A mathematical theory of cooperative communication

Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto

October 8, 2019

Abstract

Cooperative communication plays a central role in theories of human cognition, language, development, and culture, and is increasingly relevant in human-algorithm and robot interaction. Existing models are algorithmic in nature and do not shed light on the statistical problem solved in cooperation or on constraints imposed by violations of common ground. We present a mathematical theory of cooperative communication that unifies three broad classes of algorithmic models as approximations of Optimal Transport (OT). We derive a statistical interpretation for the problem approximated by existing models in terms of entropy minimization, or likelihood maximizing, plans. We show that some models are provably robust to violations of common ground, even supporting online, approximate recovery from discovered violations, and derive conditions under which other models are provably not robust. We do so using gradient-based methods which introduce novel algorithmic-level perspectives on cooperative communication. Our mathematical approach complements and extends empirical research, providing strong theoretical tools derivation of a priori constraints on models and implications for cooperative communication in theory and practice.

Significance Statement

Human are unique in our ability to accumulate knowledge quickly individually and over generations. Cooperative communication is a proposed explanation, but has only been formalized in algorithmic models, preventing a priori analysis that could allow us to determine the conditions under which such an explanation could be true. We unify existing probabilistic models of cooperative communication as instances of the more general theory of Optimal Transport, and prove results regarding the potential and limitations of existing models. Our results provide a mathematical foundation and tools for analyzing for understanding and developing richer, more precise theories of human development, cognition, language, education, and evolution.

1 Introduction

Computational-level modeling and rational analysis of cognition are first-principles approaches to developing mathematically precise models of human learning. These approaches assume that people are optimized to learn about their environment, and thus require specifying the relevant properties of the environment. If analysis of the environment does not yield strong constraints on learning, then this approach can be rightly criticized for allowing modelers to exploit unconstrained degrees to fit the data to the model [Marcus and Davis, 2013] and offer little or no theoretical insight [Jones and Love, 2011]. The existence of ambiguity about the relevant environmental features itself raises questions about whether it would ever be possible to provide a compelling, a priori
analysis of any important aspects of human cognition, an urgent goal in light of concerns about reproducibility of research findings [Collaboration et al., 2015].

One area of deep theoretical and practical importance is cooperative communication. Humans’ ability to cooperatively share information is invoked across literatures—including language, cognitive development, cultural anthropology, and robotics—to explain people’s ability to rapidly accumulate knowledge across ontogeny and phylogeny. Indeed, these accounts argue that people have evolved a specialized ecological niche [Tomasello, 1999, Boyd et al., 2011] and learning mechanisms [Csibra and Gergely, 2009, Grice et al., 1975, Sperber and Wilson, 1986], which explain our remarkable abilities to learn; however, we lack mathematical theories that would allow us to validate such claims. Cooperative communication is therefore a strong candidate for a priori theoretical analysis, and our lack of a strong theory blocks progress in multiple disciplines.

Cooperative communication also differs from other domains in which computational-level theorizing has been applied. Unlike categorization [Anderson, 1991], causal reasoning [Griffiths and Tenenbaum, 2006], and intuitive theorizing [Tenenbaum et al., 2006], analysis of cooperative communication does not hinge on analyzing true properties of the observable world. For cooperative communication, the relevant environment is people’s beliefs and actions. Thus, the required input for analysis is models of belief updating, action selection, and the application of those in models to cooperative information sharing through recursive Theory-of-Mind reasoning.

Models of belief updating and action selection, as probabilistic inference [Chater et al., 2008, Tenenbaum et al., 2011] and choice [Luce, 2012, Sutton et al., 1998] respectively, have a long history in the literature on human learning, and have recently been combined into models of cooperation and cooperative information sharing. Such models of cooperative communication have appeared in cognitive science [Shafto and Goodman, 2008a, Shafto et al., 2014], cognitive development [Eaves Jr et al., 2016, Bonawitz et al., 2011, Bridgers et al., ress], linguistic pragmatics [Goodman and Stuhlmüller, 2013], and robotics [Ho et al., 2018, Ho et al., 2016, Fisac et al., 2017]. While specific models differ in their details, none offers a strong mathematical framework that sheds light on how to understand their differences or how to understand the problem itself beyond their algorithmic prescriptions.

Among the largest questions in cooperative communication is that of common ground [Clark et al., 1991]. How do two communicators maintain enough agreement in order to successfully communicate? Models of cooperative information sharing assume that communicators know exactly what their partners’ beliefs are. However, this is mathematically impossible and practically implausible. It is therefore an open theoretical question to understand what models could explain successful cooperative communication and under what conditions.

Existing models have proposed algorithmic solutions to predicting cooperative communication. These models implement recursive reasoning common to Theory-of-Mind reasoning problems, deriving predictions from the results of approximations of such reasoning. But, they are unable to predict the outcome without simulation and are therefore unable to shed light on the statistical problem that they are solving or provide a priori constraints on conditions under which we would expect models to succeed. The lack of a mathematical theory prevents asking and answering arguably the central question of cooperative communication: how could cooperative communication succeed when partners cannot know each others’ beliefs?

In this work, our goal is demonstrate the possibility of strongly grounded, a priori analysis of cognition,
using the case of cooperative communication, and therefore providing constraints on the conditions under which cooperation can explain impressive features of human learning that inform theoretical debates across literatures. We present a computational-level analysis that results in a rigorous mathematical framework of cooperative communication as optimal transport \cite{Monge, 1781, Villani, 2008, Peyré et al., 2019}. We show previous probabilistic models to be approximations of a particular statistical inference, with associated strengths and limitations. We prove conditions under which models can address the challenge of communicating without common ground. We conclude with implications for existing debates and the broader endeavor of theory and replicability in the cognitive sciences.

2 Cooperative communication as a problem of optimal transport

Communication is a pair of processes considered between two agents, that we will refer to as a teacher and a learner, wherein the teacher selects actions and the learner draws inferences based on those data. A communicative act is a process by which data are selected to change the learner’s beliefs toward the goals of the teacher. Optimal transport provides a mathematical framework for formalizing movement of one distribution to another, and therefore a framework for modeling communication. By recasting communication as belief transport we will gain access to a rich suite of mathematical and computational techniques for understanding and analyzing the problem of cooperative communication.

2.1 Background on Optimal Transport

Optimal Transport has been discovered in many settings and fields \cite{Villani, 2008}. Examples include logistics and economics \cite{Kantorovich, 2006, Koopmans, 1949}, linear programming \cite{Dantzig, 1949}, earth mover’s distance in computer vision \cite{Brenier, 1991}. The general usefulness of optimal transport can be credited to the simplicity of the problem it solves. The original formulation, attributable to Monge \cite{Monge, 1781}, involves minimizing the effort required to move a pile of dirt from one shape to another. Where Monge saw dirt, we may see any probability distribution.

Optimal Transport is an optimization paradigm where the goal is to transform one probability distribution into another with a minimal cost. Formally, let $\mathbf{r} = (r_1, \ldots, r_n)$ and $\mathbf{c} = (c_1, \ldots, c_m)$ be probability vectors of length $n$ and $m$ respectively. A joint distribution matrix $\mathbf{P} = (P_{ij})$ of dimension $n \times m$ is called a coupling\footnote{A general definition can be made for any pair of probability measures.} of $\mathbf{r}$ and $\mathbf{c}$ if $\mathbf{P}$ has marginals $\mathbf{r}$ and $\mathbf{c}$, i.e. $\sum_{j=1}^{m} P_{ij} = r_i$ and $\sum_{i=1}^{n} P_{ij} = c_j$.

Denote the set of all couplings between $\mathbf{r}$ and $\mathbf{c}$ by $U(\mathbf{r}, \mathbf{c})$. Given an $n \times m$ non-negative cost matrix $\mathbf{C} = (C_{ij})$, **Optimal transport (OT)** is the problem of finding a coupling $\mathbf{P}^*$ that minimizes the cost of transporting $\mathbf{r}$ into $\mathbf{c}$, thus,

$$P^* = \arg \min_{\mathbf{P} \in U(\mathbf{r}, \mathbf{c})} \langle \mathbf{C}, \mathbf{P} \rangle := \arg \min_{\mathbf{P} \in U(\mathbf{r}, \mathbf{c})} \sum_{i=1}^{n} \sum_{j=1}^{m} C_{ij} \cdot P_{ij}.$$  

Here, $\langle \mathbf{C}, \mathbf{P} \rangle$ is the inner product between $\mathbf{C}$ and $\mathbf{P}$. $\mathbf{P}^*$ is called an **OT plan** and $d_{\mathbf{C}}(\mathbf{r}, \mathbf{c}) := \langle \mathbf{C}, \mathbf{P}^* \rangle$ is called the **OT distance** between $\mathbf{r}$ and $\mathbf{c}$ given cost $\mathbf{C}$. 
Recently, an approximation of OT distance, Sinkhorn distance, was proposed \cite{Cuturi, 2013}. Consider an entropy regularized optimal transport problem: for \( \lambda > 0 \), find \( P^{(\lambda)} \) such that

\[
P^{(\lambda)} = \arg\min_{P \in \mathcal{U}(r, c)} \langle C, P \rangle - \frac{1}{\lambda} \sum_{i} \sum_{j} P_{ij} \log P_{ij}.
\]

\( P^{(\lambda)} \) is called a Sinkhorn plan with parameter \( \lambda \), and \( d^{(\lambda)}_C(r, c) := \langle C, P^{(\lambda)} \rangle \) is called the Sinkhorn distance. The main result of \cite{Cuturi, 2013} states that \( P^{(\lambda)} \to P^* \) as \( \lambda \to \infty \), hence the Sinkhorn distances converge to the OT distance as \( \lambda \) increases. This regularization is both intuitive given the geometry of the OT problem and more computational efficient than evaluating OT distance directly \cite{Cuturi, 2013, Peyré et al., 2019}.

Sinkhorn plans can be computed through Sinkhorn scaling with linear convergence \cite{Knight, 2008}. \((r, c)\)-Sinkhorn scaling (SK) of a matrix \( M \) is simply the iterated alternation of row normalization of \( M \) with respect to \( r \) and column normalization of \( M \) with respect to \( c \). When marginal distributions are uniform, we sometimes call it Sinkhorn iteration. It is shown in \cite{Cuturi, 2013} that for a given cost matrix \( C \), \( P^{(\lambda)} \) can be obtained by applying \((r, c)\)-Sinkhorn scaling on \( P^{(\lambda)} \), where matrix \( P^{[\lambda]} \) is the element-wise exponential of \(-\lambda \cdot C\), thus:

\[
P^{[\lambda]} := e^{-\lambda C} = (e^{-\lambda C_{ij}})_{n \times m}.
\]

**Example 1.** Let \( r = c = (\frac{3}{8}, \frac{5}{8}) \), and the cost matrix be \( C = \left( \begin{array}{cc} \log_2 1 & \frac{1}{2} \log_2 2 \\ \frac{1}{3} \log_2 1 & \log_2 1 \end{array} \right) \). For say \( \lambda = 3 \), we may obtain \( P^{(3)} \) by applying SK scaling on \( P^{[3]} = \left( \begin{array}{cc} e^{-3 \log_2 1} & e^{-3 \frac{1}{2} \log_2 2} \\ e^{-3 \frac{1}{3} \log_2 1} & e^{-3 \log_2 1} \end{array} \right) = \left( \begin{array}{cc} 1 & 1/2 \\ 1/4 & 1 \end{array} \right) \), which proceeds as the following: first, row normalization of \( P^{[3]} \) with respect to \( r \). It can be realized as: (a) row normalizing \( P^{[3]} \) such that each row sum equals 1, this results: \( \left( \begin{array}{cc} 2/3 & 1/3 \\ 1/5 & 4/5 \end{array} \right) \); (b) multiplying the first row by 3/8 and second row by 5/8 results: \( L_0 = \left( \begin{array}{cc} 1/4 & 1/8 \\ 1/8 & 1/2 \end{array} \right) \). Then similarly, column normalization of \( L_0 \) with respect to \( c \) outputs \( T_1 = \left( \begin{array}{cc} 1/4 & 1/8 \\ 1/8 & 1/2 \end{array} \right) \). As \( L_0 = T_1 \), the SK scaling has converged with \( P^{(3)} = T_1 \). In general, many iterations may be required to reach the limit.

The convergence results regarding Sinkhorn scaling was firstly proved in \cite{Sinkhorn and Knopp, 1967}, and further developed in various fields \cite{Idel, 2016}. SK converges at a speed that is several orders of magnitude faster than other transport solvers \cite{Cuturi, 2013}. Spurred by these works in OT, as an alternative to the more conventional Kullback–Leibler divergence, Sinkhorn approximation to the Wasserstein distance—a classical family of OT distances—has been extensively applied in machine learning algorithms, for example in domain adaptation \cite{Courty et al., 2017} and training GANs \cite{Arjovsky et al., 2017}.

OT and SK scaling are widespread interest within machine learning and across domains. To our knowledge, existing applications of OT assume that the problem is one of directly transforming a distribution into another, which is importantly different from the setting in communication.

### 2.2 Cooperative communication as optimal transport

A communicative act is a transportation that aligns other agents’ beliefs on hypotheses with the acting agent’s goal. Without loss, we will say the acting agent’s goal is their own belief. Unlike most applications of OT, in
communication, beliefs about hypotheses are moved through transmission of data, the natural language between rational agents.

Theory of cooperative communication is a single problem comprised of interactions between two processes: action selection and inference. The teacher and learner have beliefs about hypotheses, which are represented as two probability distributions. The process of teaching is to select data that move the learner’s beliefs from some initial state, to a final desired state. The teacher’s selection incurs a cost. The teacher selects data that minimize cost of achieving their goal. The process of learning is then, given the data selected by the teacher, infer the beliefs of the teacher. Communication is successful when the learner’s belief, given the teacher’s data, is moved to the target distribution, subject to some costs.

**Unifying OT Framework.** Formally, each process, teaching and learning, can be modeled as a classical OT problem. Let \( \mathcal{H} \) be a hypothesis space and \( \mathcal{D} \) be a data space. Denote the common ground between agents: the shared priors on \( \mathcal{H} \) and \( \mathcal{D} \) by \( P_0(\mathcal{H}) \) and \( P_0(\mathcal{D}) \), the shared joint likelihood matrix over \( \mathcal{H} \) and \( \mathcal{D} \) by \( M \). In general, up to normalization, \( M \) is simply a non-negative matrix which also specifies the consistency between data and hypotheses. In cooperative communication, a teacher’s goal is to minimize the cost of transforming hypotheses into data points. We define the teacher’s cost matrix \( C_T = (C_{Tij})_{|D| \times |H|} \) as:

\[
C_{Tij} = -\log P_L(h_j|d_i) + S_T(d_i),
\]

where \( P_L(h_j|d_i) \) is the learner’s posterior for hypothesis \( h_j \) given data \( d_i \) and \( S_T(d_i) \) is the teacher’s expense of selecting data \( d_i \). Thus, data \( d \) is good for a teacher who wishes to communicate \( h \) if \( d \) has low selecting expense and the learner assigns a high probability to \( h \) after updating with \( d \). Symmetrically, a learner’s cost matrix \( C_L = (C_{Lij})_{|D| \times |H|} \) is defined as \( C_{Lij} = -\log P_T(d_i|h_j) + S_L(h_j) \), where \( P_T(d_i|h_j) \) is the teachers’ choice of data \( d_i \) given hypothesis \( h_j \) and \( S_L(h_j) \) is the learner’s prior on the hypothesis \( h_j \).

A **teaching plan** is a matrix \( T = (T_{ij})_{|D| \times |H|} \), where each element \( T_{ij} \) represents the probability of the teacher selecting \( d_i \) to convey \( h_j \). Similarly a **learning plan** is a matrix \( L = (L_{ij})_{|D| \times |H|} \). As an OT problem, when both \( T \) and \( L \) are couplings between \( P(\mathcal{H}) \) and \( P(\mathcal{D}) \), their efficiency can be quantified by the total cost of teaching and learning.

More generally, for arbitrary pair of communication plans (not necessary OT plans), **Cooperative Index (CI)** between \( T \) and \( L \), was introduced in [Yang et al., 2018]:

\[
CI(T, L) := \frac{1}{|H|} \sum_j \sum_i L_{ij} T_{ij},
\]

CI\((T, L)\) quantifies communication effectiveness in terms of the average probability that a hypothesis can be correctly inferred by a learner given the teacher’s selection of data. Its range is between 0 and 1.

**Remark 2.** Using Sinkhorn distance approximation as in Section 2.1, optimal cooperative communication plans can be obtained through **Sinkhorn Scaling** on the common ground \( M \) between agents as illustrated below.

\(^2\)Data, \( d_i \), are consistent with a hypothesis, \( h_j \), when \( M_{ij} > 0 \).
Example 3. For simplicity, assume zero expense of selecting data and uniform priors on both $\mathcal{D}$ and $\mathcal{H}$. A natural estimation of the learner is a naive learner who interprets data according to its joint probability with hypotheses. In this case, the teacher may approximate the learner’s posterior matrix by $L_0$, the row normalization of the joint distribution $M$. Hence the teacher’s cost matrix is $C^T = -\log L_0$. As in Eq (1), the optimal teaching plan with regularizer $\lambda$, denoted by $T^{(\lambda)}$, can be obtained by applying Sinkhorn iterations on $T^{[\lambda]}$, i.e.

$$T^{(\lambda)} = SK(T^{[\lambda]}) = SK(e^{-\lambda C^T}) = SK(L_0^{[\lambda]}),$$

where $L_0^{[\lambda]}$ represents the matrix obtained from $L_0$ by raising each element to power of $\lambda$. Symmetrically, the optimal learning plan with regularizer $\lambda$, denoted by $L^{(\lambda)}$, can be reached by Sinkhorn iteration on $L^{[\lambda]} = e^{-\lambda C^L} = T_0^{[\lambda]}$, where the learner’s approximation of teacher’s matrix is $T_0$, the column normalization of $M$.

2.3 Properties of Optimal Transport and Sinkhorn scaling

Much is known about OT and SK. Here we summarize some of their important features and establish a few technical notations, which we will use extensively in the following sections. For simplicity, we will focus on square matrices, similar analysis can be made for rectangular matrices using machinery developed in [Wang et al., 2019].

Definition 4. Let $A = (A_{ij})$ be an $n \times n$ square matrix and $S_n$ be the set of all permutations of $\{1, 2, \ldots, n\}$. For any $\sigma \in S_n$, the set $D^A_\sigma$ consists of $n$-elements $\{A_{1,\sigma(1)}, \ldots, A_{n,\sigma(n)}\}$ is called a diagonal of $A$ determined by $\sigma$. If every $A_{i,\sigma(k)} > 0$, we say that the diagonal is positive. $D^A_\sigma$ is called a leading diagonal if the product of elements on $D^A_\sigma$, $d^A_\sigma = \Pi_{i=1}^n A_{i,\sigma(i)}$, is the largest among all diagonals of $A$.

Numerous results on SK iteration have been proved. For instance, SK iteration of a square $M$ converges if and only if $M$ has at least one positive diagonal [Sinkhorn and Knopp, 1967] and the limit must be a doubly stochastic matrix, which can be written as a convex combination of permutation matrices [Dufossé and Uçar, 2016]. These will be used to analyze the dynamics of OT planing (Section 4.2). SK iteration can be viewed as a continuous map [Sinkhorn, 1972]. For positive matrices, we will illustrate, this map is in fact smooth, in particular differentiable. This allows to show that the unifying OT framework is robust to various perturbations on the common grounds and to derive precise gradient formula to recover optimal communication plans (Section 4.3).

Definition 5. Let $A, B$ be two $n \times n$ square matrices and $D^A_\sigma$ and $D^B_\sigma$ be two diagonals of $A$ determined by permutations $\sigma, \sigma'$. Denote the products of elements on $D^A_\sigma$, $D^B_\sigma$ by $d^A_\sigma, d^B_\sigma$. Then $\text{CR}(D^A_\sigma, D^B_\sigma) = d^A_\sigma / d^B_\sigma$ is called the cross-product ratio between $D^A_\sigma$ and $D^B_\sigma$. Further, let the diagonals in $B$ determined by the same $\sigma$ and $\sigma'$ be $D^B_\sigma$ and $D^B_{\sigma'}$. We say $A$ is cross-ratio equivalent to $B$, if $d^A_\sigma \neq 0 \iff d^B_\sigma \neq 0$ and $\text{CR}(D^A_\sigma, D^B_\sigma) = \text{CR}(D^B_\sigma, D^B_{\sigma'})$ holds for any $\sigma, \sigma'$.

There is a strong geometric intuition that underlies SK scaling via the cross-product ratio. Matrices converge to the same limit under SK scaling if and only if they are cross-ratio equivalent [Wang et al., 2019]. The space $\mathcal{K}(M)$ formed by all matrices with the same cross-product ratios as $M$ is a special manifold [Fienberg, 1968].
scaling moves $M$ along a path in $K$ to $M^*$ — the unique intersection between $K$ and the manifold determined by the linear marginal conditions [Fienberg et al., 1970].

**Remark 6.** Preservation of cross-product ratios over SK scaling implies that optimal communicative plans are invariant under cost matrices constructed for agents with different depths of SK. For instance in Example 3 instead of being naive, a learner could also be pragmatic who would reason about his estimation of the teacher’s reasoning and interpret data accordingly using Bayes’ rule, i.e. proportional to elements of $L_1$ which is row normalization of $T_0$. Denote the teacher’s cost matrices based on $L_0$ and $L_1$ by $C^T_0$ and $C^T_1$ respectively. Because both $L_0$ and $L_1$ are derived from $M$, they are cross-ratio equivalent. So they have the same SK limit, i.e. optimal teaching plans for both $C^T_0$ and $C^T_1$ are the same. Thus, even though the teacher’s estimation of the learner was not accurate, the teacher’s plan is still optimal. Indeed, optimal teaching plans are equivalent for any learning matrix that is cross-ratio equivalent to the common ground $M$.

Strengthened by the rich theory of OT, our framework can be used to solve much broader questions. For example, general existence of OT planning between two arbitrary probability measures are well-studied [Villani, 2008]. This provides us machinery to study cooperative communications between agents even when $H$ and $D$ are continuous spaces. Furthermore, OT plannings enjoy many other desirable features such as: the optimality passes to subsets, convexity of OT distance, which enables broader perspectives on approximate inference and computation of optimal plans.

### 3 Existing models approximate Optimal Transport

Existing models of cooperative communication are approximations of OT. We demonstrate this point by expressing representatives of three broad classes of models as OT. The first class of models [Shafto and Goodman, 2008b, Shafto et al., 2014, Shafto et al., 2012] are based on the classic Theory of Mind recursion, which compute exact answers for the case of $\lambda = 1$. The second class includes models that compute the first step of the recursion [Goodman and Stuhlmüller, 2013, Eaves Jr and Shafto, 2016] and approximate the OT solution with this probability distribution. The third class includes models that compute the first step by selecting the data that maximize the probability of the hypothesis [Hadfield-Menell et al., 2016, Ho et al., 2016, Ho et al., 2018, Fisac et al., 2017], and approximate the complete transport plan with a single data for each hypothesis. These models have characteristic strengths and limitations, which the literature has yet to explore in their fullness. After unifying these approaches as OT, we will derive and contrast these consequences, as well as expose new, yet unexplored algorithms and computational tools through which we may understand communication.

#### 3.1 Full recursive reasoning is optimal transport

Cooperative models that build on the classic Theory of Mind recursion include cooperative inference [Yang et al., 2018, Wang et al., 2019] and pedagogical reasoning [Shafto and Goodman, 2008b, Shafto et al., 2014, Shafto et al., 2012].
In this section, we will briefly review the work on cooperative inference and illustrate how Bayesian inference models fit into our unifying OT framework.

The core of cooperative inference between two agents is that the teacher’s selection of data depends on what the learner is likely to infer and vice versa. Let $P_L(h)$ be the learner’s prior of hypothesis $h \in H$, $P_T(d)$ be the teacher’s prior of selecting data $d \in D$, $P_T(d|h)$ be the teacher’s posterior of selecting $d$ to convey $h$ and $P_L(h|d)$ be the learner’s posterior for $h$ given $d$. Cooperative inference emphasizes that agents’ optimal communication plans, $T^\star = P_T(D|H)$ and $L^\star = P_L(H|D)$ should satisfy the following system of interrelated equations for any $d \in D$ and $h \in H$, where $P_L(d)$ and $P_T(h)$ are the normalizing constants:

$$
P_L(h|d) = \frac{P_T(d|h) P_L(h)}{P_T(d)} \quad P_T(d|h) = \frac{P_L(h|d) P_T(d)}{P_T(h)} \quad (5)
$$

Extending Yang et al., 2018’s results on uniform priors, we show that:

**Proposition 7.** Optimal communication plans, $T^\star$ and $L^\star$, of a cooperative inference problem with arbitrary priors denoted by $P_T(D)$ and $P_L(H)$, can be obtained through Sinkhorn scaling.

As a direct consequence, cooperative inference is a special case of the unifying OT framework with $\lambda = 1$. Let $M$ be the joint distribution, $r = P_T(D)$ be the teacher’s prior and $c = P_L(H)$ be learner’s prior. According to the proof of Proposition 7 after cooperative inference, the teacher’s posterior selection matrix $T^\star$ is the limit of $(c, r)$-SK scaling of $\tilde{L}_0 = (P_L(h_j|d_i)P_T(d_i))$. On the other hand, under the unifying OT framework, the optimal teaching plan $T^{(\lambda=1)}$ is the limit of $(c, r)$-SK scaling of $\tilde{L}_0 = (P_L(h_j|d_i)e^{S_T(d_i)})$ based on Eq (4). When the teacher’s expense $S_T(d_i)$ of selecting $d_i$ is proportional to $\log P_T(d_i)$, $T^{(1)} = T^\star$. Symmetrically, one may check the same holds for $L^{(1)} = L^\star$.

### 3.2 One-step approximate inference

Direct implementation of the recursive Theory of Mind above requires repeated computation of the normalizing constant for Bayesian inference. This is computationally challenging for large scale problems and has been argued to be algorithmically implausible as a model of human cognition. For these reasons, models including Rational Speech Act (RSA) theory Goodman and Stuhlmüller, 2013 and Bayesian Teaching Eaves Jr and Shafto, 2016 Eaves Jr et al., 2016 model cooperation as a single step of recursion. To simplify exposition, we focus on the RSA model.

RSA models the communication between a speaker and a listener, formalizing cooperation that underpins pragmatic language Goodman and Stuhlmüller, 2013 Grice et al., 1975 Levinson, 2000 Clark, 1996. A pragmatic speaker selects an utterance optimally to inform a naive listener about a world state. Whereas a pragmatic listener interprets an utterance rationally and infers the state using one step Bayesian inference. This represents a communicative process where a speaker-listener pair can be viewed as a teacher-learner pair with world states-utterances being hypotheses-data points, respectively.

---

*All proofs are included in the Appendix.*
RSA distinguishes among three levels of inference: a naive listener, a pragmatic speaker and a pragmatic listener. A naive listener interprets an utterance according to its literal meaning. That is, given a joint distribution $M$ the naive listener’s selection of $h_i$ given $d_j$ is the $ij$-th element of $L_0$ which is obtained by row normalization of $M$.

A pragmatic speaker selects an utterance to convey the state such that maximizes utility. In particular, he picks $d_i$ to convey $h_j$ by soft-max optimizing expected utility,

$$P_T(d_i|h_j) \propto e^{\alpha U(d_i;h_j)},$$

(6)

where utility is given by $U(d_i;h_j) = \log L_0(h_j|d_i) - S(d_i)$, which minimizes the surprisal of a naive listener when inferring $h_j$ given $d_i$ with an utterance cost $S(d_i)$. This formulation is the same as one step of SK iteration in OT framework (see Eq (1) and Eq (2)) where $C_T = -U(d; h)$, $\lambda = -\alpha$ as in [Goodman and Stuhlmüller, 2013].

Next, a pragmatic listener reasons about the pragmatic speaker and infers the hypothesis using Bayes rule,

$$P_L(h_j|d_i) \propto P_T(d_i|h_j)P_L(h_j),$$

(7)

Here $P_T(d_i|h_j)$ represents the listener’s reasoning on the speaker’s data selection and $P_L(h_j)$ is the learner’s prior. This is again one step recursion of OT framework of $\lambda = 1$.

As described above, teaching and learning plans in RSA are one-step approximations of the OT plans. Although limited recursion and optimization are realistic assumptions in psychology [Goodman and Stuhlmüller, 2013], in many cases, such approximations are far from optimal. For example, world states can often be organized from most abstract to least, which yield a upper triangular joint distribution matrix. Fully recursive model as cooperative inference would output a diagonal matrix as optimal plan [Yang et al., 2018], which achieves the highest efficiency, whereas cooperative index of one step approximation is much lower. Furthermore, one-step approximation plans are much more sensitive to agents’ estimation of the other agent. For instance, a pragmatic speaker’s teaching plan is tailored for a naive listener, in contrast the optimal plan obtained through fully recursion is stable for any listener derived from the same common ground as in Remark 6.

3.3 Single-step argmax approximation

This idea arises in cooperative inverse reinforcement learning, where instead of selecting acts probabilistically, the maximum probability action is always selected [Hadfield-Menell et al., 2016, Ho et al., 2016, Ho et al., 2018, Fisac et al., 2017]. In particular, [Fisac et al., 2017] introduces Pragmatic-Pedagogic Value alignment, a framework that is grounded in empirically validated cognitive models related to pedagogical teaching and pragmatic learning.

Pragmatic value alignment formalizes the cooperation between a human and a robot who perform collaboratively with the goal of achieving the best possible outcome according to an objective. The true objective however is only known to the human. The human performs pedagogical actions to teach the true objective to
the robot. After observing human’s action, the robot, who is pragmatic, updates his beliefs and perform an action that maximizes expected utility. The human, observing this action, can then update their beliefs about the robot’s current beliefs and choose a new pedagogic action. Denote actions by $d$ and objectives by $h$. We can see that when the human performs the action they act as a teacher and when robot is performing the action it is vice versa.

In particular, the pedagogic human selects an action $d_i$ to teach the objective $h_j$ according to Eq (6), where $U$ is the utility that captures human’s best expected outcome. As described in Section 3.2 this is equivalent to a single step recursion in the OT framework.

Denote robot’s prior belief distribution on the objectives by $P_R(h_j)$. The robot interprets the human’s action $d_i$ rationally and updates his beliefs about the true objective using Bayes rule as Eq (7). Then acting as a teacher, the robot chooses an action that maximizes the human’s expected utility using argmax function:

$$P_R(d_i) = \arg \max_{d_R} \sum_{d_H, h_j} U(d_R, d_H; h) \cdot P_R(h_j)$$

where, $d_R$ denotes the robot’s actions and $d_H$ denotes the human’s actions. Unlike in human teaching where the plans are chosen proportionally to a probability distribution, here the robot chooses a deterministic action using argmax function.

As described above, pragmatic-pedagogic value alignment is modeled by computing a single step of OT and selecting the action that maximizes the outcome. Unlike cooperative inference, which tends to select the leading diagonal of the common ground $M$ as $\lambda \to \infty$ (Proposition 8), pragmatic-pedagogic value alignment selects the maximal element in each column of $M$, which is not even guaranteed to form a plan to distinguish every hypothesis. As a consequence, a drawback of such argmax method is that for large hypothesis spaces, multiple hypotheses may reach argmax on the same data which lead to low communication efficiency. Further, analysis in Section 4 shows that deterministic methods as argmax are much less robust to perturbations.

4 Analyzing models of cooperative communication

With prior models unified as instances of Optimal Transport via Sinkhorn scaling, we analyze the properties of these models. We focus on two of the most important aspects: understanding the models from statistical perspective and in the context of realistic assumptions about common ground.

4.1 Full recursive reasoning is statistically and information theoretically optimal

Having demonstrated the equivalence of SK scaling and full recursive reasoning (Proposition 7), strong statistical justifications of fully Bayesian recursive reasoning follow immediately as SK scaling is optimal in the senses of entropy minimization and likelihood maximization [Csiszar, 1975, Darroch and Ratcliff, 1972, Brown et al., 1993].

Sinkhorn scaling solves entropy minimization with marginal constraints. Let $M$ be a joint distribution matrix over $D$ and $H$. Denote the set of all possible joint distribution matrices with marginals $r = P(D)$ and $c = P(H)$ by $U(r, c)$ (all couplings). Consider the question of finding the approximation matrix $P^*$ of $M$ in $U(r, c)$ that minimizes its relative entropy with $M$, i.e. $P^* = \arg \inf_{P \in U(r, c)} D_{KL}(P || M)$, where,
\[
D_{\text{KL}}(P\|M) = \sum_{i,j \in |D| \times |H|} P_{ij} \ln \frac{P_{ij}}{M_{ij}}
\]  

(8)

It is proved in for example [Csiszar, 1989, Franklin and Lorenz, 1989] that the \((r,c)\)-Sinkhorn scaling of \(M\) converges to \(P^*\) if the limit exists. We therefore directly interpret the results of fully Bayesian recursive reasoning as the communication plan with minimum discrimination information for pairs of interacting agents.

In addition, Sinkhorn scaling also arises naturally as a maximum likelihood estimation. Let \(\hat{P}\) be the empirical distribution of i.i.d. samples from a true underlying distribution, which belongs to a model family. Then the log likelihood of this sample set over a distribution \(M\) in the model family is given by \(n \cdot \sum_{ij} \hat{P}_{ij} \log M_{ij}\), where \(n\) is the sample size. Comparing with (8), it is clear that maximizing the log likelihood (and so the likelihood) over a given family of \(M\) is equivalent to minimizing \(D_{\text{KL}}(\hat{P}\|M)\). Both [Darroch and Ratcliff, 1972] and [Csiszar, 1989] show that when the model is in the exponential family, the maximum likelihood estimation of \(M\) can be obtained through SK scaling with empirical marginals.

4.2 Understanding greedy choice

As a preliminary step toward analyzing common ground, we explore the effect on the optimal plans when \(\lambda\) varies, showing that as \(\lambda \to \infty\) the solution converges to the leading diagonals of \(M\), as \(\lambda \to 0\) the solution goes to a uniform matrix, and more generally analyzing the variations on the distribution over all possible optimal plans caused by choice of \(\lambda\). To simplify notation, we assume uniform priors on \(H\) and \(D\) for our discussions.

For a given joint distribution \(M\), consider the OT problem for the teacher (similarly, for the learner). Recall that, as in Eq (4), the optimal teaching plan \(T^{(\lambda)}\) is the limit of SK iteration of \(L_0^{[\lambda]}\). Note that the limits of SK on \(L_0^{[\lambda]}\) and \(M^{[\lambda]}\) (the matrix obtained from \(M\) by raising each element to power of \(\lambda\)) are the same as they are cross-ratio equivalent (Section 2.3). Therefore to study the dynamics of \(\lambda\) regularized OT solutions, we may focus on \(M^{[\lambda]}\) and its SK limit \(M^{(\lambda)}\).

One extreme is when \(\lambda\) gets closer to zero. If \(\lambda \to 0, M^{(\lambda)}_{ij} = (M_{ij})^\lambda \to 1\) for any nonzero element of \(M\). Thus \(M^{[\lambda]}\) converges to a matrix filled with ones on the nonzero entries of \(M\), and \(M^{(\lambda)}\) converges to a uniform matrix if \(M\) has no vanishing entries. It is shown in [Wang et al., 2019] that the cooperative index (Section 2.2) attains its lower bound on uniform matrices. Hence \(M^{(\lambda)}\) reaches the lowest communicative efficiency as \(\lambda\) goes to zero.

The other extreme is when \(\lambda\) gets closer to infinity. In this case, we see that:

**Proposition 8.** \(M^{(\lambda)}\) is concentrating around the leading diagonals of \(M\) as \(\lambda \to \infty\).

This indicates that as \(\lambda \to \infty\), the number of diagonals of \(M^{(\lambda)}\) decreases. Therefore CI\((M^{(\lambda)})\) increases as \(\lambda \to \infty\) since the cooperative index of a matrix is bounded below by the reciprocal of its number of positive diagonals [Wang et al., 2019]. In particular, if \(M\) has only one leading diagonal, \(M^{(\lambda)}\) converges to a doubly stochastic matrix with only one positive diagonal, i.e. a permutation matrix. In this case, CI\((M^{(\lambda)})\) \(\to 1\), which suggests the highest communication efficiency. As pointed out in Section 3.3 the OT planning which picks the
best diagonals is notably different from the argmax selection.

In general, magnitude of $\lambda$ causes variations to the distribution over all possible optimal plans. Let $A$ be either the joint distribution $M$ or an agent’s planning matrix derived from $M$. Notice that the product of elements on a diagonal $D^A_\sigma$ of $A$ is proportional to the probability of sampling $D^A_\sigma$ from all $A$’s diagonals. Then the cross-product ratio between its two diagonals $D^A_\sigma$ and $D^{A'}_\sigma$ is precisely the ratio between probabilities of sampling $D^A_\sigma$ and $D^{A'}_\sigma$. Proposition 8 shows that the optimal plan of an agent is concentrated on diagonals of $M$. Thus, up to normalization, each $M^{(\lambda)}$ represents a distribution over all possible optimal plans (diagonals). And $M^{(\lambda=1)}$ constitute the true distribution over optimal plans derived from $M$ since $M^{(1)}$ is cross-product ratio equivalent to $M$. $M^{(\lambda\neq1)}$, in contrast, represents a distribution that either exaggerates or suppresses cross-product ratios of $M$, depending on whether $\lambda$ is greater or less than one.

### 4.3 Analyzing sensitivity to common ground

In this section, we investigate the sensitivity of the OT framework under perturbations on the common ground among agents. This ensures robustness of the inference where agents’ beliefs differ, which shows the viability of our model in practice.

First, the robustness of OT planning with a fixed regularizer $\lambda$ is considered. In this case, optimal plans are obtained through SK scaling of an initial matrix $M^{[\lambda]}$ with given marginal conditions $r$ and $c$. This can be viewed as a map, denoted by $\Phi$, from $(M, r, c)$ to the SK limit. [Wang et al., 2019] explored the sensitivity of $\Phi$ to perturbation on elements in $M$. They pointed out that $\Phi$ is continuous on $M$. In particular, they demonstrated that $\Phi$ is robust to any amount of off-diagonal perturbations on $M$.

SK scaling is also continuous on its scalars. Let $r'$ and $c'$ be vectors obtained by varying elements of $r$ and $c$ at most by $\epsilon$, where $\epsilon > 0$ quantifies the amount of perturbation. Distances between vectors or matrices are measured by $\ell^\infty$ norm (the maximum element-wise difference), e.g. $d(r', r) \leq \epsilon$. We prove that $\Phi$ is continuous on $r$ and $c$, thus the following holds:

**Theorem 9.** For any joint distribution $M$ and positive marginals $r$ and $c$, if $\Phi(M, r', c')$ and $\Phi(M, r, c)$ exist, then $\Phi(M, r', c') \to \Phi(M, r, c)$ as $r' \to r, c' \to c$.

Continuity of $\Phi$ implies that small perturbations on the joint and marginal distributions, yield close solutions for optional plans. Thus cooperative communicative actions based on the unifying OT framework are stable on variations of agents’ estimations of the common ground.

Moreover, when restricted to positive joint distribution $M$, [Luise et al., 2018] shows that $\Phi(M, r, c)$ is in fact smooth on $r$ and $c$. Built on their proof technique, we further extend the smoothness of $\Phi$ to $M$. Therefore, the following holds:

**Theorem 10.** Let $\mathcal{M} = (\mathbb{R}^+)^{|D| \times |H|}$ be the set of positive initial matrices, $\Delta^+_|D|$ and $\Delta^+_|H|$ be the set of all positive marginal distributions over $D$ and $H$ respectively. Then $\Phi: \mathcal{M} \times \Delta^+_|D| \times \Delta^+_|H| \to \mathcal{M}$ is $C^\infty$.

*General result on non-negative joint distributions is stated and proved in Appendix II.*
Remark 11. Theorem 10 guarantees that the optimal plans obtained through SK scaling are infinitely differentiable. In particular, we may explicitly derive the gradient of $\Phi$ with respect to both marginals and joint distributions. (Closed form of gradients are included in Appendix B.2.) An advantage of having these closed forms is that a fully recursive agent can quickly reconstruct a better cooperative plan using gradient descent methods once he realized the deviation from the previously assumed common ground.

Importantly, choice of $\lambda$ affects the sensitivity to violations of common ground. Without loss, assuming uniform priors on $H$ and $D$, consider the case where the teacher has the accurate $M$ and the learner’s estimation of the joint distribution $M_L$ contains additive deviation $\epsilon > 0$ on an element $M_{st}$ of $M$. When the deviation occurs on the leading diagonals, the optimal plan for the learner is the same as if they had the precise $M$ since the location of the leading diagonal is unchanged.

However, problems may occur when the deviation occurs on an element contained only in a non-leading diagonal. Intuitively, if the deviation is large enough, the rank of the diagonals in $M_L$, which determines the learner’s optimal plan, will change. This will cause a difference in two agents’ optimal plans, which would reduce the communication efficiency. Formally, let $D_\sigma$ be a leading diagonal and $D_{\sigma'}$ be a non-leading diagonal of $M$, and $d_\sigma$ and $d_{\sigma'}$ be products of their elements respectively. Further let $D_L^\sigma$ and $D_{L}^{\sigma'}$ be the corresponding diagonals in $M^L$. Then $d_L^\sigma = d_\sigma$ and $d_L^{\sigma'} = \frac{d_{\sigma'}^t}{M_{st}} \times (M_{st} + \epsilon)$. If $\epsilon > \frac{d_{\sigma'}^t}{M_{st}} \times (d_\sigma - d_{\sigma'})$, then $d_L^{\sigma'} > d_L^\sigma$, and $D_L^{\sigma'}$ will become the leading diagonal of $M^L$. Hence the learner’s optimal plan will change. In light of this, we have:

Definition 12. The stability $S$ of a joint distribution $M$ is: $S = \min_{\langle \sigma', s, t \rangle} \frac{d_{\sigma'}^t}{M_{st}} (d_\sigma - d_{\sigma'})$, where the minimum is taken over all non-leading diagonals $\sigma'$ and all $(s, t)$-entries contained only in non-leading diagonals.

Analysis in previous paragraph shows that when the deviation $\epsilon < S$, the leading diagonal in $M^L$ is unchanged no matter where the deviation arises. In this case, the learner may safely pick a sufficiently large $\lambda$. Yet when $\epsilon \geq S$, any value of $\lambda > 1$ will decrease the probability of a mutually agreed upon solution.

In the absence of strong constraints on potential violations of common ground, i.e. strong constraints on the maximum value of deviation $\epsilon$, and the number of such deviations, $\lambda = 1$ is recommended as it preserves cross-product ratios. With probability equals 1, an $n \times n$ matrix $M$ contains exactly one leading diagonal. Assume the deviation appears uniformly on each entry of $M$. Then with probability $n - 1/n$, it appears on a non-leading diagonal element. Thus for large $n$, deviation occurs almost surely on locations which could cause changes on leading diagonals, and represents deviation from the true optimal plan. Assuming independence, the vast majority of deviations would be of the unhelpful variety, thus decreasing the probability of agreement between agents about the leading diagonal. When $\lambda > 1$ decisions are more severe because the exaggeration of differences between diagonals pushes an agent’s estimation of optimal plan further away from the true optimal plan. Moreover, additive deviation can not only shuffle the rank of existing diagonals, but also introduce new diagonals. $\lambda > 1$ is much more sensitive to such a deviation comparing to $\lambda = 1$ (see further discussion in Appendix C). Thus, belief transport is most stable to violations of common ground when agents match, rather than maximize, probabilities.

The above discussion also suggests that argmax approximation method described in Section 3.3 is much
more sensitive to small perturbations. Similar to leading diagonals, location of argmax in a row or column vary non-continuously with a deviation. This may cause dramatic differences in agents’ action plans, which leads to low cooperative index. Therefore, argmax approaches do not in general yield optimal behavior.

5 Discussion and Conclusions

Computational-level and rational analyses of cognition hinge on assumptions about the structure for which the mind is optimized. When these analyses focus on properties of world—such as natural scenes or natural categories—these assumptions are hard or impossible to independently validate, which has lead to questions about the utility of the approach. When analyzing cooperative communication, the relevant structure is other people's belief updating and choices, domains for which we have strong independent theory. Given models derived from the literature, we show it is possible to unify existing algorithmic models, derive statistical interpretations, and derive a priori constraints on models by analyzing robustness of models to important, open theoretical problems. Moreover, in doing so, we expose a new algorithmic-level perspective on the implementation of cooperative communication and recovery from violations of common ground through gradient descent.

Our results clarify why, how, and under what conditions cooperative communication may facilitate learning despite violations of common ground. Why can cooperation facilitate learning? Recursive reasoning about others mental states and actions is precisely Sinkhorn scaling, which computes maximum likelihood plans for optimal transport of beliefs from one agent to another. How does cooperation succeed despite violations of common ground? Sinkhorn scaling is a continuous function, which implies that small differences in the inputs yields bounded differences between the outputs. Moreover, the smoothness property additionally guarantees the ability recover from deviations in an online fashion further increasing robustness to violations of common ground. Under what conditions is cooperative communication robust? Cooperative communication is robust to such violations precisely when the plans are based on probability matching, or at least close enough to not magnify the consequences of violations too much.

Researchers in cultural anthropology and cognitive development argue that people have evolved a specialized cultural niche and associated learning mechanisms that enables rapid accumulation of knowledge across ontogeny and phylogeny. We provide support for these claims. Specifically, cooperative communication—through the ability to reason about changes in beliefs in response to choices recursively—is a specialized adaptation to learn from other agents. Moreover, this adaptation enables effective transmission of beliefs, and hence accumulation of knowledge, through the computation of maximum likelihood plans that are robust to violations of common ground. Thus, one may theoretically transmit beliefs between agents whose beliefs are quite different, such as parents and children, speakers and listeners, teachers and learners, or across cultural groups, as is necessary to explain rapid accumulation of knowledge.

Formation and maintenance of common ground remains a formidable challenge. We have shown that cooperative communication, viewed as Optimal Transport computed through Sinkhorn scaling, has mathematical properties—as a continuous, even smooth, map—that explain how cooperative communication could succeed
in theory. Yet, in practice cooperative communication remains challenging. When communication is between teachers and learners or robots and humans the hypotheses may be organized differently or true hypothesis may not be in the hypothesis space. In education, this is because the goal is often inducing conceptual change or introducing new concepts. In robotics, the hypotheses spaces are designed for computational simplicity, rather than fidelity to humans’, and are unlikely to align cleanly or completely. These violations go beyond simple perturbations, and instead involve mismatches between the hypotheses themselves, which violate of the continuity necessary to ensure robustness.

Recent empirical results raise questions about the replicability of science across behavioral sciences [Collaboration et al., 2015]. Proposed improvements in the design and analysis of experimental results are an important toward addressing these issues. Equally important is the development of stronger, more principled approaches to theory development. While mechanisms like preregistration certainly reduce posthoc experimental and analytic degrees of freedom, they do not address the problem of how to justify hypotheses in the first place and therefore only slow down the rate of posthoc hypothesizing. Our analysis shows that it is possible to derive strong a priori predictions from first principles. Our results focus on cooperative communication, but may be extensible to Theory of Mind and other domains of reasoning that can be construed as recursive reasoning about possible plans. Moreover, the Optimal transport framework, which simply models problems of moving distributions, includes Bayesian inference as a special case, suggesting that this approach may be much more widely relevant to modeling cognition.

References

[Anderson, 1991] Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological review*, 98(3):409.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

[Bonawitz et al., 2011] Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., and Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3):322–330.

[Boyd et al., 2011] Boyd, R., Richerson, P. J., and Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, 108(Supplement 2):10918–10925.

[Brenier, 1991] Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

[Bridgers et al., ress] Bridgers, S., Jara-Ettinger, J., and Gweon, H. (in press). Young children consider the expected utility of others’ learning to decide what to teach. *Nature Human Behaviour*. 

15
[Brown et al., 1993] Brown, J. B., Chase, P. J., and Pittenger, A. O. (1993). Order independence and factor convergence in iterative scaling. *Linear algebra and its applications*, 190:1–38.

[Chater et al., 2008] Chater, N., Oaksford, M., et al. (2008). The probabilistic mind: Prospects for Bayesian cognitive science. OUP Oxford.

[Clark, 1996] Clark, H. H. (1996). *Using language*. Cambridge university press.

[Clark et al., 1991] Clark, H. H., Brennan, S. E., et al. (1991). Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

[Collaboration et al., 2015] Collaboration, O. S. et al. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

[Courty et al., 2017] Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2017). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.

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

[Csizár, 1975] Csizár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158.

[Csizsar, 1989] Csizsar, I. (1989). A geometric interpretation of darroch and ratcliff’s generalized iterative scaling. *The Annals of Statistics*, pages 1409–1413.

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

[Dantzig, 1949] Dantzig, G. B. (1949). Programming of interdependent activities: Ii mathematical model. *Econometrica, Journal of the Econometric Society*, pages 200–211.

[Darroch and Ratcliff, 1972] Darroch, J. N. and Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *The annals of mathematical statistics*, pages 1470–1480.

[Dufossé and Uçar, 2016] Dufossé, F. and Uçar, B. (2016). Notes on Birkhoff–von Neumann decomposition of doubly stochastic matrices. *Linear Algebra and its Applications*, 497:108–115.

[Eaves Jr et al., 2016] Eaves Jr, B. S., Feldman, N. H., Griffiths, T. L., and Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological review*, 123(6):758.

[Eaves Jr and Shafto, 2016] Eaves Jr, B. S. and Shafto, P. (2016). Toward a general, scaleable framework for bayesian teaching with applications to topic models. *arXiv preprint arXiv:1605.07999*.

[Fienberg, 1968] Fienberg, S. E. (1968). The geometry of an r× c contingency table. *The Annals of Mathematical Statistics*, 39(4):1186–1190.
[Fienberg et al., 1970] Fienberg, S. E. et al. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917.

[Fisac et al., 2017] Fisac, J. F., Gates, M. A., Hamrick, J. B., Liu, C., Hadfield-Menell, D., Palaniappan, M., Malik, D., Sastry, S. S., Griffiths, T. L., and Dragan, A. D. (2017). Pragmatic-pedagogic value alignment. arXiv preprint arXiv:1707.06354.

[Franklin and Lorenz, 1989] Franklin, J. and Lorenz, J. (1989). On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735.

[Goodman and Stuhlmüller, 2013] Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

[Grice et al., 1975] Grice, H. P., Cole, P., Morgan, J., et al. (1975). Logic and conversation. 1975, pages 41–58.

[Griffiths and Tenenbaum, 2005] Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, 51(4):334–384.

[Hadfield-Menell et al., 2016] Hadfield-Menell, D., Russell, S. J., Abbeel, P., and Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917.

[Ho et al., 2016] Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., and Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In *Advances in Neural Information Processing Systems*, pages 3027–3035.

[Ho et al., 2018] Ho, M. K., Littman, M. L., Cushman, F., and Austerweil, J. L. (2018). Effectively learning from pedagogical demonstrations. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

[Idel, 2016] Idel, M. (2016). A review of matrix scaling and sinkhorn’s normal form for matrices and positive maps. arXiv preprint arXiv:1609.06349.

[Jones and Love, 2011] Jones, M. and Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34(4):169.

[Kantorovich, 2006] Kantorovich, L. V. (2006). On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382.

[Knight, 2008] Knight, P. A. (2008). The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275.

[Koopmans, 1949] Koopmans, T. C. (1949). Optimum utilization of the transportation system. *Econometrica: Journal of the Econometric Society*, pages 136–146.

[Levinson, 2000] Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.
[Luce, 2012] Luce, R. D. (2012). *Individual choice behavior: A theoretical analysis*. Courier Corporation.

[Luise et al., 2018] Luise, G., Rudi, A., Pontil, M., and Ciliberto, C. (2018). Differential properties of sinkhorn approximation for learning with wasserstein distance. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 5859–5870. Curran Associates, Inc.

[Marcus and Davis, 2013] Marcus, G. F. and Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, 24(12):2351–2360.

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

[Monge, 1781] Monge, G. (1781). Memory on the theory of excavations and embankments. *History of the Royal Academy of Sciences of Paris*.

[Peyré et al., 2019] Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

[Rothblum and Schneider, 1989] Rothblum, U. G. and Schneider, H. (1989). Scalings of matrices which have prespecified row sums and column sums via optimization. *Linear Algebra and its Applications*, 114-115:737–764. Special Issue Dedicated to Alan J. Hoffman.

[Shafto and Goodman, 2008a] Shafto, P. and Goodman, N. (2008a). 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.

[Shafto and Goodman, 2008b] Shafto, P. and Goodman, N. D. (2008b). Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th annual conference of the Cognitive Science Society*, Austin, TX. Cognitive Science Society.

[Shafto et al., 2012] Shafto, P., Goodman, N. D., and Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, 7(4):341–351.

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

[Sinkhorn, 1972] Sinkhorn, R. (1972). Continuous dependence on $A$ in the DAD theorems. *Proceedings of the American Mathematical Society*, 32(2):395–398.

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

[Sperber and Wilson, 1986] Sperber, D. and Wilson, D. (1986). *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.
[Sutton et al., 1998] Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning*, volume 2. MIT press Cambridge.

[Tenenbaum et al., 2006] Tenenbaum, J. B., Griffiths, T. L., and Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in cognitive sciences*, 10(7):309–318.

[Tenenbaum et al., 2011] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

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

[Villani, 2008] Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

[Wang et al., 2019] Wang, P., Paranamana, P., and Shafto, P. (2019). Generalizing the theory of cooperative inference. *AIStats*.

[Yang et al., 2018] Yang, S. C., Yu, Y., Givchi, A., Wang, P., Vong, W. K., and Shafto, P. (2018). Optimal cooperative inference. In *AISTATS*, volume 84 of *Proceedings of Machine Learning Research*, pages 376–385. PMLR.

A Proofs of Propositions

**Proposition 7** Optimal communication plans, $T^*$ and $L^*$, of a cooperative inference problem with arbitrary priors denoted by $P_{T_0}(D)$ and $P_{L_0}(H)$, can be obtained through Sinkhorn scaling.

**Proof.** Consider cooperative inference as in (5) of the main content, we may rewrite it as follows:

\[
P_L(h|d)P_{T_0}(d) = \frac{P_T(d|h)P_{L_0}(h)P_{T_0}(d)}{P_L(d)}
\]

\[
P_T(d|h)P_{L_0}(h) = \frac{P_L(h|d)P_{T_0}(d)P_{L_0}(h)}{P_T(h)}
\]

which is equivalent to

\[
P_L(h|d)P_{T_0}(d) = \frac{P_T(d|h)P_{L_0}(h)}{P_L(d)},
\]

\[
P_T(d|h)P_{L_0}(h) = \frac{P_L(h|d)P_{T_0}(d)}{P_T(h)}.
\]

Notice that the (10) is the stable condition of Sinkhorn scaling on $\tilde{M} = P_L(h|d)P_{T_0}(d)$ with $r = P_{T_0}(D)$, $c = P_{L_0}(H)$. Hence (10) can be solved using fixed-point iteration as explored in *Shafto and Goodman, 2008a*. 
Shafto et al., 2014: for the first evaluation of the left hand side of (10a), initialize $P_L(h|d)$ by $P_{L_0}(h|d)$ which is the row normalization of the joint distribution $M = P(d,h)$ and denote $P_{L_0}(h|d) \cdot P_{T_0}(d)$ by $\tilde{T}_0$. Then the first evaluation of the left hand side of (10b), denoted by $\tilde{T}_1$, can be obtained by column normalizing $\tilde{T}_0$ with respect to $c$. Next, the second evaluation of (10a) is achieved by row normalizing of $\tilde{T}_1$ with respect to $r$, and iterate this process until convergence. This is precisely $(r,c)$-Sinkhorn scaling starting with $L_0$. Symmetrically, (10) can also be solved by $(r,c)$-Sinkhorn scaling starting with $\tilde{T}_0 = P_{T_0}(d|h) \cdot P_{L_0}(h)$.

\[\square\]

**Proposition 8.** $M^{(\lambda)}$ is concentrating around the leading diagonals of $M$ as $\lambda \to \infty$.

**Proof.** Let $D_\sigma, D_{\sigma'}$ be two diagonals of a $n \times n$ joint distribution $M$ and $d_\sigma, d_{\sigma'}$ be products of their elements respectively (Definition 3). Further, let the diagonals in $M^{(\lambda)}$ determined by the same $\sigma$ and $\sigma'$ be $D_\sigma^{(\lambda)}$ and $D_{\sigma'}^{(\lambda)}$. Their cross product ratio is denoted by $CR(D_\sigma^{(\lambda)}, D_{\sigma'}^{(\lambda)})$. If $D_{\sigma'}$ is a leading diagonal and $D_\sigma$ is not, then $d_\sigma/d_{\sigma'} < 1$, and so $CR(D_\sigma^{(\lambda)}, D_{\sigma'}^{(\lambda)}) = (d_\sigma/d_{\sigma'})^\lambda \to 0$ as $\lambda \to \infty$ (Fact A). If both $D_\sigma$ and $D_{\sigma'}$ are leading diagonals, then $d_\sigma/d_{\sigma'} = 1$, and so $CR(D_\sigma^{(\lambda)}, D_{\sigma'}^{(\lambda)}) = (d_\sigma/d_{\sigma'})^\lambda \to 1$ as $\lambda \to \infty$. We now show that for any element $M_{st}^{(\lambda)}$ of $M^{(\lambda)}$, if the corresponding element $M_{st}$ is not on a leading diagonal of $M$, then $M_{st}^{(\lambda)} \to 0$.

It is clear that if $M_{st}$ is not contained in any positive diagonal of $M$, then $M_{st}^{(\lambda)} \to 0$ as off diagonal elements vanishes along Sinkhorn iteration [Wang et al., 2019]. Now suppose that $M_{st}$ is contained in a non-leading positive diagonal determined by permutation $\sigma$. If $M_{st}^{(\lambda)}$ does not vanish, there exists an $\epsilon > 0$ such that $M_{st}^{(\lambda)} > \epsilon$ for any $\lambda$. And so $M_{st}^{(\lambda)}$ must be contained in a positive diagonal of $M^{(\lambda)}$. Without loss, we may assume $M_{st}^{(\lambda)}$ is the smallest non-vanishing element that is off leading diagonals of $M$. Then $d_\sigma^{(\lambda)} > \epsilon^n$, and so $d_\sigma^{(\lambda)}/d_{\sigma'}^{(\lambda)} > \epsilon^n$ because $d_{\sigma'}^{(\lambda)} \leq 1$ ($M^{(\lambda)}$ is doubly stochastic). This is contradiction to Fact A. Therefore, $M^{(\lambda)}$ is concentrating around the leading diagonals of $M$ as $\lambda \to \infty$.

\[\square\]

**Theorem 9** For any joint distribution $M$ and positive marginals $r$ and $c$, if $\Phi(M, r', c')$ and $\Phi(M, r, c)$ exist, then $\Phi(M, r', c') \to \Phi(M, r, c)$ as $r' \to r, c' \to c$.

**Proof.** Note that the continuity of $\Phi$ on the marginals is independent of the choice of a particular $\lambda$, we will drop the $\lambda$ for the rest of the proof to make the notation neater. Sinkhorn scaling of $M$ converges with marginal conditions $(r, c)$ and $(r', c')$ implies that $\sum_{i=1}^n r_i = \sum_{j=1}^m c_j$ and $\sum_{i=1}^n r_i' = \sum_{j=1}^m c_j'$ (see Menon and Schneider, 1969). Let $k = \sum_{i=1}^n r_i$ and $k' = \sum_{i=1}^n r_i'$. We will prove in three steps. First, we show the claim when $k = k'$. As $k = k'$, at least two elements in $r$ (or $c$) are perturbed. Without loss, we will assume that only two elements, $r_s$ and $r_t$ in $r$, are varied by amount $\epsilon$ since the general case may be treated as compositions of such. Then for $r' = (r'_1, \ldots, r'_n)$, we have $r'_s = r_s + \epsilon$, $r'_t = r_t - \epsilon$ and $r'_i = r_i$ if $i \neq s$ or $t$. Let $\Phi(M, r, c) = M^*$, $M'^{**}$ be the matrix obtained from varying the element $M_{st}$ and $M_{st}'$ of $M^*$ by $\epsilon$ and $-\epsilon$, i.e.
we are done. Then elements of $r \alpha$ has the same cross ratios of $M$ same, thus $r$ proved in [Sinkhorn, 1972].

$b \left( \Phi( M, r^e, c) \right) = \Phi( M, r^e, c) \) = M^*e$; $(d)$ holds as $d( M^*, M^e) = \epsilon$ by construction; $(e)$ holds because $\Phi$ is continuous on $M$ proved in [Sinkhorn, 1972].

Now we show the case where $k \neq k^e$, but the proportion between corresponding elements in $r$ and $r^e$ are the same, thus $r_i^e / r_i = r_j^e / r_j = \alpha$. Let $M^*o = \alpha \ast M^*$, i.e. $M^*o = \alpha \ast M^*_o$. Since $M^*o$ is $(r^e, c)$ normalized and also has the same cross ratios of $M$, $\Phi( M, r^e, c) = M^*o$. Note that $d( M^*o, M^*) \leq \epsilon$, so $\Phi( M, r^e, c) \rightarrow \Phi( M, r, c)$ as $\epsilon \rightarrow 0$.

Finally for the general case, where $k \neq k^e$ and elements of $r$ and $r^e$ are not proportional. Let $r^o = (k^e / k) \ast r$. Then elements of $r$ and $r^o$ are proportional and $\sum r_i^o = \sum r_i^e = k^e$. Thus based on the previous two cases, we have $d( \Phi( M, r, c), \Phi( M, r^e, c)) \leq d( \Phi( M, r, c), \Phi( M, r^o, c)) + d( \Phi( M, r^o, c), \Phi( M, r^e, c)) \rightarrow 0$ as $\epsilon \rightarrow 0$. Hence, we are done.

B Smoothness of $\Phi$ and its gradient

B.1 General version of Theorem 10

Enlightened by [Luise et al., 2018], we can conclude a stronger version of the smoothness of $\Phi$ in the following way:

Definition B.1. A pattern $\mathcal{P}$ is a subset of $\{1, 2, \ldots, n \} \times \{1, 2, \ldots, m \}$, and a matrix $M = (M_{ij})$ of pattern $\mathcal{P}$ is a non-negative matrix with $M_{ij} > 0$ if and only if $(i, j) \in \mathcal{P}$. In this paper, $M$ is not allowed to have a vanishing row or column.

Theorem 10 (General version of Theorem 10). Let $(\mathcal{P}, \mathcal{D})$ be a pair where $\mathcal{P}$ is a pattern, and where $\mathcal{D} \subseteq (\mathbb{R}^+)^n \times (\mathbb{R}^+)^m$ is the set consisting of vectors $(r, c) \in (\mathbb{R}^+)^n \times (\mathbb{R}^+)^m$ satisfying the equivalent conditions in Theorem 2 of [Rothblum and Schneider, 1989], in other words, pattern $\mathcal{P}$ is exact $(r, c)$-scalable. Let $M_{\mathcal{P}} = (\mathbb{R}^+)\mathcal{P}$ be the open cone of nonnegative matrices of pattern $\mathcal{P}$, then for a given $\lambda \in (0, \infty)$, $\Phi : M_{\mathcal{P}} \times \mathcal{D} \rightarrow M_{\mathcal{P}}$ is smooth.
Proof. We use the same strategy as the proof of Theorem 2 in [Luise et al., 2018]. Throughout the proof, let \( \lambda \in (0, \infty) \) be a fixed positive real number.

First we make a decomposition of \( \Phi \). This is possible because the exact scaling conditions guarantees the existence of diagonal matrices \( D_1, D_2 \) such that \( \Phi(M, r, c) = M^{(\lambda)} = D_1 M^{[\lambda]} D_2 \), equivalently, there exist vectors \( (\alpha, \beta) \in \mathbb{R}^n \times \mathbb{R}^m \) such that \( \Phi(M, r, c) = \text{diag}(e^{\lambda \alpha}) M^{[\lambda]} \text{diag}(e^{\lambda \beta}) \). The pair \((D_1, D_2)\) is unique up to a scalar \( d \in \mathbb{R}^+ \) with actions \( d (D_1, D_2) \mapsto (dD_1, d^{-1} D_2) \), thus the pair of vectors \( (\alpha, \beta) \) is unique up to a constant \( \delta : (\alpha, \beta) \mapsto (\alpha + \delta, \beta - \delta) \) (plus/minus the same number on each element of the vectors). So we may always assume that the last component of \( \beta \) vanishes, i.e., \( \beta_m = 0 \). In the following text, we use \( \tilde{\beta} \) to denote the first \( m - 1 \) components of \( \beta \), and if \( \beta \) occurs, the corresponding \( \beta \) is the vector by appending a 0 at the end of \( \tilde{\beta} \).

Then we can decompose the map \( \Psi \) into the composition of two other maps: \( \Phi = \mu \circ (\rho, \Psi) \). Here the map \( \rho : \mathcal{M}_\Phi \times \mathcal{D} \to \mathcal{M}_\Phi \) is the regularization map (regardless of the marginal conditions) \( \rho(M(r, c)) = M^{[\lambda]} \), the map \( \Psi : \mathcal{M}_\Phi \times \mathcal{D} \to \mathbb{R}^n \times \mathbb{R}^m \) maps \( (M, r, c) \) to the pair of vectors \( (\alpha, \beta) \) with \( \beta_m = 0 \) as in the above discussion (such that \( \Phi(M, r, c) = \text{diag}(e^{\lambda \alpha}) M^{[\lambda]} \text{diag}(e^{\lambda \beta}) \)), and the map \( \mu : \mathcal{M}_\Phi \times \mathbb{R}^n \times \mathbb{R}^m \to \mathcal{M}_\Phi \) is such that \( \mu(P, \alpha, \beta) = \text{diag}(e^{\lambda \alpha}) (P) \text{diag}(e^{\lambda \beta}) \). It can be easily seen that from the definitions the decomposition \( \Phi = \mu \circ (\rho, \Psi) \) is valid.

Next, having this decomposition, we just need to show that \( \mu, \rho \) and \( \Psi \) are smooth, then \( \Phi \) as the composition of smooth maps remains smooth.

(Smoothness of \( \Psi \):) We use the same strategy as Theorem 2 in [Luise et al., 2018]. Define the Lagrangian

\[
\mathcal{L}(M, r, c; \alpha, \beta) = -r^T \alpha - c^T \beta + \sum_{(i,j) \in \Psi} \frac{e^{\lambda \alpha_i} M_{ij}^\alpha e^{\lambda \beta_j}}{\lambda}.
\]

where \( \Psi(M, r, c) = (\alpha, \beta) \) optimizes \( \mathcal{L} \) for fixed \( M, r, c \) as proved in [Luise et al., 2018; Cuturi, 2013]. By smoothness of \( \mathcal{L} \) (easy to see from expression), we may conclude that \( N := \nabla_{(\alpha, \beta)} \mathcal{L} \) is \( C^k \) for any \( k \geq 0 \) and \( \nabla_{(\alpha, \beta)} \mathcal{L}(M, r, c; \Psi(M, r, c)) = 0 \) for any \( M, r, c \).

Fix \( (M_0, r_0, c_0; \alpha_0, \beta_0) \) such that \( N(M_0, r_0, c_0; \alpha_0, \beta_0) = 0 \) and \( (\beta_0)_m = 0 \). Since \( \nabla_{(\alpha, \beta)} N = \nabla_{(\alpha, \beta)} \otimes \nabla_{(\alpha, \beta)} \mathcal{L} \) is the Hessian of the strictly convex function \( \mathcal{L} \), then \( \nabla_{(\alpha, \beta)} N(M_0, r_0, c_0; \alpha_0, \beta_0) \) is invertible. Thus by Implicit Function Theorem, there exists a neighbourhood \( U \) of \((M_0, r_0, c_0)\) in \( \mathcal{M}_\Phi \times \mathcal{D} \) and a map \( \psi : U \to \mathbb{R}^n \times \mathbb{R}^m \) such that

1. \( \psi(M_0, r_0, c_0) = (\alpha_0, \beta_0) \),
2. \( \psi(M, r, c) = (\alpha, \beta) \), then the last component of \( \beta \) vanishes, \( \beta_m = 0 \), for any \( (M, r, c) \in U \),
3. \( N((M_0, r_0, c_0; \psi(M_0, r_0, c_0)) = 0 \), thus \( \psi(M, r, c) = \Psi(M, r, c) \), \forall (M, r, c) \in U \), by strict convexity of \( \mathcal{L} \) and uniqueness of \( (\alpha, \beta) \),
4. \( \psi \in C^k(U) \).

For the choice of \( k \) is arbitrary and the choice of \((M, r, c)\) as an interior point of \( \mathcal{M}_\Phi \times \mathcal{D} \) is also arbitrary,
we may see that \( \Psi \) is smooth in the interior of \( \mathcal{M}_P \times \mathcal{D} \).

In fact, we can show that \((\mathcal{M}_P \times \mathcal{D})^\circ = \mathcal{M}_P \times \mathcal{D}\), thus \( \Psi \) is smooth on \( \mathcal{M}_P \times \mathcal{D} \).

\( \mathcal{M}_P \) is isomorphic to an open subset \((\mathbb{R}^+)\mathbb{P}\) of \(\mathbb{R}^{\mathbb{P}}\). The set \( \mathcal{D} \) is a subset of \((\mathbb{R}^+)^{n+m}\), defined by finitely many equations and strict inequalities given in [Rothblum and Schneider, 1989] Theorem 2, especially part (e): for every subset \( I \subseteq \{1, 2, \ldots, n\} \) and \( J \subseteq \{1, 2, \ldots, m\} \), where \( M_{ij} = 0 \) for all \( (i, j) \in I^c \times J \) \((I^c \) is the complement of \( I \)), we have

\[
\sum_{i \in I} r_i \geq \sum_{j \in J} c_j
\]

with equality holds if and only if \( M_{ij} = 0 \) for all \( (i, j) \in I \times J^c \). The above condition means that the conditions are either equations or strict inequalities since the pattern \( \mathbb{P} \) is fixed. Among all these constraints, set of equations \( \mathcal{E} \) define a linear subspace \( V(\mathcal{E}) \) of \( \mathbb{R}^{n+m} \) and the set of strict inequalities \( \mathcal{N} \) draws an open subset \( U(\mathcal{E}, \mathcal{N}) \) on \( V(\mathcal{E}) \). And \( \mathcal{D} = (\mathbb{R}^+)^{n+m} \cap U(\mathcal{E}, \mathcal{N}) \) is open in \( U(\mathcal{E}, \mathcal{N}) \), so \( (\mathcal{D})^\circ = \mathcal{D} \).

(Smoothness of \( \rho \)): Since \( \lambda > 0 \) and for each \((i, j) \in \mathbb{P}, M_{ij} > 0\), then \( \rho \) is smooth from the smoothness of \( x^\lambda \) on \((0, \infty)\).

(Smoothness of \( \mu \)): \( \mu \) is the composition of exponential functions, multiplications and additions, all of which are smooth.

Thus \( \Phi = \mu \circ (\rho, \Psi) \) is smooth on \( \mathcal{M}_P \times \mathcal{D} \).

\[ \text{B.2 Calculation of gradient of } \Phi: \]

We make use of the decomposition \( \Phi = \mu \circ (\rho, \Psi) \) to calculate the gradient of \( \Phi \).

By implicit function theorem,

\[
(\nabla_r \Psi)_i = \frac{\partial \Psi}{\partial r_i} = -\left(\nabla_{(\alpha, \beta)} N\right)^{-1} (\nabla_r N)_i
\]

\[
= -\left(\nabla^2_{(\alpha, \beta)} \mathcal{E}\right)^{-1} (\nabla_r N)_i
\]

\[
= -\frac{1}{\lambda} \left( \text{diag}(r) M^{(\lambda)} \text{ diag}(\bar{c}) \right)^{-1} \begin{pmatrix}
(\delta_1)_n \\
\mathbf{0}_{(m-1)}
\end{pmatrix}
\]

In the last equality, the subscript col-\(i\) means the \(i\)-th column of the inverse matrix with \(1 \leq i \leq n\).
\[
(\nabla_M \Psi)_{ij} = \frac{\partial \Psi}{\partial M_{ij}} = -\left(\nabla^2_{(\alpha,\beta)} \mathcal{L}\right)^{-1} (\nabla_M \mathcal{N})_{ij}
= \frac{1}{\lambda} \left( \begin{array}{cc} \text{diag}(r) & -M^{(\lambda)} \\ M^{(\lambda)\top} & \text{diag}(\bar{e}) \end{array} \right) \cdot \lambda e^{-\lambda(\alpha_i + \beta_j)} M_{ij}^{-1} \left( \begin{array}{c} \delta_i \\ \delta_j \end{array} \right)
\]

\[
M_{ij}^{(\lambda)} \left[ \left( \begin{array}{cc} \text{diag}(r) & -M^{(\lambda)} \\ M^{(\lambda)\top} & \text{diag}(\bar{e}) \end{array} \right) \right]^{-1} + \left( \begin{array}{cc} \text{diag}(r) & -M^{(\lambda)} \\ M^{(\lambda)\top} & \text{diag}(\bar{e}) \end{array} \right)_{\text{col-}i}^{-1} \cdot \lambda e^{-\lambda(\alpha_i + \beta_j)} M_{ij}^{-1} \left( \begin{array}{c} \delta_i \\ \delta_j \end{array} \right)_{\text{col-}(\alpha+\beta)}^{-1}
\]

\(\bar{j}\) means that term does not exist if \(j = m\).

In addition, to calculate
\[
\left( \begin{array}{cc} \text{diag}(r) & -M^{(\lambda)} \\ M^{(\lambda)\top} & \text{diag}(\bar{e}) \end{array} \right) \left( \begin{array}{c} M \\ -MBD^{-1} \end{array} \right) = \left( \begin{array}{cc} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{array} \right)
\]

where \(M = (A - BD^{-1}C)^{-1}\).

For \(\rho\):
\[
\frac{\partial \rho}{\partial M_{ij}} = \lambda M_{ij}^{-1} E(i,j) \quad (11)
\]

with \(E(i,j)\) a \(n \times m\)-matrix where \(E(i,j)_{ij} = 1\) and all other entries vanish. And
\[
\nabla_{(r,c)} \rho = 0. \quad (12)
\]

For \(\mu\):
\[
\frac{\partial \mu}{\partial \alpha_i} (P, \alpha, \beta) = \lambda \text{diag}(\delta_i \lambda \alpha) P \text{diag}(\lambda \beta) = \lambda P^{*}_{(i,-)}
\]

where \(P^{*}_{(i,-)}\) is a matrix with \(i\)-th row the same as \(i\)-th row of \(P^{*}\) and vanishes elsewhere.

Similarly,
\[
\frac{\partial \mu}{\partial \beta_j} (P, \alpha, \beta) = \lambda \text{diag}(\lambda \alpha) P \text{diag}(\delta_j \lambda \beta) = \lambda P^{*}_{(-,j)}
\]

with \(j \leq m - 1\) but the size of \(P^{*}_{(-,j)}\) is still \(n \times m\).

And
\[
\frac{\partial \mu}{\partial P_{ij}} = \text{diag}(\lambda \alpha) E(i,j) \text{diag}(\lambda \beta) = \frac{P^{*}_{ij}}{P_{ij}} E(i,j)
\]

where \(P^{*}\) is the \((r,c)\)-Sinkhorn scaling limit matrix of \(P\).

Finally, we can combine all the results above to calculate the gradient of \(\Phi\). We will use \((\alpha, \beta)\) for \(\Psi\), use \(P\) for \(\rho\) when it is convenient.
\[
(\nabla_r \Phi)_t = \frac{\partial \Phi}{\partial r_t} \\
= \sum_{i,j=1}^{n,m} \frac{\partial \mu}{\partial \rho_{ij}} \frac{\partial \rho_{ij}}{\partial r_t} + \sum_{i=1}^n \frac{\partial \mu}{\partial \alpha_i} \frac{\partial \alpha_i}{\partial r_t} + \sum_{j=1}^{m-1} \frac{\partial \mu}{\partial \beta_j} \frac{\partial \beta_j}{\partial r_t} \\
= 0 + \sum_{i=1}^n \left( \nabla_s (\Phi) \right) \frac{\partial \mu}{\partial \alpha_i} + \sum_{j=1}^{m-1} \left( \nabla_t (\Phi) \right) \frac{\partial \mu}{\partial \beta_j}
\]

If we write the column \( t \) of matrix
\[
\begin{pmatrix}
\text{diag}(r) \\
M^{(\lambda)} \\
\text{diag}(c)
\end{pmatrix}
\]
with the last entry \( v_m = 0 \) then
\[
(\nabla_r \Phi)_t = -\text{diag} (u) M^{(\lambda)} - M^{(\lambda)} \text{diag}(v)
\]

To calculate \( \nabla_c \Phi \), we choose an elegant way by using the above calculations. We rewrite the map \( \Phi \) as
\[
\Phi(M, r, c) = (\Phi(M\lambda, r\lambda, c\lambda))^\top \text{ with } M\lambda = M^\top, \ r\lambda = c \text{ and } c\lambda = r.
\]

The transpose of \( M \), after regularization, scaled to \( (c, r) \) is exactly \( (M^{(\lambda)})^\top \).

So we have \( \nabla_c \Phi(M, r, c) = \nabla_{r\lambda} (\Phi(M\lambda, r\lambda, c\lambda))^\top \), thus
\[
(\nabla_c \Phi(M, r, c))_s = (((\nabla_r \Phi(M\lambda, r\lambda, c\lambda))_s)^\top \\
= -M^{(\lambda)} \text{diag} (u) - \text{diag} (v) M^{(\lambda)},
\]

where \( \begin{pmatrix} u \\ v \end{pmatrix} \) is the \( s \)-th column of matrix
\[
\begin{pmatrix}
\text{diag}(c) \\
M^{(\lambda)} \\
\text{diag}(r)
\end{pmatrix}
\]

At last,
\[
(\nabla_M \Phi)_{st} = \frac{\partial \Phi}{\partial M_{st}} \\
= \sum_{i,j=1}^{n,m} \frac{\partial \mu}{\partial \rho_{ij}} \frac{\partial \rho_{ij}}{\partial M_{st}} + \sum_{i=1}^n \frac{\partial \mu}{\partial \alpha_i} \frac{\partial \alpha_i}{\partial M_{st}} + \sum_{j=1}^{m-1} \frac{\partial \mu}{\partial \beta_j} \frac{\partial \beta_j}{\partial M_{st}} \\
= \lambda \frac{M^{(\lambda)}_{st}}{M_{st}} \left( E(s,t) - \text{diag} (u) M^{(\lambda)} - M^{(\lambda)} \text{diag} (v) \right)
\]

where \( u \in \mathbb{R}^n, v \in \mathbb{R}^m \) with the last entry \( v_m = 0 \), and
\[
\begin{pmatrix} u \\ v \end{pmatrix} = \left[ \begin{pmatrix}
\text{diag}(r) \\
M^{(\lambda)} \\
\text{diag}(c)
\end{pmatrix}\right]_{\text{col-s}}^{-1} + \left[ \begin{pmatrix}
\text{diag}(r) \\
M^{(\lambda)} \\
\text{diag}(c)
\end{pmatrix}\right]_{\text{col-(n+\ell)}}^{-1},
\]

for \( (s, t) \in \Phi \), and \( \ell \) means that term does not exist if \( t = m \).
C Further discussion on Sensitivity for large $\lambda$

Sensitivity to perturbations is a concern as $\lambda \to \infty$. Figure 1 demonstrates an example where a slight variation on the initial matrices $M_1$ and $M_2$ can result a huge difference on $M^{(\lambda)}_1$ and $M^{(\lambda)}_2$ as $\lambda$ approaches infinity. The figure plots the OT solutions $M^{(\lambda)}_i$ with the starting matrices $M_1$, $M_2$ differing from $M$ only by 2% on their $\ell^\infty$-norm. However, in this particular case, the change makes a huge difference: $M$ has two leading diagonals, while the perturbation of $M_1$ and $M_2$ on $M$ enhanced each of them, making $M_1$ and $M_2$ have only one leading diagonal. When $\lambda$ approaches zero, all products of diagonals tends to be the same, thus the curves (red for $M$, green for $M_1$ and blue for $M_2$) converges to a common limit point, the uniform matrix. But as $\lambda$ increases, the leading diagonals overwhelm others, and results in a fixed divergence on the limit when $\lambda \to \infty$. Therefore, in this case, no matter how slight the changes are, as long as they modify the set of leading diagonals, there will be a fixed difference on the limits when $\lambda \to \infty$ according to the leading diagonals. Thus, $M^{(\infty)}$ is no longer continuous on the initial matrix $M$.

In particular, as $\lambda$ increases, the cooperative index, $\text{CI}(M^{(\lambda)}_1, M^{(\lambda)}_2)$, between two agents with initial matrix $M_1$ and $M_2$ will be very small, even zero, if there is no overlapping positive element between $M^{(\lambda)}_1$ and $M^{(\lambda)}_2$ whereas $\text{CI}(M^{(1)}_1, M^{(1)}_2)$ is bounded from below by the reciprocal of the number of diagonals of $M$.

![Figure 1: Lost of Continuity when $\lambda \to \infty$](image)

**Example C.1.** Assume that the teacher has the accurate $M = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{pmatrix}$. For any $\lambda$, the optimal teaching plan $T^{(\lambda)} = I_3$. Suppose the learner gets constant noise of size 0.1 in the position of $M_{31}$. When $\lambda = 1$, the learner’s initial matrix is $L^{[\lambda=1]} = \begin{pmatrix} 1 & 3 & 0 \\ 0 & 1 & 4 \\ 0.1 & 0 & 1 \end{pmatrix}$, the corresponding optimal plan is $L^{(\lambda=1)} = \begin{pmatrix} 0.485 & 0.515 & 0 \\ 0 & 0.485 & 0.515 \\ 0.515 & 0 & 0.485 \end{pmatrix}$ and $\text{CI}(T^{(\lambda)}, L^{(1)}) = 0.485$. Similar when $\lambda = 2$, we have $L^{[\lambda=2]} = \begin{pmatrix} 1 & 9 & 0 \\ 0 & 1 & 16 \\ 0.1 & 0 & 1 \end{pmatrix}$, $L^{(\lambda=2)} = \begin{pmatrix} 0.291 & 0.709 & 0 \\ 0 & 0.291 & 0.709 \\ 0.709 & 0 & 0.291 \end{pmatrix}$ and $\text{CI}(T^{(\lambda)}, L^{(2)}) = 0.291$. Furthermore, as $\lambda \to \infty$, $L^{(\lambda)} \to \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$ and $\text{CI}(T^{(\lambda)}, L^{(\lambda)}) \to 0$. Thus, in this case communication efficiency is completely vanished due to deviations between the teacher and learner are exaggerated by greedy selection of examples.