive an upper bound for α_3. Since Φ preserves cross ratios, evaluating a cross ratio, for example $\mathrm{CR}(D_3, D_1)$, in both $\Phi(\mathbf{M}^{\epsilon})$ and $\mathbf{M}^{\epsilon}$ we have that:

$$\frac{\alpha_3(\alpha_2 + \alpha_3)}{\alpha_1^2} = \frac{d_3}{d_1} = \epsilon \cdot \frac{a_{22}a_{31}}{a_{11}a_{22}a_{33}} := \epsilon \cdot A_1 \qquad (2)$$

Since $\alpha_1 + \alpha_2 + \alpha_3 = 1$, we may assume that $\alpha_1 < 1/2$ Substituting $\alpha_2 + \alpha_3 = 1 - \alpha_1$ into Equation (2), we get $\frac{\alpha_3(1-\alpha_1)}{\alpha_1^2} = \epsilon \cdot A_1$ and this implies that:

$$\alpha_3 = \epsilon \cdot A_1 \cdot \frac{\alpha_1^2}{1 - \alpha_1} \leq \epsilon \cdot A_1 \cdot \frac{1}{2},$$

where the last '$\leq$' holds because $\frac{\alpha_1^2}{1-\alpha_1}$ reaches its maximum at $\alpha_1 = \frac{1}{2}$ for $\alpha_1 < \frac{1}{2}$. Thus, α_3 is bounded above by a constant multiple of ϵ.

The next proposition explores how sensitive Φ is to perturbations on its images. Thus, given two doubly stochastic matrices in $\mathcal{B}$, we will measure the distance between their preimages under Φ.

Proposition 26. *Let $\mathbf{L}^1, \mathbf{L}^2 \in \mathcal{B}$. If $d(\mathbf{L}^1, \mathbf{L}^2) \leq \epsilon$, for any $\mathbf{M}^1 \in \Phi_n^{-1}(\mathbf{L}^1)$ with total support, there exist a $\mathbf{M}^2 \in \Phi_n^{-1}(\mathbf{L}^2)$ and a constant C such that $d(\mathbf{M}^1, \mathbf{M}^2) \leq C \cdot \epsilon$.*

In fact, restricting to matrices with total support, Φ can be amended into a homeomorphism (see Supplemental Materials). Viewing SK iteration as a representation selecting process, the homeomorphic property of Φ indicates that such process preserves important information needed to reconstruct the original model.

5 Lower Bound for CI

Cooperative Index measures the effectiveness of the cooperative communication. However, for a given consistency matrix $\mathbf{M}$, in order to calculate $\mathrm{CI}(\mathbf{M})$ one needs to obtain $\Phi(\mathbf{M})$ by *SK iterations*, which sometimes can be an expensive process. We provide bounds on $\mathrm{CI}(\mathbf{M})$ that do not require computing SK.

.

First, we derive a uniform bound for $\mathrm{CI}(\mathbf{M})$ which only depends on the size of $\mathbf{M}$.

Proposition 27. *For an $n \times n$ matrix $\mathbf{M}$, $\mathrm{CI}(\mathbf{M}) \geq \frac{1}{n}$ with the equality when $\mathbf{M}$ is uniformly distributed.*

Proof. Let $\mathbf{M}^* = \left(m_{ij}^*\right)_{n \times n}$ be the limit of $\mathbf{M}$ under SK iteration. By Generalized Mean Inequality, we have $\left(\frac{\sum_{ij} m_{ij}^*}{n^2}\right)^2 \leq \frac{\sum_{ij}(m_{ij}^*)^2}{n^2}$. Since $\mathbf{M}^*$ is doubly stochastic, we have $\sum_{ij} m_{ij}^* = n$ and it follows that $\sum (m_{ij}^*)^2 \geq 1$. Therefore, $\mathrm{CI}(\mathbf{M}) = \frac{\sum (m_{ij}^*)^2}{n} \geq \frac{1}{n}$. $\square$

[4]The authors would like to thank Yue Yu for pointing out counterexamples.

Notice that, as the size of $\mathbf{M}$ increases the above bound is not effective. However, the number of positive diagonals a matrix consists can be small regardless of its size. Next, we provide another lower bound for $\mathrm{CI}(\mathbf{M})$ that depends only on the number of positive diagonals.

Proposition 28. *For an $n \times n$ matrix $\mathbf{M}$ with d positive diagonals,* $\mathrm{CI}(\mathbf{M}) \geq 1/d$.

Proof. Since $\mathbf{M}$ is a square matrix, the limit of SK iteration is a unique doubly stochastic matrix $\mathbf{M}^*$. Therefore, by Birkhoff-von Neumann theorem $\mathbf{M}^* = \sum_{i=1}^{d} \theta_i P_i$ and by Definition 3 we have:

$$\begin{aligned}\mathrm{CI}(\mathbf{M}) &= \frac{1}{n}\mathbf{M}^* \odot \mathbf{M}^* = \frac{1}{n}\left(\sum_{i=1}^{d}\theta_i P_i\right) \odot \left(\sum_{i=1}^{d}\theta_i P_i\right) \\ &\geq \frac{1}{n}\sum_{i}^{d}\theta_i P_i \odot \theta_i P_i \overset{(1)}{=} \sum_{i}^{d}\theta_i^2 \overset{(2)}{\geq} \frac{1}{d}\end{aligned}$$

Equality (1) holds because each P_i is a permutation matrix and so $P_i \odot P_i = n$. Inequality (2) is obtained from Generalized Mean Inequality as $\sum_{i=1}^{d} \theta_i = 1$. □

Such a bound makes sense because CI measures the effectiveness of the cooperative communication. Each diagonal is a representation for communication. $\mathrm{CI}(\mathbf{M})$ decreases as the number of diagonals increases. The optimal $\mathrm{CI}(\mathbf{M})$ is achieved when $\mathbf{M}$ has only one diagonal, i.e.$\mathbf{M}$ is upper triangular up to permutation.

Example 29. Consider $\mathbf{M}$ in Example 22. We have $n = 3, d = 2$ and $\mathrm{CI}(\mathbf{M}) = \left(0.5^2 \times 4 + 1\right)/3 = 2/3 > 1/2 = 1/d > 1/3 = 1/n$.

Above example shows that when the number of diagonals is small, Proposition 28 provides a good bound. However, counting the number of diagonals of an $n \times n$ matrix can also be computationally expensive. Next, we provide a much more accessible bound.

Definition 30. An $n \times n$ matrix A is **indecomposable** if there exists no permutation matrices P and Q such that $PAQ = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22}\end{pmatrix}$ where, A_{11} and A_{22} are square submatrices.

Proposition 31. *For any $n \times n$ matrix $\mathbf{M}$,* $\mathrm{CI}(\mathbf{M}) \geq \frac{1}{\eta - 2n + \tau + 1}$*, where η is the number of positive elements and τ is the number of indecomposable components.*

Proof. Let $\mathbf{M}^*$ be the SK limit of $\mathbf{M}$ and η^* and τ^* be the number of positive elements and the number of indecomposable components in $\mathbf{M}^*$, respectively. Then according to (Brualdi, 1982), $\mathbf{M}^*$ has a Birkhoff-von Neumann decomposition with k permutation matrices, where $k \leq \eta^* + \tau^* - 2n + 1$. Further note that $\eta + \tau \leq \eta^* + \tau^*$. Hence, similarly as in the proof of Proposition 28, we have that $\mathrm{CI}(\mathbf{M}) = \mathrm{CI}(\mathbf{M}^*) \geq \frac{1}{k} \geq \frac{1}{\eta^* + \tau^* - 2n + 1} \geq \frac{1}{\eta + \tau - 2n + 1}$. □

Example 32. Consider an $n \times n$ matrix $\mathbf{M}$ of the form

$$\begin{pmatrix} * & & & & & & * \\ & \ddots & & & & \iddots & \\ & & * & * & * & & \\ & & * & * & * & & \\ & & * & * & * & & \\ & \iddots & & & & \ddots & \\ * & & & & & & * \end{pmatrix}$$

where, any * is a positive number and the rest are zeros. Notice that, it quickly becomes challenging to count d when n is large.

When $n = 5$, we have $\eta = 13, \tau = 2, d = 12$, and so $\mathrm{CI}(\mathbf{M}) \geq 1/(\sigma + \tau - 2n + 1) = 1/6 > 1/d$.

6 Connections to other work

Geometric interpretation. Cooperative inference is intuitive given the geometric interpretation of SK iteration, which has been long known and favored in the study of contingency tables (Fienberg et al., 1970; Borobia and Cantó, 1998). Each joint distribution matrix $\mathbf{M} = (m_{ij})$ of dimension $u \times v$ can be viewed as a point in the $(uv-1)$-dimension simplex, $\mathcal{S}_{uv} = \{(m_{11}, \ldots, m_{uv}) : m_{ij} \geq 0, \sum_{ij} m_{ij} = 1\}$. In (Fienberg, 1968), the author showed that, in $\mathcal{S}_{uv}$, positive[5] matrices with the same cross-product ratios[6] form a special case of determinantal manifold $\mathcal{H}$, which is studied in (Room, 1938). In particular, for the 2×2 case, authors of (Fienberg et al., 1970) built a homeomorphic map from $\mathcal{H}$ to the unit square and illustrated the convergence path of successive SK iterations in the unit square. Similarly, non-negative joint-distribution matrices with the same pattern locate on a lower-dimension face of $\mathcal{S}_{uv}$ and matrices with the same cross ratios form a further subspace.

Optimal transport. Choosing a suitable distance to compare probabilities is a key problem in statistical machine learning. When the probability space is a metric space, optimal transport distances (earth mover's in computer vision) define a powerful geometry to compare probabilities (Villani, 2008). Optimal transport distances are a fundamental family of distances for probability measures and histograms of features. (Cuturi, 2013) proposed a new family of optimal transport distances, **Sinkhorn distance**, that look at transport problems from a maximum entropy perspective. The resulting optimum is a proper distance which can be computed through Sinkhorn's matrix scaling. Let C be the cost matrix, $\mathbf{r}$ and $\mathbf{c}$ be the marginal distributions for a given optimal transport problem. The matrix $\mathbf{M}^*$ that optimizes the Sinkhorn distance can be obtained by applying $(\mathbf{r}, \mathbf{c})$-scalar SK iteration on $\mathbf{M} = e^{-\lambda \cdot C}$, where λ is the regularization parameter. Cuturi (2013) proved that this Sinkhorn algorithm can be computed at a speed that is several orders of magnitude faster than that of transport solvers.

[5] Every element is positive.

[6] Two matrices with the same cross-product ratios must be cross ratio equivalent (Definition 16).

Optimal transport with Sinkhorn distance provides a powerful tool for domain adaptation. Courty et al. (2015) proposed the following method: first link two domains based on prior knowledge (build an initial cost matrix C); then learn an optimal distribution matrix $\mathbf{M}^*$ (w.r.t, Sinkhorn distance) from one domain to the other by applying scalar SK iteration on $\mathbf{M} = e^{-\lambda \cdot C}$. If certain transports should never happen, i.e. elements of C are allowed to be ∞, then the corresponding $\mathbf{M}$ will be a sparse matrix. Remark 13 notes that scalar SK iteration of a sparse $\mathbf{M}$ may not converge and the convergence criteria can not be easily verified. Whereas, in the case that both domains have uniform marginal distributions, Theorem 12 guarantees the existence of the optimal distribution matrix $\mathbf{M}^*$ for any choice of the cost matrix. In the sparse case, the convergence rate can be further sped up by first identifying and removing off-diagonal elements, i.e. turning $\mathbf{M}$ into $\overline{\mathbf{M}}$, then applying SK iteration on $\overline{\mathbf{M}}$ as in Remark 15. More importantly, our results in Section 4 capture the essential features of the Sinkhorn distance approach. Proposition 19 implies that cost matrices that are cross ratio equivalent lead to the same optimal transport. Proposition 23 indicates that optimal distribution matrix $\mathbf{M}^*$ is continuous to the choice of regularization parameter λ, which can be used to discretize the range of λ.

Importance Sampling Cooperative inference can be interpreted as selection of optimal distributions for importance sampling. A straightforward view is to consider Equation (1). Given a joint distribution $\mathbf{M}$, let $\mathbf{T}^1$ be the column normalization of $\mathbf{M}$. The ij^{th}-element $P_{\mathbf{T}^1}(d_i|h_j)$ of $\mathbf{T}^1$ can be viewed as the teacher's initial probability of selecting d_i to convey h_j. Once d_i is observed, the learner needs to sample a concept to match d_i. Assume that the learner's prior on $\mathcal{H}$ is uniform. To minimize the variance, the learner should sample from the optimal distribution: $P_{\mathbf{L}^1}(h_j|d_i) = \frac{P_{\mathbf{T}^1}(d_i|h_j)}{\sum_j P_{\mathbf{T}^1}(d_i|h_j)}$. Thus the optimal learner's matrix $\mathbf{L}^1$ is the row normalization of $\mathbf{T}^1$. Similarly, based on $\mathbf{L}^1$, to reduce variance, the teacher should sample according to the matrix $\mathbf{T}^2$, the column normalization of $\mathbf{L}^1$. This alternating process is precisely SK iteration. So, the solution of cooperative inference is not only the stable limit of a sequence of optimal distributions for individual d and h, but also the only doubly stochastic matrix cross ratio equivalent to $\mathbf{M}$.

A more subtle and interesting version of importance sampling is also achieved by cooperative inference. Let $\mathbf{M}$ be an $n \times n$ joint distribution. Suppose that the teacher aims to convey the whole set of n concepts simultaneously. To do so, the teacher must teach n different data points at once—one data point per concept. This is equivalent to picking a map from $\mathcal{D}$ to $\mathcal{H}$, i.e. a permutation $\sigma \in S_n$ as $|\mathcal{D}| = |\mathcal{H}| = n$. Then $P_T(\mathcal{D}_\sigma|\mathcal{H}_T) = \Pi_i P_T(d_{\sigma(i)}|h_i)$ is the probability that the teacher picks σ to teach and $P_L(\mathcal{H}_L|\mathcal{D}_\sigma) = \Pi_i P_L(h_i|d_{\sigma(i)})$ is the probability that given σ, the learner's inference completely matches the teacher's intention. Therefore, in order to efficiently estimate the communication accuracy, $P(\mathcal{H}_L|\mathcal{H}_T) = \sum_{\sigma \in S_n} P_L(\mathcal{H}_L|\mathcal{D}_\sigma) P_T(\mathcal{D}_\sigma|\mathcal{H}_T)$, one must sample permutations that make large positive contributions to the summation. Such an importance sampling is attained by Cooperative inference for the following reasons. (1) SK iteration completely removes the probability of sampling off-diagonal elements. Thus a σ will be sampled only if it could lead to a prefect teaching. (2) Beichl and Sullivan (1999) proved that the limit of SK iteration maximizes entropy for doubly stochastic matrices that have the same pattern as $\mathbf{M}$, and further they showed that this is the ideal property for sampling positive diagonals.

Other connections. See supplemental materials for pointers to other connections.

7 Conclusion

Cooperative inference holds promise as a theory of human-human, human-machine, and machine-machine information sharing. An impediment to realizing this promise is the lack of foundational results related to convergence, robustness, and effectiveness. We have addressed each of these limitations, including specific results showing the convergence of Cooperative Inference via SK iteration for any rectangular matrix, equivalence classes of models in terms of their cross-ratios, continuity of SK iteration which implies stability to perturbation, and several different bounds on the effectiveness of Cooperative Inference that can be derived from the original model. We also demonstrated connections and implications through geometric interpretations of Cooperative inference, optimal transport, and importance sampling. Important open questions include developing methods for modifying machine learning models to increase the efficacy and furthering our understanding of the representational implications of Cooperative Inference.

Acknowledgements

This material is based on research sponsored by the Air Force Research Laboratory and DARPA under agreement number FA8750-17-2-0146. 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, and NSF NSF-1549981.

References

Isabel Beichl and Francis Sullivan. Approximating the permanent via importance sampling with application to the dimer covering problem. *Journal of computational Physics*, 149(1):128–147, 1999.

Michael W Berry, Murray Browne, Amy N Langville, V Paul Pauca, and Robert J Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis*, 52(1):155–173, 2007.

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.

Alberto Borobia and Rafael Cantó. Matrix scaling: A geometric proof of sinkhorn's theorem. *Linear algebra and its applications*, 268:1–8, 1998.

Richard A Brualdi. Notes on the birkhoff algorithm for doubly stochastic matrices. *Canadian Mathematical Bulletin*, 25(2):191–199, 1982.

Daphna Buchsbaum, Alison Gopnik, Thomas L Griffiths, and Patrick Shafto. Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, 120(3):331–340, 2011.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *arXiv preprint arXiv:1507.00504*, 2015.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

Fanny Dufossé and Bora Uçar. Notes on birkhoff–von neumann decomposition of doubly stochastic matrices. *Linear Algebra and its Applications*, 497:108–115, 2016.

Andrew L Dulmage and Nathan S Mendelsohn. Coverings of bipartite graphs. *Canadian Journal of Mathematics*, 10(4):516–534, 1958.

Baxter S Eaves Jr, Naomi H Feldman, Thomas L Griffiths, and Patrick Shafto. Infant-directed speech is consistent with teaching. *Psychological review*, 123 (6):758, 2016.

Stephen E Fienberg. The geometry of an r× c contingency table. *The Annals of Mathematical Statistics*, 39(4):1186–1190, 1968.

Stephen E Fienberg et al. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.

Jaime F Fisac, Monica A Gates, Jessica B Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S Shankar Sastry, Thomas L Griffiths, and Anca D Dragan. Pragmatic-pedagogic value alignment. *arXiv preprint arXiv:1707.06354*, 2017.

Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Dave Golland, Percy Liang, and Dan Klein. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics, 2010.

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, Peter Cole, Jerry Morgan, et al. Logic and conversation. *1975*, pages 41–58, 1975.

Leonid Gurvits. Classical deterministic complexity of edmonds' problem and quantum entanglement. In *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 10–19. ACM, 2003.

Hyowon Gweon, Hannah Pelton, Jaclyn A Konopka, and Laura E Schulz. Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, 132(3):335–341, 2014.

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.

Mark K Ho, Michael L Littman, Fiery Cushman, and Joseph L Austerweil. Effectively learning from pedagogical demonstrations. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2018.

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

Daniel Lassiter and Noah D Goodman. Adjectival vagueness in a bayesian model of interpretation. *Synthese*, 194(10):3801–3836, 2017.

MV Menon and Hans Schneider. The spectrum of a nonlinear operator associated with a matrix. *Linear Algebra and its applications*, 2(3):321–334, 1969.

Todd K Moon, Jacob H Gunther, and Joseph J Kupin. Sinkhorn solves sudoku. *IEEE Transactions on Information Theory*, 55(4):1741–1746, 2009.

EE Osborne. On pre-conditioning of matrices. *Journal of the ACM (JACM)*, 7(4):338–345, 1960.

Oliver Pretzel. Convergence of the iterative scaling procedure for non-negative matrices. *Journal of the London Mathematical Society*, 2(2):379–384, 1980.

Thomas Gerald Room. *The geometry of determinantal loci*, volume 1. The University Press, 1938.

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, Baxter Eaves, Daniel J Navarro, and Amy Perfors. Epistemic trust: Modeling children's reasoning about others' knowledge and intent. *Developmental science*, 15(3):436–447, 2012a.

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, 2012b.

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.

Laura Shneidman, Hyowon Gweon, Laura E Schulz, and Amanda L Woodward. Learning from others and spontaneous exploration: A cross-cultural investigation. *Child Development*, 87(3):723–735, 2016.

Richard Sinkhorn. Continuous dependence on a in the dad theorems. *Proceedings of the American Mathematical Society*, 32(2):395–398, 1972.

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

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

Richard Sinkhorn and Paul Knopp. Problems involving diagonal products in nonnegative matrices. *Transactions of the American Mathematical Society*, 136:67–75, 1969.

George W Soules. The rate of convergence of sinkhorn balancing. *Linear algebra and its applications*, 150: 3–40, 1991.

Michael Tomasello. *Why we cooperate*. MIT Press, Cambridge, MA, 2009.

Helge Tverberg. On sinkhorn's representation of nonnegative matrices. *Journal of Mathematical Analysis and Applications*, 54(3):674–677, 1976.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Wai Keen Vong, Ravi B Sojitra, Anderson Reyes, Scott Cheng-Hsin Yang, and Patrick Shafto. Bayesian teaching of image categories. In *Proceedings of the Annual Conference of the Cognitive Science Society*, 2018.

Scott Cheng-Hsin Yang and Patrick Shafto. Explainable artificial intelligence via bayesian teaching. In *NIPS 2017 workshop on Teaching Machines, Robots, and Humans*, 2017.

Scott Cheng-Hsin Yang, Yue Yu, Arash Givchi, Pei Wang, Wai Keen Vong, and Patrick Shafto. Optimal cooperative inference. In *AISTATS*, volume 84 of *Proceedings of Machine Learning Research*, pages 376–385. PMLR, 2018.