

ERIC SCHULZ
HARVARD UNIVERSITY
52 OXFORD STREET, ROOM 295.08
CAMBRIDGE, MA 02138, UNITED STATES
ERICSCHULZ@FAS.HARVARD.EDU

**PRÉCIS FOR
“TOWARDS A UNIFYING THEORY OF
GENERALIZATION”**

January 6, 2018

University College London
Date of award: August 29, 2017

INTRODUCTION

“Because any object or situation experienced by an individual is unlikely to recur in exactly the same form and context, psychology’s first general law should, I suggest, be a law of generalization” Roger Shepard argued in 1987. Thirty years later, Gershman and Daw (2017) point out that current theories of reinforcement learning cannot account for the fact that the “data are relatively sparse and, indeed, precisely the same situation may never be encountered twice”. Whereas Shepard proposed an exponential similarity function driving generalization in psychological space, Gershman & Daw propose a non-parametric generalization mechanism guiding reinforcement learning. Yet, the overarching question remains the same: How do humans and other animals know what to think or do in novel situations?

In my thesis, I have developed an empirically grounded and theoretically refined account of human generalization. The main premise is that generalization is mostly achieved through function learning, i.e., the human ability to learn a mapping from inputs to outputs that generalizes to novel input-output pairs. I use Gaussian Process regression (Rasmussen and Williams, 2006), a non-parametric Bayesian approach that acts as a universal function approximator, as a model that can capture human generalization with mathematical elegance and computational ease.

Across 14 experiments and using extensive computational modeling, I show that this model generates testable predictions about human preferences over different levels of complexity, provides a window onto compositional inductive biases, and—combined with an optimistic yet efficient information sampling strategy—guides human decision making in complex environments. Beyond that, I show how insights gained from human behavior can be used to build more adaptive Bayesian optimization routines. The thesis greatly improves our understanding of the interplay between functional knowledge and psychological similarity, and provides a plausible model of generalization across a large variety of psychological domains.

CHAPTER 2-3: FROM GENERALIZATION TO FUNCTION LEARNING

Historically, function learning and generalization have always been close kin. In particular, the first studies on human function learning (Carroll, 1963) were put forward precisely to show that Shepard’s theory of generalization, which assumed that reactions will be the same as to the closest previously experienced stimuli, is too simplistic. Sometimes completely novel behavior can emerge, which in Carroll’s view could only be explained by functional

extrapolation. Lucas et al. (2015) showed that Gaussian Process regression can explain many of the behavioral phenomena observed in human function learning and the theory presented in my thesis builds on and extends those results.

Gaussian process regression is used to model general functions f relating input \mathbf{x} to output y as

$$y = f(\mathbf{x}) + \epsilon \quad \epsilon \sim N(0, \sigma_\epsilon^2) \quad (1)$$

A prior distribution over these functions is defined as Gaussian Process:

$$f \sim GP(m, k), \quad (2)$$

where m is a mean function (normally set to $m = 0$ without loss of generality) and k is a kernel (or covariance) function specifying the similarity between between inputs. A frequently chosen kernel is the *Radial Basis Function* (RBF) kernel:

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\lambda}\right), \quad (3)$$

where λ governs how quickly correlations between points \mathbf{x} and \mathbf{x}' decay towards zero as their distance increases. Figure 1 shows gradients of functional generalization resulting from different values of the length-scale parameter. From a psychological perspective, λ could also be called the “gradient of functional generalization” as it determines how quickly the correlation between points decays as their distance grows, and thereby the functional similarity of different points given their mutual distance. This is similar to Shepard’s gradient of generalization defining how likely stimuli are to produce the same reaction as they are further away from another stimulus. However, the framework is more general; choosing other kernels such as a linear or periodic kernel results in dramatically different gradients of generalization. Moreover, different kernels can be combined additively or multiplicatively, allowing the composition of complex generalization gradients from simpler building blocks.

CHAPTER 4: FROM THEORY TO BEHAVIOR

Imagine having to learn a relation between an input and output from discrete observations. In one situation, the relation is relatively complex but the observations relatively precise, in the other situation, the relation is relatively simple but the observations noisy. Which one would you choose? By mathematically deriving learning curves for different kernels,

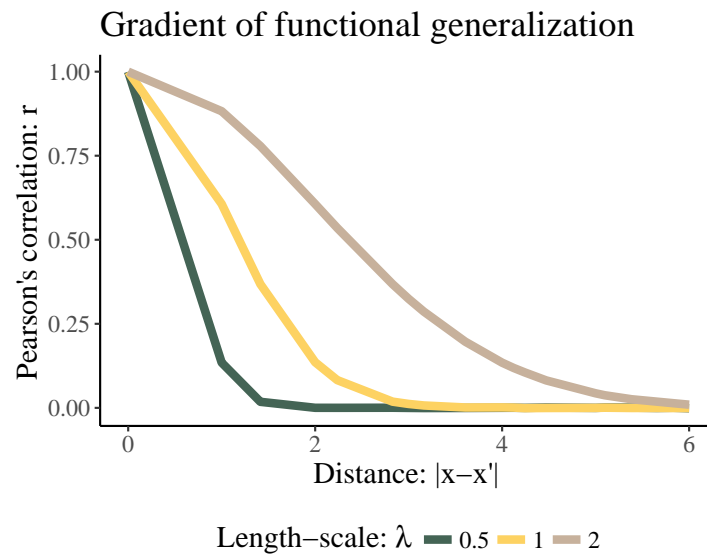


Figure 1: Gradient of functional generalization. Larger values for λ lead to correlations that decay more slowly as the distance between two stimuli increases. For example, the outputs for two stimuli that differ by a distance of 2 are almost independent if $\lambda = 0.5$, but are on average correlated with $r = 0.5$ if $\lambda = 2$. Thus, larger λ -values lead to larger gradients of functional generalization.

the theory of Gaussian Process regression provides guidance on what a rational decision maker would prefer. Learning curves relate the expected generalization error of a model to the amount of training data. They can be seen as a mathematical expression of a function’s *predictability*, given assumptions about priors over functions, the noise process, and the distribution of inputs. Intuitively, a function exhibits a higher predictability if it is easier to predict the output of randomly chosen new input points. It can be shown (see Sollich, 1999) that for kernels such as the RBF, the generalization error scales as:

$$\mathcal{E}(n) \propto \left(\frac{\sigma_\epsilon^2}{n} \right)^{-\frac{2p-1}{2p}}, \quad (4)$$

where p is the function’s smoothness (i.e., the function’s wiggleness, which corresponds to its complexity), n is the sample size and σ_ϵ^2 is the attached observation noise. Thus, learning curves make clear predictions about predictability, namely that smoothness is more important than noise, i.e. complex but deterministic systems are easier to predict than simple but noisy ones. We tested this prediction in a task where participants rated the predictability of different functions with varying smoothness, noise level, and the sample size. The results are shown in Figure 2. We found that the smoothness of a function exerted a stronger influence on predictability than noise or sample size, consistent with the theoretically derived

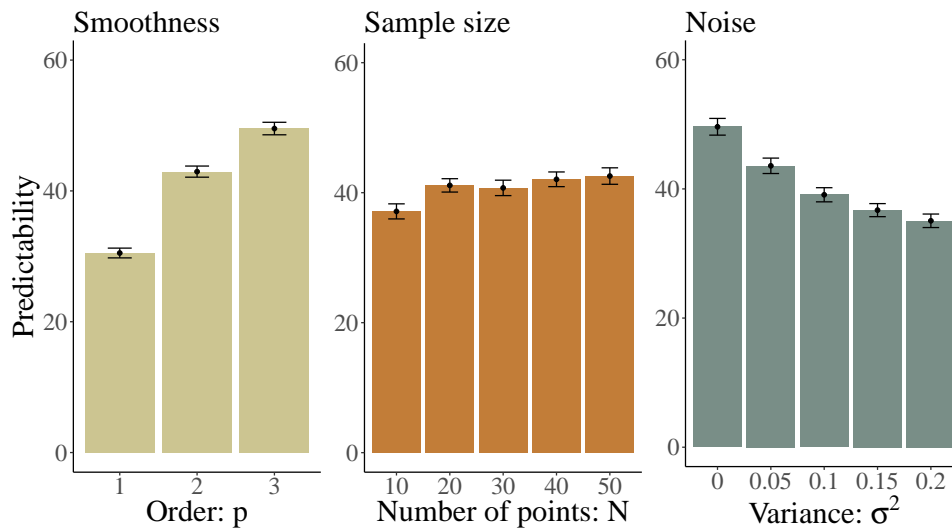


Figure 2: Results of predictability experiment. Smoothness is more important than sample size and noise. All effects go in the right direction, as predicted by the derived learning curves.

learning curves. This means that a smooth but noisy function is perceived as more predictable than a complex but near-deterministic function. Moreover, a function’s overall predictability was the best predictor of participants’ judgments. This shows a remarkable correspondence between mathematical predictions of rational function learning models and human behavior.

CHAPTER 5: A COMPOSITIONAL THEORY OF GENERALIZATION

In the longest chapter of my thesis, I develop a compositional theory of generalization and test this theory across 8 experimental paradigms. The underlying question of this theory is how people are able to generalize as effectively as they do. Our account is based on the assumption that they do so through compositional inductive biases, i.e. priors over different structural forms that they can combine and reuse, thereby creating potentially infinite means of generalization from a finite set of building blocks (Duvenaud et al., 2013). We used three base components: a RBF, a periodic, and a linear kernel, which can be combined by either adding or multiplying them together in order to create new kernels with which Gaussian Process regression can be performed (see Figure 3). We compare this model with a simple Radial Basis Function kernel, which can learn any smooth function without structural assumptions, and a non-parametric kernel, which can learn any structure implicitly by approximating the gradient of functional generalization using a mixture of Gaussians (Wilson and Adams, 2013).

In the first set of experiments, participants were shown different pattern completions

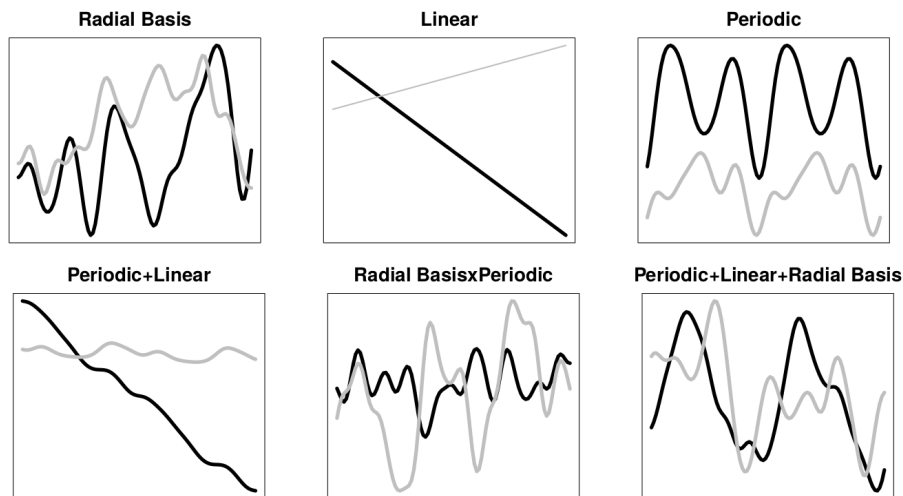


Figure 3: Compositional kernel. **Top:** Base kernels. **Bottom:** Examples of compositions.

generated by the different models and had to choose the one they thought best completed a given function. We found that participants chose the compositional pattern completions more frequently if the underlying functions were compositional, but even if they were sampled from the non-parametric kernel, thus showing a strong preference for compositional completions. The second series of experiments assessed participants’ distributions over posterior predictive compositions by letting them accept or reject proposed pattern completions, thereby creating a Monte Carlo Markov chain over people’s preferred compositional completions (Sanborn and Griffiths, 2008). In a first experiment, we found that, on average, participants’ posterior compositions converged to the true generating compositional distribution. A second experiment used real-world data and found that their posterior compositions are highly correlated with the posteriors found by our model. Next, we let participants complete patterns manually by drawing lines into a cloud of dots, where the dots were sampled from the compositional kernel. The results revealed that participants’ drawings are best described by the compositional model, especially during extrapolation judgments.

For later experiments, we generated two sets of functions: 15 sampled from a compositional and 15 from a non-parametric kernel. We made sure that these two sets were similar on as many time series measurements (wavelet difference, information contained in their spectral density, and so forth) as possible to further pit the compositional and the non-parametric model against each other. In an experiment following the predictability paradigm introduced in Chapter 3, we found that compositional functions were perceived as more predictable than non-compositional functions and that a compositional model of predictability provided the best trial-by-trial predictions of participants’ judgments. In another experiment, we assessed

participants' learning behavior in a traditional function learning paradigm where participants were sequentially presented with input points and asked to predict the output, followed by feedback about the true output. Again, we found that participants were able to learn compositional functions more easily. Importantly, a compositional model of function learning lead to the best descriptions of their predictions for both interpolation and extrapolation trials. When we overlaid dots over either compositional or non-compositional functions in a numerosity experiment, we found that the number of dots overlaid over compositional structure were systematically more underestimated than dots overlaid over non-compositional structure, in line with the intuition that points embedded in structure are more readily perceived as part of an overall whole. Within a change detection paradigm, we randomly selected and swapped points of either compositional or non-compositional functions and found that participants detected change more easily if it occurred for compositional functions. Moreover, we built and tested a Gaussian Process-based model of change detection that uses the optimized parameters of a kernel to evaluate the probability that change has occurred at test and found that a compositional model of change detection described participants' responses best. Finally, we developed a short-term memory task in which participants were sequentially shown a varying number of different functions and afterwards had to indicate whether or not a probe function had been presented in the sequence. Again, our results revealed that compositional functions were better remembered and more easily detected as having been within in the sequence. Moreover, we built a Gaussian Process-based model of short-term memory which memorizes functions by using optimized parameters for the different kernels and then decides whether or not a probe was in the sequence by averaging over previously encoded kernels. Using this approach, we found that a compositional model of short-term memory described participants' responses in our task best.

Taken together, the results presented in Chapter 5 present strong evidence that human function learning and generalization is based on compositional representations.

CHAPTER 6: GENERALIZATION AND EXPLORATION IN VAST SPACES

Chapter 6 introduces a novel paradigm in which participants not only have to learn about an underlying function, but have to use this knowledge to choose optimal inputs that provide the highest outputs. This leads to an exploration-exploitation dilemma: participants have to both take exploratory actions to gain knowledge about the underlying function and exploitative actions to produce high rewards. We assess how people optimize unknown functions in uni-dimensional (Experiment 1) and two-dimensional (Experiment 2) grid worlds, where

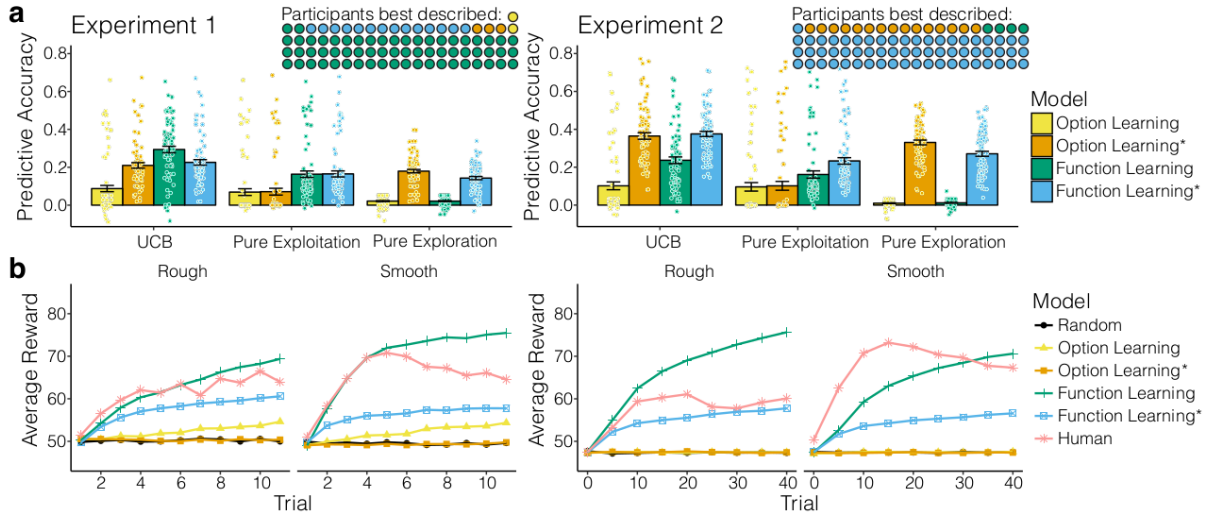


Figure 4: **a)** Cross-validated predictive accuracy of each model, with bars indicating the group mean (\pm SEM). Each individual participant is shown as a single dot, with the number of participants best described shown as an icon array. Asterisks (*) indicate a localized variant of the Option Learning or Function Learning models. **b)** Averaged learning curves of participants and models (combined with UCB-sampling only) simulated over 10,000 replications using sampled participant parameter estimates. Learning curves are separated by environment (smooth vs. rough).

the underlying functions were sampled from a Gaussian Process. We used extensive model comparison to assess the out-of-sample predictive performance of 27 models for each participant individually. Extensive model and parameter recovery studies ensured that all models and their parameters were identifiable from the data. The best overall model was based on a Gaussian Process regression (parameterized by a RBF-kernel) and a decision strategy defines the utility u of an action through an explicit trade-off between expected reward $\mu(x)$ and its attached uncertainty $\sigma(x)$:

$$u(x) = \alpha\mu(x) + \beta\sigma(x), \quad (5)$$

This decision strategy is also called Upper Confidence Bound (UCB) sampling and currently the only strategy with competitive performance guarantees (Srinivas et al., 2010).

Figure 4 shows the result of the predictive model comparison for the Gaussian Process functions learning model as compared to a simple mean tracking model which does not generalize over different options. Both models were additionally tested with either a pure exploration ($\alpha = 0$) or pure exploitation ($\beta = 0$) decision strategy and we also built localized versions of the models to better account for participants’ tendency to sample locally from the last revealed output. Throughout both experiments, the function learning models outperformed the option learning models and the upper confidence bound sampling decision strategy provided the best predictions (see Fig. 4A). Moreover, the final models set up with

participants’ parameter estimates lead to human-like task performance (see Fig. 4B). Taken together, the study shows that participants are well-predicted by a meaningful and robust model which generalizes over different options and solves the exploration-exploitation trade-off by optimistically inflating expectations by their attached uncertainties. Interestingly, when we assessed the estimated length-scale λ of our final model, we found that participants’ had a systematic tendency to under-generalize, with smaller estimates of λ than used to generate the (rough or smooth) environments in our task.

CHAPTER 7: CONTEXTUAL MULTI-ARMED BANDITS

Chapter 7 tried to push the limits of the human ability to generalize in experiential decision-making tasks even further. We developed a new paradigm called the contextual multi-armed bandit task in which the reward associated to options depends on the changing values of a set of context variables. In the task, participants chose between the same four planets (the options) to mine for emeralds (the rewards), after being informed of the current state of the environment on three environmental elements (the context). As each planet reacted differently to the different elements and elements were changing from trial to trial, participants had to learn and optimize the underlying functions mapping elements to a planet’s number of emeralds while gaining as many emeralds as possible.

We ran three experiments assessing participants’ behavior over 150 trials within a discrete (binary features of the elements, underlying function linear), a continuous-linear (continuous features, underlying function linear), and a continuous-nonlinear (continuous features, function sampled from a Gaussian Process) contextual multi-armed bandit. Within all experiments, participants were able to learn, improving their score over time and sampling the best option on a given trial more frequently over time. Comparing different non-generalizing (mean tracker, Kalman filter) and generalizing (linear regression, Gaussian Process) models in terms of their participant-wise one-step-ahead predictive performance, we found that the Gaussian Process parameterized with an RBF kernel and combined with a UCB decision strategy lead to the best predictive performance in all experiments.

Figure 5 shows the resulting model performance and retrieved parameter estimates for each experiment. We can see that the predictive performance (R^2) was higher for the simpler discrete contextual multi-armed bandit than for the more complex experiments with continuous features. Moreover, participants explored more in easier experiments (higher β -values) and assumed that there was more noise in more difficult experiments (higher σ). The estimated length-scale λ did not differ significantly between experiments. Instead, the

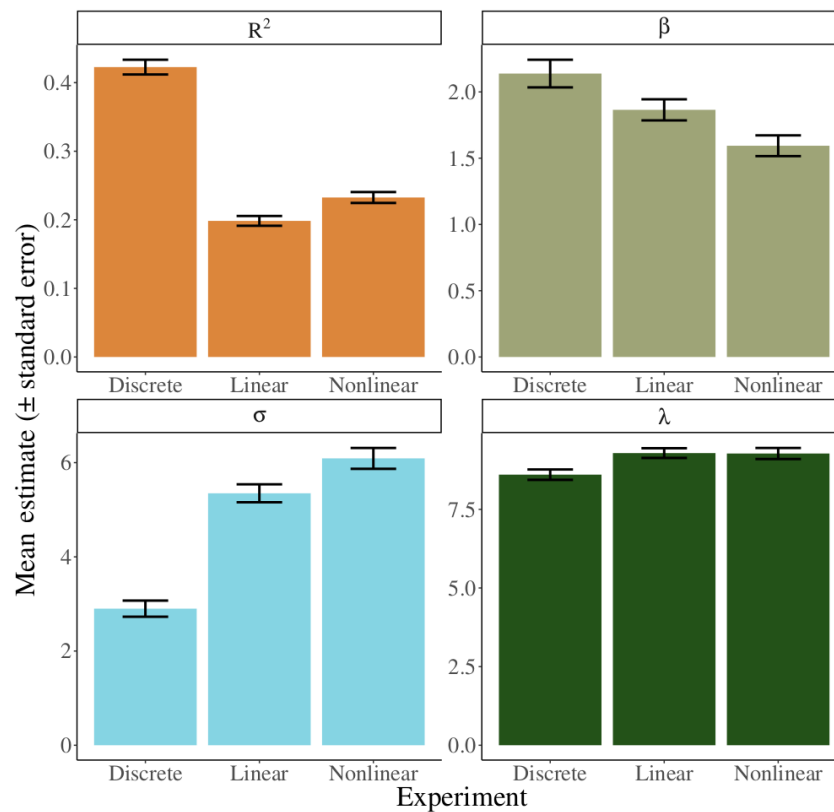


Figure 5: Results of the Gaussian Process-UCB sampling model across all 3 experiments. Easier experiments lead to higher predictive performance (R^2), more exploration (β), and lower perceived noise levels (σ^2). The extent of generalization is constant across experiments (λ), and shows clear under-generalization in 2 out of 3 Experiments.

retrieved values again suggested that participants did not generalize as much as possible, leading to lower than expected estimates of the extent of generalization.

CHAPTER 8: FROM BEHAVIOR TO THEORY

In all of the exploration-exploitation experiments, we observed a robust tendency to under-generalize, given the actual smoothness (λ) of the environment. Subsequent simulations revealed that this is likely to be adaptive: under-generalization largely leads to better performance than over-generalization—and remarkably—is sometimes even better than an exact match between the extent of one’s generalization and the underlying structure of the environment. Our simulations consisted of generating search environments by sampling from a GP prior specified using a *teacher* length-scale (λ_0), and then simulating search in this environment using the Gaussian Process-Upper Confidence Bound sampling model, where

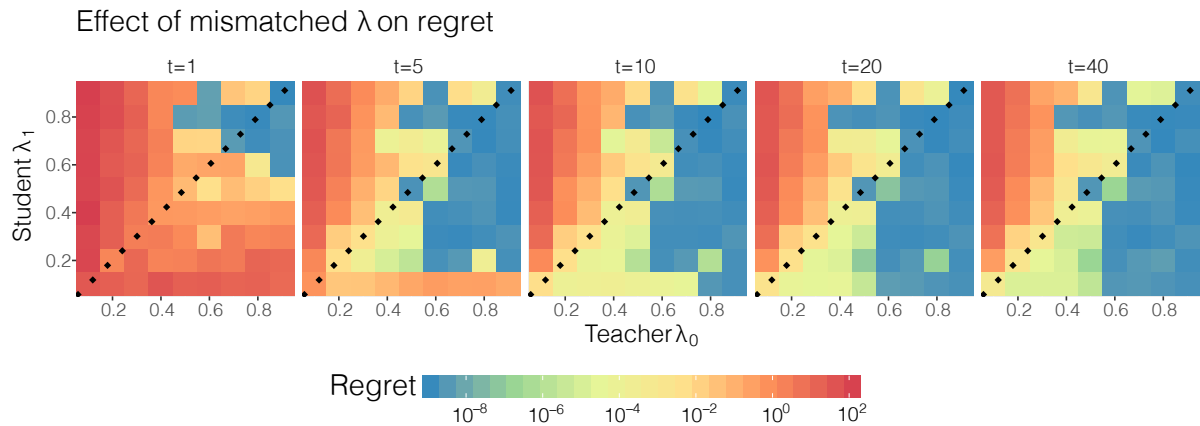


Figure 6: Mismatched length-scale (λ) simulation results. The teacher length-scale λ_0 is on the x-axis and the student length-scale λ_1 is on the y-axis. The teacher λ_0 values were used to generate environments, while the student λ_1 values were used to parameterize the Function Learning-UCB model to simulate search performance. The dotted lines show where $\lambda_0 = \lambda_1$ and mark the difference between under-generalization and over-generalization, with points below the line indicating under-generalization. Performance is measured by regret, i.e. the difference between the current output and the best possible output.

the GP of the function learning component was specified by a *student* length-scale (λ_1).

We used a set-up regularly used by the machine learning community (Metzen, 2016) and ran a simulation with continuous bivariate inputs and using every combination between $\lambda_0 = \{0.1, 0.2, \dots, 1\}$ and $\lambda_1 = \{0.1, 0.2, \dots, 1\}$. We found that under-generalization largely leads to better performance than over-generalization. This effect is more pronounced over time t , whereby a mismatch in the direction of under-generalization recovers over time (higher scores for larger values of t). This is not the case for a mismatch in the direction of over-generalization, which continues to produce low scores. Estimating the best possible alignment between λ_0 and λ_1 to produce the highest score revealed that underestimating λ_0 by an average of about 0.21 produces the best scores over all scenarios. These simulation results show that the systematically lower estimates of λ do not suggest a bias in human behavior—but instead—can sometimes lead to enhanced performance. Under-generalization, as it turns out, might not be a bug but rather a feature of human behavior. Beyond that, under-generalization can also lead to better performance of frequently used Bayesian optimization routines as we have replicated this effect across many different decision strategies used by the machine learning community.

CONCLUSION

Situations requiring generalization are ubiquitous. Whether we infer abstract patterns in the environment or decide where to dine tonight, there is little doubt that generalization is a powerful mechanism to generate expectations in novel situations, and a core ingredient of human intelligence. The controversial question is how we achieve this and what a unifying theory of human generalization might look like.

In my thesis, I have proposed to use Gaussian Process regression as model of, and a method to investigate, human generalization. The results of my experiments showed that generalization is compositional, that it guides decision making through complex environments, and that the human tendency to under-generalize can sometimes be beneficial rather than harmful. Thus, the gradient of generalization is not necessarily the result of inferences someone makes, but a core ingredient of those inferences. My thesis demonstrates the ability of a single theoretical framework to capture a wide range of empirical phenomena concerning the nature of generalization. The framework led to directly testable predictions about the perception of generalization but also to surprising results regarding under-generalization that can be beneficial for actual Bayesian optimization applications. Doing so, I have combined psychology and machine learning in order to advance both fields by further moving towards a unifying theory of generalization.

Hopefully, this theory will continue to produce interdisciplinary insights. One way this can be accomplished is by developing a compositional theory of optimization as well as integrating function learning with the broader framework of program induction; both are topics I am currently working on.

Bibliography

- Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Report Series*, 1963(2).
- Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1166–1174.
- Gershman, S. J. and Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68:101–128.
- Lucas, C. G., Griffiths, T. L., Williams, J. J., and Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, 22(5):1193–1215.
- Metzen, J. H. (2016). Minimum regret search for single-and multi-task optimization. *arXiv preprint arXiv:1602.01064*.
- Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.
- Sanborn, A. and Griffiths, T. L. (2008). Markov chain monte carlo with people. In *Advances in Neural Information Processing Systems*, pages 1265–1272.
- Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.
- Sollich, P. (1999). Learning curves for gaussian processes. In *Advances in Neural Information Processing Systems*, pages 344–350.
- Srinivas, N., Krause, A., Kakade, S., and Seeger, M. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. International Conference on Machine Learning (ICML)*.
- Wilson, A. and Adams, R. (2013). Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1067–1075.