A Mathematical Theory of Relational Generalization in the Face of Exceptions

Luke Cheng (lc3616@columbia.edu), Samuel Lippl (sl4742@columbia.edu)
Center for Theoretical Neuroscience, Columbia University, New York, USA

Abstract

Relational reasoning is a cornerstone of higher-order cognition in humans and animals, enabling zero-shot generalization to novel situations using rules like transitivity. In the real world, agents need to flexibly decide when to apply these rules and when to learn exceptions from them. It has remained unclear how standard learning systems can accomplish this. To investigate this topic, we introduce a new task paradigm: transitive inference with exceptions. This requires subjects to infer an ordered relation and generalize using the transitive rule but also requires them to memorize a certain violation to this rule. We use a standard statistical learning system to understand the minimal inductive biases necessary to perform this task. Intriguingly, these models can generalize where possible and memorize exceptions where necessary. However, successful generalization depends on their representational geometry: an overly conjunctive representation yields a systematic pattern of errors in generalization. Ultimately, we introduce a novel task paradigm for understanding relational reasoning in the real world, explain how a standard learning system can generalize on this task, and make systematic predictions for human behavior.

Keywords: relational cognition; transitive inference; rule learning

Introduction

Humans and animals can use rules like transitivity to solve relational problems (for example, if $A > B$ and $B > C$, then $A > C$) (Halford et al., 2010). Yet, the real world exhibits many exceptions to these simple rules. Exceptions in rule learning offer a way to mechanistically break down a more complex reasoning process. In behavioral studies, exceptions drove computational models of human behavior to construct exception representations that were both differentiated from normal items and integrated based on shared characteristics (Xie & Mack, 2024). Although learning systems need suitable "relational inductive biases" to learn such rules for generalization (Battaglia et al., 2018), it's currently unclear how they can both flexibly learn relational rules and memorize exceptions to those rules.

To investigate this problem, we introduce a new task paradigm, transitive inference (TI) with exceptions, and analyze how a standard statistical learning model (a kernel model) behaves in this task. Transitive Inference (TI) is a classical relational reasoning task that tests the ability to generalize a relational order across objects (McGonigle & Chalmers, 1977). Subjects learn from a set of adjacent pairwise comparisons for which there is an implicit rank order ($A > B$, $B > C$, ..., $F > G$). After training, subjects are tested on their performance on non-adjacent pairs (e.g. AD; see Fig. 1). In our new task, subjects are trained on an additional violation to the transitive rule (e.g. $E > C$). This creates a non-transitive ordering enclosed by the exception.

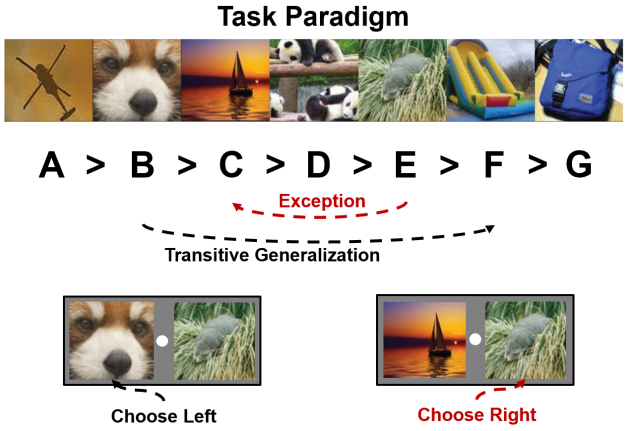


Figure 1: TI with exceptions paradigm. Training cases consist of adjacent items (e.g. $A > B$, $B > C$) plus one exception pair (e.g. $E > C$).

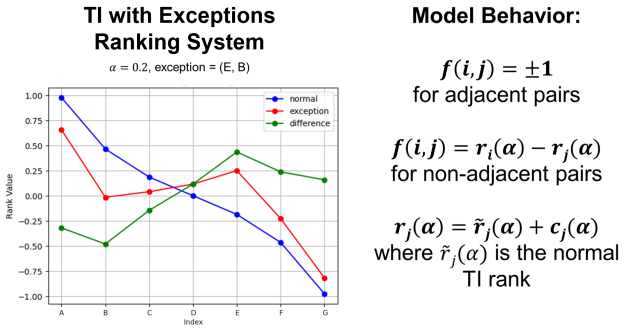
Methods

We employ a minimal models approach to investigate our question. We assume that each input is represented by a one-hot vector (X) for each of the two items. The simplest model, a direct linear readout, can only encode transitive relations (Lippl et al., 2024) and is therefore unable to memorize the exception. Thus, we increase the complexity by one step, considering a random weights neural network with a fixed hidden representation $g(X,Y)$ and a learned linear readout $w \circ g(X,Y)$. We assume that $f(X,Y) > 0$ implies $X > Y$ and $f(X,Y) < 0$ implies $X < Y$. On the training set, we train the model to output $f(X,Y) = 1$ if $X > Y$ and $f(X,Y) = -1$ if $Y > X$, using gradient descent over mean squared error. At convergence, these models identify the readout weights capturing the training set with minimal L^2-norm, a standard statistical inductive bias (Gunasekar et al., 2018). Finally, we assume that larger margins (i.e. higher magnitude of the output) indicate greater performance, as measured by e.g. reaction time or accuracy.

Previous work has shown that in the infinite width limit, the trial-by-trial similarity fully determines model behavior. The representational similarity of trials $\langle g(X,Y), g(X',Y') \rangle$ converges to simply 3 values, depending on whether trials are identical ($X = X'$ and $Y = Y'$), overlapping ($X = X'$ or $Y =$

Y'), or distinct ($X \neq X'$ and $Y \neq Y'$). The specific values are determined by the network architecture (Lippl et al., 2024).

Results



Model Behavior:

$\boldsymbol{f(i,j) = \pm 1}$
for adjacent pairs

$\boldsymbol{f(i,j) = r_i(\alpha) - r_j(\alpha)}$
for non-adjacent pairs

$\boldsymbol{r_j(\alpha) = \tilde{r}_j(\alpha) + c_j(\alpha)}$
where $\tilde{r}_j(\alpha)$ is the normal TI rank

Figure 2: Emergent ranking system learned by the model. ($\alpha = 0.2$, exception pair of BE).

Standard statistical learning models implement memorization of the exception and a flexible transitive inductive bias

Lippl et al. (2024) previously found that this standard statistical learning model generalizes on the standard TI task by learning an implicit and emergent ranking system. Here we find that the model also learns such a ranking system on TI with exceptions, determining its test output for non-adjacent items by subtracting the ranks for each item: $f(X,Y) = r(X) - r(Y)$. This ranking system majorly depends on α, the conjunctivity factor. Intuitively, α encodes whether the network ranges from fully "itemic" ($\alpha = 0$) to fully conjunctive ($\alpha = 1$).

Comparing the ranking systems on TI vs TI with exceptions shows an additional anti-tonic ranking as a result of the exception (Fig. 2). Analysis shows that this second ranking system is always counter to the original TI ranking system, which predicts that TI with exceptions will always be a more difficult task (i.e. a task yielding smaller margins).

By varying the conjunctivity factor, we can determine when generalization first fails for a fixed list length and exception position. For example, through numerical simulations, on an item list of 7 and the exception pair of CE, we find that an $\alpha < 0.3$ is necessary for generalization (see Fig. 3). Intriguingly, our analysis shows that increasing the generalizable list length (i.e. the number of items outside the exception length) greatly increases the difficulty of the task compared to increasing the exception length.

Model predictions for human and animal behavior

Ultimately, our model and analytic expression allow us to make predictions on human and animal behavior on this new task. Since adding an exception decreases the margin across trials, TI with exceptions will always have a higher task difficulty than without exceptions. Additionally, due to the reversal in the

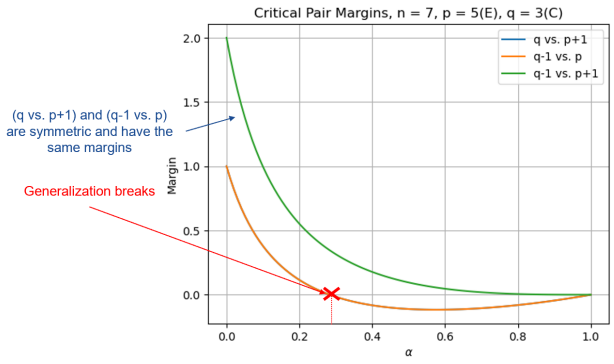


Figure 3: We can determine the smallest α for which the model fails to generalize, as a general measure of task difficulty. This is determined using critical pairs, the first pair of items which fail. The model will generalize between $0 < x < \alpha_{breaking}$. The exception pair is PQ.

rank representation, items later in the hierarchy that are enclosed by the exception will have comparatively higher ranks. This is another direct violation of the "symbolic distance" effect, similar to the violation caused by the "memorization" effect in the original TI task (Vasconcelos, 2008). Notably, this added complexity points to an area of the list that could plausibly develop a more complex separate representation to deal with non-transitivity.

Discussion

Our work contributes to a systematic understanding of relational generalization in the face of exceptions — an important yet understudied ability. We found that despite their simplicity, standard statistical learning models are able to generalize successfully on this task by memorizing violations to transitive rules while also flexibly adapting its own transitive rule. By deriving exact analytical solutions, we pinpoint the mechanism by which these models accomplish this. In particular, this analysis highlights that successful generalization depends on their representational geometry. We also make several systematic behavioral predictions that should be tested in human experiments. Finally, we have used generic statistical learning models, drawing a potential connection to behavioral paradigms investigating probabilistic statistical learning (as these same models could be used to model those paradigms (Jäkel et al., 2009; Seger and Peterson, 2013; Willmore et al., 2010)). Future work should investigate whether representation learning mechanisms can overcome the model's sensitivity and test these model predictions in human behavioral studies.

References

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani,

A., Allen, K., Nash, C., Langston, V., ... Pascanu, R. (2018, October 17). Relational inductive biases, deep learning, and graph networks. https://doi.org/10.48550/arXiv.1806.01261

Gunasekar, S., Lee, J., Soudry, D., & Srebro, N. (2018). Characterizing implicit bias in terms of optimization geometry [ISSN: 2640-3498]. *Proceedings of the 35th International Conference on Machine Learning*, 1832–1841. Retrieved June 11, 2025, from https://proceedings.mlr.press/v80/gunasekar18a.html

Halford, G. S., Wilson, W. H., & Phillips, S. (2010). Relational knowledge: The foundation of higher cognition. *Trends in Cognitive Sciences*, *14*(11), 497–505. https://doi.org/10.1016/j.tics.2010.08.005

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*(9), 381–388. https://doi.org/10.1016/j.tics.2009.06.002

Lippl, S., Kay, K., Jensen, G., Ferrera, V. P., & Abbott, L. F. (2024). A mathematical theory of relational generalization in transitive inference. *Proceedings of the National Academy of Sciences*, *121*(28), e2314511121. https://doi.org/10.1073/pnas.2314511121

McGonigle, B. O., & Chalmers, M. (1977). Are monkeys logical? *Nature*, *267*(5613), 694–696. https://doi.org/10.1038/267694a0

Seger, C. A., & Peterson, E. J. (2013). Categorization = decision making + generalization. *Neuroscience & Biobehavioral Reviews*, *37*(7), 1187–1200. https://doi.org/10.1016/j.neubiorev.2013.03.015

Vasconcelos, M. (2008). Transitive inference in non-human animals: An empirical and theoretical analysis. *Behavioural Processes*, *78*(3), 313–334. https://doi.org/10.1016/j.beproc.2008.02.017

Willmore, B. D. B., Prenger, R. J., & Gallant, J. L. (2010). Neural representation of natural images in visual area v2. *Journal of Neuroscience*, *30*(6), 2102–2114. https://doi.org/10.1523/JNEUROSCI.4099-09.2010

Xie, Y., & Mack, M. L. (2024). Reconciling category exceptions through representational shifts. *Psychonomic Bulletin & Review*, 1–13. https://doi.org/10.3758/s13423-024-02501-8