Maximum Entropy Function Learning

Simon Segert (ssegert@princeton.edu)
Princeton Neuroscience Institute
Princeton,NJ

Jonathan Cohen (jdc@princeton.edu)
Princeton Neuroscience Institute
Princeton,NJ

bstract

Understanding how people generalize and extrapolate from limited amounts of data remains an outstanding challenge. We study this question in the domain of scalar function learning, and propose a simple model based on the Principle of Maximum Entropy (Jaynes, 1957). Through computational modeling, we demonstrate that the theory makes two specific predictions about peoples' extrapolation judgments, that we validate through experiments. Moreover, we show that existing Gaussian Process models of function learning cannot account for these effects.

Keywords: Learning; Pattern recognition; Computational Modeling

Introduction

One of the most impressive aspects of human intelligence is the ability to detect and extrapolate a variety of abstract patterns that commonly occur in the world. Here, for tractability, we focus on patterns that can be expressed using one-dimensional functions, the class of which is still sufficiently rich to encompass many real-world tasks, such as deciding whether to invest in a certain stock, or predicting how hard to hit a golf ball so that it will travel a certain distance. Work on *function learning* (McDaniel & Busemeyer, 2005; Delosh, Busemeyer, & McDaniel, 1997; Bott & Heit, 2004) has catalogued the form of several inductive biases that people employ in this setting, including a preference for positive linear forms (Kwantes & Neal, 2006) or compositional construction from a small number of simple forms ("atoms;" Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017). It is an outstanding open question whether this collection of biases is sufficient to fully characterize peoples' extrapolation judgements of scalar functions.

Most current models of function extrapolation cast it as a Bayesian inference problem (Lucas, Grrifiths, Williams, & Kalish, 2015; Schulz et al., 2017; Wilson, Dann, Lucas, & Xing, 2015), the solution to which can be expressed using the mathematical formalism of Gaussian Processes (GPs) (Rasmussen & Williams, 2006). In contrast, we propose a novel, basic inductive bias in the context of function extrapolation, the form of which is based on the Principle of Maximum Entropy ("MaxEnt;" Jaynes, 1957). Put simply, we posit that peoples' extrapolations arise as samples from the "maximally indeterminate" distribution that is consistent with certain observed structural features of the function. This proposal is supported by the key role that the MaxEnt principle has played in other contexts within cognitive science. For example, an early such application was the work of Myung (Myung, 1994; Myung & Shepard, 1996), showing that certain classical categorization models could be seen as special cases of the MaxEnt principle. The principle has also been used to derive the functional form of psychoeconomic weighting functions (Bhui & Gershman, 2018), and to provide a normative account of many perceptual capacity limitations (Frankland, Webb, & Cohen, 2021).

Here, we leverage a classic theorem of Burg (1967) to show that MaxEnt functional extrapolation takes a relatively simple and psychologically plausible form. To test this hypothesis, we derive two specific behavioral predictions from it, that we term *roughness calibration* and *typicality preference*. In a pair of experiments, we test whether people show these effects when making extrapolation judgements, and contrast these with predictions of GP-based models in our experimental setting. We find that the data are consistent with predictions of the MaxEnt model and not existing GP-based models. In light of these results, we discuss how the MaxEnt model and GP model may potentially complement each other in a Resource Rational analysis of function learning (Lieder & Griffiths, 2020).

Maximum Entropy extrapolation

In typical function extrapolation experiments (McDaniel & Busemeyer, 2005; Delosh et al., 1997; Schulz et al., 2017), the experimenter chooses some function $f : \mathbb{R} \to \mathbb{R}$, and the participant is shown the values of f in a bounded region of the domain. The participant is then tasked with inferring the values of the function at points that lie outside of the region. For simplicity, we make the additional assumption that the points are equally spaced along the x axis; this permits rescaling the domain, if necessary, so that the points lie on the positive integers. Thus, the participant is shown the values $f\ 1), f\ 2), \quad , f\ N)$, and then is tasked with inferring the values $f\ N+1), \quad , f\ N+M)$.

This may be naturally cast as a problem of probabalistic inference, in which the target of the inference is the distribution of future values of f, conditional on the observed ones. In general, MaxEnt posits that the preferred solution to such an inference problem takes the form that requires the least amount of additional information to satisfy the constraints imposed by the data, among all those that are consistent with those constraints (Jaynes, 1957); this can be thought of as making the smallest representational commitment needed to

In J. Culbertson, A. Perfors, H. Rabagliati ⊠ V. Ramenzoni (Eds.), *Proceedings of the ⊠⊠th Annual Conference of the Cognitive Science Society*.

accommodate the data, so that new information can be accommodated with a minimum of disruption to existing representations. More formally, the inference may be cast as a constrained maximization problem, with the entropy of the distribution being the objective, and the observed values of certain statistics (e.g. means or covariances) being the constraints. For the case in which the data correspond to values of a scalar function, Burg's theorem implies that the solution to the optimization takes a remarkably simple form, as shown below:

Theorem 1 *(Burg, 1967) Let* $\hat{Y}_1, \quad ,\hat{Y}_k)$ *be an observation of a time series*[1]*, and let let* $\quad_j = \sum_i \hat{Y}_i\hat{Y}_{i+j}, 1 \leq j \leq L$ *be the empirical lagged covariances, for some fixed L. The maximal entropy distribution of* $Y_1, \quad ,Y_k)$*, subject to the constraints* $\sum_i Y_iY_{i+j} = \quad_j$ *is given by*

$$Y_i = \sum_{j=1}^{L} w_j Y_{i \quad j} + \varepsilon_i$$

for some choice of coefficients w_i*. Here* ε_i *are iid Gaussian variables with mean* 0 *and variance* σ^2*.*

(For a proof, see Cover & Thomas, 1991, Section 12.6.) Burg's Theorem thus suggests a natural way to estimate the values of f on points outside the domain of definition, in both generative and evaluative paradigms. In the generative case, the value at $N+1$ is estimated by applying the autoregressive weights to the last L observations, $\hat{f} \; N+1) := \sum_{j=1}^{L} w_j f \; N \quad j+1)$ (optionally with normal additive random noise). We may then append this value to the vector of observed values and repeat the process iteratively, obtaining estimates of $\hat{f} \; N+k)$ for arbitrarily large k. In the evaluative case (e.g., multiple choice), the participant must judge the quality of a proposed completion $\; \hat{f} \; N+1), \quad \hat{f} \; N+M))$ of f. This may be done in a in a manner that uses the same machinery as the generative case; we describe this process in greater detail further below.

Gaussian Process models of function learning

The MaxEnt model differs from approaches to modeling function learning that use Gaussian Processes (Schulz et al., 2017; Lucas et al., 2015; Wilson et al., 2015). In this section, we briefly outline the key assumptions of the GP framework.

Formally, a Gaussian process (GP) defines a probability distribution on the space of all possible one-dimensional functions $f : \mathbb{R} \to \mathbb{R}$ (Rasmussen & Williams, 2006). This assumes that, for any finite set of observations of values of the function $\; x_1, f \; x_1)), \quad , \; x_n, f \; x_n))$, that elements of the vector $\; f \; x_1), \quad ,f \; x_n))$ are distributed according to a zero-mean multivariate Gaussian, with the covariance structure satisfying $K \; x_1,x_2) = Cov \; f \; x_1), f \; x_2))$, where K is a referred to as a *kernel function*.

Following the previous convention, we assume that the x_i values are always given on the set of positive integers, $x_i = i$. In this case, a GP with kernel function K is equivalent to a mean-zero multivariate normal distribution with covariance $C^K_{ij} = K \; i,j)$. Thus, given the first N values of the function $f \; 1), \quad ,f \; N)$, and a kernel function K, the posterior over later values $\; f \; N+1), \quad ,f \; N+M))$ is given by conditionalizing the distribution $\mathcal{N} \; 0, C^K)$ on the values of the first N components. The posterior is also Gaussian, the mean and variance of which can be explicitly expressed in terms of C^K and the vector $\; f \; 1), \quad ,f \; N))$ (Lucas et al., 2015; Rasmussen & Williams, 2006).

In the context of function learning, kernel functions are used to encode an agent's prior beliefs about likely kinds of functions. Given such priors, the problem of inferring the values of a function at a new set of points can be recast as a form of rational probabilistic inference (Lucas et al., 2015). In order to complete the specification of such a model, it is necessary to posit a specific kernel or collection thereof.

common choice is the Radial Basis Function (RBF) kernel, defined by the formula $K^{rbf}_{\sigma} \; x,y) = e^{\quad x \quad y)^2/\sigma}$ (Lucas et al., 2015; Schulz et al., 2017). Samples from this distribution do not generally satisfy any global parametric form, but rather obey a form of local statistical regularity: the curves are smooth, but tend to "meander" randomly due to the lack of long-range correlations. This kernel has thus been used as a model for a basic smoothness prior. Other kinds of kernels, that encode more specific forms of structure, such as linearity or periodicity, are also possible (Schulz et al., 2017).

final type of kernel, which is common in machine learning applications but less common in studies of human function learning, is the Matern kernel (Rasmussen & Williams, 2006). This kernel is a generalization of the RBF, which can generate curves that are locally "rough". The kernel contains an additional hyperparameter ν that controls this degree of roughness (see Figure 1). We use this kernel to generate stimuli for Experiment 1. This kernel has been previously used in human function learning experiments by Schulz, Tenenbaum, Reshef, Speekenbrink, and Gershman (2015). In our experiments, all stimuli were generated by sampling from Gaussian kernels, either RBF or Matern.

Modeling of multiple choice extrapolations

Following (Schulz et al., 2017), the experiments we report below used a multiple-choice (evaluative) extrapolation paradigm. In this paradigm, participants are shown a static graphical representation of a prompt curve $y_{prompt} = \; f \; 1), \quad ,f \; N))$ as well as of several candidate completions, and tasked with selecting the most plausible completion. We denote the ith candidate completion by y_i, which is represented as a vector of potential values of f at the unseen points $N+1, \quad ,N+M$. In all experiments, we take $N = M = 100$.

GP model

Following previous accounts of GP inference (Schulz et al., 2017; Duvenaud, Lloyd, Grosse, Tenenbaum, & Ghahramani,

[1] Note that a time series is merely a notational variant of the ordered vector of values $\; f \; 1), f \; 2), \quad ,f \; N))$ of a function as formulated above.

2013), we assume that the GP agent makes a multiple choice decision using the following two-step process:

1. Given y_{prompt}, estimate the kernel most likely to have generated it: $\hat{K} = argmax_K P_K\ \ y_{prompt})$

2. Evaluate the posterior likelihood of each completion $P_{\hat{K}}\ \ y_i|y_{prompt})$, using the kernel inferred in step 1.

Here, P_K denotes the probability distribution defined by the kernel K. We then assume that the choice probabilities are determined by the conditional likelihood of each candidate. Thus, given prompt y_{prompt} and candidate completions y_i, we assume that the GP agent selects choice i with probability

$$p_i^{GP} \propto P_{\hat{K}}\ \ y_i|y_{prompt})^{\gamma}$$

where $\gamma > 0$ is an inverse temperature parameter. We treat γ as a participant-specific hyperparameter, and fit it to each participant using maximum likelihood. Finally, for simplicity and conceptual clarity, we will assume that the inference in step 1 is perfectly accurate. That is, if y_{prompt} was generated from some kernel K, then the GP agent can exactly recover K after seeing y_{prompt}. Thus the GP agent acts as an Ideal Observer.

MaxEnt model

For this model, rather than relying on specific prior distributions, the choice probabilities are based on the simple iterative extrapolation process implied by Burg's theorem. When presented with a prompt and set of candidate completions, we assume that the MaxEnt agent first fits the parameters w, σ of an autoregressive linear model on y_{prompt}. The agent then uses these parameters to evaluate the plausibility of a given completion. The form of the MaxEnt solution gives a natural way to make such judgments. Let $\{r_{ij}\}_j$ denote the set of residuals (i.e., prediction errors) along the candidate completion y_i with respect to the fitted regression model. That is, $r_{ij} = y_i[j] \quad \sum_{k=1}^{L} w_k y_i[j \quad k]$ (where $y_i[j]$ denotes the jth entry of the vector y_i). Burg's theorem implies that *if the completion was generated by the same process as the prompt, then the residuals are independent samples from* $N\ \ 0, \sigma)$. Conversely, systematic departure of the distribution of residuals from $N\ \ 0, \sigma)$ indicates that the completion is unlikely to have been generated by the same process as generated the prompt. ccordingly, a simple way to assess the fit is to compute the empirical mean $\hat{\mu}_i = \quad {}_j r_{ij}$ and standard deviation $\hat{\sigma}_i = \quad \overline{\quad {}_j\ r_{ij} \quad \hat{\mu}_i)^2}$ of the residuals, and compare them to the expected values 0 and σ. The further these values deviate from 0 and σ, respectively, the less likely it is that the completion y_i was generated from the same process as y_{prompt}. Based on this, we assume that the MaxEnt agent makes its decision using the following three-step process:

1. Compute the regression parameters w, σ on the prompt curve

2. For each candidate completion y_i, compute the regression residuals $\{r_{ij}\}_j$ along y_i, as well as the mean $\hat{\mu}_i$ and standard deviation $\hat{\sigma}_i$ of these residuals.

3. Select the choice i with probability

$$p_i^{MaxEnt} \propto KL\ \ N\ \ \hat{\mu}_i, \hat{\sigma}_i)||N\ \ 0, \sigma))\ \ ^{\gamma}$$

where KL denotes the Kullback-Liebler divergence [2] between the two Gaussian distributions, and $\gamma > 0$ is an inverse temperature parameter . The exponent is negative because high values of KL correspond to poor fits, and thus to low choice probabilities. This model thus contains two participant-specific hyperparameters: γ and the window length L. We fit these two parameters independently for each participant using maximum likelihood. ccordingly, we report model comparison results via BIC in which the MaxEnt model is considered to have two parameters, and the GP to have one. [3]

Experiment 1: Roughness calibration

The previous function learning literature has focused primarily on smooth curves (McDaniel & Busemeyer, 2005; Delosh et al., 1997), or ones with piecewise discontinuities (Wilson et al., 2015). Here we test the hypothesis that people also have consistent preference patterns for rough curves, and that the pattern of responses to both smooth and rough curves can be explained by the MaxEnt model. More specifically, it predicts that people will be sensitive to the degree of *local* roughness along a curve, and will tend to prefer completions that are calibrated (i.e., matched) to the prompt in this regard.

We used the Matern kernel to generate multiple choice extrapolation problems, in which the roughness of the prompt curve, as well as the roughness of the individual completions, were independently varied. More specifically, to generate a prompt curve, we first sampled $\nu_0 \in \{\ 25, 1, 4\}$, and then sampled from the corresponding Matern kernel $y_{prompt} \sim K_{\nu_0}^{matern}$. We then generated three candidate choice curves y_i, defined as samples from the three posterior distributions $y_i \sim P_{K_{\nu_i}^{matern}}\ \ \cdot|y_{prompt})$, where $\nu_1 = \ 25, \nu_2 = 1$ and $\nu_3 = 4$.

n example is shown in Figure 1. We generated a total of 15 stimuli for each prompt roughness level, for a total of 45 stimuli. Participants were instructed to choose the completion they judged to be the best completion of the prompt. The order of prompt curves was randomized for each participant, and the order in which the three candidate completions were displayed was randomized within each trial.

52 participants were recruited from Prolific (`www.prolifi . o`). fter exclusion of participants who failed attention check trials, $N = 43$ participants remained (15 female, mean age=27 0 ± 10 3). Participants were paid $1.59 for their participation. On average, the experiment took 4.5 minutes to complete.

[2] Using KL divergence is not essential, and other metrics could also be used without changing the model's qualitative behavior.

[3] Strictly speaking, this is an accommodation that disadvantages the MaxEnt. Since one of the parameters of the MaxEnt model is discrete, the parameter space of this model is geometrically a finite set of disjoint one-dimensional intervals. Thus, arguably, the MaxEnt model should be considered as having only one free parameter, since finitely many small intervals can be concatenated into one large interval.

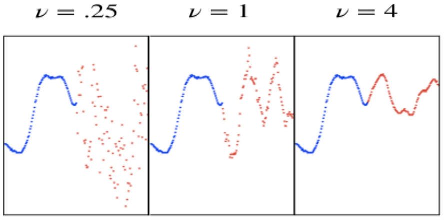

Figure 1: Example stimulus in Experiment 1. The prompt curve (blue) was in this case sampled from a Matern kernel with $\nu = 1$. The plot labels indicate the ν corresponding to each respective completion. In this case, the MaxEnt model would prefer the middle completion (assuming a small L), while the GP would prefer the right one.

Behavioral results

On average, participants selected the completion that was matched to the prompt in terms of roughness in 79 percent of trials, suggesting a strong preference for roughness-calibrated completions. This preference was strongest when the prompt curves were maximally smooth (i.e. $\nu_0 = 4$).

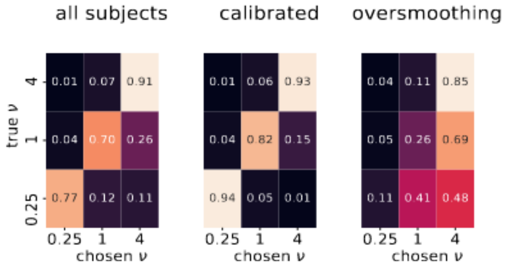

Figure 2: Empirical confusion matrices. The left matrix is the average response over all participants, while the other two are averages only over participants classified as calibrated or oversmoothing, respectively (see text for explanation)

We next examined individual differences in response patterns by performing K-means clustering on the confusion matrix of each participant. The best fit (according to a silhouette score (Rousseeuw, 1987)) was obtained for $K = 2$, with the two groups corresponding to qualitatively different response patterns. The first group, that we refer to as *calibrated* participants, chose the completion with the noise level matched to the prompt on nearly all trials. This group comprised 34/43 subjects. In contrast, the second group, that we refer to as *oversmoothing* participants, consistently chose completions that were smoother than the prompt (or matched it, in the case of maximally smooth prompts). This group comprised the remaining 9/43 participants. In Figure 2 we show the empirical confusion matrices averaged across all participants and for each of the two groups individually.

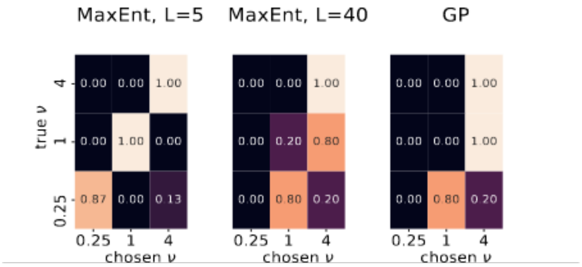

Figure 3: Model response patterns. For simplicity, in all models $\gamma = \infty$ (corresponding to a deterministic argmax decision rule). In MaxEnt models, L is window length (see text).

Model results

Before presenting formal model fits, we first consider the qualitative behavior of the models. In Figure 3 we plot the confusion matrices for the MaxEnt model for both large and small L values; the figure also shows the confusion matrix for the GP model, which has no free parameters after fixing the value of γ. The results show that the MaxEnt model can exhibit both calibrated response patterns (for low values of L) as well as oversmoothing response patterns (for higher values of L), thus offering a potential account of individual variability. In contrast, the GP model is constrained to consistently oversmooth relative to the prompt (i.e. chooses curves with larger values of ν), and thus can account only for the less common of the two observed behavioral patterns.

Figure 4 shows the fits of the MaxEnt model in greater detail. s expected, calibrated participants exhibit smaller L values, whereas oversmoothing participants exhibit larger values. Thus the L parameter can account for this difference in response patterns.

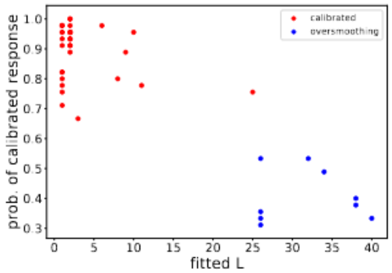

Figure 4: Fitted window length parameters, and calibration probabilities (i.e. average diagonal entry of the confusion matrix) for each participant. Large L values correspond to low calibration probabilities, with the majority of participants having high calibration. The colors indicate the cluster assignment of each participant.

These observations are supported by formal model fits. When fitted to each individual participant,the MaxEnt model attains a log-likelihood (llh) value of 0 502 ± 0 259 (mean ± std), whereas the GP model attains a fit of 0 729 ± 0 147.

random guessing model attains log $1/3 =$ 1 099. dditionally, the MaxEnt model provides a better MLE fit to 41 of 43 participants than the GP model, and attains a lower BIC value on 35 of 43 participants.

These results may seem surprising: Why couldn't the GP model "pick out the right" completion, given that it has access to the ground truth generative distribution of each prompt curve? This is because, regardless of the value of ν, the posterior *mean* closely approximates a horizontal line, with the likelihood of a given completion being inversely related to the deviations from this mean. Since, in general, smooth candidate curves will tend to deviate less from a line than will rough candidate curves, it is likely that a *given* smooth curve may have a higher posterior probability than a *given* rough curve, despite the fact that the *set of all* rough curves may have far higher posterior probability than the *set of all* smooth curves.

The MaxEnt also exhibits a tendency to oversmooth, but only for larger values of L, and for a different reason. When L is very large, the autoregressive model will overfit to the prompt curve, and the estimated σ value will be too small, relative to the intrinsic volatilty along the curve. Since the decisions are made by comparing the residual variance along each candidate to the fitted residual variance σ along the prompt, this model will thus have a bias towards smoother completions when L is large. Psychologically, a large L may correspond to an increased reliance on the "global character" of the function at the expense of local features such as roughness.

Experiment 2: Typicality preference

Previous work on function learning has focused on curves generated from an RBF kernel (e.g., Lucas et al., 2015; Schulz et al., 2017). The MaxEnt model predicts that, in this more restricted case, completions that look "representatitve" of the posterior distribution will be seen as more plausible completions than will the mode of that distribution. This contrasts with the GP model which predicts the opposite, since the mode has (by definition) a higher posterior likelihood value than any other possible completion.

To generate stimuli, we sampled prompt curves from an RBF, and generated two choices for each curve: the first was the posterior mode (with respect to the underlying RBF kernel), and the second was an unbiased sample from the posterior distribution. We refer to these as "modal" and "typical" completions, respectively. 64 participants were recruited from Prolific using the same criteria and payments as Experiment 1. fter excluions, N=52 participants remained (24 female, mean age=32 3 ± 10 6). On average the experiment took 5.2 minutes to complete .

Behavioral results

Overall, the typical completion was chosen on 74.9 percent of all trials. 43/53 participants chose typical completions on the majority of trials they performed; and for 47/50 stimuli, the typical completion was chosen by the majority of participants. Furthermore, as in Experiment 1, there was considerable individual variability in response patterns, with the proportion of trials for which the typical completion was selected ranging from .38 to 1.0 across participants.

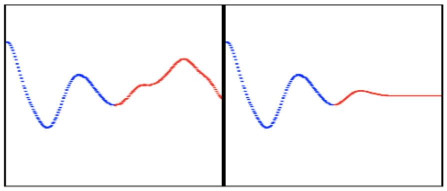

Figure 5: Example stimulus from Experiment 2. typical completion is on the left, and the modal completion is on the right. Note the regression to the mean in the latter.

Model results

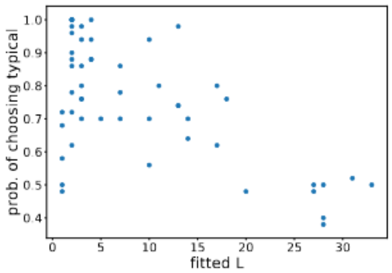

Figure 6: Individual response patterns in Experiment 2. Fitted L value of each participant tracks the proportion of trials on which participants chose the typical completion.

The MaxEnt model exhibits a strong bias for selecting the typical completions, similar to what is observed in the empirical data. In contrast, the GP model provides a poor qualitative fit. This is because it cannot assign $>$ 50 percent probability to the typical completion for any stimulus, regardless of the value of the inverse temperature parameter γ. Thus, the GP agent assigns a higher choice probability to the *modal* completion on *all* trials. This contrasts with the empirical observation that the typical completion was preferred for the vast majority (94 percent) of stimuli.

In formal model fits, the MaxEnt model attains a llh of 0 450 ± 222 on an average participant, compared to 692 ± 006 for the GP model which is not appreciably different than a random guessing model (log $1/2 =$ 693). The MaxEnt model also provides a higher MLE for every participant individually, and exhibits a lower BIC for 41 out of 52 participants. Finally, Figure 6 also shows that the fitted L value for each participant closely tracks the participant's preference for typical completions, with smaller L values corresponding to greater preference for typical completions.

Related work

Classic studies of function learning (McDaniel & Busemeyer, 2005; Delosh et al., 1997; Bott & Heit, 2004) have focused on peoples' ability to learn curves generated by simple parametric families, such as linear, quadratic or sinusoidal. Little and Shiffrin (2016) considered a more complex set of functions by using polynomials of varying degrees, and found that peoples' preferred completions were biased towards lower dimensional polynomials. In (Lucas et al., 2015), it was shown that many phenomena could be explained using the Gaussian Process framework. This work relied on a prespecified set of kernel functions, raising the general question of what class of kernels should be considered. Wilson et al. (2015) explored the question of discovering an appropriate kernel function using unsupervised methods. Schulz et al. (2017) argued that the class of kernels is constrained by a principle of compositionality, in which more complex kernels are constructed through combinations of simpler atomic ones.

lthough, to our knowledge, no previous studies have specifically tested the MaxEnt model (or a version thereof), several recent studies have generated data relevant to its assumptions, and the design of the experiments used here. For example, Gelpi, Saxena, Lifchits, Buchsbaum, and Lucas (2021) used an active function learning paradigm, in which only parts of the entire function were shown, and participants had to choose locations at which to query the value of the function. They found that people tended to prefer an even spacing of such points along the x-axis, consistent with our assumption of evenly-spaced observations. Furthermore, León-Villagrá, Preda, and Lucas (2018) incorporated explicit memory constraints into the GP modeling framework. Their strategy for this was very similar to the autoregressive form of the MaxEnt model implemented here, in that they considered only the rightmost previously encountered k points at any given location along the x axis. The autoregressive model also bears a notable similarity to the nearest-neighbor heuristics used in the context of graph structured spaces by Wu, Schulz, and Gershman (2021). dditionally, in a machine learning context, Segert and Cohen (2021) showed how to use the MaxEnt principle for unsupervised learning of "intuitive functions" (that is, smooth functions with simple underlying statistical structure, similar to those considered in our experiments). But, they considered only synthetic data and did not model peoples' extrapolation judgements directly.

Discussion

We have presented a model of function extrapolation that is based on the Principle of Maximum Entropy. The model is based on a general inductive bias that is not specific to function learning, and provides a simple algorithmic process (viz., linear autoregression) by which extrapolation judgments are made. We derived two behavioral predictions from the model: roughness calibration and typicality preference, and showed that both of these effects are robustly present in human judgments. Furthermore, we showed that the L hyperparameter in the MaxEnt model (that determines the autoregressive window length) can explain participant-level variation in response patterns, with larger values of L corresponding to a preference for smoother completions.

The MaxEnt model contrasts with existing models of function learning, based on Bayesian inference in the space of curves with a Gaussian Process prior. The latter has proven to be a useful and conceptually clarifying framework, providing one possible normative benchmark against which to evaluate performance. However, its use has focused on computational-level accounts of the problem. Furthermore, the resulting posterior computation, despite having a closed-form mathematical solution, is computationally intensive, generally requiring the inversion of a large matrix (Rasmussen & Williams, 2006). It is unclear how people might perform this computation, or what approximation they might use; though it is possible that identifying such an approximation might also account for the experimental results we report here.

The MaxEnt approach is also grounded in a potentially normative account of function learning, in which the MaxEnt principle can be thought of as a form of regularization that protects against overfitting the data (and thereby maximizing the potential for generalization and/or future flexibility in learning). Thus, in line with recent work on Resource Rationality (Lieder & Griffiths, 2020; Dasgupta, Schulz, Tenenbaum, & Gershman, 2020), the MaxEnt model may complement the GP perspective by providing a simple and psychologicaly plausible algorithm for performing inference over the space of functions.

It is possible that a GP model could account for the experimental data, by using some kernel besides the veridical one (a "human kernel"). But since the true kernel already oversmooths, such a model could only account for calibrated response patterns if the putative kernel placed most of its probability mass on *rougher* curves-and would thus be very different from previously proposed "human kernel" candidates (Schulz et al., 2017; Wilson et al., 2015).

t the same time, the MaxEnt model has several limitations, that suggest directions for future work. First, it is primarily a model of extrapolation, and it is not immediately clear how it can be applied to interpolation tasks. Second, our implementation assumed observation points to be equally spaced, which is not always borne out in real functions. Finally, one desirable property of GP models, which we have not utilized here, is their ability to quantify uncertainty in a natural way-it remains to be further explored how to best accomplish this within the MaxEnt framework.

In summary, the MaxEnt perspective on function learning, in addition to its success in fitting empirical data, provides several theoretical advantages: leveraging Burg's theorem, it provides a solution to inference problems using a tractable algorithm that is simple and parsimonious, relying only on a domain-general inductive bias. The algorithm provides the basis for a mechanistic, process-level understanding of human function learning, while the generic nature of the inductive bias suggests that the insights obtained in this domain may apply more broadly to other domains requiring generalization and/or extrapolation.

cknowledgements

The authors are grateful to Mayank grawal, Zack Dulberg, Steven Frankland, and Taylor Webb for helpful discussions and feedback. SS is supported by a T32 Training Grant in Computational Neuroscience (T32MH065214). This project was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation.

References

Bhui, R., & Gershman, S. (2018). Decision by sampling implements efficient coding of psychoeconomic functions. *Psychological Review*, 985-1001.

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

Burg, J. (1967). Maximum entropy spectral analysis. In *Proceedings of the 37th meeting of the society of exploration geophysicists.*

Cover, T., & Thomas, J. (1991). *Elements of information theory*. John Wiley and Sons.

Dasgupta, I., Schulz, E., Tenenbaum, J., & Gershman, S. (2020). theory of learning to infer. *Psychological Review*, *127*, 412-441.

Delosh, E., Busemeyer, J., & McDaniel, M. (1997). Extrapolation: the sine qua non for abstraction in function learning. *J Exp. Psychology*, *23*.

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., & Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *International conference on machine learning* (p. 1166-1174).

Frankland, S., Webb, T., & Cohen, J. (2021). No coincidence, george: Capacity-limits as the curse of compositionality. *psyarxiv*.

Gelpi, R., Saxena, N., Lifchits, G., Buchsbaum, D., & Lucas, C. (2021). Sampling heuristics for active function learning. *Psy rXiv*.

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, 620-630.

Kwantes, P., & Neal, . (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

León-Villagrá, P., Preda, I., & Lucas, C. (2018). Data availability and function extrapolation. In *Proceedings of the cognitive science society.*

Lieder, F., & Griffiths, T. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Little, D., & Shiffrin, R. (2016). Simplicity bias in the estimation of causal functions. In *Cogsci* (p. 1157-1162).

Lucas, C., Grrifiths, T., Williams, M., & Kalish, M. (2015). rational model of function learning. *Psychonomic Bulletin and Review*, *22*, 1193-1215.

McDaniel, M., & Busemeyer, J. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psych Bull Rev.*, *12*, 24-42.

Myung, J. (1994). Maximum entropy interpretation of decision bound and context models of categorization. *Journal of Mathematical Psychology*, *38*, 335-365.

Myung, J., & Shepard, R. (1996). Maximum entropy inference and stimulus generalization. *Journal of Mathematical Psychology*, *40*, 342-347.

Rasmussen, C., & Williams, C. (2006). *Gaussian processes for machine learning*. MIT Press.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and pplied Mathematics*, *20*, 53–65.

Schulz, E., Tenenbaum, J., Reshef, D., Speekenbrink, M., & Gershman, S. (2015). ssessing the perceived predictability of functions. In *Cogsci.*

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44-79.

Segert, S., & Cohen, J. (2021). self-supervised framework for function learning and extrapolation. *arXiv:2106.07369*.

Wilson, ., Dann, C., Lucas, C., & Xing, E. (2015). The human kernel. In *dvances in neural information processing systems* (p. 2864-62).

Wu, C. M., Schulz, E., & Gershman, S. J. (2021). Inference and search on graph structured spaces. *Computational Brain Behavior*, *4(2)*, 125-147.