Searching for rewards in graph-structured spaces

Charley M. Wu (cwu@mpib-berlin.mpg.de)
Max Planck Institute for Human Development, Berlin, Germany

Eric Schulz (ericschulz@fas.harvard.edu)
Harvard University, Cambridge, MA, USA

Samuel J. Gershman (gershman@fas.harvard.edu)
Harvard University, Cambridge, MA, US

Abstract

How do people generalize and explore structured spaces? We study human behavior on a multi-armed bandit task, where rewards are influenced by the connectivity structure of a graph. A detailed predictive model comparison shows that a Gaussian Process regression model using a diffusion kernel is able to best describe participant choices, and also predict judgments about expected reward and confidence. This model unifies psychological models of function learning with the Successor Representation used in reinforcement learning, thereby building a bridge between different models of generalization.

Keywords: Generalization; Graph structures; Exploration-Exploitation; Gaussian Process; Successor Representation

Introduction

From social networks to subway maps, many decision-making environments can be described using graph structures, where relationships are defined based on transition structure rather than comparing features. Here, we propose the diffusion kernel as a similarity metric for functions on graphs, which combined with the Gaussian Process (GP) framework, allows us to make Bayesian predictions about unobserved nodes. Using a graph-correlated bandit task, we study how people generalize and search for rewards in structured spaces. We show that the GP model best predicts choices, produces human-like learning curves, and predicts judgments about expected reward and confidence for unobserved nodes. Overall, these results extend the scope of previous theories of generalization in spatial (Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018; Schulz, Wu, Ruggeri, & Meder, 2018) and conceptual domains (Wu, Schulz, Garvert, Meder, & Schuck, 2018; Stojic, Schulz, Analytis, & Speekenbrink, 2018) to structured spaces.

Generalization on graph structures

We can specify a graph $G = (\mathcal{S}, \mathcal{E})$ with nodes $s_i \in \mathcal{S}$ and edges $e_i \in \mathcal{E}$ to represent a structured state space (Fig. 1a). Nodes represent states and edges represent allowed transitions. For now, we assume that all edges are undirected (i.e., if $x \to y$ then $y \to x$). The connectivity structure of the graph determines which states are accessible from a given prior state, and is often described using the graph Laplacian L:

$$L = D - A \tag{1}$$

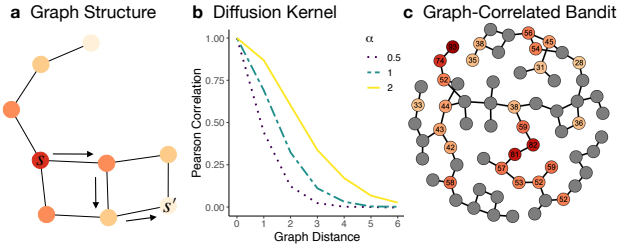

Figure 1: Inference over graphs. **a**) An example of a graph structure, where nodes represent states and edges indicate the transition structure. **b**) A diffusion kernel is a similarity metric between nodes on a graph, allowing us to generalize the value of unobserved nodes based on the assumption that correlations of rewards decays as an exponential function of the graph-distance between two nodes. The diffusion parameter (α) governs the rate of decay. **c**) Screenshot of our graph-correlated bandit task, where each node is an arm of a bandit with rewards correlated across the graph structure.

where A is the adjacency matrix and D is the degree matrix. Each element $a_{ij} \in A$ is 1 when nodes i and j are connected, and 0 otherwise, while the diagonals of D describe the number of connections of each node. The graph Laplacian can also describe graphs with weighted edges, where D becomes the weighted degree matrix and A becomes the weighted adjacency matrix.

The diffusion kernel

The diffusion kernel (DF; Kondor & Lafferty, 2002) defines a similarity metric $k(s, s')$ between any two nodes based on the matrix exponentiation of the graph Laplacian:

$$k(s, s') = \exp(\alpha L). \tag{2}$$

Intuitively, the diffusion kernel assumes that rewards diffuse along the graph similar to a heat diffusion process (i.e., by assuming a continuous random walk), with closely connected nodes assumed to have similar values. The parameter α models the level of diffusion, where $\alpha \to 0$ assumes complete independence between nodes, while $\alpha \to \infty$ assumes all nodes are perfectly correlated.

Gaussian Process regression

From the similarity metric defined by the diffusion kernel (Eq. 2), we use Gaussian Process (GP) regression (Rasmussen & Williams, 2006) to perform Bayesian inference on graph structures. A GP defines a distribution over functions $f : \mathcal{S} \to \mathbb{R}^n$ that map the state space $\mathcal{S}$ to real-valued scalar

outputs (e.g., rewards). Functions are modeled as a random draw from a multivariate normal distribution:

$$f \sim \mathcal{GP}(m,k), \tag{3}$$

where $m(s)$ is a mean function specifying the expected output of s, and $k(s,s')$ encodes prior assumptions about the underlying function. We use the diffusion kernel (Eq. 2) to represent covariance based on the connectivity structure of the graph (see Smola & Kondor, 2003; Kemp & Tenenbaum, 2009, for alternative implementations).

Given some observations $\mathcal{D}_t = \{\mathbf{s}_t, \mathbf{y}_t\}$ of observed rewards $\mathbf{y}_t$ at states $\mathbf{s}_t$, we can compute the posterior distribution $p(f(s_*)|\mathcal{D}_t)$ for any target state s_*. The posterior is a normal distribution with mean and variance defined as:

$$m(s_*|\mathcal{D}_t) = \mathbf{k}_{t,*}^\top (\mathbf{K}_t + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{y}_t \tag{4}$$

$$v(s_*|\mathcal{D}_t) = k(s_*,s_*) - \mathbf{k}_{t,*}^\top (\mathbf{K}_t + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{k}_{t,*}, \tag{5}$$

where $\mathbf{K}_t$ is the $t \times t$ covariance matrix evaluated at each pair of observed inputs, and $\mathbf{k}_{t,*} = [k(s_1,s_*),\ldots,k(s_t,s_*)]$ is the covariance between each observed input and the target input s_*, and σ_ε^2 is the noise variance. Thus, for any node in the graph, we can make Bayesian predictions about the expected reward $m(s_*|\mathcal{D}_t)$ and the uncertainty $v(s_*|\mathcal{D}_t)$ attached to the prediction.

Experiment: Graph-correlated bandit

We designed a task where rewards were defined by the connectivity structure of a graph. Participants searched for rewards by clicking the nodes of a graph, where connections between nodes influenced rewards. This provided a correlated reward structure allowing for similarity-based generalization to aid in search, but where similarity was defined based on connectivity structure rather than perceptual features.

Methods

Participants and design. We recruited 100 participants on Amazon MTurk (requiring 95% approval rate and 100 previously completed HITs). Two participants were excluded because of missing data, leading to a total sample size $N = 98$ ($M_{age} = 34.3$; $SD = 8.7$; 32 female). Participants were paid \$2.00 for completing the task and earned an additional performance contingent bonus of up to \$3.00. Overall, the task took 7.2 $\pm$ 3.3 minutes and participants earned \$4.32 $\pm$ \$0.24 on average.

Materials and procedure. Participants were instructed to earn as many points as possible by clicking on the nodes of a graph (Fig. 1c). Expected rewards were defined by a graph-correlated structure, such that connected nodes generated similar rewards. Along with instructions, participants were shown four fully revealed graphs to familiarize them with the reward structure and answered three comprehension questions before starting the task.

The task was performed over 10 rounds, each corresponding to a different graph structure (8x8 lattice graphs with 40% edges randomly pruned). In each round, participants were initially shown a single randomly revealed node, and had 25 clicks to either explore unrevealed nodes or to reclick previously observed nodes. Each clicked node displayed the numerical value (most recent observation if multiple selections) and a color aid, where darker colors corresponded to larger rewards. Participants were informed about their performance after each round as a percentage of the best possible score (w.r.t. the global optimum). The final performance bonus (up to \$3.00) was also calculated based on this percentage, averaged over all rounds.

Judgments. Participants were informed that the last round was a "bonus round", where the goal of maximizing points remained the same. However, after 20 clicks, participants were shown a series of 10 unrevealed nodes and asked to make judgments about the expected reward and their confidence. After making the judgments, participants were forced to choose one of the 10 options, and then the task was completed as normal. Behavioral and modeling results exclude the bonus round, except for the analyses of the judgment data.

Results

Participants achieved higher rewards over successive trials ($r = .93$, $p < .001$, $BF > 100$; Fig. 2a) and decisively outperformed a random baseline ($t(97) = 29.6$, $p < .001$, $d = 3.0$, $BF > 100$). We found no influence of round number on performance ($r = .49$, $p = .182$, $BF = 1$), indicating that the fully revealed environments in the instructions and comprehension questions were sufficient for conveying the goal of the task and the correlational structure.

Model Comparison

We used computational modeling to make predictions about choices and participants' judgments in order to understand how subjects reasoned about graph-correlated environments. Models were fit using leave-one-round-out cross validation, and then compared using the summed out-of-sample prediction accuracy of the left-out rounds. Altogether, we compared five different models corresponding to different strategies for generalization and exploration (see below).

Each model computes a value for each option $q(s)$, which is then transformed into a probability distribution using a softmax choice rule $P(s_i) = \exp(q(s_i)/\tau)/\sum_j \exp(q(s_j)/\tau)$, where the temperature parameter τ is a free parameter controlling the level of random exploration. In addition, all models use a stickiness parameter ω that adds a bonus onto the value of the most recently chosen option and is a common feature of reinforcement learning models (e.g., Gershman, Pesaran, & Daw, 2009).

Gaussian process with diffusion kernel. The Gaussian Process (GP) model uses the diffusion kernel (Eq. 2) to make predictive generalizations about rewards, where we fit α as a free parameter defining the extent to which generalizations

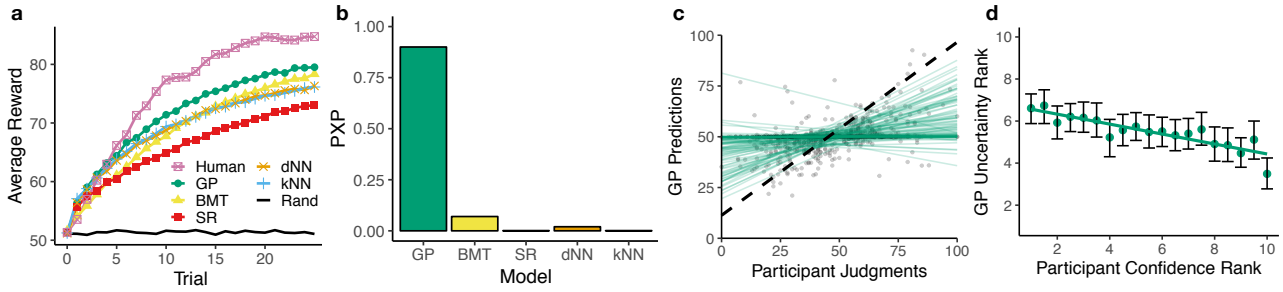

Figure 2: Results. **a**) Participant performance (pink) compared against models simulated by sampling (with replacement) from participant parameter estimates (10k replications). The black line provides a random baseline. **b**) Model comparison. Y-axis shows the protected exceedance probability (PXP), describing prevalence of each model in the population (corrected for chance) based on the out-of-sample predictive accuracy. **c**) The correspondence between each bonus round judgment and GP model predictions (using median per participant parameter estimates from other rounds). Each dot is one data point and each green line is a linear regression at the individual level. Dashed black line shows the fixed effect of a group-level mixed effects regression (with participant as a random effect). **d**) The correspondence between rank ordered (per participant) confidence judgments and model uncertainty estimates. Dots and error bars show the mean and 95% CI, while the colored lines represent a linear regression.

diffuse along the graph structure. To model how participants balance exploiting high value rewards with exploring highly uncertain options, we use upper confidence bound (UCB) sampling (Auer, 2002):

$$q_{\text{UCB}}(s) = m(s) + \beta\sqrt{v(s)}, \tag{6}$$

where the exploration bonus β is a free parameter governing the level of exploration directed towards uncertain options.

Bayesian mean tracker. The Bayesian mean tracker (BMT) is a prototypical reinforcement learning model that can be interpreted as a Bayesian variant of the Rescorla-Wagner model (Rescorla & Wagner, 1972; Gershman, 2015). The BMT also produces normally distributed predictions of reward $m(s)$ and $v(s)$ for each node, but are learned independently without generalization. Predictions of unobserved nodes defaulted to a prior of $m_0 = 50$ and $v_0 = 500$. The BMT has the error variance σ_ε^2 as a free parameter, which can be interpreted as inverse sensitivity. Smaller values result in larger updates to the learned mean $m(s)$ and larger reductions of uncertainty $v(s)$. The BMT also uses UCB as a sampling strategy, along with stickiness and a softmax choice rule.

Successor representation. The successor representation (SR; Dayan, 1993) is a reinforcement learning model that performs generalization based on building a predictive map of the connection structure. The successor representation matrix $M(s,s') = (I - \gamma T)^{-1}$ models the similarity of node s to node s' based on future expected state occupancy, where we assume a random walk policy by setting the transition matrix T to the row normalized graph Laplacian $T = I - D^{-1}L$. The extent of generalization is governed by the temporal discount parameter γ, which we treat as a free parameter.

While the SR has theoretical equivalencies to the diffusion kernel (Stachenfeld, Botvinick, & Gershman, 2014; Machado et al., 2018), there are practical differences when computed on finite graphs and also by modeling the extent of generalization using the temporal discount rate γ rather than the diffusion parameter α. Additionally, the SR only makes predictions about expected value

$$m(s) = \sum_{s'} M(s,s')R(s'), \tag{7}$$

where $R(s')$ is the observed reward at state s'. Because there are no uncertainty estimates, the SR does not implement any directed sampling using UCB. Instead, we set $q(s) = m(s)$ and apply stickiness along with a softmax choice rule.

Nearest neighbors models. In addition to reinforcement learning models, we also consider two simple nearest neighbor averaging models. The d-nearest neighbors (dNN) model estimates expected reward for unobserved node by averaging the rewards of all observed nodes within a distance of d. The k-nearest neighbors (kNN) model estimates expected reward by averaging the observed rewards for the k nearest nodes, including all ties. Both d and k are estimated as free parameters. For predictions where no observed nodes satisfied the averaging rule (i.e., all observations were too far away), we defaulted to an expected value of $m(s) = 50$. Both dNN and kNN also apply stickiness and use a softmax choice rule.

Model results

We compared model in terms of predictive accuracy using out-of-sample loss. Figure 2b shows the relative performance of each model in terms of the protected exceedance probability (PXP), which is a Bayesian model selection framework for estimating the probability that a given model is more prevalent in the population than all others, corrected for chance (Rigoux, Stephan, Friston, & Daunizeau, 2014). Overall, the GP had the highest predictive accuracy, with an exceedance probability of PXP=.90.

We also simulated the behavior of each model by sampling (with replacement) from the set of participant parameter estimates (10k samples) and computing the average learning curves (Fig. 2a). Although all models under-performed compared to the human curves, the GP had the closest match.

Judgments. To provide additional support for our modeling results, we predicted participant judgments in the bonus round

using parameters estimated over all rounds except the bonus round. Comparing participant and model predictions about expected reward, the GP had the highest average correlation ($r = .41$; Fig. 2c), which was better than the dNN (comparing Z-transformed correlation coefficients: $t(97) = 3.0$, $p = .004$, $d = 0.2$, $BF = 7$), and kNN models ($t(97) = 3.0$, $p = .003$, $d = 0.2$, $BF = 8$), but equally good as the SR ($t(97) = 0.1$, $p = .901$, $d = 0.0$, $BF = .11$). This is intuitive, because the GP and the SR should generate close-to-equivalent mean predictions in our task. Correlations are undefined for the BMT, since it invariably makes the same prediction.

Additionally the GP uncertainty predictions were predictive of participant confidence ratings (Fig. 2d; see also Wu, Schulz, & Gershman, 2019). Using mixed effects regression to predict the raw confidence judgment (Likert scale 1-11) using the GP uncertainty estimate as a fixed effect and participant as a random effect, we find higher GP uncertainty estimates predicted lower confidence ratings ($\beta = -.30$, $t(414) = -5.7$, $p < .001$, $BF > 100$). The BMT assumes the same level of uncertainty for all unobserved nodes (i.e., making no predictions about confidence), while none of the other models represent uncertainty.

Conclusion

We studied how people generalize in structured spaces, where the transition structure rather than the singular stimuli features define the distribution of rewards in the environment, extending previous work (Wu, Schulz, Speekenbrink, et al., 2018). We find that a Gaussian process (GP) model using the diffusion kernel is able to capture how people use generalization to guide search in structured environments. The GP provides the best predictive accuracy of choices, produces similar learning curves to human performance, and can robustly predict judgments about expected reward and confidence. While the SR matches the GP in terms of the correspondence between participant judgments and model predictions, it performed less well in predicting choices and in simulating human-like learning curves. Thus, while there is a theoretical equivalency between the SR and the diffusion kernel, the ability to estimate uncertainty within the GP framework gives it a clear advantage in describing search behavior.

Acknowledgments

ES is supported by the Harvard Data Science Initiative. This material is based upon work supported by the Office of Naval Research (N00014-17-1-2984).

References

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, *3*, 397–422.

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*, e1004567.

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, *29*, 13524–13531.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning* (pp. 315–322).

Machado, M. C., Rosenbaum, C., Guo, X., Liu, M., Tesauro, G., & Campbell, M. (2018). Eigenoption discovery through the deep successor representation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Rasmussen, C. E., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press: Cambridge, MA.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studiesrevisited. *Neuroimage*, *84*, 971–985.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2018). Searching for rewards like a child means less generalization and more directed exploration. *bioRxiv preprint*.

Smola, A. J., & Kondor, R. (2003). Kernels and regularization on graphs. In *Learning theory and kernel machines* (pp. 144–158). Springer.

Stachenfeld, K. L., Botvinick, M., & Gershman, S. J. (2014). Design principles of the hippocampal cognitive map. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2528–2536).

Stojic, H., Schulz, E., Analytis, P. P., & Speekenbrink, M. (2018). Its new, but is it good? how generalization and uncertainty guide the exploration of novel options. *PsyArXiv*.

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). Connecting conceptual and spatial search via a model of generalization. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1183–1188). Austin, TX: Cognitive Science Society.

Wu, C. M., Schulz, E., & Gershman, S. J. (2019). Generalization as diffusion: human function learning on graphs. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society.*

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*, 915–924.