Generalized Belief Transport

Junqi Wang
Department of Math & CS
Rutgers University
Newark, NJ, 07102
junqi.wang@rutgers.edu

Pei Wang
Department of Math & CS
Rutgers University
Newark, NJ, 07102
peiwang@rutgers.edu

Patrick Shafto
Department of Math & CS
Rutgers University
Newark, NJ, 07102
shafto@rutgers.edu

Abstract

Human learners have ability to adopt appropriate learning approaches depending on constraints such as prior on the hypothesis, urgency of decision, and drift of the environment. However, existing learning models are typically considered individually rather than in relation to one and other. To build agents that have the ability to move between different modes of learning over time, it is important to understand how learning models are related as points in a broader space of possibilities. We introduce a mathematical framework, Generalized Belief Transport (GBT), that unifies and generalizes prior models, including Bayesian inference, cooperative communication and classification, as parameterizations of three learning constraints within Unbalanced Optimal Transport (UOT). We visualize the space of learning models encoded by GBT as a cube which includes classic learning models as special points. We derive critical properties of this parameterized space including proving continuity and differentiability which is the basis for model interpolation, and study limiting behavior of the parameters, which allows attaching learning models on the boundaries. Moreover, we investigate the long-run behavior of GBT, explore convergence properties of models in GBT mathematical and computationally, document the ability to learn in the presence of distribution drift, and formulate conjectures about general behavior. We conclude with open questions and implications for more unified models of learning.

Learning and inference are subject to internal and external constraints. Internal constraints include the availability of relevant prior knowledge. External constraints include the availability of time to accumulate evidence versus the need make the best decision now or environmental non-stationarity. Standard models of machine learning tend to view different constraints as different problems, which impedes development of unified learning agents.

These internal and external constraints map onto classic dichotomies in machine learning. Availability of prior knowledge maps onto the Frequentist-Bayesian dichotomy in which the latter uses prior knowledge as a constraint on posterior beliefs, while the former does not. Within Bayesian theory, a classic debate pertains to uninformative, or minimally informative, settings of priors [Jeffreys, 1946, Robert et al., 2009]. Availability of time to accumulate evidence informs the use of generative versus discriminative approaches [Ng and Jordan, 2001], and static or drift/dynamic models [Dagum et al., 1992, Murphy, 2002]. Combining constraints on probability of beliefs and costs of data models cooperative communication [Wang et al., 2020b].

Learning agents must interpolate between modes of reasoning as necessary given the constraints of the moment. Imagine observing an agent behaving in an environment. As an observer, one may wish to learn about the environment from the agent's actions. However, any inferences depend on one's model of the agent and their constraints. How is the agent updating their beliefs? Do they have stable goals, or are they changing over time? Perhaps the agent is selecting actions to communicate what they know? In order to draw inferences over these possibilities, one must parameterize the space,

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

ideally in such a way one could optimize over the possibilities. Indeed, in order to implement these possibilities, the agent *themself* must parameterize the space in order to interpolate between classic dichotomies such as Bayesian and frequentist, static and dynamic environments, and helpful versus neutral agent, given constraints.

We introduce Generalized Belief Transport (GBT), based on Unbalanced Optimal Transport (Sec. 1), which paramterizes and interpolates between known reasoning modes (Sec. 2.2), with four major contributions. First, we prove continuity in the parameterization and differentiability on the interior of the parameter space (Sec. 2.1). Second, we analyze the behavior under variations in the parameter space (Sec. 2.3). Third, we study sequential learning, where learners may (not) track the empirically observed data frequencies in (Sec. 3). Fourth, we investigate predictive performance under environmental drift (Sec. 4).

Notations. $\mathbb{R}_{\geq 0}$ denotes the non-negative reals. Vector $\mathbf{1} = (1, \ldots, 1)$. The i-th component of vector v is $v(i)$. $\mathcal{P}(A)$ is the set of probability distributions over A. For a matrix M, M_{ij} represents its (i,j)-th entry, $M_{(i,\_)}$ denotes its i-th row, and $M_{(\_,j)}$ denotes its j-th column. Probability is $\mathbb{P}(\,\cdot\,)$.

1 Learning as a problem of unbalanced optimal transport

Consider a general learning setting: an agent, which we call a **learner**, updates their belief about the world based on observed data subject to constraints. There is a finite set $\mathcal{D} = \{d^1, \ldots, d^n\}$ of all possible data, that defines the interface between the learner and the world. The world is defined by a true hypothesis h^*, whose meaning is captured by a probability mapping $\mathbb{P}(d|h^*)$ onto observable data. For instance, the world can either be the environment in classic Bayesian inference [Murphy, 2012] or a **teacher** in cooperative communication [Wang et al., 2020b].

A learner is equipped with a set of hypotheses $\mathcal{H} = \{h^1, \ldots, h^m\}$ which may *NOT* contain h^*; an initial belief on the hypotheses set, denoted by $\theta_0 \in \mathcal{P}(\mathcal{H})$; and a non-negative cost matrix $C = (C_{ij})_{m \times n}$, where C_{ij} measures the underlying cost of mapping d^i into h^j [1]. The cost matrix can be derived from other matrices that record the relation between $\mathcal{D}$ and $\mathcal{H}$, such as likelihood matrices in classic Bayesian inference or consistency matrices in cooperative communication (see details in Section 2.2).This setting reflects an agent's learning constraints: pre-selected hypotheses, and the relations between them and the communication interface (data set).

A learner observes data in sequence. At round k, the learner observes a data d_k that is sampled from $\mathcal{D}$ by the world according to $\mathbb{P}(d|h^*)$. Then the learner updates their beliefs over $\mathcal{H}$ from θ_{k-1} to θ_k through a *learning scheme*, where $\theta_{k-1}, \theta_k \in \mathcal{P}(\mathcal{H})$. For instance, in Bayesian inference, the learning scheme is defined by Bayes rule; while in discriminative models, the learning scheme is prescribed by a code book.

The learner transforms the observed data into a belief on hypotheses $h \in \mathcal{H}$ with a minimal cost, subject to appropriate constraints, with the goal of learning the exact map $\mathbb{P}(d|h^*)$. We can naturally cast this learning problem as Unbalanced Optimal Transport.

1.1 Unbalanced Optimal Transport

Unbalanced Optimal Transport (UOT), introduced by Liero et al. [2018], is a generalization of (entropic) Optimal Transport [Villani, 2008, Cuturi, 2013, Peyré and Cuturi, 2019], that relaxes the marginal constraints. Formally, for non-negative scalar parameters $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$, the *UOT plan* is,

$$P^\epsilon(C, \eta, \theta) = \underset{P \in (\mathbb{R}_{\geq 0})^{n \times m}}{\arg\min} \{\langle C, P\rangle - \epsilon_P H(P) + \epsilon_\eta \mathrm{KL}(P\mathbf{1}|\eta) + \epsilon_\theta \mathrm{KL}(P^T\mathbf{1}|\theta)\}. \tag{1}$$

Here, $\langle C, P\rangle = \sum_{i,j} C_{ij}P_{ij}$ is the inner product between C and P, $H(P) = -\sum_{ij} P_{ij}(\log P_{ij} - 1)$ is the *entropy* of P, and $\mathrm{KL}(\mathbf{a}|\mathbf{b}) := \sum_i (a_i \log(a_i/b_i) - a_i + b_i)$ is the Kullback–Leibler divergence between vectors. It is shown in Chizat et al. [2018] that UOT plans can be solved efficiently via Algorithm 1 : Given a cost C, P^ϵ can be obtained by applying (η, θ, ϵ)-unbalanced Sinkhorn scaling on $K^\epsilon := e^{-\frac{1}{\epsilon_P}C} = (e^{-\frac{1}{\epsilon_P}C_{ij}})_{m \times n}$, with convergence rate $\tilde{\mathcal{O}}(\frac{mn}{\epsilon_P})$ [Pham et al., 2020].

[1] To guarantee the hypotheses are distinguishable, we assume that C does not contain two columns that are only differ by an additive scalar.

Proposition 1. *The UOT problem with cost matrix C, marginals θ, η and parameters $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$ generates the same UOT plan as the UOT problem with tC, θ, η, $t\epsilon = (t\epsilon_P, t\epsilon_\eta, t\epsilon_\theta)$ for any $t \in (0, \infty)$. Therefore, the analysis on ϵ and $t\epsilon$ are the same for general cost C.*

Thus a positive common factor on $C, \epsilon_P, \epsilon_\eta, \epsilon_\theta$ does not affect the solution of Eq. (1). Therefore, for the later analysis, we fix $\epsilon_P = 1$ unless otherwise stated.

Framework: Generalized Belief Transport (GBT). Learning, efficiently transport one's belief with constraints, is naturally a UOT problem, i.e. a *Generalized Belief Transport*. Each round, a learner, defined by a choice of $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$, updates their beliefs as follows. Let η_{k-1}, θ_{k-1} be the learner's estimations of the data distribution and the belief over hypotheses $\mathcal{H}$ after round $k-1$, respectively. At round k, the learner first improves their estimation of the mapping between $\mathcal{D}$ and $\mathcal{H}$, denoted by M_k, through solving the UOT plan Eq. (1) with $(C, \eta_{k-1}, \theta_{k-1})$, i.e. $M_k = P^\epsilon(C, \eta_{k-1}, \theta_{k-1})$. Then with data observation d_k, the learner updates their beliefs over $\mathcal{H}$ using corresponding row of M_k, i.e. suppose $d_k = d^i$ for some $d^i \in \mathcal{D}$, the learner's belief θ_k is defined to be the row normalization of the i-th row of M_k. Finally, the learner updates their data distribution to η_k by increment of the i-th element of η_{k-1}, see Algorithm 2.

Algorithm 1 Unbalanced Sinkhorn Scaling

input: C, θ, η, $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$, N stopping condition ω
output: $P^\epsilon(C, \eta, \theta)$
initialize: $\mathbf{K} = \exp(-\epsilon_P C)$, $\mathbf{v}^{(0)} = \mathbf{1}_m$
while $k < N$ **and not** ω **do**

$$\mathbf{u}^{(k)} \leftarrow \left(\frac{\eta}{\mathbf{K}\mathbf{v}^{(k-1)}}\right)^{\frac{\epsilon_\eta}{\epsilon_\eta + \epsilon_P}},$$

$$\mathbf{v}^{(k)} \leftarrow \left(\frac{\theta}{\mathbf{K}^T\mathbf{u}^{(k)}}\right)^{\frac{\epsilon_\theta}{\epsilon_\theta + \epsilon_P}}$$

end while
$P^\epsilon(C, \eta, \theta) = \text{diag}(u)\mathbf{K}\text{diag}(v)$

Algorithm 2 Generalized Belief Transport

input: C, θ_0, η_0, h^*, N, data sampler τ based on $\mathbb{P}(d|h^*)$, stopping condition ω
output: M, θ
initialize: $k \leftarrow 1$
while $k < N$ **and not** $\omega(\theta)$ **do**
  $M \leftarrow P^\epsilon(C, \eta_{k-1}, \theta_{k-1})$
  get data d^i sampled from τ
  $\eta_k \leftarrow update(\eta_{k-1}, d^i)$ via update rule
  $\mathbf{v} \leftarrow M_{(i,\_)}$
  $\theta_k \leftarrow \mathbf{v}/\sum_{h\in\mathcal{H}} \mathbf{v}(h)$
  $k \leftarrow k + 1$
end while

2 Generalized Belief Transport

Many learning models—including Bayesian inference, Frequentist inference, Cooperative learning, and Discriminative learning—are unified under our GBT framework under choice of ϵ. In this section, we focus on the single-round behavior of the GBT model, i.e., given a pair of marginals (θ, η), how different learners update beliefs with a single data observation. We first visualize the entire learner set as a cube (in terms of parameters), see Figure 1. Then, we study the topological properties of the learner set through continuous deformations of parameters ϵ. In particular, we show that existing models including Bayesian inference, cooperative inference and discriminative learning are learners with parameters $(1, 0, \infty)$, $(1, \infty, \infty)$ and $(0, \infty, \infty)$ respectively in our UOT framework.

2.1 The parameter space of GBT model

The space of learners in GBT are parameterized by three regularizers for the underlying UOT problem (1): ϵ_P, ϵ_η and ϵ_θ, each ranges in $[0, \infty)$. Therefore, the constraint space for GBT is $\mathbb{R}^3_{\geq 0}$, with the standard topology. When C, θ and η are fixed (assume $\eta \in \mathbb{R}^m_{>0}$), the map $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta) \mapsto (P^\epsilon)$ bears continuous properties:

Proposition 2. [2] *The UOT plan P in Equation* (1)*, as a function of ϵ, is continuous in $(0, \infty) \times [0, \infty)^2$. Furthermore, P is differentiable with respect to ϵ in the interior of its domain $\mathbb{R}^3_{\geq 0}$.*

Continuity on ϵ provides the basis for interpolation between different learning agents. The proof of Proposition 2 also implies the continuity on η and θ. Further, towards the boundaries of the parameter space (where theories like Bayesian, Cooperative Communication live in), we show:

[2] Proofs of all claims are included in the Supplementary Materials.

Proposition 3. *Let s_P, s_η, $s_\theta \geq 0$ be arbitrary finite numbers, the following holds:*

(1) The limit of P^ϵ exists as ϵ approaches $(\infty, s_\eta, s_\theta)$. In fact, $\lim_{\epsilon \to (\infty, s_\eta, s_\theta)} P^\epsilon_{ij} = 1$ for all i, j.

(2) As $\epsilon \to (s_P, \infty, s_\theta)$, P^ϵ converges to the solution to

$$\min \langle C, P \rangle - s_P H(P) + s_\theta KL(P^T \mathbf{1} | \theta), \text{ with constraint } P\mathbf{1} = \eta,$$

(3) Similarly, as $\epsilon \to (s_P, s_\eta, \infty)$, P^ϵ converges to the solution to

$$\min \langle C, P \rangle - s_P H(P) + s_\eta KL(P\mathbf{1} | \eta), \text{ with constraint } P^T \mathbf{1} = \theta.$$

(4) And when $\epsilon \to (s_P, \infty, \infty)$, the matrix P^ϵ converges to the EOT solution:

$$\min \langle C, P \rangle - s_P H(P), \text{ with constraints } P^T \mathbf{1} = \theta \text{ and } P\mathbf{1} = \eta.$$

(5) When $\epsilon \to (\infty, \infty, s_\theta), (\infty, s_\eta, \infty)$ or (∞, ∞, ∞), the limit does not exist, but the directional limits can be calculated.

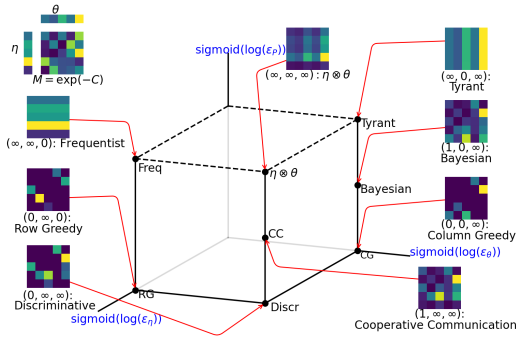


Figure 1: The parameter space $\mathcal{S}$ of GBT. Parameters $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$ can take the value ∞, rendering the corresponding regularization to a strict constraint. The two dashed edges with $\epsilon_P = \infty$ are not generally well-defined since the limits do not exist. The vertices corresponding to $\theta \otimes \eta$, Frequentist ($\eta \otimes \mathbf{1}$) and $\mathbf{1} \otimes \theta$ are the limits taken along the vertical edges. Given (C, θ, η) as shown in the left corner, each colored map plots each GBT learner (differ by constraints)'s estimation of the mapping between hypotheses and data (UOT plan).

The parameter space for GBT with its boundaries can be visualized in Fig. 1. Proposition 3 implies that the parameter space is $\mathcal{S} = [0, \infty]^3 \backslash (\{(\infty, \infty, x) : x \in [0, \infty]\} \cup \{(\infty, x, \infty) : x \in [0, \infty]\})$. In Fig. 1, segment $[0, \infty)$ is mapped to $[0, 1)$ by $\text{sigmoid}(\log(x))$. Then boundaries are added to the image cube $[0, 1)^3$. The dashed lines on top of the cube indicates limits that do not exist.

2.2 Special points in the parameter space

Bayesian Inference. Given observed data, a Bayesian learner (BI) [Murphy, 2012] derives posterior belief $\mathbb{P}(h|d)$ based on prior belief $\mathbb{P}(h)$ and likelihood matrix $\mathbb{P}(d|h)$, according to the Bayes rule. Intuitively, due to soft time constraint ($\epsilon_P = 1$), a Bayesian learner is a generative agent who puts a hard internal constraint on their prior belief ($\epsilon_\theta = \infty$), and omits the estimated data distribution η in the learning process, ($\epsilon_\eta = 0$).

Corollary 4. *Consider a UOT problem with cost $C = -\log \mathbb{P}(d|h)$, marginals $\theta = \mathbb{P}(h)$, $\eta \in \mathcal{P}(\mathcal{D})$. The optimal UOT plan $P^{(1, \epsilon_\eta, \epsilon_\theta)}$ converges to the posterior $\mathbb{P}(h|d)$ as $\epsilon_\eta \to 0$ and $\epsilon_\theta \to \infty$. Thus, Bayesian inference is a special case of GBT with $\epsilon = (1, 0, \infty)$.*

Frequentist Inference. A frequentist updates their belief from data observations by increasing the corresponding frequencies of datum. Intuitively, a frequentist is an agent who puts a hard constraint on the data distribution η ($\epsilon_\eta = \infty$), and omits prior knowledge θ ($\epsilon_\theta = 0$) in a learning process without time constraint ($\epsilon_P = \infty$). Formally we show:

Corollary 5. *Consider a UOT problem with $\theta \in \mathcal{P}(\mathcal{H})$, $\eta = \mathbb{P}(d)$. The optimal UOT plan $P^{(\epsilon_P, \infty, 0)}$ converges to $\eta \otimes \mathbf{1}$ as $\epsilon_P \to \infty$. Frequentist Inference is a special case of GBT with $\epsilon = (\infty, \infty, 0)$.*

Cooperative Communication. Two cooperative agents, a teacher and a learner, are considered in Yang et al. [2018], Wang et al. [2020b], Shafto et al. [2021]. Cooperative learners (CI) draw inferences about hypotheses based on which data would be most effective for the teacher to choose Given a data observation, a cooperative learner derives an optimal plan $L = \mathbb{P}(\mathcal{H}, \mathcal{D})$ based on a prior belief $\mathbb{P}(h)$, a shared data distribution $\mathbb{P}(d)$ and a matrix M specifies the consistency between data and hypotheses (such as M_{ij} records the co-occurrence of d^i and h^j). Intuitively, a cooperative learner is also a generative agent who puts hard constraints on both data and hypotheses ($\epsilon_\eta = \infty, \epsilon_\theta = \infty$), and aims to align with the true belief asymptotically, ($\epsilon_P = 1$). Thus we show:

Corollary 6. *Let cost $C = -\log M$, marginals $\theta = \mathbb{P}(h)$ and $\eta = \mathbb{P}(d)$. The optimal UOT plan $P^{(1,\epsilon_\eta,\epsilon_\theta)}$ converges to the optimal plan L as $\epsilon_\eta \to \infty$ and $\epsilon_\theta \to \infty$. Cooperative Inference is a special case of GBT with $\epsilon = (1, \infty, \infty)$, which is exactly entropic Optimal Transport [Cuturi, 2013].*

Discriminative learning. A discriminative learner decodes an uncertain, possibly noise corrupted, encoded message, which is a natural bridge to information theory [Cover, 1999, Wang et al., 2020b]. A discriminative learner builds an optimal map to hypotheses $\mathcal{H}$ conditioned on observed data $\mathcal{D}$. The map is perfect when, for all messages, encodings are uniquely and correctly decoded. Intuitively, a discriminative learner aims to quickly build a deterministic code book (implies $\epsilon_P = 0$) that matches the marginals on $\mathcal{H}$ and $\mathcal{D}$. We show that discriminative learner is GBT with $\epsilon = (0, \infty, \infty)$:

Corollary 7. *Consider a UOT problem with cost $C = -\log \mathbb{P}(d, h)$, $m = n$, and marginals $\theta = \eta$ are uniform. The optimal UOT plan $P^{(\epsilon_P,\epsilon_\eta,\epsilon_\theta)}$ approaches to a diagonal matrix as $\epsilon_\eta, \epsilon_\theta \to \infty$ and $\epsilon_P \to 0$. In particular, discriminative learner is a special case of GBT with $\epsilon = (0, \infty, \infty)$, which is exactly classical Optimal Transport [Villani, 2008].*

Many other interesting models are unified under GBT framework as well. GBT with $\epsilon = (0, \infty, 0)$ denotes Row Greedy learner which is widely used in Reinforcement learning community [Sutton and Barto, 2018]; $\epsilon = (\infty, \infty, \infty)$ yields $\eta \otimes \theta$ which is independent coupling used in χ^2 [Fienberg et al., 1970]; $\epsilon = (\epsilon_P, \epsilon_\theta, \infty)$ is used for adaptive color transfer studied in [Rabin et al., 2014]; and $\epsilon = (0, \epsilon_\theta, \epsilon_\eta)$ is UOT without entropy regularizer developed in [Chapel et al., 2021]. Other points in the GBT parameter space are also of likely interest, past or future.

2.3 General properties on the transportation plans

The general GBT framework builds a connection between the above theories, and the behavior of theory varies according to the change of parameters. In particular, each factor of $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$ expresses different constraints of the learner. Given (C, θ, η) as shown in the top-left corner of Fig. 1, we plot each learner's UOT plan with darker color representing larger elements.

ϵ_P controls a learner's learning horizon. When $\epsilon_P \to 0$, a learner's UOT plan is concentrated on a clear leading diagonal which allows them to make fast decisions. This corresponding to agents who are under the time pressure of making immediate decision, i.e. discriminative learner, or row greedy learner on the bottom of the cube (Fig. 1). Most of the time, one datum is enough to identify the true hypothesis and convergence is achieved within every data observation. When $\epsilon_P \to \infty$, GBT converges to a reticent learner, such as learners on the top of the cube. Data do not constrain the true hypothesis, and learners draw their conclusions independent of the data. In between, GBT provides a generative (probabilistic) learner. When $\epsilon_P = 1$, we have Bayesian learner and Cooperative learner, for whom data accumulate to identify the true hypothesis in a manner broadly consistent with probabilistic inference, and consistency is asymptotic.

ϵ_η controls a learner's knowledge on the data distribution η. When $\epsilon_\eta \to \infty$, GBT converges to a learner who is aware of the data distribution and reasons about the observed data according to the probabilities/costs of possible outcomes. Examples include the Discriminative and Cooperative learners on the front of the cube. When $\epsilon_\eta \to 0$, GBT converges to a learner who updates their belief without taking η into consideration, such as Bayesian learners on the back of the cube, and the Tyrant who does not care about data nor cost and is impossible to be changed by anybody.

ϵ_θ controls the strength of the learner's prior knowledge. When $\epsilon_\theta \to 0$, GBT converges to learners who utilizes no prior knowledge. Hence, they do NOT maintain beliefs over $\mathcal{H}$, and draws their conclusions purely on the data distribution, such as a Frequentist learner on the left of the cube. When $\epsilon_\theta \to \infty$, GBT converges to a learner who enforces a strict prior such as Bayesian, Cooperative and Discriminative learners on the right of the cube. In particular, we show that:

Proposition 8. *In GBT with $\epsilon_\theta = \infty$, cost C and current belief θ. The learner updates θ with UOT plan in the same way as applying Bayes rule with likelihood from $P^\epsilon(C, \eta, \theta)$, and prior θ.*

3 Sequential GBT - Static

The sequential GBT captures the asymptotic behavior of a learning problem (C, θ_0, h^*). Static world where there exists a fixed true hypotheses h^* is considered in this section. Data is sampled from

$\eta = \mathbb{P}(d|h^*)$ (not necessarily related to some $h \in \mathcal{H}$). Then the learner follows GBT with cost C, and parameter ϵ, starts with a prior θ_0, then in each round k, applies GBT with η_{k-1} and θ_{k-1} to generate θ_k.

We investigate two cases. In the Preliminary sequential model (**PS**), we assume $\eta_k = \eta$ for all k. In practice, often a learner does not have access η. Instead, in each round the learner may choose to use the current observed data distribution $\eta_k(d)$ as an estimation of η, Thus we study the Real sequential model (**RS**) where $\eta_k \xrightarrow{a.s.} \eta$.

In statistics, a model is said to be consistent when, for every fixed hypothesis $h \in \mathcal{H}$, the model's belief θ over the hypotheses set $\mathcal{H}$ converges to δ_h in probability as more data are sampled from $\eta = \mathbb{P}(d|h)$. Such consistency has been well studied for Bayesian Inference since Bernstein and von Mises and Doob [Doob, 1949], and recently demonstrated for Cooperative Communication [Wang et al., 2020a]. However, the challenge arises when one tries to learn a h^* that is not contained in the pre-selected hypothesis space $\mathcal{H}$. It is not clear which $h \in \mathcal{H}$ is the 'correct' target to converge to.

In this section, we demonstrate GBT's ability of learning new hypothesis. Analogize to consistency, the properties are stated directly in the language of posterior sequence $(\Theta_k)_{k=1}^{\infty}$ as random variables, focusing on whether the sequence converges (and in which sense), and how conclusive (how likely to a stable new hypothesis is learned) the sequence is.

For a learning problem (C, θ_0, h^*), results in this section are organized based on different ϵ_θ values.

Conclusive and Bayesian-style: $\epsilon_\theta = \infty$. These learners are located on the right side of Cube Fig. 1. Many well-studied learners are in this class: Bayesian, Cooperative, Discriminative, Row Greedy etc. According to Prop 8, learners in this class perform "Bayesian" style learning.

When $\epsilon_\eta = 0$, i.e. the learners who update their belief without considering data distribution, (**PS**) and (**RS**) are essentially the same. The following holds:

Theorem 9 ([Doob, 1949],[Wang et al., 2020a])**.** *In GBT sequential model (both (**PS**) and (**RS**)) with $\epsilon = (\epsilon_P, 0, \infty)$ where $\epsilon_P \in (0, \infty)$, the sequence Θ_k converges to some δ_h almost surely, h is the closest column of e^{-C/ϵ_P} to η in the sense of KL-divergence.*

When $\epsilon_\eta = \infty$, the models (**PS**) and (**RS**) present slightly different behaviors.

Theorem 10 (PS). *When $\epsilon_\eta = \epsilon_\theta = \infty$, for the **PS** problem belief random variables of a GBT learner $(\Theta_k)_{k\in\mathbb{N}}$ converge to the random variable Y in probability, where $Y = \sum_{h\in\mathcal{H}} \theta_0(h)\delta_h$ and Y is supported on $\{\delta_h\}_{h\in\mathcal{H}}$ with $\mathbb{P}(Y = \delta_h) = \theta_0(h)$ for $\epsilon_\eta = \epsilon_\theta = \infty$ and $\epsilon_P \in (0, \infty)$.*

Corollary 11. *Given a fixed data sequence d_i sampled from η, if θ_k converges to δ_{h^j}, then the j-th column of M_k converges to η.*

Thus a GBT learner, with access to the data distribution and using strict marginal constraints, converges to the true hypothesis mapping η with probability 1. Moreover, the probability of which $h \in \mathcal{H}$ is shaped into η is determined by their prior θ_0. That is, GBT learners converge to the truth by reforming one of their original hypotheses into the true hypothesis.

Proposition 12. *When $\epsilon_\eta = \epsilon_\theta = \infty$, for the (**RS**) problem, the belief random variables of a GBT learner $(\Theta_k)_{k\in\mathbb{N}}$ satisfies that for any $s > 0$, $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(\Theta(h) > 1-s) = 1$. As a consequence, M_k as the transport plan has a dominant column (h^j) with total weights $> 1-s$, and $|(M_k)_{ij} - \eta_k(i)| < s$.*

In fact, as long as the sequence of η_k as random variables converges to η in probability, the above proposition holds. The limit $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(\Theta(h) > 1 - s)$ measures how conclusive the model is.

In contrast with standard Bayesian or other inductive learners, Proposition 12 shows that a GBT learner is able to learn *any* hypothesis mapping $\eta = \mathbb{P}(d|h^*)$ up to a given threshold s with probability 1. In addition to unifying disparate models of learning, GBT enables a fundamentally more powerful approach to learning by empirically monitoring the data marginal.

Fig. 2 illustrates convergence over learning problems and episodes. In each bar, we sample 100 learning problems (C, θ_0, h^*) from Dirichlet distribution with hyperparameters the vector $\mathbf{1}$. Then we sample 1000 data sequences (episodes) of maximal length $N = 10000$. The learner learns with Algo. 2 where the stopping condition ω is set to be $\max_{h\in\mathcal{H}} \theta(h) > 1 - s$ with $s = 0.001$. The

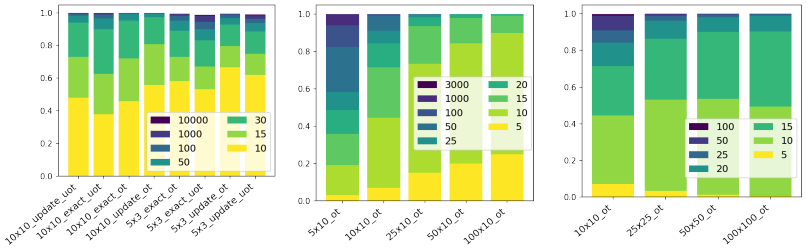


Figure 2: Evidence of general consistency: we plot the percentage of episodes that reaches a threshold (0.999) by round number (in colors of the bars). Each bar represents a size of matrix, for each bar 100 matrices were randomly sampled, and 1000 rounds were simulated per matrix. "exact" means learner uses $\eta_k = \eta$, (**PS**), "update" means learner uses statistics on current data in the episode (**RS**). "uot" takes $\epsilon = (1, 40, 40)$ and "ot" comes with exact and $\epsilon = (1, \infty, \infty)$.

y-axis in the plots represents the percentage of total episode converged. The color indicates in how many rounds the episode converges. For instance, in the bar corresponding to '10×10_update_uot', with 10 data points (yellow portion), about 50% episodes satisfy the stopping condition.

The first plot shows results for 10×10 and 5×3 matrices. The second plot shows results for rectangular matrices of dimension $m \times 10$ with m ranges in $[5, 10, 25, 50, 100]$. The third plot shows results for square matrices of dimension $m \times m$ with m ranges in $[10, 25, 50, 100]$. Here 'exact' and 'update' indicate the problem is (**PS**) or (**RS**), respectively. For parameters, *uot* represents the parameter choice $(\epsilon_P = 1, \epsilon_\theta = \epsilon_\eta = 40)$ vs. *ot* represents the parameter choice $(\epsilon_P = 1, \epsilon_\theta = \epsilon_\eta = \infty)$.

The first plots demonstrates that learners that do not have access to the true hypothesis (empirically builds estimation of η) learn faster than learners who have full access. The second plot indicates with a fixed number of hypotheses, learning is faster when the dimension of $\mathcal{D}$ increases. The third plot shows that the GBT learner scales well with the dimension of the problem.

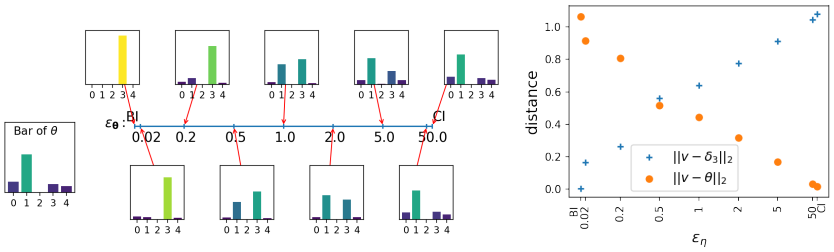


Figure 3: Left: Behavior of models spanning the line segment between BI and CI. With $\epsilon_P = 1$ and $\epsilon_\theta = \infty$, when ϵ_η varies from 0 to ∞, the theory changes from BI to CI. Each bar graphs the Monte-Carlo result of 400,000 teaching sequences, we empirically observe that the coefficients $a(h)$ of the limit in terms of $\sum_{h\in\mathcal{H}} a(h)\delta_h$ changes from BI to CI continuously from $\delta(h^3)$ by Bernstein-von Mises to $\theta_0(h)$ by Theorem 10. Right: the Euclidean distances of each coefficient $a(h)$ to BI result (blue crosses), and to CI result (orange dots).

Then we study the learners that interpolate between Bayesian and Cooperative learners (located on the line connecting CI and BI in Fig 1). Consider a fixed learning problem (C, θ_0, h^*). Consistency of Bayesian inference states that asymptotically, the learner Bayesian converges to a particular hypothesis $h_b \in \mathcal{H}$ almost surely where h_b is the hypothesis closest to h^* under KL divergence. Theorem 10 indicates that a GBT cooperative learner modifies one of the hypotheses into h^* in probability 1. The probability of h^j converges to h^* is determined by $\theta_0(h^j)$.

In Fig. 3, we study the asymptotic behavior of the learners corresponding to $\epsilon = (1, \epsilon_\eta, \infty)$, with $\epsilon_\eta \in \{0, 0.02, 0.2, 0.5, 1, 2, 5, 50, \infty\}$. We sample a learning problem with a dimension 5×5 from Dirichlet distribution with hyperparameters the vector $\mathbf{1}$. Each learner $\epsilon = (1, \epsilon_\eta, \infty)$ is equipped with a fixed C, θ_0 and $\eta_k = \eta$ for all k. We run $400,000$ learning episodes per learner, and plot their convergence summary in the bar graph. A continuous transition from a Bayesian learner to a cooperative learner can be empirically observed: the coefficients $a(h)$ of the limit in terms of $\sum_{h\in\mathcal{H}} a(h)\delta_h$ changes from $\delta(h^3)$ by Bernstein-von Mises to $\theta_0(h)$ by Theorem 10.

From the previous empirical results, we conclude the following conjecture:

Conjecture 13. *When $\epsilon = (\epsilon_P, \epsilon_\eta, \infty)$, where $\epsilon_P \in (0, \infty)$, the sequence of posteriors Θ_k from generic C, η, θ and ϵ as random variables satisfy $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(|\Theta_k(h) - 1| < e) = 1$ for any $e > 0$.*

Further, we pick out those episodes with $\theta_N(h) > 0.95$, plot the values $\mathbb{E}_{\theta_N(h)>0.95}[\ln \theta_k(h) - \ln(1 - \theta_k(h))]$ for each h against k in Fig. 4. Near linear relations are observed away from the first several rounds and before the values reaches the precision threshold. These are empirical estimates of the rate of convergence.

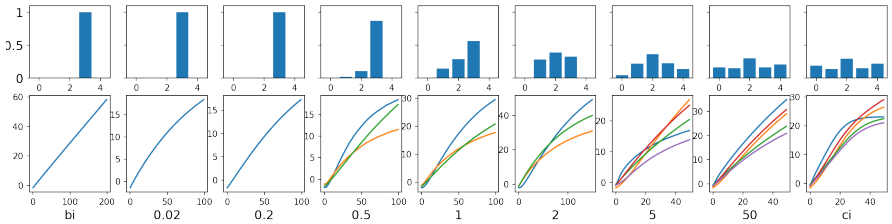


Figure 4: Top: For a learning problem C, behaviors of 9 different learners with $\epsilon_P = 1$, $\epsilon_\theta = \infty$ and various ϵ_η (denoted in figure) on conclusion distributions, $a(h)$ in bar graph, plots below bars are estimated convergence rates $\mathbb{E} \ln(\theta_k(h)/(1 - \theta_k(h)))$ averaged on episodes converging to h, one curve per hypothesis.

Inconclusive and independent: $\epsilon_\theta = 0$. The following holds for both (**PS**) and (**RS**):

Proposition 14. *For $\epsilon = (\epsilon_P, \epsilon_\eta, 0)$ with $\epsilon_P \in (0, \infty)$, as $\eta_k \to \eta$ almost surely, the sequence Θ_k of posteriors as a sequence of random variables converges in probability to variable Θ, where $\mathbb{P}(\Theta = \mathbf{v}^i) = \eta(i)$ and $\mathbf{v}^i = P_{(i,\_)} / \left(\sum_{j=1}^m P_{ij}\right)$ and $P = P^\epsilon(C, \eta, \theta)$. Therefore, for any $s > 0$, $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(|\Theta_k(h) - 1| < s) = 0$ for generic (for all but in a closed subset) cost C and η, θ.*

With $\epsilon_\theta = 0$, the constraint on column-sum (ϵ_η-term) fails to affect the transport plan, thus the Θ_k's in the sequence are independent from each other, in contrast that in all other cases the adjacent ones are correlated via a nondegenerate transition distribution. The independence makes the sequence of posterior-samples in one episode behave totally random, thus rarely converge as points in $\mathcal{P}(\mathcal{H})$. Furthermore, when consider the natural coupling (Θ_{k-1}, Θ_k) from Markov transition measure for $\epsilon_\theta = 0$ (which is independent), $\mathbb{E}\left(|\Theta_{k-1} - \Theta_k|^2\right)$ converges to the variance $Var(\eta)$. In contrast, for $\epsilon_\theta = \infty$, $\mathbb{E}\left(|\Theta_{k-1} - \Theta_k|^2\right)$ converges to 0 if Conj. 13 holds.

$\epsilon_\theta \in (0, \infty)$: partially conclusive. From Conj. 13 and Prop. 14, together with the continuity of the transition distribution on ϵ, we conjecture the following continuity on conclusiveness:

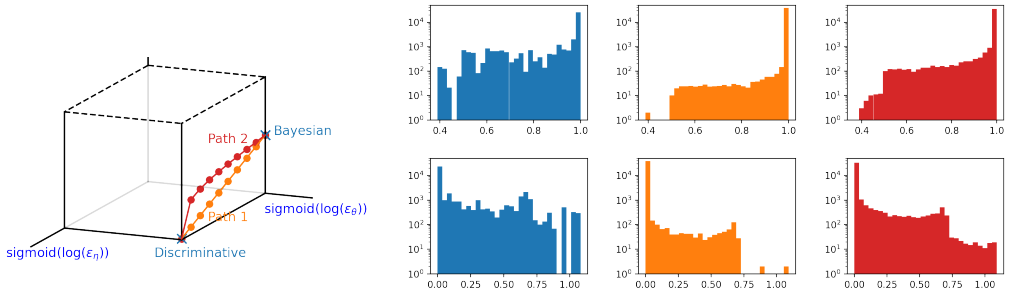
Conjecture 15. *For both (**PS**) and (**RS**) models, when $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$ with $\epsilon_P, \epsilon_\theta \in (0, \infty)$, the posterior sequence Θ_k from generated from generic C, η, θ and ϵ satisfy that $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(|\Theta_k(h) - 1| < s) = L$ exists, and $L \in (0, 1)$, for any $s > 0$.*

An Exploration on Interpolation It is popular in the state of art machine learning models that an agent learns probabilistically, but makes decisions greedily. This heuristic represents a path where a big leap on the cube was taken at the last step. An interesting question is under what circumstances this is optimal, what are the trade-offs, and under what conditions smoother trajectories are preferable. Instead of the giant leap, small steps along two paths are explored in the cube.

Our exploration takes a slightly different situation where the last step is replaced by a discriminative learning strategy $\epsilon = (0, \infty, \infty)$. We choose a matrix randomly of shape 4×4 (see Supplementary for detail), set total steps or total data points taught $N = 10$, and uniform prior $\theta \in \mathcal{P}(\mathcal{H})$ on hypotheses. Three learners are postulated: Blue performs Bayesian in first 9 steps and Discriminative in the last. Orange and Red follows two different interpolation curves with the same endpoints on the cube drawn in Fig. 5a. The curves are line and parabola segments on the cube sides.

Results of the three learners are shown in Fig. 5b. We sample h^* uniformly from the 4 columns of (the column-normalized) M, 40000 repeats for each learner. Conclusiveness (minimal l^1 distance between posterior and a 1-hot vector) and posterior entropy are plotted as histograms. The results show that the smoother path may lead to a more conclusive posterior. Numerical results: Conclusiveness of Blue: mean 0.9406, standard deviation 0.1300. Conclusiveness of Orange: mean 0.9964, standard deviation

0.0327. Conclusiveness of Red: mean 0.9834, standard deviation 0.0676. Furthermore, compared with a sudden jump, gradual interpolations have lower entropy. Numerical results: entropy of Blue: mean 0.1261, standard deviation 0.2435, entropy of Orange: mean 0.0079, standard deviation 0.0629; entropy of Red: mean 0.0388, standard deviation 0.1336.



(a) Three learners (b) Posteriors and Entropy distributions

Figure 5: (a): Baseline Blue and two learners Orange and Red following corresponding interpolation paths. (b): Results of the three learners over 40000 repeats. Top: conclusiveness, the frequency distribution of maximal posterior component. Bottom: entropy distribution.

From this experiment, in the 10-sample learning, two smoother learners behave more conclusive in their posterior with smaller posterior entropy. Meanwhile, we still know very little about the interpolation behavior, such as which path works better, how to distribute vertices on the curve, etc.

4 Sequential GBT - Dynamic

While static models are frequently studied, in many cases world changes dynamically. In this section, we take a first step in this direction by exploring the sequential behaviors of GBT learners assuming the world changes periodically. Demonstrate that unlike existing learners, GBT learner is capable of detecting the non-static property of a given problem.

Let integer $p > 0$ be the period, given a set of true hypotheses distributions $\vec{\eta} = \{\eta_0, \eta_1, \ldots, \eta_{p-1}\} \subseteq \mathcal{P}(\mathcal{D})$, datum d_t is sampled from $d_{k\%p}$ where $k\%p$ represents the remainder of k under division by p.

Proposition 16. *For a Bayesian learner, the posterior sequence $\{\Theta_i\}$ converges almost surely to the average of true hypotheses $\overline{\eta} = \frac{1}{p}\sum_{k=0}^{p-1}\eta_k$.*

For random variables of learner's posterior sequence $(\Theta_k)_{k=1}^{\infty}$, group them by period, we denote $\vec{\Theta}_t := (\Theta_{tp}, \Theta_{tp+1}, \ldots, \Theta_{(t+1)p-1})$. Here k represents the time step, t denotes the period index.

Proposition 17. *For ϵ in the interior of the cube, for (**PS**) problem, the sequence $\{\vec{\Theta}_t\}$ (random variables over $\mathcal{P}(\mathcal{H})^p$) form a time-homogeneous Markov chain. For (**RS**) problem, $\{(\vec{\Theta}_t, \frac{1}{pt}\sum_{k=0}^{p-1} t\eta_k)\}$, the random variable sequence producing samples $\{(\vec{\theta}_t, \frac{1}{pt}\sum_{k=0}^{pt-1}\delta_{d_k})\}$, forms a Markov chain.*

Next we compare different learners' behavior empirically for (**RS**) problem. For visualization, M is taken of shape 3×3, thus $\mathcal{P}(\mathcal{D})$ and $\mathcal{P}(\mathcal{H})$ are both of dimension 2. If the Markov chain defined in Prop. 17 stabilizes, $\mathbb{E}[\Theta_k]$ will be periodic, matching the pattern of $\vec{\Theta}_t$. In fact, the period could be p, or a factor of p, or stabilizes where the period can be considered as 1. Thus we analyze $\vec{\eta}$ in $\mathcal{P}(\mathcal{D})$ and $\mathbb{E}[\Theta_k]$ in $\mathcal{P}(\mathcal{H})$, obtained from Monte-Carlo sampling along certain amount of episodes.

Fig. 6 (a) assume the true hypothesis travel along the triangular path connecting the 3 columns of M (shown in blue crosses). We found that GBT learners with ϵ in the interior of the cube (general GBT) produce a posterior path of period p, while the posteriors of Bayesian and SCBI learners tend to converge (Fig. 6 (b-e)). Thus a general GBT learner can naturally detect the periodicity of the world. We simulate up to $k = 400$ steps and 10240 repeats for each learner.

Moreover, we discovered that a general GBT learner's posteriors converge to a curve whose area is proportional to the area of the path of true hypotheses. In Fig. 7, as the path of true hypotheses vary by radius and shapes, the ratio (shown as slope) between both areas tends to be the same, it is 0.1620 in (a) and 0.1616 in (d), which suggests that this ratio is independent from the path of $\vec{\eta}$.

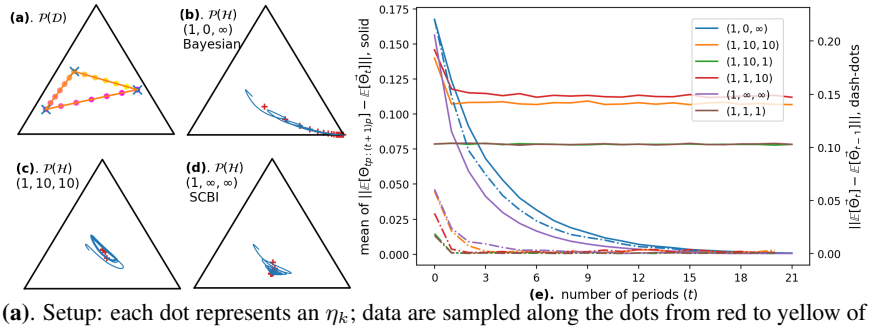


Figure 6: **(a)**. Setup: each dot represents an η_k; data are sampled along the dots from red to yellow of period 18, M has 3 columns represented by the blue crosses. **(b-d)**. Bayesian, general GBT, SCBI learners, resp., blue curve shows $\mathbb{E}[\Theta_k]$ and red crosses are $\mathbb{E}[\vec{\Theta}_t]$ (mean of 18 consecutive $\mathbb{E}[\Theta_k]$'s). **(e)**. for 6 different learners (shown in colors), plot (1) the averaged distance between $\mathbb{E}[\Theta_k]$ and its center v.s. number of periods, in solid lines and left y-ticks, and (2) the step-length of $\mathbb{E}[\vec{\Theta}_t]$ between consecutive periods in dash-dots and right y-ticks.

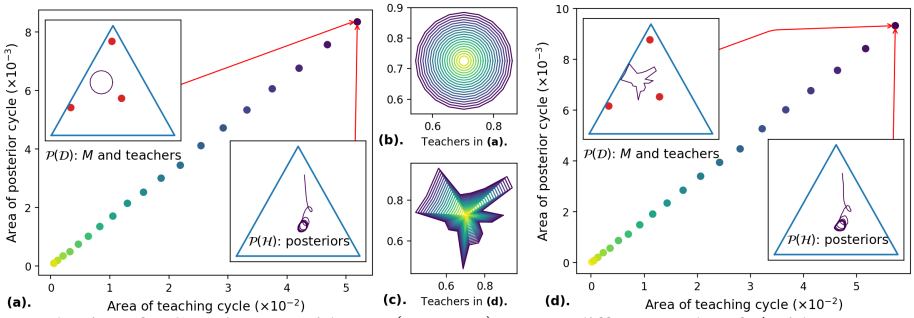


Figure 7: Behavior of a GBT learner with $\epsilon = (1, 10, 10)$ on two different paths of $\vec{\eta}$ with $p = 20$, tested in 300 steps and 10240 episodes. M is fixed and represented by the red dots. Learner's posteriors form roughly periodic paths, small panels on corners of (a, d), plot path of $\vec{\eta}$ and posterior paths, ratio between their enclosed areas are shown in yellow to blue dots. (b, c) shows the 20 concentric similar paths that $\vec{\eta}$ follow. Colors are matched between paths and corresponding area ratios.

Related Work. Prior work defines and outlines basic properties of Unbalanced Optimal Transport [Liero et al., 2018, Chizat et al., 2018, Pham et al., 2020]. Bayesian approaches are prominent in machine learning [Murphy, 2012] and beyond [Jaynes, 2003, Gelman et al., 1995]. There is also research on cooperative learning [Wang et al., 2019, 2020b,a] see also [Liu et al., 2021, Yuan et al., 2021, Zhu, 2015, Liu et al., 2017, Shafto and Goodman, 2008, Shafto et al., 2014, Frank and Goodman, 2012, Goodman and Frank, 2016, Fisac et al., 2017, Ho et al., 2018, Laskey et al., 2017]. Discriminative learning is the reciprocal problem in which one sees data and asks which hypothesis best explains it [Ng and Jordan, 2001, Mandler, 1980]. We are unaware of any work that attempts to unify and analyze the general problem of learning in which each of these are instances.

5 Conclusions

We have introduced Generalized Belief Transport (GBT), which unifies and parameterizes classic instances of learning including Bayesian inference, Cooperative Inference, and Discrimination, as Unbalanced Optimal Transport (UOT). We show that each instance is a point in a continuous, differentiable on the interior, 3-dimensional space defined by the regularization parameters of UOT. Moreover, to demonstrate general GBT's capacity of supporting generalized learning, we prove and illustrate asymptotic consistency and estimate rates of convergence, including convergence to hypotheses with zero prior support, and ability of gripping dynamic of the world. In summary, GBT unifies very different modes of learning, yielding a powerful, general framework for modeling learning agents.

Acknowledgments and Disclosure of Funding

This research was supported in part by DARPA awards HR0011-23-9-0050 and W912CG22C0001 and NSF MRI 2117429 to P.S.

References

Laetitia Chapel, Rémi Flamary, Haoran Wu, Cédric Févotte, and Gilles Gasso. Unbalanced optimal transport through non-negative penalized linear regression. *Advances in Neural Information Processing Systems*, 34:23270–23282, 2021.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

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

Paul Dagum, Adam Galper, and Eric Horvitz. Dynamic network models for forecasting. In *Uncertainty in artificial intelligence*, pages 41–48. Elsevier, 1992.

Joseph L Doob. Application of the theory of martingales. *Le calcul des probabilites et ses applications*, pages 23–27, 1949.

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.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 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.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

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

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3): 969–1117, 2018.

Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In *International Conference on Machine Learning*, pages 2149–2158. PMLR, 2017.

Weiyang Liu, Zhen Liu, Hanchen Wang, Liam Paull, Bernhard Schölkopf, and Adrian Weller. Iterative teaching by label synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

George Mandler. Recognizing: The judgment of previous occurrence. *Psychological review*, 87(3): 252, 1980.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

Khiem Pham, Khang Le, Nhat Ho, Tung Pham, and Hung Bui. On unbalanced optimal transport: An analysis of sinkhorn algorithm. In *International Conference on Machine Learning*, pages 7673–7682. PMLR, 2020.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE international conference on image processing (ICIP)*, pages 4852–4856. IEEE, 2014.

Christian P Robert, Nicolas Chopin, and Judith Rousseau. Harold jeffreys's theory of probability revisited. *Statistical Science*, 24(2):141–172, 2009.

Patrick Shafto and Noah D. Goodman. Teaching games: Statistical sampling assumptions for learning in pedagogical situations. In *Proceedings of the 30th annual conference of the Cognitive Science Society*, Austin, TX, 2008. Cognitive Science Society.

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.

Patrick Shafto, Junqi Wang, and Pei Wang. Cooperative communication as belief transport. *Trends in cognitive sciences*, 25(10):826–828, 2021.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

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

Junqi Wang, Pei Wang, and Patrick Shafto. Sequential cooperative bayesian inference. In *International Conference on Machine Learning*, pages 10039–10049. PMLR, 2020a.

Pei Wang, Pushpi Paranamana, and Patrick Shafto. Generalizing the theory of cooperative inference. *AIStats*, 2019.

Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 33, 2020b.

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.

Luyao Yuan, Dongruo Zhou, Junhong Shen, Jingdong Gao, Jeffrey L Chen, Quanquan Gu, Ying Nian Wu, and Song-Chun Zhu. Iterative teacher-aware learning. *Advances in Neural Information Processing Systems*, 34:29231–29245, 2021.

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

Generalized Belief Transport: Supplementary Material

Junqi Wang
Department of Math & CS
Rutgers University
Newark, NJ, 07102
junqi.wang@rutgers.edu

Pei Wang
Department of Math & CS
Rutgers University
Newark, NJ, 07102
peiwang@rutgers.edu

Patrick Shafto
Department of Math & CS
Rutgers University
Newark, NJ, 07102
shafto@rutgers.edu

1 Additional Materials

Cooperative Communication. *Cooperative communication* formalizes a single problem comprised of interactions between two processes: teaching and learning. The teacher and learner have beliefs about hypotheses, which are represented as 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 process of learning is then, given the data selected by the teacher, infer the beliefs of the teacher. The teacher's selection and learner's inference incur costs. The agents minimize the cost to achieve their goals. Communication is successful when the learner's belief, given the teacher's data, is moved to the target distribution.

Formally, denote the common ground between agents: the shared priors on $\mathcal{H}$ and $\mathcal{D}$ by $\mathbb{P}(h)$ and $\mathbb{P}(d)$, the shared initial matrix over $\mathcal{D}$ and $\mathcal{H}$ by M of size $|\mathcal{D}| \times |\mathcal{H}|$. In general, up to normalization, M is simply a non-negative matrix which also specifies the consistency between data and hypotheses[1]

In cooperative communication, a learner's goal is to minimize the cost of transforming the observed data distribution $\mathbb{P}(\mathcal{D})$ to the shared prior over hypotheses $\mathbb{P}(\mathcal{H})$. A learner's cost matrix $C^L = (C^L_{ij})_{|\mathcal{M}|\times|\mathcal{H}|}$ is defined as $C^L_{ij} = -\log M$. A *learning plan* is a joint distribution $L = (L_{ij})$, where $L_{ij} = P_L(d_i, h_j)$ represents the probability of the learner inferring h_j given d_i. It is proved in [Wang et al., 2019] that:

Proposition S.1. *Optimal cooperative communication plans, L, is the EOT plan with cost C^L and marginals being $\eta = \mathbb{P}(d)$ and $\theta = \mathbb{P}(h)$.*

2 Proofs

Proposition 1. *The UOT problem with cost matrix C, marginals θ, η and parameters $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$ generates the same UOT plan as the UOT problem with tC, θ, η, $t\epsilon = (t\epsilon_P, t\epsilon_\eta, t\epsilon_\theta)$ for any $t \in (0, \infty)$.*

Proof. Consider that the UOT problem solution is

$$P^\epsilon(C, \eta, \theta) = \underset{P \in (\mathbb{R}_{\geq 0})^{n\times m}}{\arg\min} \{\langle C, P\rangle - \epsilon_P H(P) + \epsilon_\eta \mathrm{KL}(P\mathbf{1}|\eta) + \epsilon_\theta \mathrm{KL}(P^T\mathbf{1}|\theta)\}. \quad (1)$$

[1] Data, d_i, are consistent with a hypothesis, h_j, when $M_{ij} > 0$.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Algorithm 1 Unbalanced Sinkhorn Scaling

input: C, θ, η, $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$, N stopping condition ω
initialize: $\mathbf{K} = \exp(-\epsilon_P C)$, $\mathbf{v}^{(0)} = \mathbf{1}_m$
while $k < N$ **and not** ω **do**
  $\mathbf{u}^{(k)} \leftarrow (\frac{\eta}{K\mathbf{v}^{(k-1)}})^{\frac{\epsilon_\eta}{\epsilon_\eta+\epsilon_P}}$, $\mathbf{v}^{(k)} \leftarrow (\frac{\theta}{K^T\mathbf{u}^{(k)}})^{\frac{\epsilon_\theta}{\epsilon_\theta+\epsilon_P}}$
end while
output: $M = \text{diag}(u)K\text{diag}(v)$

where the objective function is linear on C and ϵ.

$$\begin{aligned} P^{t\epsilon}(tC,\eta,\theta) &= \underset{P\in(\mathbb{R}_{\geq 0})^{n\times m}}{\arg\min} \{\langle tC, P\rangle - t\epsilon_P H(P) + t\epsilon_\eta \mathbf{KL}(P\mathbf{1}|\eta) + t\epsilon_\theta \mathbf{KL}(P^T\mathbf{1}|\theta)\} \\ &= \underset{P\in(\mathbb{R}_{\geq 0})^{n\times m}}{\arg\min} \; t\cdot\{\langle C, P\rangle - \epsilon_P H(P) + \epsilon_\eta \mathbf{KL}(P\mathbf{1}|\eta) + \epsilon_\theta \mathbf{KL}(P^T\mathbf{1}|\theta)\} \\ &= P^\epsilon(C,\eta,\theta). \end{aligned} \tag{2}$$

□

Proposition 2. *The UOT plan P in Equation 1, as a function of ϵ, is continuous in $(0,\infty)\times[0,\infty)^2$. Furthermore, P is differentiable with respect to ϵ in the interior.*

Proof. For simplicity, in this proof, for a vector v, we use both v_i and $v(i)$ to represent a component of v.

By definition, the UOT plan P minimizes the objective function $\Omega(P;\epsilon) = \langle C, P\rangle - \epsilon_P H(P) + \epsilon_\eta KL(P\mathbf{1}|\eta) + \epsilon_\theta KL(P^T\mathbf{1}|\theta)$. Since Ω is a strict convex function on P, there is only one minimal P. So the UOT plan P is the solution to $\nabla_P\Omega = 0$. From a direct calculation,

$$(\nabla_P\Omega)_{ij} = C_{ij} + \epsilon_P \ln P_{ij} + \epsilon_\eta(\ln(\sum_{k=1}^m P_{ik}) - \ln\eta(i)) + \epsilon_\theta(\ln(\sum_{k=1}^n P_{kj}) - \ln\theta(j))$$

and

$$(\nabla_P^2\Omega)_{ijkl} = \frac{\epsilon_P\delta_{ik}\delta_{jl}}{P_{ij}} + \frac{\epsilon_\eta\delta_{ik}}{\sum_{t=1}^m P_{it}} + \frac{\epsilon_\theta\delta_{jl}}{\sum_{t=1}^n P_{tj}}.$$

As we assume that $P_{ij} > 0$ for all i, j, all the terms above are well-defined. Besides, $\nabla_P\Omega$ is C^1 on η, θ and ϵ. Therefore, we can show $P^\epsilon(C,\eta,\theta)$ is continuous not only on ϵ but also on η and θ after checking Hessian. From implicit function theorem, if we show the above Hessian is invertible for $\epsilon_P > 0$, then the results of the proposition are true. Equivalently, it suffices to show that $\det H \neq 0$ where matrix H is the flattened $\nabla_P^2\Omega$ by mapping $(i,j,k,l) \mapsto (im+j, km+l)$.

Invertibility of H. Let $\mathbf{r}$ be the vector of reciprocals of row sums of P, i.e., $r_i = 1/\left(\sum_j P_{ij}\right)$, and similarly, let $\mathbf{c}$ be the vector of reciprocals of column sums of P, i.e., $c_j = 1/\left(\sum_i P_{ij}\right)$. Then

$$(\nabla_P^2\Omega)_{ijkl} = \frac{\epsilon_P\delta_{ik}\delta_{jl}}{P_{ij}} + \epsilon_\eta\delta_{ik}r_i + \epsilon_\theta\delta_{jl}c_j.$$

Let ϕ be the map $(i,j) \mapsto (im+j)$, then ϕ induces a reshaping of P to a vector of size mn, denoted by P^ϕ. When there is no ambiguity, we may omit the ϕ superscript.

Further define p^ϕ as a vector of dimension mn where $p_k^\phi = \epsilon_P/P_k^\phi$. By definition, $H^\phi = \epsilon_P(diag(p^\phi)) + \epsilon_\eta\mathbb{1}_m \otimes (diag(\mathbf{r})) + \epsilon_\theta(diag(\mathbf{c})) \otimes \mathbb{1}_n$ where $\mathbb{1}_k$ is the $k\times k$ matrix of ones, and $A\otimes B$ is Kronecker product (tensor product of matrices). Decompose $H = D + G$ where $D = \epsilon_P(diag(p^\phi))$ and $G = \epsilon_\eta\mathbb{1}_m \otimes (diag(\mathbf{r})) + \epsilon_\theta(diag(\mathbf{c})) \otimes \mathbb{1}_n$.

From now on, we may use P-row, P-column to represent i, j style indices, and G-row, G-column or simply row/column to represent those of G, or the ones in range $[1, mn]$. D is diagonal, and $\det G = 0$. Furthermore,

($*$) any row or column of G with index k can be represented by an entry position (i,j) of P by inverse of ϕ, and any rows of indices k_1,k_2,k_3,k_4 corresponding to $(i_1,j_1),(i_1,j_2),(i_2,j_1),(i_2,j_2)$ (i.e., determined as intersections of two P-rows and two P-columns) is linearly dependent: $G_{(k_1,\_)}+G_{(k_4,\_)}-G_{(k_2,\_)}-G_{(k_3,\_)}=\mathbf{0}$, we denote this property as ($*$).

Structure of $\det H$: Let $D=diag(p_1,p_2,\ldots,p_{mn})$, then $\det H$ is a polynomial on p_k's with constant term 0. Each term in $\det H$ is of form $f(\mathcal{I})\left(\prod_{k\notin\mathcal{I}}p_k\right)$ for each subset $\mathcal{I}\subseteq\{1,2,\ldots,mn\}$, and the coefficient $f(\mathcal{I})=\det G_{(\mathcal{I},\mathcal{I})}$ where $G_{(\mathcal{I},\mathcal{I})}$ is the submatrix with lines of indices not in $\mathcal{I}$, i.e., the entries of $G_{(\mathcal{I},\mathcal{I})}$ are of the form G_{ij} with $i\in\mathcal{I}$ and $j\in\mathcal{I}$.

Next we show that $f(\mathcal{I})$ is nonnegative for all $\mathcal{I}$, then with $p_k>0$ for all k, we can conclude that $\det H>0$. Since $\mathcal{I}\subseteq\{1,2,\ldots,mn\}$, $\phi^{-1}(\mathcal{I})\subseteq\{1,2,\ldots,n\}\times\{1,2,\ldots,m\}$, and ϕ is a bijection, we may not distinguish $\mathcal{I}$ from $\phi^{-1}(\mathcal{I})$, in order to make the statement neater.

1. [Operation-($*$) on $\mathcal{I}$]: We want to investigate the operations on $\mathcal{I}$ producing a subset $\mathcal{J}$ such that $f(\mathcal{I})=f(\mathcal{J})$. By the properties of determinant, ($*$) induces one operation: when $\mathcal{I}$ containing 4 integer pairs which can form the vertices of a rectangle, $f(\mathcal{I})=0$. Moreover, for any k_1,k_2,k_3,k_4 such indices in ($*$), we can generate row $G_{(k_4,\_)}$ by $G_{(k_4,\_)}=G_{(k_2,\_)}+G_{(k_3,\_)}-G_{(k_1,\_)}$, then if $\{k_1,k_2,k_3\}\subseteq\mathcal{I}$, we can build $G_{(k_4,\_)}$ on any $G_{(k_i,\_)}$, thus the determinant $\det G^{row}_{(\mathcal{I},\mathcal{I})}=\pm\det G_{(\mathcal{I},\mathcal{I})}$ (positive for k_2 and k_3, negative for k_1). Similarly, if we follow the same operation on columns, we have $\det G^{col}_{(\mathcal{I},\mathcal{I})}=\pm\det G_{(\mathcal{I},\mathcal{I})}$. And when doing both, $\det G^{col\cdot row}_{(\mathcal{I},\mathcal{I})}=\det G_{(\mathcal{I},\mathcal{I})}$. Therefore, we know that if $k_1,k_2,k_3\in\mathcal{I}$, and $\mathcal{J}=\{k_4\}\cup\mathcal{I}\backslash\{k_i\}$ for any $i=1,2,3$, then $f(\mathcal{I})=f(\mathcal{J})$. Such operations changing $\mathcal{I}$ to $\mathcal{J}$ is denoted by operation-$*$. In short, an operation-$*$ moves an end of a small "L-shaped" set of 3 pairs along a P-row or a P-column, producing another L-shaped set of 3 pairs.

2. [Regularized form of $\mathcal{I}$, and decomposition of nondegenerate regularized form $\mathcal{I}^\sharp$ into L-shaped subsets]: Once $\mathcal{I}$ or any $\mathcal{J}$ equivalent to $\mathcal{I}$ via operations-$*$ contains 4 pairs satisfying condition ($*$), $f(\mathcal{I})=0$, then we call $\mathcal{I}$ degenerate. In decomposing $\mathcal{I}$, when we find it degenerate, we stop since $f(\mathcal{I})$ is known.

We decompose $\mathcal{I}$ as set of pairs inductively in the following way before stopping. Start with any $(i,j)\in\mathcal{I}$, we look for pairs of form (i,l) and (k,j) in $\mathcal{I}$, adding them into the subset $A_{(i,j)}$ containing (i,j). Then check the degeneracy, by looking for whether $\mathcal{I}$ contains a point (k,l) with $(i,l),(j,k)\in A_{(i,j)}$, whenever $\mathcal{I}$ is degenerate, we stop since $f(\mathcal{I})=0$. Next we enlarge $A_{(i,j)}$ by changing the set $\mathcal{I}$ to a regularized form using operation-$*$'s. For each (k,l) with $(i,l)\in A_{(i,j)}$, then (k,j) can be constructed on (k,l) via an operation-$*$ with (i,j) and (i,l). Thus we modify $\mathcal{I}$ into $\mathcal{J}=(i,l)\cup\mathcal{I}\backslash(k,l)$ that $f(\mathcal{I})=f(\mathcal{J})$, and adding (i,l) into set $A_{(i,j)}$. Similar process can be done for those $(k,l)\in\mathcal{I}$ with $(k,j)\in A_{(i,j)}$.

After regularizing $\mathcal{I}$ and enlarging $A_{(i,j)}$ to maximum about (i,j), we get a regularized form $\mathcal{J}$ of $\mathcal{I}$, with $f(\mathcal{I})=f(\mathcal{J})$, and a component $A_{(i,j)}$ of L-shape. The set of $\mathcal{J}\backslash A_{(i,j)}$ has no elements of form (k,l) with $(i,l)\in A_{(i,j)}$ or $(k,j)\in A_{(i,j)}$, as they are already moved to $A_{(i,j)}$ by operation-$*$. Therefore, $\mathcal{J}\backslash A_{(i,j)}$ is supported on a rectangular region by deleting all P-rows $(k,\_)$'s and P-columns $(\_,l)$'s where k,l's occur in $A_{(i,j)}$.

Repeating the L-shaped component construction above for $\mathcal{J}\backslash A_{(i,j)}$, we can transform $\mathcal{I}$ into a regularized form (not unique or standard) $\mathcal{I}^\sharp$ and we have a decomposition $\mathcal{I}^\sharp=\bigcup A_{(i_t,j_t)}$ into L-shaped components, which do not intersect with each other. The name "regularized form" is given to the transformed set with a L-shaped decomposition, and since only operation-$*$ is applied, $f(\mathcal{I})=f(\mathcal{I}^\sharp)$.

3. [Properties between the L-shaped subsets:] For each $\mathcal{I}$ which we did not conclude $f(\mathcal{I})=0$ in the last step, we get $\mathcal{I}^\sharp$ and a decomposition $\mathcal{I}^\sharp=\bigcup_{t\in T}A_t$ into L-shaped subsets.

The construction of components A_t induces such a property: for two distinct components A_t there is no elements $(i,j)\in A_t$ and $(k,l)\in A_s$, in normal words, the A_t occupies certain P-rows and P-columns which is distinct from those of A_s.

For (i,j) and (k,l) with $i\neq k$ and $j\neq l$, $G_{im+j,km+l}=0$ from the formula that $G_{im+j,km+l}=\epsilon_\eta r_i\delta_{ik}+\epsilon_\theta c_j\delta_{jl}$. Therefore, the decomposition $\mathcal{I}^\sharp=\bigcup_{t\in T}A_t$ induces a decomposition of matrix

$G_{(\mathcal{I}^\sharp,\mathcal{I}^\sharp)}$ into blockwise diagonal matrix

$$\begin{bmatrix} G_{A_1,A_1} & 0 & \dots & 0 \\ 0 & G_{A_2,A_2} & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & G_{A_t,A_t} \end{bmatrix} \tag{3}$$

So for a decomposition $\mathcal{I}^\sharp = \bigcup_{t\in T} A_t$, we have $f(\mathcal{I}^\sharp) = \prod_{t\in T} f(A_t)$.

4. [$f(A)$ for an L-shaped component]: The last part is to show $f(A) > 0$ for all L-shaped components. Recall that $G_{im+j,km+l} = \epsilon_\eta r_i \delta_{ik} + \epsilon_\theta c_j \delta_{jl}$, so for A an L-shaped component with s P-rows and t P-columns, $G_{(A,A)}$ in general is of form

$$G_{(A,A)} = \begin{bmatrix} r_1+c_1 & \dots & r_1 & r_1 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_1 & \dots & r_1+c_{t-1} & r_1 & 0 & \dots & 0 \\ r_1 & \dots & r_1 & r_1+c_t & c_t & \dots & c_t \\ 0 & \dots & 0 & c_t & c_t+r_2 & \dots & c_t \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & c_t & c_t & \dots & c_t+r_s \end{bmatrix} \tag{4}$$

Recall the formula $\det \begin{bmatrix} E & B \\ C & D \end{bmatrix} = \det(E)\det(D - CE^{-1}B)$ and the matrix determinant lemma

$$\det(diag(c) + r\mathbf{1}\mathbf{1}^T) = (1 + r\mathbf{1}^T diag(c)^{-1}\mathbf{1})\det(diag(c)) = \prod c_i(1 + \sum(r/c_i)).$$

If $s = 1$ or $t = 1$, the determinant of $G_{(A,A)}$ can be calculated directly by the matrix determinant lemma above.

If $s > 1$ and $t > 1$, we cut Eq. (4) into 4 blocks $\begin{bmatrix} E & B \\ C & D \end{bmatrix}$ where E contains the upper left $t \times t$ part, B is zero but the last row, C is zero but the last column, D is a matrix in a similar form as E.

According to the characters of B, C stated above, it can be found that $CE^{-1}B = c_t^2 \mathbf{1} E_{t,t}^{-1} \mathbf{1}^T$ which is an $s \times s$-matrix. The entry $E_{t,t}^{-1} = \det E_{(1:t-1,1:t-1)} / \det E$ where $E_{(1:t-1,1:t-1)}$ is the matrix E without the last row and last column, moreover, $E_{t,t}^{-1} = \left(\prod_1^{t-1} c_i(1 + \sum_1^{t-1}(r_1/c_i))\right) / \left(\prod_1^{t} c_i(1 + \sum_1^{t}(r_1/c_i))\right) = \dfrac{1 + \sum_1^{t-1}(r_1/c_i)}{c_t(1 + \sum_1^{t}(r_1/c_i))} < 1/c_t$. Therefore, $CE^{-1}B = \lambda \mathbf{1}\mathbf{1}^T$ with $\lambda < c_t$ and $D - CE^{-1}B = diag(r_{2:s}) + (c_t - \lambda)\mathbf{1}\mathbf{1}^T$, whose determinant is positive according to the matrix determinant lemma.

As a consequence, $\det G_{(A,A)} > 0$ for each L-shaped components A. So combining the discussions in [1-4], we have $\det H = \det(D + G) > 0$.

Then the implicit function theorem implies the differentiability of P^ϵ on ϵ. □

Proposition 3. *For any finite s_P, s_η, $s_\theta \geq 0$, the limit of P^ϵ exists as ϵ approaches to $(\infty, s_\eta, s_\theta)$. In fact, $\lim_{\epsilon\to(\infty,s_\eta,s_\theta)} P^\epsilon_{ij} = 1$ for all i, j (Limit 1). Moreover, P^ϵ converges to the solution to*

$$\min\langle C, P\rangle - s_P H(P) + s_\theta KL(P^T\mathbf{1}|\theta), \text{ with constraint } P\mathbf{1} = \eta, \tag{5}$$

as $\epsilon \to (s_P, \infty, s_\theta)$ (Limit 2). Similarly, P^ϵ converges to the solution to

$$\min\langle C, P\rangle - s_P H(P) + s_\eta KL(P\mathbf{1}|\eta), \text{ with constraint } P^T\mathbf{1} = \theta, \tag{6}$$

as $\epsilon \to (s_P, s_\eta, \infty)$ (Limit 3). And in the case when $\epsilon \to (s_P, \infty, \infty)$, the matrix P^ϵ converges to the EOT solution (Limit 4):

$$\min\langle C, P\rangle - s_P H(P), \text{ with constraints } P^T\mathbf{1} = \theta \text{ and } P\mathbf{1} = \eta. \tag{7}$$

When $\epsilon \to (\infty, \infty, s_\theta), (\infty, s_\eta, \infty)$ or (∞, ∞, ∞), the limit does not exist, but the directional limits can be calculated..

Proof. Recall that $H(P) = -\sum_{ij}(P_{ij}\ln P_{ij} - P_{ij})$, $(\nabla_P H)_{ij} = -\ln P_{ij}$, and $H(P)$ is strictly concave, therefore H has a unique maximum mn at $P_{ij} = 1$, denoted by $\mathbb{1}$. Similarly, $KL(a|b) = \sum_i (a_i(\ln a_i - \ln b_i) - a_i + b_i)$, $\nabla_a KL(a|b)_i = \ln a_i - \ln b_i$, KL is strictly convex, therefore KL has a minimum 0 at $a_i = b_i$ for all i.

Limit 1. Shown by contradiction: When $\epsilon \to (\infty, s_\eta, s_\theta)$, suppose the limit $\lim_{\epsilon\to(\infty,s_\eta,s_\theta)} P^\epsilon_{ij}$ for some (i,j) does not exist, or is not 1. Thus there is $e > 0$ that, for any $\delta > 0$ and $N > 0$, there exists a parameter $\epsilon_1 = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$ such that $\epsilon_P > N$, $|\epsilon_\eta - s_\eta| < \delta$ and $|\epsilon_\theta - s_\theta| < \delta$, satisfying $|P^\epsilon_{ij} - 1| > e$.

However, for any $0 < e < 1/2$, let $\delta = 1$, let $E = (1+e)\ln(1+e) - (1+e) + 1 > 0$, $\min\Omega(P;\epsilon) \le \Omega(\mathbb{1};\epsilon) < C$ for some $G > 0$ where $(\mathbb{1})_{ij} = 1$ for all (i,j), and any $\epsilon \in \{(\epsilon_P, \epsilon_\eta, \epsilon_\theta) : s_\eta/2 < \epsilon_\eta < 3s_\eta/2, s_\theta/2 < \epsilon_\theta < 3s_\theta/2, \}$. So there is a $N > 0$ such that $NE > G + \max_{ij} C_{ij} + mn + L$ where $L = -\inf\{\epsilon_\eta KL(P\mathbf{1}|\eta) + \epsilon_\theta KL(P^t\mathbf{1}|\theta)\}$, meaning those P with $|P_{ij} - 1| > e$ for some (i,j) is not minimizing Ω.

The contradiction indicates that $\lim_{\epsilon\to(\infty,s_\eta,s_\theta)} P^\epsilon_{ij} = 1$ for all i, j.

Limit 2 & 3: The situation of $\epsilon_\theta \to \infty$ and $\epsilon_\eta \to \infty$ are similar, so we only prove for $\epsilon_\theta \to \infty$ case. Let $\hat{P}$ denote the solution to Eq. (6).

Let $\hat{P}$ be the solution to the optimization with constraints. We first show that $\lim_{\epsilon\to(s_P,s_\eta,\infty)} \sum_{k=1}^n P^\epsilon_{kj} = \theta_j$.

This is similar to limit 1. Suppose the limit either does not exist or is not θ_j, then there exists an $e > 0$ such that for any $N > 0$, $\delta > 0$, there exists $\epsilon_\theta > N$, $|\epsilon_\eta - s_\eta| < \delta$ and $|\epsilon_P - s_P| < \delta$, such that

$$\left|\sum_{k=1}^n P^\epsilon_{kj} - \theta_j\right| > e \tag{8}$$

for some j. Thus $KL((P^\epsilon)^T\mathbf{1}|\theta) > E$ for some $E > 0$. Consider that $\langle C, P\rangle \ge 0$, $H(P) \ge -mn$ and $KL(P\mathbf{1}|\eta) \ge 0$ are lower bounded, we can take sufficiently large N such that the P^ϵ satisfying Eq. (8) satisfy $\Omega(P^\epsilon;\epsilon) > \Omega(\hat{P};\epsilon)$, making P^ϵ fail to optimize $\Omega(\cdot;\epsilon)$, which is a contradiction. Thus we have $\lim_{\epsilon\to(s_P,s_\eta,\infty)} \sum_{k=1}^n P^\epsilon_{kj} = \theta_j$.

For each $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$, let θ^ϵ denote the $(P^\epsilon)^T\mathbf{1}$, then for any ϵ, the solution P^ϵ is also the solution to

$$\min_P \langle C, P\rangle + \epsilon_P H(P) + \epsilon_\eta KL(P\mathbf{1}|\eta), \text{ with constraint } P^T\mathbf{1} = \theta^\epsilon. \tag{9}$$

Denote $\Phi(P, \epsilon_P, \epsilon_\eta) := \langle C, P\rangle + \epsilon_P H(P) + \epsilon_\eta KL(P\mathbf{1}|\eta)$ When $\epsilon_P \in (0, \infty)$, the new objective function $\Phi(P, \epsilon_P, \epsilon_\eta)$ is continuous on P and ϵ_P,ϵ_η, and each minimization problem gets a unique solution since the objective function is strictly convex. Therefore, the limit $\lim_{\epsilon\to(s_P,s_\eta,\infty)P^\epsilon} = \widehat{P}$. We show this via contradiction:

Suppose the opposite, there exists some $\xi > 0$ such that $||P^\epsilon - \hat{P}||_2 > \xi$ for ϵ arbitrarily close to (s_P, s_η, ∞). Let

$$\alpha := \inf_{P^T e=\theta, ||P-\hat{P}||_2>\xi} \Phi(P, s_P, s_\eta) - \Phi(\hat{P}, s_P, s_\eta),$$

$\alpha > 0$ since the minimum $\hat{P}$ is unique and the objective is strictly convex. The sets $P^T e = \theta^\epsilon$ is compact since it is closed and bounded, so there exists bounds $b = (b_1, b_2, b_3)$ for $\epsilon = (\epsilon_P, \epsilon_\eta, \epsilon_\theta)$ such that in the bound where $|\epsilon_P - s_P| < b_1$, $|\epsilon_\eta - s_\eta| < b_2$ and $\epsilon_\theta > b_3$, $\max\Phi(P, s_P, s_\eta) - \Phi(P^\sharp, \epsilon_P, \epsilon_\eta) < \alpha/3$ for P with $P^T e = \theta$ and $P^\sharp$ its Euclidean projection to $\{P^T e = \theta^\epsilon\}$, and $\max\Phi(P, \epsilon_P, \epsilon_\eta) - \Phi(P^\flat, s_P, s_\eta) < \alpha/3$ for P with $P^T e = \theta^\epsilon$ and $P^\flat$ its Euclidean projection to $\{P^T e = \theta\}$.

Let ϵ be a parameter in the above bound b to (s_P, s_η, ∞), where $P = argmin_{P^T e=\theta^\epsilon}\Phi(P, \epsilon_P, \epsilon_\eta)$ is ξ far from $\hat{P}$. Then $\Phi(P, \epsilon_P, \epsilon_\eta) > \Phi(P^\flat, s_P, s_\eta) - \alpha/3 > \Phi(\hat{P}, s_P, s_\eta) + 2/3\alpha > \Phi(\hat{P}^\sharp, \epsilon_P, \epsilon_\eta) + \alpha/3 > \Phi(\hat{P}^\sharp, \epsilon_P, \epsilon_\eta)$, which is a contradiction to the assumption that P is the argmin.

Limit 4: Similar to the previous two limits, we can say that $\lim_{\epsilon\to(s_P,\infty,\infty)} \sum_{k=1}^n P^\epsilon_{kj} = \theta_j$ and $\lim_{\epsilon\to(s_P,\infty,\infty)} \sum_{k=1}^m P^\epsilon_{ik} = \eta_i$. Then the problem becomes the EOT problem, which has a unique solution.

Boundaries at $\epsilon_\eta = 0$ or $\epsilon_\theta = 0$: It is simple to check the continuity when $\epsilon_\eta \to 0$ or $\epsilon_\theta \to 0$. From Prop. 2, the continuity and differentiability hold for $\epsilon_\eta \to 0$ or $\epsilon_\theta \to 0$ when $\epsilon_P > 0$.

Nonexistence of the limits when $\epsilon_P, \epsilon_\eta \to \infty$, and directional limits: Let a sequence $\epsilon_1, \epsilon_2, \dots$ where $\epsilon_i = (\epsilon_P^i, \epsilon_\eta^i, \epsilon_\theta^i)$ satisfy $\lim \epsilon_P^i = \lim \epsilon_\eta^i = \infty$ and $\lim(\epsilon_\eta^i/\epsilon_P^i) = t$, then the limit P of P^ϵ satisfy $P_{ij} = t(\ln c_j - \ln n)/(t+1)$, since the limit minimizes the following objective function

$$H(P) + tKL(P\mathbf{1}|\eta).$$

The reason is, as $\sum \eta_i = 1$, $H(P)$ and $KL(P\mathbf{1}|\eta)$ cannot vanish for the same P, thus the minima of objective function approaches to infinity, therefore the finite terms $\langle C, P\rangle$ and $\epsilon_\theta KL(P^T\mathbf{1}|\theta)$ tend to have no effect on the minimal point P as $\epsilon_P, \epsilon_\eta$ increases to infinity.

A direct consequence of the above discussion is, when t changes, the limits P of those sequences changes, which indicates that the limit of P^ϵ as $\epsilon \to (\infty, \infty, s_\theta)$ fails to exist. And similar situation happens when $\epsilon \to (\infty, s_\eta, \infty)$

Nonexistence of the limits when $\epsilon_P, \epsilon_\eta, \epsilon_\theta \to \infty$, and directional limits : Similar to the discussions above, let the sequence $\epsilon_1, \epsilon_2, \dots$ where $\epsilon_i = (\epsilon_P^i, \epsilon_\eta^i, \epsilon_\theta^i)$ satisfy $\lim_{i\to\infty} \epsilon_i = (\infty, \infty, \infty)$. Further let $\lim(\epsilon_\eta^i/\epsilon_P^i) = u$, $\lim(\epsilon_\theta^i/\epsilon_P^i) = w$, then P^{ϵ_i} converges to the solution to the problem

$$H(P) + uKL(P1|\eta) + wKL(P^T1|\theta),$$

which could be considered as another UOT problem with cost function constantly 0.

□

Corollary 4. *Consider a UOT problem with cost $C = -\log \mathbb{P}(d|h)$, marginals $\theta = \mathbb{P}(h)$, $\eta \in \mathcal{P}(\mathcal{D})$. The optimal UOT plan $P^{(1,\epsilon_\eta,\epsilon_\theta)}$ converges to the posterior $\mathbb{P}(h|d)$ as $\epsilon_\eta \to 0$ and $\epsilon_\theta \to \infty$. Bayesian inference is a special case of GBT with $\epsilon = (1, 0, \infty)$.*

Proof. As direct application of Limit 3 of Proposition 3, we only need to show that the optimal plan $P^{(1,0,\infty)}$ is propositional to the posterior $\mathbb{P}(h|d)$.

$$P^{(1,0,\infty)} = \underset{P\in U(\theta)}{\arg\min}\, K(P) := \underset{P\in U(\theta)}{\arg\min}\ \{\langle C, P\rangle - H(P)\}. \tag{10}$$

where $U(\theta) = \{P \in \mathcal{M}(D \times H)|P^T\mathbf{1} = \theta\}$.

Let $\boldsymbol{\lambda} \in \mathbb{R}^{+m}$, consider the corresponding Lagrangian problem:

$$L(P, \boldsymbol{\lambda}) := \langle C, P\rangle - H(P) + \langle \boldsymbol{\lambda}, (P^T\mathbf{1} - \theta)\rangle$$

Partial derivatives $\partial_{P_{ij}} = 0$ and $\partial_{\lambda_j} L = 0$ result the following system of equations:

$$\log P_{ij} - \log \mathbb{P}(d_i|h_j) + \lambda_j = 0 \quad \sum_i P_{ij} - \mathbb{P}(h_j) = 0 \tag{11}$$

Calculation shows that the solution to Equation 11 is $P_{ij} = \frac{\mathbb{P}(d_i|h_j)\mathbb{P}(h_j)}{\sum_i \mathbb{P}(d_i|h_j)} = \mathbb{P}(d_i|h_j)\mathbb{P}(h_j) \propto \mathbb{P}(h_j|d_i)$. Hence the proof is completed. □

Corollary 5. *Consider a UOT problem with $\theta \in \mathcal{P}(\mathcal{H})$, $\eta = \mathbb{P}(d)$. The optimal UOT plan $P^{(\epsilon_P,\infty,0)}$ converges to $\eta \otimes \mathbf{1}$ as $\epsilon_P \to \infty$. Frequentist Inference is a special case of GBT with $\epsilon = (\infty, \infty, 0)$.*

Proof. As direct application of Proposition 3, we only need to show that $P^{(\infty,\infty,0)} = \eta \otimes \mathbf{1}$. Notice that the limit problem

$$P^\epsilon(C, \eta, \theta) = \underset{P\in(\mathbb{R}_{\geq 0})^{n\times m}}{\arg\min}\ \{\langle C, P\rangle - \epsilon_P H(P) + \epsilon_\eta \mathrm{KL}(P\mathbf{1}|\eta) + \epsilon_\theta \mathrm{KL}(P^T\mathbf{1}|\theta)\}. \tag{12}$$

as $\epsilon \to (\infty, \infty, 0)$ along ϵ_P-up direction is equivalent to

$$P^{(\infty,\infty,0)} = \underset{P \in (\mathbb{R}_{\geq 0})^{n \times m}}{\arg\min} H(P), \text{ with constraint } P\mathbf{1} = \eta \tag{13}$$

Hence $P^{(\infty,\infty,0)} = \eta \otimes \mathbf{1}$. □

Corollary 6. *Let cost $C = -\log M$, marginals $\theta = \mathbb{P}(h)$ and $\eta = \mathbb{P}(d)$. The optimal UOT plan $P^{(1,\epsilon_\eta,\epsilon_\theta)}$ converges to the optimal plan L as $\epsilon_\eta \to \infty$ and $\epsilon_\theta \to \infty$. Cooperative Inference is a special case of GBT with $\epsilon = (1, \infty, \infty)$, which is exactly entropic Optimal Transport [Cuturi, 2013].*

Proof. According to proposition 1, $L = P^{(1,\infty,\infty)}$, and the convergence result is a direct application of Limit 4 of Proposition 3 □

Corollary 7. *Consider a UOT problem with cost $C = -\log \mathbb{P}(d, h)$, $m = n$, and marginals $\theta = \eta$ are uniform. The optimal UOT plan $P^{(\epsilon_P,\epsilon_\eta,\epsilon_\theta)}$ approaches to a diagonal matrix as $\epsilon_\eta, \epsilon_\theta \to \infty$ and $\epsilon_P \to 0$. In particular, discriminative learner is a special case of GBT with $\epsilon = (0, \infty, \infty)$, which is exactly classical Optimal Transport [Villani, 2008].*

Proof. Limit 4 of Proposition 3 implies the convergence of $P^{(\epsilon_P,\epsilon_\eta,\epsilon_\theta)} \to P^{(0,\infty,\infty)}$ as $\epsilon_\eta, \epsilon_\theta \to \infty$ and $\epsilon_P \to 0$. When $m = n$, $P^{(0,\infty,\infty)}$ is a permutation matrix is the result of Wang et al. [2020b][Proposition 8].

□

Proposition 8. *In GBT with $\epsilon_\theta = \infty$, cost C and current belief θ. The learner updates θ with UOT plan in the same way as applying Bayes rule with likelihood from $P^\epsilon(C, \eta, \theta)$, and prior θ.*

Proof. From the GBT algorithm (Algorithm 1 in the main text), for a general data point d^i chosen, the GBT takes the vector normalization of some row P^ϵ, i.e., $\theta' = P^\epsilon_{(i,\_)}/(\sum_j P^\epsilon_{ij})$.

On the other hand, when we apply Bayes rule to P^ϵ, prior is $\theta = \mathbb{P}(h)$, likelihood $\mathbb{P}(d|h)$ is the column normalization of P^ϵ, satisfying $\mathbb{P}(d^i|h^j) = P^\epsilon_{ij}/(\sum_i P^\epsilon_{ij}) = P^\epsilon_{ij}/\theta_j$. The last equality is because $\theta(i) = \sum_j P^\epsilon_{ij}$ when $\epsilon_\theta = \infty$. So the posterior $\mathbb{P}(h|d^i)$ is the vector normalization of $\mathbb{P}(d^i|h)\mathbb{P}(h)$, by $\mathbb{P}(d^i|h^j)\mathbb{P}(h^j) = P^\epsilon_{ij}/\theta_j * \theta_j = P^\epsilon_{ij}$. Therefore, $\mathbb{P}(h^j|d^i) = \theta'(h^j)$. □

Now, we introduce some notations will be used in the following proofs.

Notations. Denote the set of all possible belief by $\Delta = \mathcal{P}(\mathcal{H})$. Distribution of Θ_k is denoted by μ_k. We only consider the case where no two hypotheses are the same in $\mathcal{H}$. Hence we make the following assumption that columns of $\exp(-\epsilon_P C)$ are not differ by a multiplicative scalar, i.e. columns of C are not differ by an additive scalar.

Lemma S.2. *For $\epsilon = (\epsilon_P, \infty, \infty)$, $\epsilon_P \in (0, \infty)$, given cost C with initial belief $\theta_0 \in \mathcal{P}(\mathcal{H})$ and fixed teaching and learning distribution $\eta_k = \eta \in \mathcal{P}(\mathcal{D})$ for all k, then the belief random variables $(\Theta_k)_{k\in\mathbb{N}}$ have the same expectation on h: $\mathbb{E}_{\Theta_k}[\theta(h)] = \theta_0(h)$.*

Proof. We start the proof by showing $\mathbb{E}_{\Theta_k}[\theta(h)] = \mathbb{E}_{\Theta_{k-1}}[\theta(h)]$ for $k \geq 1$. Notice that given cost C and data marginal η, an observed data $d \in \mathcal{D}$ and UOT planning uniquely determines a map from a learner's initial belief θ_{k-1} to one's posterior belief θ_k. Denote this map by $T_d : \theta_{k-1} \mapsto \theta_k$. Let the distribution of Θ_{k-1} over $\mathcal{P}(\mathcal{H})$ be μ_{k-1}, denote its support by S_{k-1}. Then the following holds:

$$\begin{aligned}
\mathbb{E}_{\Theta_k}[\theta(h^j)] &= \sum_{\theta\in S_{k-1}} \mu_{k-1}(\theta) \sum_{d^i\in\mathcal{D}} \eta^i T_{d^i}(\theta)(h^j) = \sum_{\theta\in S_{k-1}} \mu_{k-1}(\theta) \sum_{d^i\in\mathcal{D}} \eta^i \frac{M_k(i,j)}{\eta^i} \\
&= \sum_{\theta\in S_{k-1}} \mu_{k-1}(\theta) \sum_{d^i\in\mathcal{D}} M_k(i,j) = \sum_{\theta\in S_{k-1}} \mu_{k-1}(\theta)\theta(h^j) = \mathbb{E}_{\Theta_{k-1}}[\theta(h)]
\end{aligned}$$

Hence $\mathbb{E}_{\Theta_k}[\theta(h)] = \mathbb{E}_{\Theta_{k-1}}[\theta(h)] = \cdots = \mathbb{E}_{\Theta_0}[\theta(h)] = \theta_0(h)$.

□

Theorem 10 (PS). *Consider a learning problem with initial belief $\theta_0 \in \mathcal{P}(\mathcal{H})$, and the true hypothesis h^* defined by $\eta \in \mathcal{P}(\mathcal{D})$. If the learner's data distribution $\eta_k = \eta$, then belief random variables $(\Theta_k)_{k\in\mathbb{N}}$ converge to the random variable Y in probability, where $Y = \sum_{h\in\mathcal{H}} \theta_0(h)\delta_h$ and Y is supported on $\{\delta_h\}_{h\in\mathcal{H}}$ with $\mathbb{P}(Y = \delta_h) = \theta_0(h)$ for $\epsilon_\eta = \epsilon_\theta = \infty$ and $\epsilon_P \in (0, \infty)$.*

Proof. Step 1: First, we show the following claim inspired the proof proposition 5.1 in Wang et al. [2020a]

Claim: $\lim_{k\to\infty} \mu_k(\Delta_\epsilon) = 0$, for any $\epsilon > 0$, where $\Delta_\epsilon \coloneqq \{\theta \in \Delta : \theta(h) \leq 1 - \epsilon, \forall h \in \mathcal{H}\}$.

Assume the claim does not hold, then there exists $\alpha > 0$ and a subsequence $(\mu_{k_i})_{i\in\mathbb{N}}$ such that $\mu_{k_i}(\Delta_\epsilon) > \alpha$ for all i.

Let the center of Δ be u, we define $L(\mu) \coloneqq \mathbb{E}_\mu f(\theta)$, where $f(\theta) = \|\theta - u\|_2^2$, ($f$ may also be chosen as entropy $H(\theta)$). Then $L(\mu_{k+1}) = \mathbb{E}_{\mu_k}(\mathbb{E}_{d\sim\eta} f(T_d(\theta)))$.

Notice that f is strictly convex, by Jensen's inequality,

$$\mathbb{E}_{d\sim\eta} f(T_d(\theta)) \overset{(a)}{\geq} f(\mathbb{E}_{d\sim\eta} T_d(\theta)) \overset{(b)}{=} f(\theta) \tag{14}$$

Here (b) holds because:

$$\mathbb{E}_{d\sim\eta} T_d(\theta) \overset{(c)}{=} \sum_{d^i\in\mathcal{D}} \eta^i \cdot (M_k(i,\_)/\eta^i) = \sum_{d^i\in\mathcal{D}} M_k(i,\_) \overset{(d)}{=} \theta \tag{15}$$

$(c), (d)$ hold since M_k has marginals η, θ.

Moreover, equality holds in (a) if and only if $T_d(\theta) = \theta$ for all $d \in \mathcal{D}$. Thus rows of M_k are the same up to a scalar. This implies either (1) only one column of M_k is none zero, thus $\Theta_k \equiv \delta_h$ for some h or (2) M_k has at least two columns are differed by a scalar.

In the case of (1), if $\theta_0 \neq \delta_h$, $\Theta_k \equiv \delta_h$ is contradict to Lemma 8. Otherwise, $Y = \delta_h$, the result holds. In the case of (2), according to Wang et al. [2019], M_k is cross-ratio equivalent to $\exp(-\epsilon_P C)$, hence $\exp(-\epsilon_P C)$ has two columns differ by a multiplicative scalar, contradict to the assumption.

Thus for any $\theta \in \Delta_\epsilon$, $\mathbb{E}_{d\sim\eta} f(T_d(\theta)) > f(\theta)$. Therefore $L(\mu_{k+1}) > L(\mu_k)$ for any k.

Moreover, notice that Δ_ϵ is compact, there is a lower bound $\beta > 0$, such that $\mathbb{E}_{d\sim\eta} f(T_d(\theta)) - f(\theta) > \beta$ for all $\theta \in \Delta_\epsilon$. Therefore:

$$\begin{aligned} L(\mu_{k_i+1}) &= \mathbb{E}_{\theta_{k_i+1}\in\Delta_\epsilon}(\mathbb{E}_{d\sim\eta} f(T_d(\theta))) + \mathbb{E}_{\theta_{k_i+1}\in\Delta\setminus\Delta_\epsilon}(\mathbb{E}_{d\sim\eta} f(T_d(\theta))) \\ &> \mathbb{E}_{\theta_{k_i}\in\Delta_\epsilon}(f(\theta)) + \mathbb{E}_{\theta_{k_i}\in\Delta\setminus\Delta_\epsilon}(f(\theta)) + \alpha * \beta \\ &= L(\mu_{k_i}) + \alpha * \beta. \end{aligned} \tag{16}$$

Thus $L(\mu_{k_i+s}) > L(\mu_{k_i}) + s * \alpha * \beta \to \infty$ as $s \to \infty$. On the other hand, by definition, $f(\theta)$ is bounded above by the diameter of Δ under l^2 norm, so $L(\mu)$ is also bounded above. Contradiction! Therefore, the Claim holds.

Step 2. We show $\lim_{k\to\infty} \mathbb{P}(\Theta_k \in \Delta^h_{1-\epsilon}) = \lim_{k\to\infty} \mu_k(\Delta^h_{1-\epsilon}) = \theta_0(h)$, for all $h \in \mathcal{H}$ where $\Delta^h_{1-\epsilon} \coloneqq \{\theta \in \Delta : \theta(h) > 1 - \epsilon\}$.

For a fixed $h \in \mathcal{H}$, we have:

$$\begin{aligned} \theta_0(h) \overset{(a)}{=} \mathbb{E}_{\Theta_k}(\theta(h)) &\overset{(b)}{=} \mathbb{E}_{\theta_k\in\Delta^h_{1-\epsilon}}(\theta(h^j)) + \mathbb{E}_{\theta_k\in\Delta^u_{1-\epsilon}}(\theta(h)) + \mathbb{E}_{\theta_k\in\Delta_\epsilon}(\theta(h)) \\ &\overset{(c)}{\leq} \mu_k(\Delta^h_{1-\epsilon}) \cdot 1 + \mu_k(\Delta^u_{1-\epsilon}) \cdot \epsilon + \mu_k(\Delta_\epsilon) \cdot 1 \\ &= \mu_k(\Delta^h_{1-\epsilon}) + \epsilon + \mu_k(\Delta_\epsilon) \end{aligned}$$

where $\Delta^u_{1-\epsilon}$ denotes the union of all the other corners of Δ, i.e. $\Delta^u_{1-\epsilon} \coloneqq \cup_{h'\in\mathcal{H}\setminus h} \Delta^{h'}_{1-\epsilon}$. Here (a) is direct application of Lemma 8; (b) holds since $\Delta = \Delta^h_{1-\epsilon} \cup \Delta^u_{1-\epsilon} \cup \Delta_\epsilon$. (c) holds because in general

$\theta(h^j) < 1$, and $\theta(h^j) < \epsilon$ for any $\theta \in \Delta^u_{1-\epsilon}$. Therefore, $0 \leq \theta_0(h) - \mu_k(\Delta^h_{1-\epsilon}) \leq \epsilon + \mu_k(\Delta_\epsilon) \to \epsilon$ as $k \to \infty$ hold for any choice of ϵ. Pick a sequence of $\epsilon \to 0$, we have that $\lim_{k\to\infty} \mu_k(\Delta^h_{1-\epsilon}) = \theta_0(h)$.

Hence combining results from Step 1 and Step 2, we have shown Θ_k converges to Y in probability: $\mathbb{P}(|\Theta_k - Y| > \epsilon) \leq \mu_k(\Delta_\epsilon) + \sum_{h\in\mathcal{H}}(\theta_0(h) - \mu_k(\Delta^h_{1-\epsilon})) \to 0$ as $k \to \infty$. Hence the proof is completed. □

Corollary 11. *Given a fixed data sequence d_i sampled from η, if θ_k converges to δ_{h^j}, then the j-th column of M_k converges to η.*

Proof. For $\epsilon > 0$, there exists $N > 0$ such that $\theta_k(h^j) > 1 - \epsilon$ for any $k > N$. So $\sum_{j'\neq j} M_k(i, j') < \epsilon$ for any $d_i \in \mathcal{D}$, on the other hand $\sum_{j'} M_k(i, j') = \eta_i$. This implies that $\eta_i - \epsilon < M_k(i, j) < \eta_i$, so $M_k(i, j) \to \eta_i$ as $\epsilon \to 0$. Therefore the j-th column of M_k converges to η. □

Proposition 12. *Consider a learning problem with cost C, initial belief $\theta_0 \in \mathcal{P}(\mathcal{H})$, the true hypothesis h^* defined by $\eta \in \mathcal{P}(\mathcal{D})$. If the learner updates the estimation η_k with observed data (sampled from η) as stated above, then belief random variables $(\Theta_k)_{k\in\mathbb{N}}$ satisfies that for any $s > 0$, $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(\Theta(h) > 1 - s) = 1$. As a consequence, M_k as the transport plan has a dominant column (h^j) with total weights $> 1 - s$, and $|(M_k)_{ij} - \eta_k(i)| < s$. In fact, as long as the sequence of η_k as random variables converges to η in probability, the above proposition holds.*

Proof. The proof is similar to Step 1 of Theorem 10. The major difference is that data are sampled from η in each step, whereas the learner only has an estimation η_k at round k. Therefore, under current condition, equality (b) of Eq 14 need to be modified as following:

$$\mathbb{E}_{d\sim\eta} T_d(\theta_k) = \sum_{d^i\in\mathcal{D}} \eta^i \cdot (M_k(i, \_)/\eta^i_k) = \sum_{d^i\in\mathcal{D}} M_k(i, \_) \cdot \frac{\eta^i}{\eta^i_k} = \theta_k \odot \mathbf{v_k}. \tag{17}$$

where $\mathbf{v_k} = (\frac{\eta^i}{\eta^i_k})$ is a vector of the size of the data set $\mathcal{D}$, and $\odot$ represents element-wise product. Hence $\mathbb{E}_{d\sim\eta} f(T_d(\theta_k)) = f(\theta_k \odot \mathbf{v_k})$ holds for all $\theta_k \in \Delta$. Since $\eta_k \to \eta$ as $k \to \infty$. For any $\alpha * \beta > 0$, there exists $N > 0$ such that for $k > N$, $|1 - \frac{\eta^i}{\eta^i_k}| < \sqrt{\frac{\alpha*\beta}{2n}}$. Hence: $|f(\theta_k \odot \mathbf{v_k}) - f(\theta_k)| \leq \frac{\alpha*\beta}{2}$. Then corresponding to Eq 16, for $k_i > N$, we have:

$$\begin{aligned}
L(\mu_{k_i+1}) &= \mathbb{E}_{\theta_{k_i+1}\in\Delta_\epsilon}(\mathbb{E}_{d\sim\eta} f(T_d(\theta))) + \mathbb{E}_{\theta_{k_i+1}\in\Delta\setminus\Delta_\epsilon}(\mathbb{E}_{d\sim\eta} f(T_d(\theta))) \\
&> \mathbb{E}_{\theta_{k_i}\in\Delta_\epsilon}(f(\theta_k \odot \mathbf{v_k})) + \mathbb{E}_{\theta_{k_i}\in\Delta\setminus\Delta_\epsilon}(f(\theta_k \odot \mathbf{v_k})) + \alpha * \beta \\
&> \mathbb{E}_{\theta_{k_i}\in\Delta_\epsilon}(f(\theta_k)) + \mathbb{E}_{\theta_{k_i}\in\Delta\setminus\Delta_\epsilon}(f(\theta_k)) - \frac{\alpha*\beta}{2} + \alpha * \beta \\
&= L(\mu_{k_i}) + \frac{\alpha*\beta}{2}.
\end{aligned}$$

Hence the contradiction on the upper bound of $L(\mu_{k_i+1})$ still holds, which shows the claim that: $\lim_{k\to\infty} \mu_k(\Delta_\epsilon) = 0$. So $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(\Theta(h) > 1 - s) = 1$. The proof for the second part of the proposition follows exactly as Corollary 11. □

Proposition 14. *For $\epsilon = (\epsilon_P, \epsilon_\eta, 0)$ with $\epsilon_P \in (0, \infty)$, as $\eta_k \to \eta$ almost surely, the sequence Θ_k of posteriors as a sequence of random variables converges in probability to variable Θ, where $\mathbb{P}(\Theta = \mathbf{v}^i) = \eta(i)$ and $\mathbf{v}^i = P_{(i,\_)}/\left(\sum_{j=1}^m P_{ij}\right)$ and $P = P^\epsilon(C, \eta, \theta)$. Therefore, for any $s > 0$, $\lim_{k\to\infty} \sum_{h\in\mathcal{H}} \mathbb{P}(|\Theta_k(h) - 1| < s) = 0$ for generic (for all but in a closed subset) cost C and η, θ.*

Proof. First, $\epsilon_\theta = 0$ means that $P^\epsilon(C, \eta, \theta)$ is independent of θ. Therefore, $M_k = P^\epsilon(C, \eta_k, \theta)$ and has a limit $P^\epsilon(C, \eta, \theta)$, regardless of the concrete posterior θ_k. From construction of GBT, the posterior Θ_k is determined by $\mathbb{P}(\Theta_k = \mathbf{w}^i_k) = \eta(i)$ where $\mathbf{w}^i_k = (M_k)_{(i,\_)}/\sum_{j=1}^m (M_k)_{ij}$. Given the coupling (Θ_k, Θ) by setting only $\mathbb{P}(\Theta_k = \mathbf{w}^i_k, \Theta = \mathbf{v}^i) = \eta(i)$ for each i, we may calculate $\mathbb{P}(|\Theta_k - \Theta| < s)$ converge to 1 as M_k converge to $P^\epsilon(C, \eta, \theta)$.

For generic C, η, θ, the probability of $P^\epsilon(C,\eta,\theta)$ having a row with only one nonzero entry is 0.

□

Remark: As $\eta_k \to \eta$ almost surely, for any $e>0$, there exists $N>0$, such that, when $k>N$, the probability of having η_k e-close to η is 1. Thus in almost all episodes, with generic C,η,θ, when e is small enough, for any $||\eta'-\eta|| < e$ (using $p-\infty$ norm, same for below), the row-normalized (to $\mathbb{1}_n$) UOT plans

$$\max_i ||P_r^\epsilon(C,\eta',\theta)_{(i,\_)} - P_r^\epsilon(C,\eta',\theta)_{(i,\_)}|| < \frac{1}{4}\min_{i,j}||P_r^\epsilon(C,\eta,\theta)_{(i,\_)} - P_r^\epsilon(C,\eta,\theta)_{(j,\_)}||$$

where P_r^ϵ is the row normalization of P^ϵ.

Therefore, for such e, we may find an $N>0$ such that for any $k,k'>N$, $P_r^\epsilon(C,\eta_k,\theta) \neq P_r^\epsilon(C,\eta'_k,\theta)$. However, for generic η, say, no entry of η is 0, $||\theta_k-\theta'_k|| <$ when $k,k'>N$ and $d_k \neq d_{k'}$. Thus the posterior sequence of almost every episode fails to converge.

The original statement of the following Proposition is problematic, we changed the statement accordingly.

Proposition 16. *For a Bayesian learner, the posterior sequence $\{\Theta_k\}$ converges almost surely to δ_h where $h = \arg\min_{h'\in\mathcal{H}} KL(\overline{\eta}|M_{(\_,h')})$ and $M = e^{-C/\epsilon_P}$, $\overline{\eta} = \frac{1}{p}\sum_{k=0}^{p-1}\eta_k$.*

Proof. Based on the proof of Prop. 9, the behavior of the posterior sequence is determined by teaching data governed by the Central Limit Theorem.

We calculate $log\left(\Theta_k(h')/\Theta_k(h)\right)$ for any $h' \neq h$. With a tuple $(d_0,d_1,\ldots,d_k)$ of data points sampled from $\vec{\eta}$ periodically,

$$\begin{aligned}\log\left(\theta_k(h')/\theta_k(h)\right) &= \sum_{s=0}^{k}\left(\log(M_{(d_s,h')}) - \log(M_{(d_s,h)})\right)\\ &= \sum_{d\in\mathcal{D}}\lambda_d(\log(M_{(d,h')}) - \log(M_{(d,h)})\\ &= t\left(KL(\lambda|M_{(\_,h')}) - KL(\lambda|M_{(\_,h)})\right). \end{aligned} \tag{18}$$

where λ is the empirical distribution of the data points $(d_0,d_1,\ldots,d_k)$.

According to the central limit theorem, the teacher following $\vec{\eta}$ produces a sequence with associated empirical distribution $\overline{\eta}$ almost surely. Thus the posterior sequence converges to δ_h with h of the greatest KL-divergence. □

Proposition 17. *For ϵ in the interior of the cube, for (**PS**) problem, the sequence $\{\vec{\Theta}_t\}$ (random variables over $\mathcal{P}(\mathcal{H})^p$) form a time-homogeneous Markov chain. For (**RS**) problem, $\{(\vec{\Theta}_t, \frac{1}{pt}\sum_{k=0}^{p-1} t\eta_k)\}$, the random variable sequence producing samples $\{(\vec{\theta}_t, \frac{1}{pt}\sum_{k=0}^{pt-1}\delta_{d_k})\}$, forms a Markov chain.*

Proof. Define $\Phi_t = (\vec{\Theta}_t, \frac{1}{pt}\sum_{k=0}^{p-1} t\eta_k)$, whose sample is $\phi_t = (\vec{\theta}_t, \lambda_t)$ where $\vec{\theta}_t = (\theta_{(t-1)p}, \theta_{(t-1)p+1},\ldots,\theta_{tp-1})$ and λ_t is the empirical (statistical) distribution of the set of taught data points $\{d_0,d_1,\ldots,d_{tp-1}\}$.

Since in (**PS**) problem, θ_k is determined by θ_{k-1}, a fixed η and d_{k-1}, via the UOT solution. Thus, Θ_k depends on Θ_{k-1} only. So, $\vec{\Theta}_t$ depends only on $\vec{\Theta}_{t-1}$, showing that $\vec{\Theta}_t$ is time-homogeneous Markovian.

For (**RS**) problem, λ_t is determined by λ_{t-1} and the sample $(d_{(t-1)p}, d_{(t-1)p+1},\ldots,d_{tp-1})$ from $\vec{\eta}$, and $\vec{\theta}_t$ is determined by $\vec{\theta}_{t-1}$ (in fact, just the last element $\theta_{(t-1)p-1}$) and λ_{t-1}. Therefore, we get the Markovianess. □

3 Additional Simulations

Interpolation between learning models can be investigated properly under GBT. Human learners appear to be capable of moving between different learning models gradually. Consider an individual at a carnival who is playing a game. At each of 10 trials, a bit of information is provided, but the available reward decreases. The individual has a pool of tickets with which they can bet on the outcome at each trial. The question is how the individual should update their beliefs in order to maximize their rewards. On the first trial, their belief update, in order to accurately reflect the evidence, should follow Bayes rule. However, for the last trial, one should focus bets on the most probable outcome in order to maximize chances for rewards, that is, their beliefs should be optimized for discriminating among the possible outcomes. GBT offers a coherent way of interpolating between these two approaches to provide candidate strategies on the intermediate steps. Such situations are common where there is an explicit constraint on the time horizon after which point no further evidence can be obtained, and there are incentives to act early, rather than to wait until evidence has fully accumulated; for example, identifying dangerous situations (tiger or not? poisonous or not?).

We now demonstrate how continuity of GBT (section 3.1) allows one to gradually interpolate between Bayesian and discriminative learning over steps (rather than a sharp switch).

3.1 Simulation Setup

Suppose a learner who observes data sampled from a true hypothesis $\mathbb{P}(d|h^*)$, and needs to make a conclusion on whether h^* is one of the hypotheses in $\mathcal{H}$ within a fixed number N of observations.

Here we compare a baseline learner who utilizes Bayesian inference ($\epsilon = (1, 0, \infty)$) on the first $N-1$ observations, and switch to discriminative learning ($\epsilon = (0, \infty, \infty)$) on the last observation, against learners who interpolate from Bayesian to discriminative learning gradually along a sequence of models on curves in GBT. Two curves along with intermediate models are shown red and orange in Figure 1.

We take a random sampled M of shape 4×4 as an example,

$$M = \begin{bmatrix} 0.225779 & 0.014886 & 0.433787 & 0.050735 \\ 0.613779 & 0.322347 & 0.172658 & 0.109262 \\ 0.069799 & 0.620178 & 0.29083 & 0.243635 \\ 0.090643 & 0.042588 & 0.102725 & 0.596368 \end{bmatrix}.$$

Thus $|\mathcal{H}| = |\mathcal{D}| = 4$. Set $N = 10$ and start from uniform $\theta = (0.25, 0.25, 0.25, 0.25)$.

Simulation details: We perform 40000 trials in total. For each trial s (or say each episode), we uniformly sample $X_s \in \mathcal{P}(\mathcal{H})$, and let the true hypothesis h^* be a normalized (thus a distribution) column of M, uniformly sampled from the 4 columns. While teaching the episode, in each round, we sample a hypothesis $h \in \mathcal{H}$ following X_s, then sample a data d following the column of M corresponding to d. During inference, we set η_k by counting the frequency of each $d \in \mathcal{D}$ (starting from 1 to avoid 0 in η_k) and then normalize, as stated in (**RS**) model in Sec. 3.

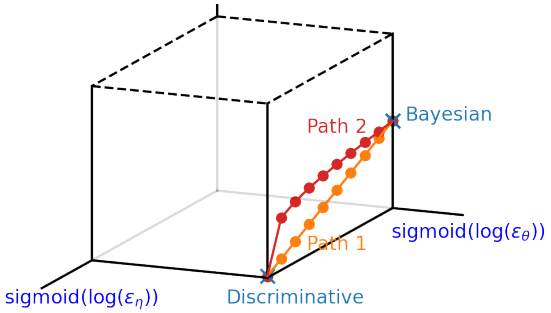


Figure 1: Baseline (sharp change) and two paths we follow on the parameter space of GBT.

3.2 Results

Following paths shown in Fig. 1, for baseline (blue, left), path 1 (orange, middle), and path 2 (red, right), the distribution of maximal component of each posterior at round 10 are shown in histograms of 30, and the entropy of these posteriors are plotted in the lower three figures.

Conclusiveness (minimal l^1 distance between posterior and a 1-hot vector) and posterior entropy are plotted as histograms. The results show that the smoother path may lead to a more conclusive posterior. Numerical results: Conclusiveness of Blue: mean 0.9406, standard deviation 0.1300. Conclusiveness of Orange: mean 0.9964, standard deviation 0.0327. Conclusiveness of Red: mean 0.9834, standard deviation 0.0676. Furthermore, compared with a sudden jump, gradual interpolations have lower entropy. Numerical results: entropy of Blue: mean 0.1261, standard deviation 0.2435, entropy of Orange: mean 0.0079, standard deviation 0.0629; entropy of Red: mean 0.0388, standard deviation 0.1336.

Thus learning tends to be more conclusive along these paths. Here conclusiveness means that the ability of getting a conclusion (one component of the posterior eventually becoming dominant). Furthermore, the entropy distributions shown in the lower figures also illustrate this point, as compare to baseline, gradual interpolations have lower entropy.

It is necessary to consider that, the two paths and interpolations are chosen for demonstration purpose, by no means they are optimal. However, we believe GBT is capable of facilitating exploration of such optimization.

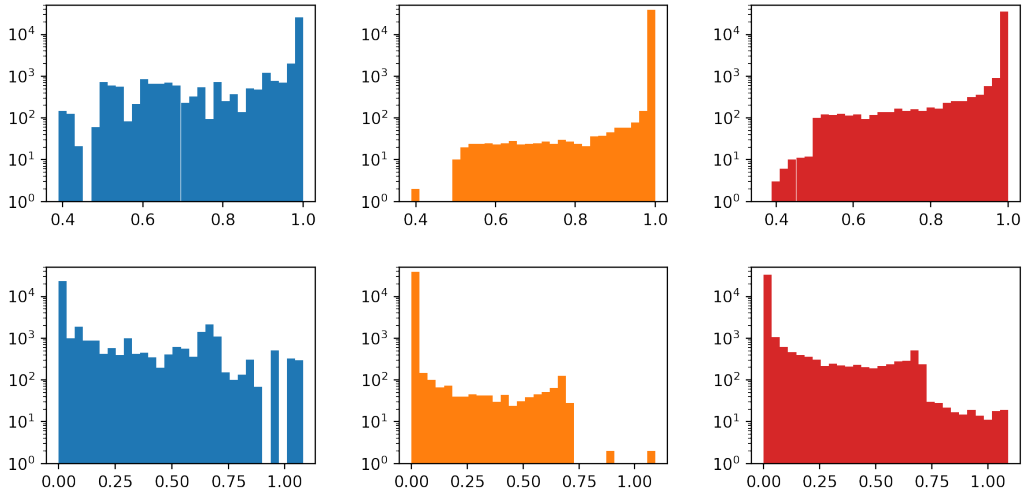


Figure 2: Results. Upper: distribution of maximal component of posterior. Lower: Entropy distribution of posteriors. Left: baseline. Middle: along path 1. Right, along path 2.

3.3 Sequential GBT: Dynamic

There are some more data from simulation, all with M of size 3×3, exploring the effects of varying ϵ and choosing different M.

We first investigate the behavior when $\epsilon = (1, \epsilon_\eta, \epsilon_\theta)$ where $\epsilon_\eta, \epsilon_\theta \in [0, \infty)$. We choose a grid $(10^{-2}, 10^{-1}, \ldots, 10^9)^2$ and measure asymptotic diverging distance $\frac{1}{p}\sum_{k=(t-1)p}^{tp-1} ||\mathbb{E}[\Theta_k] - \mathbb{E}[\vec{\Theta}_t]||_2$ at each point in the grid, where M and the circular teaching path is shown in Fig. 3 (a), and the result is shown in (b). The "asymptotic" value is the average of last 5 periods in the 15 period simulations where $t = 15$, period $p = 20$ and total steps $k = 300$ in each episode (empirically, the last 5 periods are usually stable enough to represent the asymptotic situation). The mean of 10240 episodes are taken to estimate the expectation of Θ_k and $\vec{Theta}_t$. We see a higher contribution of ϵ_θ than ϵ_η in controlling the posteriors' converging either to a point or to an attractive curve.

Next, we choose a set of M randomly, and set the teacher teaching along a circle of period 20. We are interested in the relation between the matrix and the area ratio (posterior loop divided by the teaching

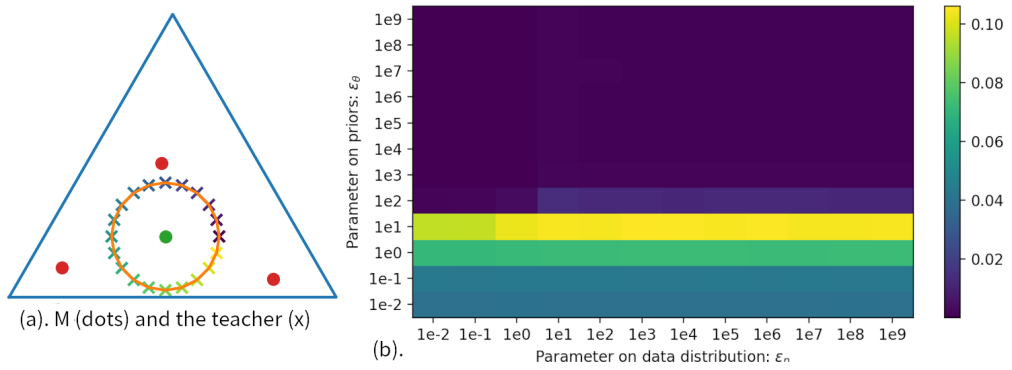


Figure 3: Influence of ϵ on the average distance between stable posterior and the Euclidean barycenter of each posterior period. (a) The setup, M and the teaching path. (b). the result. With the asymptotic average diverging distance of each period represented by colors of each cell, and the parameter ϵ represented by positions, it can be seen from the figure that when ϵ_θ is large, the average posterior tends to converge and fail to detect the periodicity of teacher. The most sensitive ϵ occurs when $\epsilon_\theta \approx 10$.

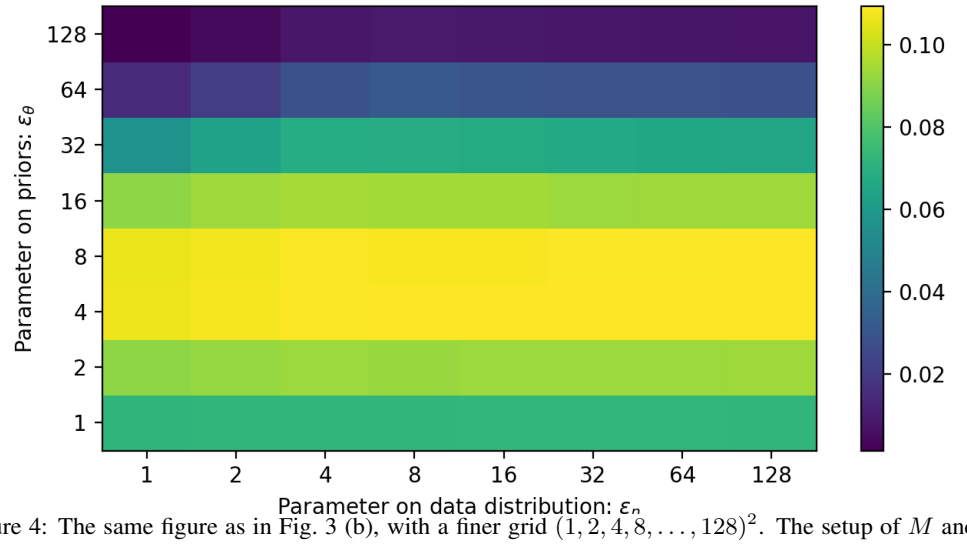


Figure 4: The same figure as in Fig. 3 (b), with a finer grid $(1, 2, 4, 8, \ldots, 128)^2$. The setup of M and the teacher stays the same as in Fig. 3 (a)

loop). In Fig. 5, the area ratio roughly follows a linear relation to the area of the 3 columns of M (equivalently, a constant times $\det(M)$). A linear regression with the R value 0.997 shows that the slope is 0.318 and the intercept is 0.005.

References

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

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

Junqi Wang, Pei Wang, and Patrick Shafto. Sequential cooperative bayesian inference. In *International Conference on Machine Learning*, pages 10039–10049. PMLR, 2020a.

Pei Wang, Pushpi Paranamana, and Patrick Shafto. Generalizing the theory of cooperative inference. *AIStats*, 2019.

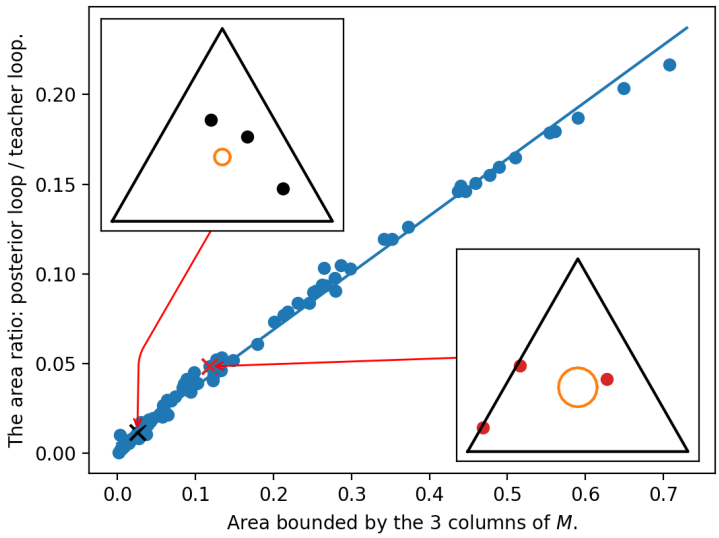


Figure 5: Area ratio v.s. the area bounded by 3 columns of M. $\epsilon = (1, 1, 1)$. There are 69 points plotted. The small figures show the setup — matrix M and the teaching path — for two of the points in the plot.

Pei Wang, Junqi Wang, Pushpi Paranamana, and Patrick Shafto. A mathematical theory of cooperative communication. *Advances in Neural Information Processing Systems*, 33, 2020b.