Optimal and heuristic teaching in vast concept spaces

Maya Malaviya (maya.malaviya@nyu.edu)[1] **& Mark K. Ho (mark.ho@nyu.edu)**[1]
[1]Department of Psychology, New York University

Abstract

Humans are remarkably adaptive instructors who adjust advice based on their estimations about a learner's prior knowledge and current goals. Inspired by prior work in rational pedagogy, we model teachers that reason about how learners will update their beliefs when given different examples, and thereby select examples that minimize expected learner error. We demonstrate that Bayesian non-parametric approaches can characterize teaching strategies in continuous domains, where traditional rational teaching models are intractable. We compare human teaching choices against our model and a variety of heuristics. Our model explains significant variance in human choices beyond a mixture of heuristics, and provides insight insight into how teachers formulate pedagogical guidance in computationally tractable ways, even in vast spaces.

Keywords: pedagogy; social learning; Gaussian processes; function learning

Introduction

Learners recognize when they are in a pedagogical scenario and are able to infer more information as a result, even when given very little data (Shafto et al., 2014). Teachers also rely on this situational understanding when choosing advice, tailoring their choices to assist a learner's beliefs and goals (Rafferty et al., 2011, 2015), even from a young age (see natural pedagogy, Gweon (2021)). However, the majority of previous task paradigms examine pedagogy in small and discrete domains, like categorization or feature-learning (Bridgers et al., 2020; Sumers et al., 2021). Other work, which models teacher choices in more complex tasks like arithmetic and the number game (Rafferty et al., 2011) only allow teachers to choose from a small pool of discrete actions. To our knowledge, few works in cognitive psychology attempt to model teaching in domains with a continuous space of possible examples and vast hypothesis spaces for learners.

We develop a novel approach that combines ideas from the cognitive science of communication (M. Carroll et al., 2019; Goodman et al., 2008; Shafto et al., 2014) with tools from Bayesian non-parametrics (Rasmussen, 2003; Williams & Rasmussen, 1995). In particular, we build on work in *rational teaching* (Shafto et al., 2014) and take inspiration from *rational speech act theory* (Goodman & Frank, 2016; Shafto et al., 2012) that formalizes people's intuitive capacity to teach others new concepts by providing maximally informative examples, demonstrations, and explanations. Standard approaches to rational teaching are only tractable when the teacher's action space is discrete, finite, and relatively small (Bridgers et al., 2020; Rafferty et al., 2011; Sumers et al., 2021), making it unclear how pre-existing accounts could possibly scale to real-world domains where the space of decisions and concepts is vast and unbounded (e.g., medicine, robotics, driving).

Though pedagogical research has emphasized teaching in discrete settings, a paradigm in cognitive psychology called *function learning* has characterized how people learn and extrapolate mappings from inputs to outputs, including support for continuous settings (J. D. Carroll, 1963; DeLosh et al., 1997; Koh & Meyer, 1991). One method of particular interest is a visual regression task. For example, in Schulz et al. (2017), human participants observed an image with a few dots placed along a domain, then drew a line that represented the function which they believed had produced those dots (Figure 1B).

Function learning in this continuous setting has been modeled using a tool called Gaussian process (GP) regression, which is a kernel-based method capable of representing the theoretically infinite space of function hypotheses in an interpretable manner. Gaussian processes provide a flexible framework that allow generalization to a continuous domain given a finite set of known points. GPs can capture both smooth, similarity-based generalization (through kernels that privilege local interpolation, e.g., k_{sim} in Figure 2) and rule-based extrapolation (through kernels that contain structural assumptions, such as linearity or periodicity k_{per} in Figure 2). Therefore, GPs unify earlier perspectives in the function learning literature by using kernels to express behavior consistent with *both* associative models and rule-based models (see Griffiths et al., 2008; Schulz et al., 2017). Furthermore, GP regression with compositional kernels accurately predict human patterns of function hypothesis generation and learning (Lucas et al., 2015; Schulz et al., 2017). But, in all previous experiments, points were assumed to be selected randomly from a hidden function; this body of research has not examined the role of pedagogy in shaping the learner's representations.

In this paper, we combine rational teaching with Gaussian processes, to formulate a model of *function teaching*. A teacher is tasked with selecting data points from a function to give the learner, and therefore infers what points would best help the learner estimate the true function. We model the learner that is estimating a function from observed data points as a GP. The learner's GP has parameters that define

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

the structure of the functions they are expecting to learn. A rational learner fits their conception of which hyperparameters are most probable using the data they have observed. We focus on a model of the learner that contains a mixture of different parameters to represent different structures of functions. The teacher decides what data points to give the learner by 1) simulating how the learner GP will update to a posterior when given a candidate set of points and 2) comparing that posterior to the true function they are attempting to communicate. As such, the teacher considers the utility of candidate point sets using a negative mean squared error ($-MSE$) function between the estimated learner GP and the true function. With this method, we can score the effectiveness of teacher examples given a learner representation.

This approach naturally handles continuous representations. In operating over continuous functions rather than discrete concepts, we can apply our framework to continuous embeddings. Cognitive psychology has a long tradition of representing psychological similarity and generalization in terms of embedding spaces (Attneave, 1950; Shepard, 1987). Therefore, learning in continuous spaces is of interest for the modeling of concept learning and categorization. In addition, embeddings from images and text are the foundation of many modern machine learning systems (Goodfellow et al., 2016; Lecun et al., 1998). By modeling continuous embeddings, we open the door to rational teaching in settings where traditional discrete approaches are infeasible. Thus, this work can also serve as a basis for developing scalable approaches that allow AI systems with embedding spaces to efficiently and intuitively communicate with humans (Collins et al., 2024; Sucholutsky et al., 2023).

Ultimately, this model allows us to generate new hypotheses about human behavior in the difficult task of distilling knowledge to guide a learner by selecting positive examples. In this paper, we report teaching behavior from simple heuristics, our GP teacher, and humans. We perform a model comparison and show that, while some heuristics are very important for behavior, our posited model explains significant variance in human teacher decisions.

Computational Framework

Gaussian Processes

A Gaussian process (GP) defines a distribution over functions, parameterized by a mean function μ, which specifies the expected output function, and kernel function k which specifies the covariance of outputs. Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a function drawn from a $\mathcal{GP}$. Consider an example $\mathcal{GP}$ parameterized by a periodic kernel function k_{per}, a standard option for capturing functions that repeat themselves, with parameters p (periodicity), σ ("amplitude"; average distance from the mean of the mean function), ℓ (within-period smoothness). We can then represent the $\mathcal{GP}$ like so:

$$f \sim \mathcal{GP}(\mu, k) \qquad k_{\text{per}}(\mathbf{x_i}, \mathbf{x_j}) = \sigma^2 \left(-\frac{2\sin^2(\pi|\mathbf{x_i}, \mathbf{x_j}|)/p}{\ell^2} \right) \tag{1}$$

Kernels A kernel function defines the similarity between two inputs; this means the kernel specifies the kinds of function patterns the GP prior considers plausible before seeing any data. Then, once data points are observed, the kernel function dictates how to use that data to make inferences about other areas of the function. Two $\mathcal{GP}$s with different kernel functions will make different generalizations, even when given the same data (see Figure 2).

Gaussian Process Regression We have discussed a representation of GPs, but not how to update them. Gaussian process regression updates a prior over functions to a posterior given observed data (Rasmussen, 2003; Williams & Rasmussen, 1995). For inputs $\mathbf{x} = (x_1, \ldots, x_n)$ with outputs $\mathbf{y} = (y_1, \ldots, y_n)$, the GP prior expresses a multivariate normal distribution:

$$\mathbf{y} \sim \mathcal{N}\big(\mu(\mathbf{x}), K(\mathbf{x}, \mathbf{x}) + \sigma^2 I\big), \tag{2}$$

where $\mu(\mathbf{x})$ is the mean function (often initialized to 0), and $K(\mathbf{x}, \mathbf{x})$ is the covariance matrix with entries defined by the kernel function applied to each pair of elements in $\mathbf{x}$, i.e., $k(x_i, x_j) \forall i, j \in \mathbf{x}$. σ^2 is the observation noise.

GP Learner Model

The learner observes a set of data points $(\mathbf{x}, \mathbf{y})$ and map them to a distribution over possible functions that could have generated those points. The generalization target of the learner (the domain over which they must learn the function) is notated as $\mathbf{x'} = (x'_1, ..., x'_l)$. Equation 3 represents the prediction of a simple learner's GP; it gives a probability distribution of values for $\mathbf{y_i}'$ at any value of $\mathbf{x_i}'$ given a kernel k.

$$P_L = P(\mathbf{y'}|\mathbf{x}, \mathbf{y}, \mathbf{x'}, k) \tag{3}$$

How does the learner decide upon a kernel k in the first place? Equation 4 represents the inference problem of selecting a kernel to use in response to observed data $(\mathbf{x}, \mathbf{y})$.

$$P(\mathbf{y'}, k_\theta|\mathbf{x}, \mathbf{y}, \mathbf{x'}) = P(\mathbf{y'}|\mathbf{x}, \mathbf{y}, \mathbf{x'}, k_\theta) P(k_\theta|\mathbf{x}, \mathbf{y}) \tag{4}$$

A more complex learner can represent multiple hyperparameter sets, k_θ. In this case, the learner is modeled as a mixture of $\mathcal{GP}$s, each corresponding to a candidate kernel k_i. This formulation (Equation 5) allows the learner to maintain uncertainty over which kernel best explains the data. Upon observing data points (from a teacher or prior knowledge), the learner performs a Bayesian update on each GP independently, and then assigns weights to the individual components based on their likelihood of having generated the data. The resulting weighted mixture is the learner's posterior distribution over functions given the data points.

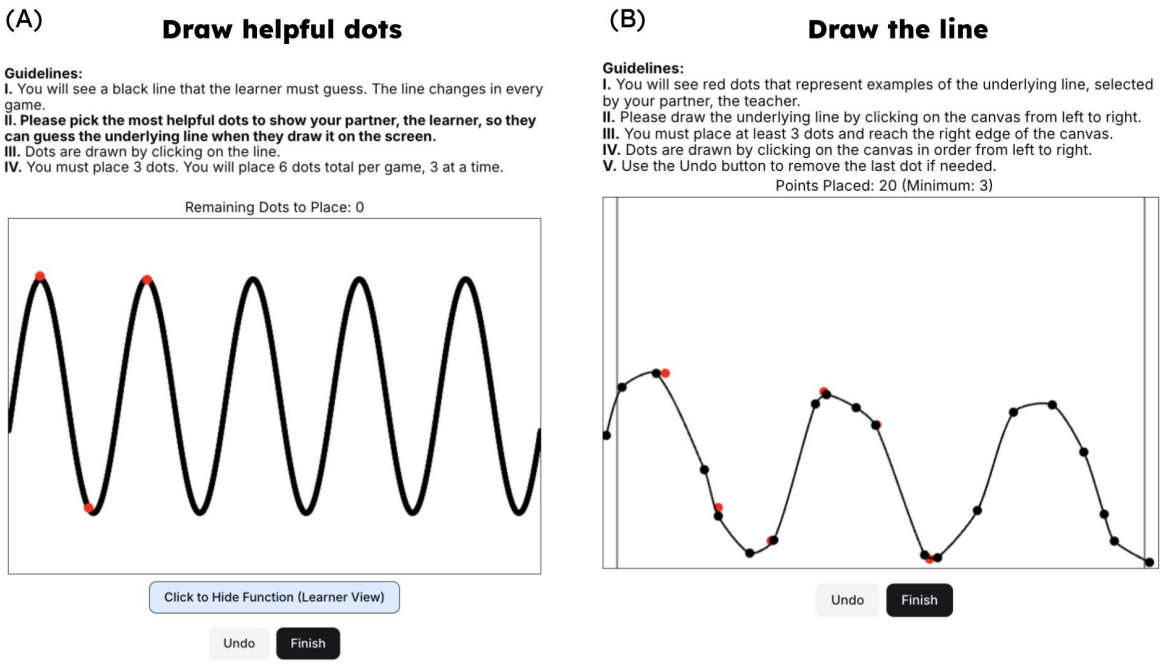

Figure 1: (A) Teacher view. The teacher is presented with a function. They place a limited number of example points such that their partner, another participant, will be able to draw an accurate representation of the function. (B) Learner view. The learner is presented with red dots from the function. They place their best guess of the underlying function that produced those red dots. The small black dots are the points that the learner placed, and the black line is a Bezier interpolation between them.

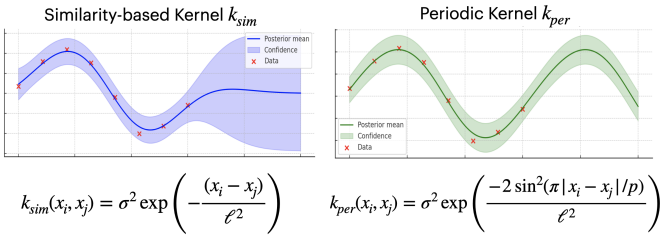

Figure 2: Two GPs with different kernels, trained on the same data points (red xs). Left: similarity-based kernel encodes that the closer x_i and x_j are to each other along the domain, the more similar y_i and y_j estimates will be. Right: periodic kernel, encodes that if you know x_i, y_i, you can infer y_j when $x_j = x_i + cp$ where $c \in \mathbb{Z}$. This kernel can make estimates even in areas without observed data, because of the assumption of an underlying repeating structure.

$$P(\mathbf{y}'|\mathbf{x}, \mathbf{y}, \mathbf{x}') = \sum_{\theta} P(k_\theta|\mathbf{x}, \mathbf{y})\ P(\mathbf{y}'|\mathbf{x}, \mathbf{y}, \mathbf{x}', k_\theta) \quad (5)$$

An important requirement of this approach is specifying in advance a finite set of kernels and associated hyperparameter ranges that define the hypothesis space. This means that if there is a true optimal k^*_θ but it is not enumerated in the set, you are not guaranteed an accurate representation. The choice of kernel hyperparameters is often regularized by a prior distribution $P(k_\theta)$. For instance, our learner incorporates a log-normal prior on the length-scale ℓ to reflect an inductive bias that functions are generally smooth, not discontinuous. We also use a prior on the variance parameter σ^2 to avoid solutions with vanishing or exploding amplitudes. These priors prevent overfitting and also encode psychological assumptions about which functions are more plausible (Griffiths et al., 2008; Rasmussen, Williams, et al., 2006).

One limitation of this approach is that many GPs are updated in this learner model, even if some are unlikely given the observed data. Such a representation is not as compressed as a human representation could be. But, these updates can be computed in parallel and thus can be efficient in practice. Conceptually, this representation provides richer coverage of possible learner hypotheses than a single-kernel model, which is especially valuable since teachers cannot assume that learners know the "true" kernel a priori.

GP Teacher Model

The teacher's goal is to select a set of teaching points $\mathbf{x}, \mathbf{y}$ so that the learner forms an accurate model $\mathbf{y}'$ of the true underlying function $\mathbf{y}^*$. The teacher knows the ground-truth function values $\mathbf{y}^*$ over the generalization domain $\mathbf{x}'$. Notably, for our formalization, the teacher is assumed to know the learner's candidate kernels and inference process P_L. The learner's posterior predictive distribution $P_L(\mathbf{y}'|\mathbf{x}, \mathbf{y}, \mathbf{x}')$ is given by Equation 5.

The teacher utility function evaluates a candidate teaching set $(\mathbf{x}, \mathbf{y})$ based on the expected squared error between the learner's predictions $\mathbf{y}'$ and the true function $\mathbf{y}^*$:

$$U_T(\mathbf{x}, \mathbf{y}) = -\mathbb{E}_{P_L}\left[\sum_{i=1}^{l}(y_i' - y_i^*)^2\right] \tag{6}$$

Using the bias-variance decomposition for mean-squared error (Geman et al., 1992), we can show that the teacher is sensitive both to the variance of the learner's predictions and the bias of their mean away from the truth. We can vary the relative importance of each term with the parameter λ.

$$U_T(\mathbf{x}, \mathbf{y}) = -\sum_{i=1}^{l}\Big(\underbrace{\lambda(\mathbb{E}[y_i'] - y_i^*)^2}_{\text{Bias}^2} + \underbrace{(1-\lambda)\mathbb{E}[(y_i' - \mathbb{E}[y_i'])^2]}_{\text{Variance}}\Big) \tag{7}$$

Heuristic Teaching Models

To investigate whether human teaching behavior could be explained by more simple rules, we built other candidate utility functions for teaching point sets.

Spread Out Points This utility encourages teaching points to be spread across the domain by maximizing the log-distances between all pairs of points.

$$U_{\text{spread}}(\mathbf{x}) = \sum_{i<j}\log|x_i - x_j| \tag{8}$$

Extrema This utility encourages teaching points at extrema of the function, such as local minima, maxima, or the edges of the domain.

$$U_{\text{extrema}}(\mathbf{x}) = \sum_{i=1}^{n}\log P_{\text{extrema}}(x_i \mid C, \sigma) = -\sum_{i=1}^{n}\frac{d_i^2}{2\sigma^2} \tag{9}$$

where C is the set of critical points and $d_i = \min_{c\in C}|x_i - c|$ is the distance between a teaching point x_i and its nearest critical point.

High Slope This utility encourages teaching points at high rates of change in the function, such as local minima, maxima, or the edges of the domain.

$$U_{\text{slope}}(\mathbf{x}) = \sum_{i=1}^{n}\log P_{\text{slope}}(x_i \mid \mathcal{S}, \sigma) = -\sum_{i=1}^{n}\frac{d_i^2}{2\sigma^2} \tag{10}$$

where $d_i = \min_{s\in\mathcal{S}}|x_i - s|$ is the distance from teaching point x_i to its nearest steep point, and $\mathcal{S}$ is the set of points where the absolute slope exceeds a threshold τ.

Center Bias This heuristic encourages teaching points near the center of the function, as opposed to the edges.

$$U_{\text{center}}(\mathbf{x}) = \sum_{i=1}^{n}\log P_{\text{center}}(x_i) = -\sum_{i=1}^{n}\frac{(x_i - x_{\text{center}})^2}{2\sigma^2} \tag{11}$$

where $x_{\text{center}} = \frac{x_{\min}+x_{\max}}{2}$ is the center of the domain.

Experiment

Approach

We designed a synchronous, dyadic teaching task (Figure 1) to investigate how people engage in sophisticated pedagogical reasoning when selecting examples to teach continuous functions. Rather than studying teaching with hypothetical learners, we paired up participants to interact as learners and teachers, allowing teachers to observe the consequences of their example selections and adapt over time. This paradigm also allowed us to test whether teachers' example selections are better explained by a rational model of pedagogical reasoning or by simpler heuristics. We collected choices across 20 functions (Figure 3) that vary in their compatibility with the learner model's hypothesis space (similarity, Matern, and periodic kernels). This variety of functions can probe whether teachers flexibly adjust their strategies based on function structure.

Methods

Participants Participants were $n = 82$ adults in $n = 41$ pairs (38 men, 44 women, mean age = 41) recruited from the Prolific platform in exchange for monetary compensation. They had to be at least 18 years old, reside in the United States, have completed more than 100 studies with a >95% acceptance rate, and be accessing Prolific from a desktop or laptop computer.

Procedure Participants completed the experiment from their web browsers. The experiment was coded using the online experimental design framework Smile (with underlying languages JavaScript and Vue). We custom-coded trials that allowed participants to draw points and lines on a canvas, adapted from Schulz et al. (2017).

Participants read instructions and completed a quiz about the structure of the experiment. One they passed the quiz, or took it twice and got the correct answers, they were placed into a waiting room. If another participant was available at the same time, they would be matched and randomly assigned the role of learner or teacher. In dyads, participants completed 20 trials. Each trial consisted of two rounds. In round 1, teachers would see a function and then choose three points to place to communicate that function to the learner. Then, the learner would guess a line by drawing it on the screen's canvas. In round 2, teachers would see the same function, their three points, and the learner's guess. They would then decide three more points to give to the learner. The learner would get these points, make a final guess, and then both teacher and learner would see the final guess. Both participants earned a score based on learner accuracy:

$$\text{score} = \sqrt{\frac{1}{2}\sum_{i=1}^{2}(\mathbf{y}_i - \mathbf{y}_i^*)^2} \tag{12}$$

Participants saw this score after each trial, and were issued bonus compensation on a sliding scale ($0–$6) based on the cumulative score.

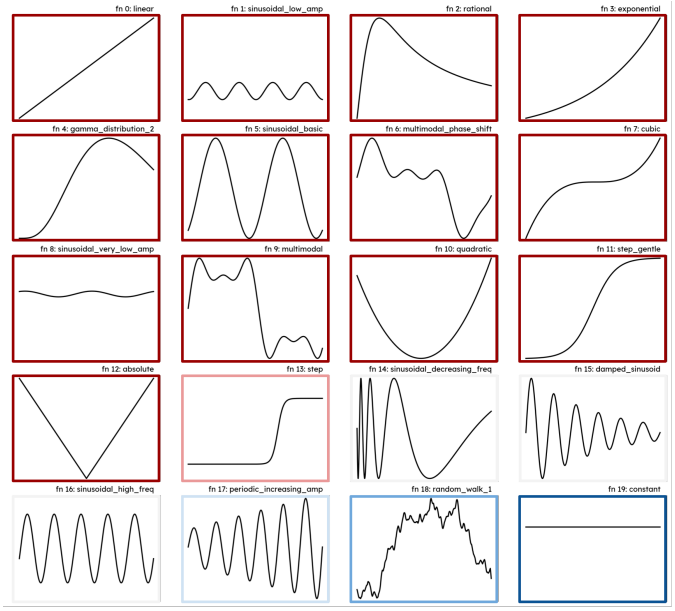

Figure 3: The 20 functions in the experiment. Functions were chosen to span across being well-described by kernels in our learner model (smooth, periodic), and being not as easily captured (periodic with changing amplitudes or frequencies, sigmoid, absolute). Outline color denotes β_{teaching} in the same scale as Figure 4.

Results

To assess whether the teaching model explains point selection beyond simpler heuristic factors, we compared nested models that scored point set utility in a likelihood ratio test. The base model predicted weights for spread, extrema, slope, and center contributions to choices, and the full model added a weight for the GP teacher utility.

Combined Models for Model Comparison To evaluate the unique influence of teaching utility on teacher choices, we fit a combined continuous choice model that assigns a likelihood to combinations of teaching points according to:

$$P(\mathbf{x}) \propto \exp\left(\sum_{h \in \mathcal{H}_{\text{full}}} \beta_h \, U_h(\mathbf{x})\right) \tag{13}$$

where $\mathcal{H}_{\text{full}} = \{\text{spread, extrema, slope, center, teaching}\}$ is the set of model components as defined by the Equations 8–11. We also define a nested model which is a special case of the model in Equation 13 in which $\beta_{\text{teaching}} = 0$ (i.e., $\mathcal{H}_{\text{lesion}} = \{\text{spread, extrema, slope, center}\}$).

We fit β in two ways: per-participant (aggregating across all functions, what mixture of strategies do participants use?) and per-function (aggregating across all participants, do some functions evoke specific teaching strategies?).

Participant-Level Modeling Mean β weights across participants were, in descending order: $\beta_{\text{spread}} = 2.15$ (SD = 1.28), $\beta_{\text{teaching}} = 0.15$ (SD = 0.73), $\beta_{\text{center}} = -0.94$ (SD = 0.54), $\beta_{\text{slope}} = -2.72$ (SD = 0.90), $\beta_{\text{extrema}} = -0.94$ (SD = 0.50). At the individual level, adding the GP teacher utility significantly improved model fit for 11 of 41 participants at $p < .05$ (mean LR statistic = 3.16, median p = .32). At the population level, teaching utility significantly improved model fit; summing individual LR statistics across participants resulted in $\chi^2(41) = 129.64$, ($p < .001$). In addition, we fit λ from Equation 7 per participant and found a mean of $\lambda = 0.9$, indicating that human teachers were more sensitive to the mean prediction (bias term) of the learner, yet still considerate of learner uncertainty.

Function-Level Modeling Mean β weights across functions were, in descending order,: $\beta_{\text{spread}} = 1.73$ (SD = 1.37), $\beta_{\text{teaching}} = 1.72$ (SD = 2.37), $\beta_{\text{extrema}} = -0.17$ (SD = 5.04), $\beta_{\text{center}} = -0.92$ (SD = 1.44), $\beta_{\text{slope}} = -3.98$ (SD = 1.58). At the single function level, adding the GP teacher utility significantly improved model fit for 13 of 20 functions at $p < .05$ (mean LR statistic = 15.41, median p = 0.005). At the aggregate level, teaching utility significantly improved model fit; summing individual LR statistics across functions resulted in $\chi^2(20) = 308.25$, ($p < .001$). Example trials are visualized in Figure 5. Individual weights for each function are visualized in Figure 4. Most functions have significant β for teaching and spread strategies; those that do not are likely explained by placing points around extremas, for functions such as 14–17.

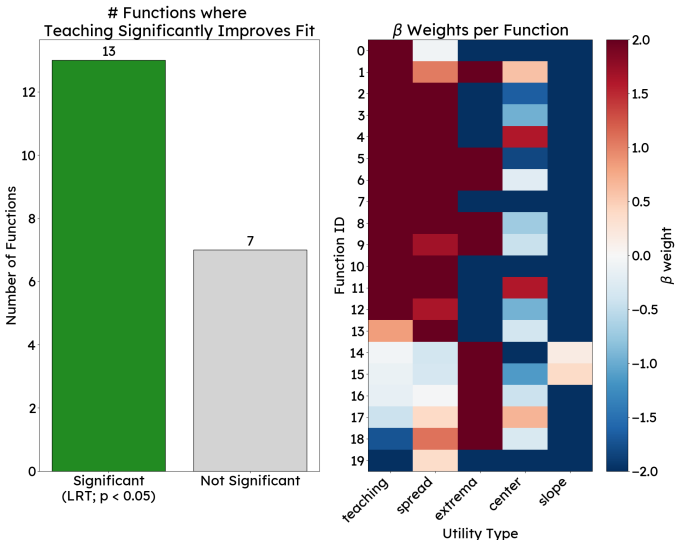

Figure 4: Left: The GP teacher utility significantly improved model fit for 13 of 20 functions at $p < .05$. Right: β weights for each function. Positive weights indicate the strategy was likely used for selecting teaching points. Negative weights indicate the strategy was actively avoided (e.g,, if center is negative for a function, participants tended to cluster points near the edges). Functions are sorted by β_{teaching} descending.

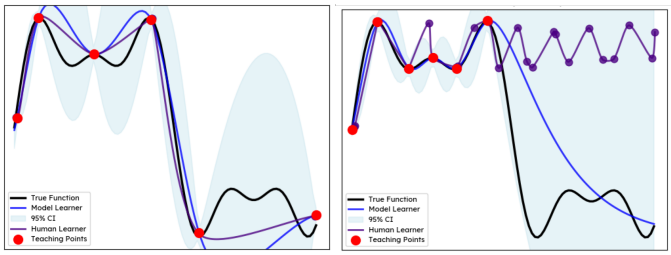

Figure 5: Teaching trials for function 9, which had high β_{teaching} and β_{extrema}. We see that the model learner (blue) is able to approximate this function and has a reasonable confidence interval. Left: Human teaching points (red) convey rapid change in the middle of the function, as well as local extrema. Both human (purple) and model learner make accurate guesses. Right: A worse set of teaching points does not communicate important elements of the function, and the human learner guesses that the points are from a periodic function.

Discussion

We generalized computational models of pedagogy to vast, theoretically infinite hypothesis spaces for both learners and teachers. Specifically, we considered a function learning paradigm where learners must infer underlying patterns from sparse examples. Teaching in this function domain presents the challenge of reasoning about what learners might believe across a large space of possibilities, and learners must extract meaningful structure from very limited data. Our approach demonstrates that leveraging Gaussian process models of teaching can scale to these complex settings while capturing patterns in human pedagogical behavior.

Limitations

Kernel selection and function coverage. Our teaching model does not significantly improve fit for every function. In particular, sinusoidal and variations on sinusoidal functions are better explained by the extrema heuristic. We hypothesize this occurs because the experimental functions were not drawn from the GP prior's hyperparameter distribution, making the learner model less well-suited to capture these specific functions. Since the learner model is a critical component of teacher utility, misspecification here can hamper teaching performance. Kernel selection remains a known challenge for GP models, though efforts to automate this process are ongoing (Abdessalem et al., 2017; Zhao et al., 2024). Future experiments could control for this by testing on functions explicitly parameterized by the GP mixture prior or employing these automated kernel search techniques.

Learner model assumptions. Our learner model represents a large set of hyperparameters across multiple kernels; while this is computationally feasible due to parallelization in our model code, it may not reflect how human learners actually represent functions. Human learners could instead maintain a limited set of kernels with hyperparameters that they refine over time. Additionally, our GP learner does not incorporate constraints that humans likely used in this task, such as the belief that the underlying functions would not exceed the boundary of the outlined canvas. Our model's mean functions and confidence intervals span all of $\mathbb{R}$. Future work could impose more human-like boundary value constraints on the GP (Gulian et al., 2022; Swiler et al., 2020) to better capture human-like priors.

Non-pragmatic learners. While our framework is inspired by Theory of Mind and rational speech acts, our current learner does not leverage the pedagogical context. Prior work emphasizes that learners in teaching settings reason pragmatically about why a teacher selected specific examples (Shafto et al., 2014). Extending our model to include pragmatic learners would require modifying the learner posterior to weight hypotheses by the probability that a teacher would select given examples under each hypothesis. This pragmatic extension could enable recursive reasoning between teacher and learner and bring learner inferences closer to human performance.

Non-responsive teachers. In our experiments, teachers see a learner's first guess before providing final teaching points, creating an opportunity for generating examples that directly respond to or incorporate the learner's estimate. However, our current model assumes teachers select points without inferring the learner's actual beliefs from their guess. A more sophisticated teacher would *infer* the learner's mean function and kernel from their estimate, then select corrective examples. We hypothesize that responsive teachers would show points highly unlikely under the learner's current (mistaken) representation, signaling that they should update their hyperparameter distribution (de la Torre, 2009; Griffiths & Tenenbaum, 2005; Ross & Andreas, 2024). Modeling this feedback loop is an important direction for future work.

Future Directions

Future work could investigate function teaching beyond visual functions, in different modalities (e.g., motor learning, auditory learning, abstract causal reasoning). Additionally, we could generalize our formalism to multiple output dimensions to model teaching in complex embedding spaces with just a few examples. Our experimental paradigm only considers teaching via examples, but we could also consider linguistic guidance from teachers. This is of interest because abstract descriptions of underlying rules can sometimes communicate a concept more clearly than demonstrative examples (Sumers et al., 2023). In fact, related prior work investigated how people describe functions with language, and found that certain types of functions are more easy to describe with language than others (Schulz et al., 2020). Ultimately, we hope that this paradigm can provide insight into computationally tractable methods for teaching and reasoning about complex, vast domains. By demonstrating that rational pedagogy can scale to continuous hypothesis spaces, we move towards understanding how humans effectively communicate knowledge with sparse, selective examples in the rich conceptual spaces they navigate daily.

References

Abdessalem, A. B., Dervilis, N., Wagg, D. J., & Worden, K. (2017). Automatic kernel selection for gaussian processes regression with approximate bayesian computation and sequential monte carlo. *Frontiers in Built Environment*, *3*, 52.

Attneave, F. (1950). Dimensions of similarity. *The American journal of psychology*, *63*(4), 516–556.

Bridgers, S., Jara-Ettinger, J., & Gweon, H. (2020). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, *4*(22), 144–152. https://doi.org/10.1038/s41562-019-0748-6

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*(2), i–144. https://doi.org/10.1002/j.2333-8504.1963.tb00958.x

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., & Dragan, A. (2019). On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, *32*.

Collins, K. M., Sucholutsky, I., Bhatt, U., Chandra, K., Wong, L., Lee, M., Zhang, C. E., Zhi-Xuan, T., Ho, M., Mansinghka, V., et al. (2024). Building machines that learn and think with people. *Nature human behaviour*, *8*(10), 1851–1863.

de la Torre, J. (2009). Dina model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*(1), 115–130. https://doi.org/10.3102/1076998607309474

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1–58.

Goodfellow, I., Bengio, Y., & Courville, A. (2016, November). *Deep learning* [Google-Books-ID: omivDQAAQBAJ]. MIT Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154. https://doi.org/10.1080/03640210701802071

Griffiths, T., Lucas, C., Williams, J., & Kalish, M. (2008). Modeling human function learning with gaussian processes. *Advances in neural information processing systems*, *21*.

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

Gulian, M., Frankel, A., & Swiler, L. (2022). Gaussian process regression constrained by boundary value problems. *Computer Methods in Applied Mechanics and Engineering*, *388*, 114117.

Gweon, H. (2021). Inferential social learning: Cognitive foundations of human social learning and teaching. *Trends in Cognitive Sciences*, *25*(10), 896–910. https://doi.org/10.1016/j.tics.2021.07.008

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *17*(5), 811–836. https://doi.org/10.1037//0278-7393.17.5.811

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215. https://doi.org/10.3758/s13423-015-0808-5

Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2011). Faster teaching by pomdp planning. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education* (pp. 280–287). Springer. https://doi.org/10.1007/978-3-642-21869-9_37

Rafferty, A. N., LaMar, M. M., & Griffiths, T. L. (2015). Inferring learners' knowledge from their actions. *Cognitive Science*, *39*(3), 584–618. https://doi.org/10.1111/cogs.12157

Rasmussen, C. E. (2003). Gaussian processes in machine learning. In *Summer school on machine learning* (pp. 63–71). Springer.

Rasmussen, C. E., Williams, C. K., et al. (2006). *Gaussian processes for machine learning* (Vol. 1). Springer.

Ross, A., & Andreas, J. (2024, May). Toward In-Context Teaching: Adapting Examples to Students' Misconceptions [arXiv:2405.04495 [cs]]. https://doi.org/10.48550/arXiv.2405.04495

Schulz, E., Quiroga, F., & Gershman, S. J. (2020). Communicating compositional patterns. *Open Mind*, *4*, 25–39.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79. https://doi.org/10.1016/j.cogpsych.2017.11.002

Shafto, P., Goodman, N. D., & Frank, M. C. (2012). Learning from others: The consequences of psychological reasoning for human learning. *Perspectives on Psychological Science*, *7*(4), 341–351. https://doi.org/10.1177/1745691612448481

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. https://doi.org/10.1016/j.cogpsych.2013.12.004

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., et al. (2023). Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*.

Sumers, T. R., Ho, M. K., Hawkins, R. D., & Griffiths, T. L. (2023). Show or tell? exploring when (and why) teaching with language outperforms demonstration. *Cognition*, *232*, 105326. https://doi.org/10.1016/j.cognition.2022.105326

Sumers, T. R., Ho, M. K., Hawkins, R. D., Narasimhan, K., & Griffiths, T. L. (2021). Learning rewards from linguistic feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(7), 6002–6010. https://doi.org/10.1609/aaai.v35i7.16749

Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C., & Jakeman, J. D. (2020). A survey of constrained gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, *1*(2).

Williams, C., & Rasmussen, C. (1995). Gaussian processes for regression. *Advances in neural information processing systems*, *8*.

Zhao, S., Lu, J., Yang, J., Chow, E., & Xi, Y. (2024). Efficient two-stage gaussian process regression via automatic kernel search and subsampling. https://arxiv.org/abs/2405.13785