Data Availability and Function Extrapolation

Pablo León Villagrá Irina Preda Christopher G. Lucas
Informatics Forum, 10 Crichton Street, EH8 9AB,
Edinburgh, United Kingdom

Abstract

In *function learning* experiments, where participants learn relationships from sequentially-presented examples, people show a strong tacit expectation that most relationships are linear, and struggle to learn and extrapolate from non-linear relationships. In contrast, experiments with similar tasks where data are presented simultaneously – typically using scatter plots – have shown that human learners can discover and extrapolate from complex non-linear trends. Do people have different expectations in these task types, or can the results be attributed to effects of memory and data availability? In a direct comparison of both paradigms, we found that differences between task types can be attributed to data availability. We show that a simple memory-limited Bayesian model is consistent with human extrapolations for linear data for both high and low data availability. However, our model underestimates the participants' ability to infer non-monotonic functions, especially when data is sparse. This suggest that people track higher-order properties of functions when learning and generalizing. **Keywords:** function learning, function estimation, resource rationality

Introduction

Many everyday situations require us to make predictions with very limited information. Consider a situation in which you are in a holiday flat and want to use the heater. You haven't used it before but you will have a general idea of how long it will take to heat up. Furthermore, it is natural to assume the relationship between setting and resulting changes in temperature – the temperature should change in continuous fashion and increase until it reaches the selected temperature. In addition, given few, seemingly unrelated bits of information, like for example the state of other appliances, you can generalize and update your beliefs.

Forming generalizations requires acquiring a representation of how known features relate to unknown targets and contrasting this relation with potential alternatives. Humans exhibit a remarkable ability to perform these generalizations and much research in the cognitive sciences and artificial intelligence has centered on understanding or reproducing these phenomena. One of the most prominent fields of generalization research, categorization, has focused on situations where the known characteristics are assumed features of some entity and the target is the category of the entity (a categorical label). Function learning takes a more general perspective on the generalization target and allows for continuous values. Both categorization and function learning research share fundamental questions about how humans acquire productive representations of these relationships and what sort of representations allow generalization. For instance, what kind of relationships can humans learn and generalize from, and how are extrapolations reflective of particular human biases? In *function learning* experiments, where experimental participants learn relationships from sequentially-presented points, people show a strong bias toward inferring linear functions (Brehmer, 1976; DeLosh, Busemeyer, & McDaniel, 1997; Kalish, Lewandowsky, & Kruschke, 2004). They learn linear relationships more quickly (Brehmer, 1974; Byun, 1995) and have difficulty making non-linear and especially non-monotonic extrapolations (Brehmer et al., 1985; Byun, 1995; Bott & Heit, 2004; Kalish, 2013). This has led to the development of models that attach a special representational status to linear relationships (DeLosh et al., 1997; Kalish et al., 2004), or assume that people have a strong inductive bias favoring linearity (Brehmer et al., 1985). In contrast, when data are presented simultaneously, usually as scatter plots (which we will call *function estimation* tasks) human learners can discover and extrapolate from complex non-linear trends (Wilson et al., 2015; Schulz et al., 2017; Lucas et al., 2015; Little & Shiffrin, 2009). How do we reconcile these experimental results?

One possibility is that people respond to these presentation modes in different ways, for reasons that may be perceptual, cognitively innate, or experience-dependent. An alternative possibility is that the same inductive biases and cognitive processes support both *function learning* and *function estimation*, and differences between these tasks can be attributed to differences in their memory demands.

In *function learning* experiments participants have to maintain learned data in memory and update and evaluate the appropriateness of a representation against alternatives, whereas *function estimation* allows an effortless recall of the data. We hypothesize that, due to these differences, many participants in *function learning* tasks only maintain sparse representations of the data. Given that only a subset of the data is maintained, extrapolations will resemble inductive biases in the absence of data. In contrast, having all data visually available, as in *function estimation*, allows to counteract inductive biases and facilitates extrapolations resembling richer functions.

Experiment

We set up an experiment to contrast extrapolations in *function learning* and *function estimation*. To distinguish the contribution of presentation from memory requirements imposed by the experiment, we introduced a new experimental condition that shared presentational-, but not memory-related characteristics with *function estimation*. In this new condition data was presented as scatter plots, but data points disappeared from display immediately after submission. Since the condition exhibits similar characteristics to classical *function learning* tasks we predicted that extrapolations should more closely resemble *function learning* conditions, as participants will have to rely on recollection of the presented data for their extrapolations. We will refer to the scatter plot condition presenting the full data as Scatter$^+$ and the new condition as Scatter$^-$ throughout this paper. We will refer to the traditional *function learning* conditions as Bar.

Participants

We recruited 322 participants via Amazon's Mechanical Turk service. Participants received 0.4$ for participation and took an average of 8 minutes to complete the experiment. Participants were randomly assigned to one of the 9 conditions $\{\mathrm{f}_{\mathrm{lin}}$, f_{x^2}, $\mathrm{f}_{\mathrm{cos}}$ $\} \times \{$Scatter$^+$, Scatter$^-$, Bar $\}$, as described below.

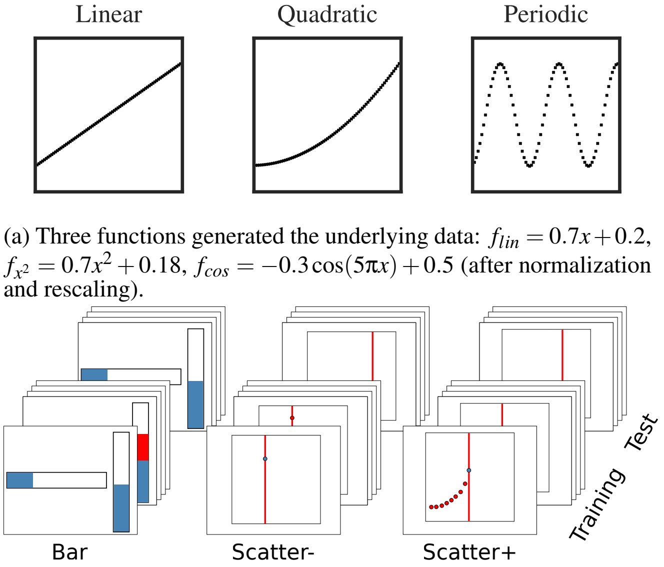
Material

The data presented in the experiment was generated by one of three functions: linear ($\mathrm{f}_{\mathrm{lin}}$), quadratic ($\mathrm{f}_{x^2}$) or periodic ($\mathrm{f}_{\mathrm{cos}}$). These functions and their parametrization were designed such as to allow for informative error patterns resulting from human inductive biases. Since previous research reported strong biases for linear functions with 0 intercept and $\frac{1}{1}$ slope (we will refer to this function as $f(x) = x$), we selected a shallower positive slope. The quadratic function was relatively flat in the training block to test if participants would revert to linearity or choose non-linear alternatives. Finally, a periodic function, was used to evaluate if participants were able to extrapolate in non-monotonic fashion. To allow space for extrapolation beyond the function ranges we normalized the data to span $(0,1)$ in both x, y and then rescaled and centered such as to span $\frac{1}{2}$ of the y-axis. For the full set of materials after transformation, see Figure 1a.

Procedure

Participants were instructed that they would be presented with data and that given their understanding of the relationship in the data they had to predict new values. Then participants proceeded to a block of training trials (the training block), which provided participants with feedback about the true function.

Training Block In the Scatter$^+$ and Scatter$^-$ conditions the current test value (x) was marked with a red line spanning the whole vertical range. Participants were prompted to select a y value by clicking on the line. Once selected, the input value was highlighted with a blue point. Selected points could be updated by reselecting a y value. The selected values were submitted by pressing the space key. In the Bar condition current x values were presented as the width of a bar on the left of the screen and participants selected values by selecting the height of a bar on the right. As in the Scatter$^+$ and Scatter$^-$ conditions, participants could readjust these values. In all conditions x values were presented sequentially in ascending order. If the selected y value was within the error margin (± 0.05 of the true y), the true value was shown for 600ms in red. Afterwards, a message indicated that the choice was correct and the remaining number of trials was shown. If the selected value was not inside the margin the message indicated an unsuccessful submission. Then the selected value was removed and participants had to resubmit. After erroneous submissions the true y was displayed as a red bar (Bar) or a red dot (Scatter$^+$, Scatter$^-$). Participants had to resubmit values until a admissible y was chosen. Participants received 40 points in total during the training block.



(a) Three functions generated the underlying data: $f_{lin} = 0.7x + 0.2$, $f_{x^2} = 0.7x^2 + 0.18$, $f_{cos} = -0.3\cos(5\pi x) + 0.5$ (after normalization and rescaling).

(b) Procedure for Bar, Scatter$^-$ and Scatter$^+$ conditions.

Figure 1: Participants were randomly assigned to one of the 9 experimental conditions. All participants performed a training block consisting of 40 value pairs with feedback followed by a test block of 40 extrapolations without feedback.

Test Block The test block followed the same procedure as the training block, but no feedback was provided. After submitting 40 values in the test block, participants concluded the experiment by submitting an optional short survey. For the full procedure see Figure 1b.

Results

Functions and Presentation Form

Consistent with previous findings, mean absolute error (MAE) in the test block was largest for $\mathrm{f}_{\mathrm{cos}}$, MAE = 0.24, SD = 0.1, n = 108. Errors for f_{x^2} and $\mathrm{f}_{\mathrm{lin}}$ were small, with $\mathrm{f}_{\mathrm{lin}}$ exhibiting the smallest error, $\mathrm{MAE}_{x^2} = 0.14$, $SD_{x^2} = 0.09$, n_{x^2} = 106, $\mathrm{MAE}_{lin(x)} = 0.11$, $SD_{lin(x)} = 0.1$, $n_{lin(x)} = 108$.

The errors in the presentation conditions were compatible with our hypothesis, with Scatter$^+$ lowest, MAE = 0.1, SD = 0.1, n = 106 and Scatter$^-$ and Bar at similar, higher levels, $\mathrm{MAE}_{\mathrm{Bar}} = 0.19$, $SD_{\mathrm{Bar}} = 0.12$, $n_{\mathrm{Bar}} = 110$, $\mathrm{MAE}_{\mathrm{Scatter^-}} = 0.19$, $SD_{\mathrm{Scatter^-}} = 0.11$, $n_{\mathrm{Scatter^-}} = 106$. For all errors in the subgroups of function and presentation conditions see Figure 2.

Data Availability and Presentation

To assess the effect of data availability (DA, a binary variable denoting if the condition was Scatter$^+$, or either Scatter$^-$

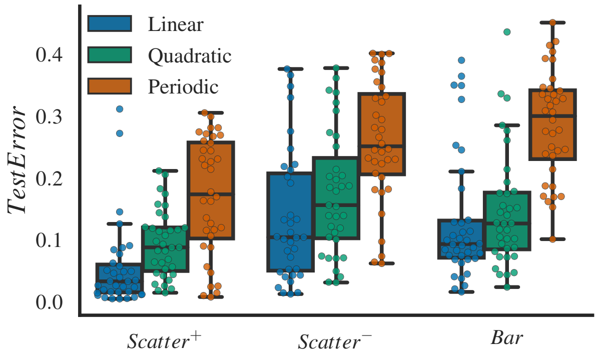


Figure 2: In all presentation conditions participants exhibited the lowest errors for linear, followed by quadratic and periodic functions. Boxplots display first, second (median) and third quartiles. Whiskers show the ±1.5 interquartile range (IQR). Each point represents the MAE of one participant.

or Bar) and function (f_{lin}, f_{x^2}, f_{cos}) on errors, while controlling for the effect of presentation (*Scatter* denoting if the presentation condition was either Scatter$^+$ or Scatter$^-$, or Bar), we fitted a generalized linear model (GLM): $Y_{MAE} \sim \beta_0 + \beta_f * (\beta_{Scatter} + \beta_{DA})$. The GLM was specified as a Gaussian with identity link function and allowed for interactions between *Scatter* and function as well as *DA* and function.

Table 1: Results of the GLM model assessing if function type (f_{lin}, f_{x^2}, f_{cos}), presentation (*Scatter*), or data availability (*DA*) were predictive of MAE in the test block. The f_{cos} condition had a significant positive effect on MAE. In addition, having all data available (*DA*, corresponding to condition Scatter$^+$) had a significant, small negative effect.

	β	SE	z	$P > \|z\|$	95%CI
β_0	0.13	0.02	8.78	$p < 0.001$	$0.1, 0.16$
$f_{cos(x)}$	0.15	0.02	7.28	$p < 0.001$	$0.11, 0.2$
f_{x^2}	0.02	0.02	0.75	0.45	$-0.03, 0.06$
Scatter	$\beta < 0.01$	0.02	0.22	0.82	$-0.04, 0.05$
DA	-0.08	0.02	-3.75	$p < 0.001$	$-0.13, -0.04$
Scatter$f_{cos(x)}$	-0.03	0.03	-1.02	0.3	$-0.01, 0.03$
Scatterf_{x^2}	0.02	0.03	0.75	0.46	$-0.04, 0.08$
DA$f_{cos(x)}$	-0.01	0.03	-0.24	0.8	$-0.07, 0.05$
DA.f_{x^2}	$\beta < 0.01$	0.03	0.04	0.97	$-0.06, 0.06$

In concordance with previous findings, f_{cos} had a significant positive effect on error. As expected by our hypothesis, data availability (*DA*) had a significant small effect on error, but presentation (*Scatter*) was not significant. No other main effect and none of the interaction terms had a significant effect. For the full GLM results, see Table 1. For all extrapolations performed by the participants, see Figure 3.

Learning a Function or Minimizing Error

The results are consistent with our hypothesis that differences in *errors* are attributable to differences in data availability. However, a stronger test of our hypothesis lies in the *patterns* of extrapolations that people make. Do these patterns differ systematically between presentation conditions, or are differences explainable in terms of condition-independent biases? In the final section we will explore how differences in availability, imposed by our experimental design reflect in the participants extrapolations. To analyze these extrapolations we compared human extrapolations to two Bayesian models, one exhibiting low available data and one considering all available data.

Modeling Function Extrapolations

The computational problem faced in extrapolation tasks consists in determining new values y_{n+1} for test values x_{n+1}, conditional on previously learned $\mathbf{x}_n, \mathbf{y}_n$ and a prior belief $p(f)$ over possible functions. We will adopt a Gaussian process perspective on regression, an approach that has been applied successfully in previous function learning research (Lucas et al., 2015; Schulz et al., 2017).

A Gaussian process specifies a distribution over functions $f(x) \sim GP(\mu, k)$, where $\mu(x) = E[f(x)]$ and k is the covariance kernel $k(x, x') = cov(f(x), f(x'))$. The kernel specifies how much values of x' depend on the other values x and specifies a similarity measure over x. We assume that two sets of priors can capture participant extrapolations in our study — a prior over kernel types describing the space of possible functions $f_i \sim \mathcal{F}$, and a prior for individual kernel parameters θ_{f_i}.

Human Function Priors

To specify a plausible prior over functions $\mathcal{F}$ we closely followed Lucas et al. (2015). We used the same prior probabilities for functions $\mathcal{F}$, favoring $f(x) = x$ (*Linear*$^+$) over negative linear functions (*Linear*$^-$), and linear functions over other monotonic functions (*RBF*, the radial basis function kernel). Since our experiment included periodic data that we did not want to exclude a priori, we added a periodic kernel (*Periodic*) with good coverage over the range of x, y. We chose a low prior weight for the periodic to account for the difficulty in learning non-monotonic functions (Bott & Heit, 2004; Kalish, 2013). For a full list of parameter priors θ, see Table 2, for samples of the prior functions, see Figure 4.

With the priors $\mathcal{F}$ and θ we can express the task faced by our participants in general terms:

$$p(y_{n+1}|x_{n+1}\mathbf{x}_n, \mathbf{y}_n, f) = \int_f p(y_{n+1}|x_{n+1}, y, f) p(f|\mathbf{x}, \mathbf{y}) df \tag{1}$$

Given appropriate priors and Equation 1 a variety of human inductive biases can be accounted for, from strong biases for $f(x) = x$, to results in iterated learning experiments (Lucas et al., 2015).

However, this model assumes that all previously encountered data, $\mathbf{x}, \mathbf{y}$, are equally available and inform posterior inference. In some function learning experiments, where participants repeat training until they achieve a very low error rate, these assumptions may be appropriate. In other contexts, including many sequential function learning problems in the natural world, it is less plausible.

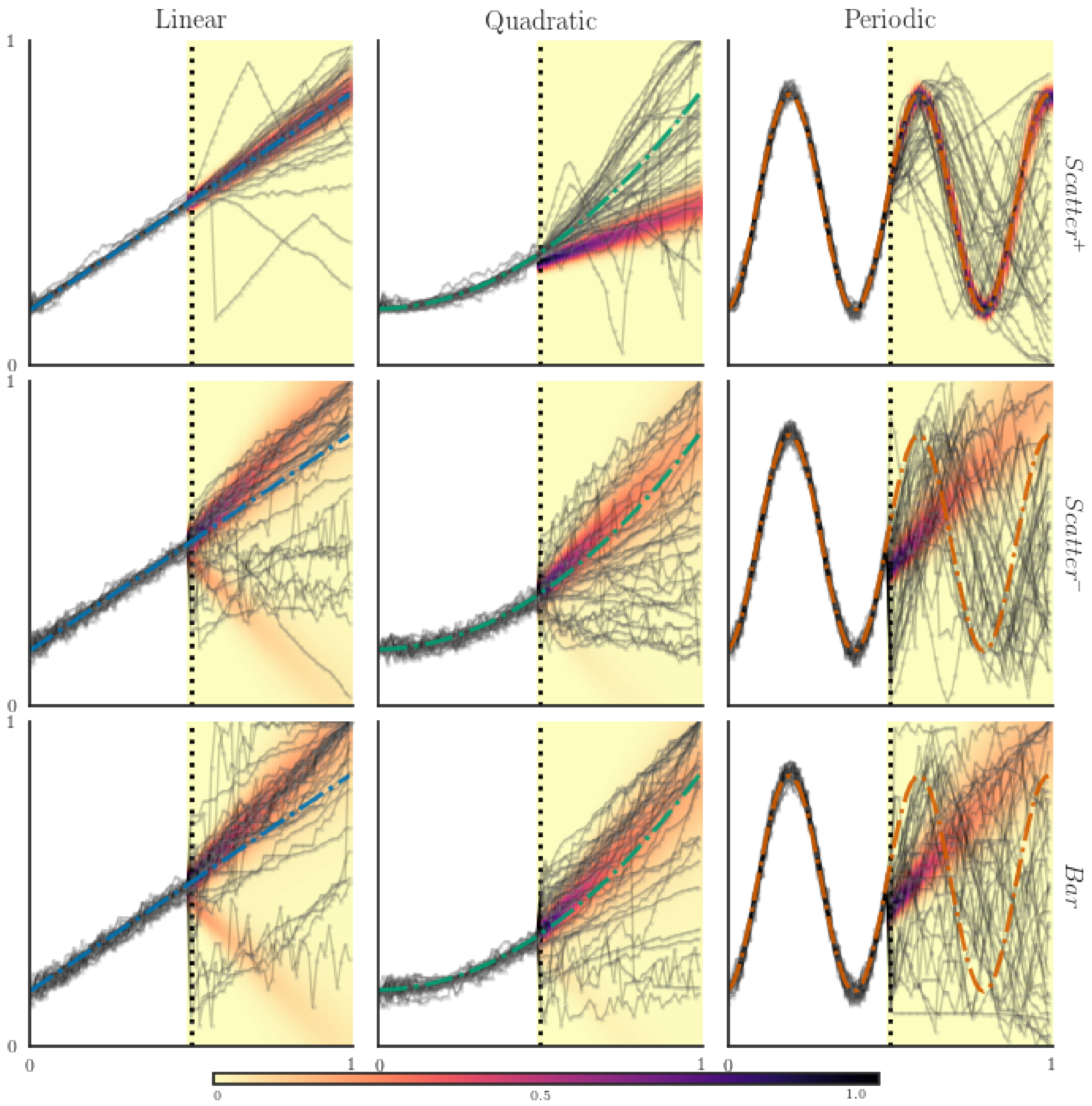


Figure 3: All participant extrapolations (gray lines) in the 9 experimental conditions. Submissions within the admissible error for the training block are displayed on the left-hand side of the dotted vertical line. Extrapolations for the test block are displayed on the right-hand side. Function conditions are presented by column, presentation conditions by row. As background color the posterior densities of the models described in the model section, darker colors correspond to higher posterior density. According to our hypothesis, participants in the Scatter$^+$ condition kept a full representation of the training data available, corresponding to the full model (top row). In the two bottom rows the model density is conditional on only the last 5 training points, corresponding to our sparse model.

Modeling Data Availability

We contrasted the predictions of a model trained on the full dataset (the 40 training points) with a model that had only a sparse set of data available. As a first approximation of the effect of data availability we assumed that only the last $k = 5$ points in the training block were available in the Bar and Scatter$^-$ conditions. While the amount of data underlying participants' extrapolations might differ systematically, our analysis is not particularly sensitive to the size of the subset. In general, larger subsets will emphasize the training data,

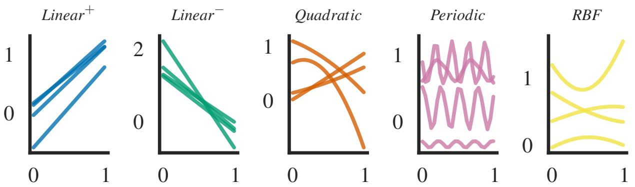


Figure 4: Four samples for each of the functions constituting $\mathcal{F}$. $\mathcal{F}$ consisted of a linear kernel biased towards $f(x) = x$, a negative linear, a quadratic, a periodic and a *RBF* kernel. All kernels had additional intercept terms. The distribution over functions $\mathcal{F}$ was chosen to closely match Lucas et al. (2015) and was proportional to $8, 1, 0.1, 0.01, 0.01$.

Table 2: Priors used to specify the two models. All models had a fixed noise variance of 0.0025, that matched the admissible error in the test set. The lengthscale of the RBF kernel was θ_l, while θ_π specified the period of the periodic kernel. The curvature of the quadratic kernel was specified by θ_c.

	σ^2	Intercept	Slope	θ_c	θ_l	θ_π
Linear$^+$	$Exp(\frac{1}{6})$	$N(0, \frac{1}{2})$	$N(1, \frac{1}{10})$	–	–	–
Linear$^-$	$Exp(\frac{1}{6})$	$N(1, \frac{1}{2})$	$N(\text{-}1, \frac{1}{10})$	–	–	–
Quadratic	$Exp(\frac{1}{6})$	$N(\frac{1}{2}, 1)$	$N(0, 1)$	$N(0, 2)$	–	–
Periodic	$Exp(\frac{1}{6})$	$N(\frac{1}{2}, 1)$	–	–	$N(1, \frac{1}{4})$	$N(\frac{1}{2}, \frac{1}{4})$
RBF	$Exp(\frac{1}{6})$	$N(\frac{1}{2}, 1)$	–	–	$N(1, \frac{1}{4})$	–

while smaller sets will result in posteriors emphasizing prior inductive biases, since the likelihood of the data plays a diminished role. For the posterior probability for functions for both models see Figure 5.

We compared both models in terms of their ability to account for characteristic biases in human function learning as well as differences between the extrapolations for Scatter$^-$ and Bar and Scatter$^-$. To evaluate our models, we classified participants' extrapolations in the test block as either belonging to full or sparse experimental conditions according to the likelihood of the models (trained on the training block). Then we contrasted this classification with the true experimental condition. For confusion matrices for this classification procedure see Figure 6. For examples of the classified extrapolations see Figure 7.

Model Results

Both sparse and full models captured the strong inductive biases for positive linear functions. Furthermore, our sparse model predicted the strong inductive bias for $f(x) = x$ in Scatter$^-$ and Bar conditions, aligning well with the participants' data (see Figure 3).

For f_{x^2} both full and sparse models reflected the strong prior for positive linearity. As a result the full model did not capture the extrapolations of participants in Scatter$^+$. While the model extrapolated in linear fashion from the available data, participants performed steeper, quadratic-like extrapolations (see Figure 3). The sparse model was more predictive of the participants extrapolations in Scatter$^-$ conditions, extrapolating in steep linear fashion. For $\mathrm{f}_{\cos}$ the sparse model did not capture the participants extrapolations well (see Figure 6). While the model did favor positive linearity and extrapolated accordingly, many participants exhibited non-monotonic, high variance extrapolations (see Figure 3). In contrast, the full model captured the highly periodic extrapolations in the Scatter$^+$ condition and closely resembled human extrapolations.

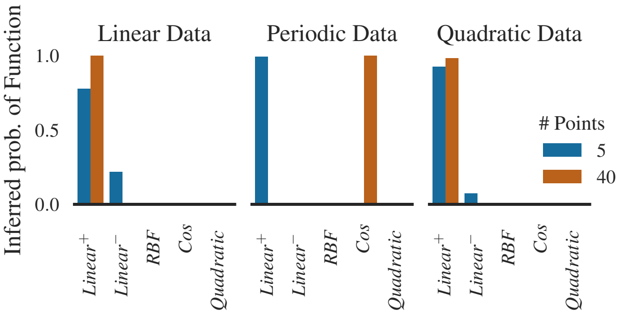


Figure 5: The inferred posterior function probability for each function condition for both sparse (5 points) and full (40) models. The full model assigned high probability to the true underlying function for linear and periodic data. For quadratic data it favored linear functions, reflecting the prior and the seemingly linear training data. The sparse model generally reflected the prior.

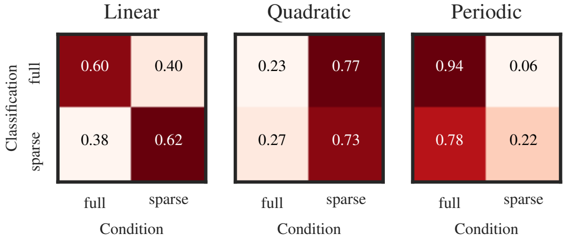


Figure 6: We contrasted our classification with the true experimental conditions. Our classification captured the effect of data availability for $\mathrm{f}_{\mathrm{lin}}$. However, it exhibited systematic misclassification for f_{x^2} and $\mathrm{f}_{\cos}$. In f_{x^2} we were able to classify participants as belonging to Scatter$^-$ or Bar, but failed to recognize Scatter$^+$. In $\mathrm{f}_{\cos}$ our procedure misclassified participants in the sparse conditions, but captured extrapolations in the Scatter$^+$ condition.

Discussion

We hypothesized that differences between *function learning* and *function estimation* experiments can be attributed to participants having direct access to all data points in the latter. More precisely, we sought to test the idea that the same inductive biases are at work in both settings, but that the reduced access to data in function learning designs causes these

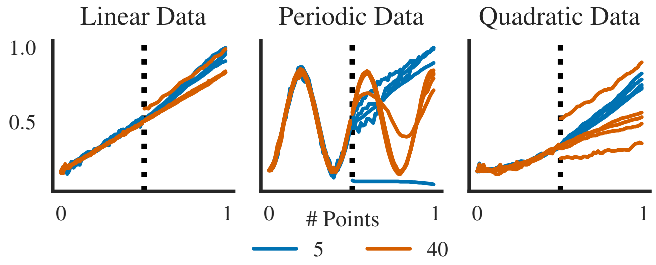


Figure 7: Five extrapolations with the highest categorization scores in each function condition. We categorized participants' extrapolations by contrasting the likelihood of our full and sparse models.

biases to play a stronger role in shaping participants' extrapolations. As we anticipated, participants' behavior in both Scatter$^-$ and Bar was almost indistinguishable, demonstrating the same qualitative patterns, which were clearly different from those in the Scatter$^+$ conditions.

However, we found mixed support for the more detailed hypotheses reflected in our Bayesian model. Behavior in all linear function conditions was as we predicted, with participants in Scatter$^-$ and Bar conditions tending to extrapolate according to the $f(x) = x$ function that past research has shown in favored a priori, rather than the true function.

In the quadratic conditions, our model captured participants' behavior in the Scatter$^-$, but not in Scatter$^-$ and Bar conditions, where participants were *more* likely than the model to infer a non-linear relationship. There are many possible explanations, one of which is that our simplistic assumptions about participants' memory failed to capture the loss of precision in the locations of points.

Perhaps the most interesting deviation between the model's predictions and participants' judgments is in the $\mathrm{f}_{\mathrm{cos}}$ conditions. Contra the model's predictions – as well as our expectations – individual participants were quick to infer non-monotonic functions even in the Scatter$^-$ and Bar conditions. This also admits several explanations, but one intriguing possibility is that people are better at tracking high-level, qualitative properties of functional relationships than the details of those relationships' parameterizations.

If higher-level properties allow for non-monotonic extrapolations, how do humans acquire these representations and in which situations do they prove beneficial? One could make an argument for cognitive economy – coarse-grained representations and extrapolations might serve a learner's goals well enough, while maintaining detailed task-specific representations is an intractable or sub-optimal policy for a resource-limited agent. We are currently exploring the relevance of resource-efficient non-parametric models to human behavior in these tasks, where representational complexity scales with an agent's goals and the complexity of the task (for a related result in categorization, see Fischer and Holt (2017)). Adding a computational-level perspective to these questions allows us to characterize these interrelations more precisely and can highlight general similarities between fields like categorization and function learning (Lucas et al., 2015; Jäkel, Schölkopf, & Wichmann, 2008).

References

Bott, L., & Heit, E. (2004). Nonmonotonic extrapolation in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 38.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Brehmer, B. (1976). Learning complex rules in probabilistic inference tasks. *Scandinavian Journal of Psychology*, *17*(1), 309–312.

Brehmer, B., Alm, H., & Warg, L.-E. (1985). Learning and hypothesis testing in probabilistic inference tasks. *Scandinavian journal of psychology*, *26*(1), 305–313.

Byun, E. (1995). *Interaction between prior knowledge and type of nonlinear relationship on function learning* (Unpublished doctoral dissertation). Purdue University.

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.

Fischer, H., & Holt, D. V. (2017). When high working memory capacity is and is not beneficial for predicting nonlinear processes. *Memory & cognition*, *45*(3), 404–412.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008, Apr 01). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256–271.

Kalish, M. L. (2013). Learning and extrapolating a periodic function. *Memory & Cognition*, *41*(6), 886–896.

Kalish, M. L., Lewandowsky, S., & Kruschke, J. K. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*(4), 1072.

Little, D. R., & Shiffrin, R. (2009). Simplicity bias in the estimation of causal functions. In *Proceedings of the cognitive science society* (Vol. 31).

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.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*.

Wilson, A. G., Dann, C., Lucas, C., & Xing, E. P. (2015). The human kernel. In *Advances in neural information processing systems* (pp. 2854–2862).