A mathematical theory of relational generalization in transitive inference

Samuel Lippl[1,2,3,*], Kenneth Kay[1,2,4], Greg Jensen[1,3,5], Vincent P. Ferrera[1,3,6], L.F. Abbott[1,2,3]

1 Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, NY,
2 Center for Theoretical Neuroscience, Columbia University, NY,
3 Department of Neuroscience, Columbia University Medical Center, NY,
4 Grossman Center for the Statistics of Mind, Columbia University, NY,
5 Department of Psychology at Reed College, OR,
6 Department of Psychiatry, Columbia University Medical Center, NY.

* Correspondence: samuel.lippl@columbia.edu

Abstract

Humans and animals routinely infer relations between different items or events and generalize these relations to novel combinations of items ("compositional generalization"). This allows them to respond appropriately to radically novel circumstances and is fundamental to advanced cognition. However, how learning systems (including the brain) can implement the necessary inductive biases has been unclear. Here we investigated transitive inference (TI), a classic relational task paradigm in which subjects must learn a relation ($A > B$ and $B > C$) and generalize it to new combinations of items ($A > C$). Through mathematical analysis, we found that a broad range of biologically relevant learning models (e.g. gradient flow or ridge regression) perform TI successfully and recapitulate signature behavioral patterns long observed in living subjects. First, we found that models with item-wise additive representations automatically encode transitive relations. Second, for more general representations, a single scalar "conjunctivity factor" determines model behavior on TI and, further, the principle of norm minimization (a standard statistical inductive bias) enables models with fixed, partly conjunctive representations to generalize transitively. Finally, neural networks in the "rich regime," which enables representation learning and often leads to better generalization, deviate in task behavior from living subjects and can make generalization errors. Our findings show systematically how minimal statistical learning principles can explain the rich behaviors empirically observed in TI in living subjects, uncover the mechanistic basis of transitive generalization in standard learning models, and lay out a formally tractable approach to understanding the neural basis of relational generalization.

1 Introduction

Humans and animals have a remarkable ability to generalize to circumstances that are radically different from their prior experience. They are able to do so by conceptualizing prior experience in terms of components that can be re-configured in a novel situation [**"compositional generalization,"** 1–6]. An important instance of inherently compositional learning is **"relational cognition"** [7], which draws connections between different events or items. For example, subjects can reason about social relationships between individuals they have never seen interact [8–10], take a novel route between familiar locations [11], or apply familiar tools to novel problems [12]. Accordingly, relational cognition and compositional generalization have been implicated in a variety of fields concerned with learning, including social cognition [13], spatial navigation [14], and logical and causal cognition [15].

Despite the wide-ranging importance of relational cognition and compositional generalization, it remains unclear how living subjects, and learning systems more generally, implement these abilities. To generalize from their limited experience, subjects (whether they are humans, animals, or learning models) need an "inductive bias": a disposition towards certain behaviors among the many that are consistent with past experience [16]. Much attention has been devoted to clarifying suitable inductive biases on standard statistical tasks, which require generalization to nearby data points ("near transfer" or "in-distribution generalization") [17–22]. An important instance of such a "statistical inductive bias" is norm minimization, which selects the model parameters with the smallest weight norm and is at the core of many successful learning models [23–25]. Both theoretical and practical insights suggest that such a solution is likely to generalize well in distribution [26, 27]. In contrast, our understanding of how learning models perform relational and compositional tasks, which require generalization to radically different circumstances ("far transfer," [28] "out-of-distribution generalization," [29] or, more specifically, "generalization to the unseen" [30]), has been

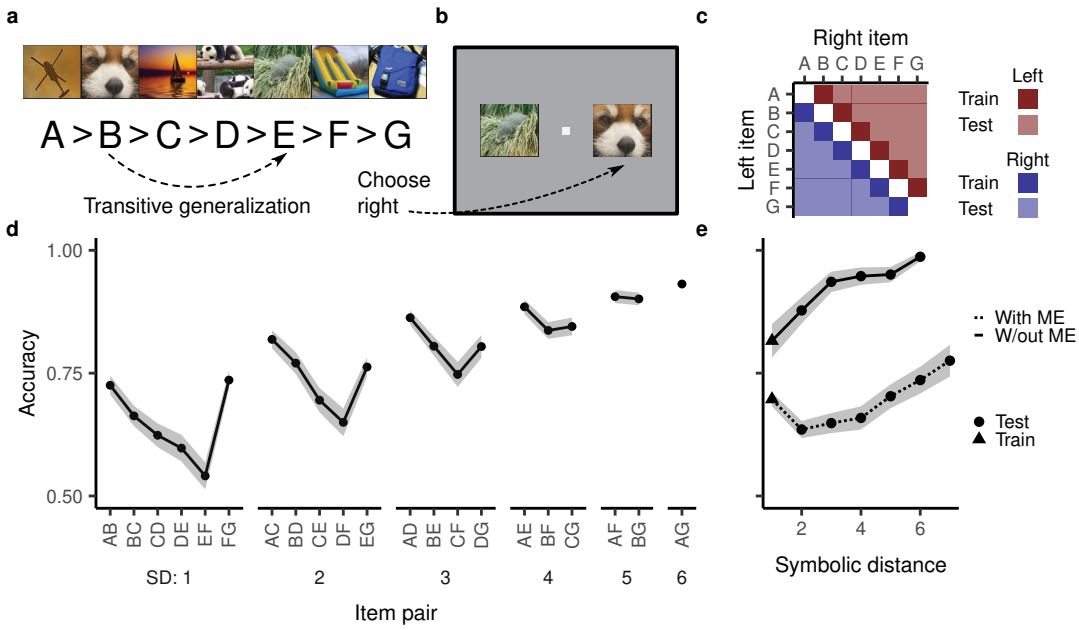

Figure 1: Transitive inference and behavioral patterns observed in subjects performing this task. **a**, Example stimuli taken from Jensen *et al.* [32]. An example generalization ($B > E$) is highlighted. **b**, In a given trial, the subject is asked to choose between the two presented items. They are rewarded if the chosen item is "larger" according to the underlying hierarchy. This panel depicts a test trial where successful performance would consist in picking the item on the right. **c**, Schematic of the training and test cases. **d**, Example accuracy on all training and test cases (in this case by rhesus macaques). Terminal item, symbolic distance, and asymmetry effect are apparent in the plot. In the subsequent figures, we leave off the item pair labels but use the same ordering. Symbolic distance (SD) is the separation in the rank hierarchy. The data is reproduced from [32]. **e**, Symbolic distance effect with and without a memorization effect (ME). Data without ME reproduced from [33] and data with ME reproduced from [32]. Shaded regions in both panels indicate mean $\pm$ one standard error.

much more limited [31]. Addressing this gap is essential to understanding how living subjects perform these tasks.

To investigate this question, we studied transitive inference [TI; Fig. 1a; 34–37], a classical cognitive task that tests if humans or animals can generalize transitively. In this task, subjects are presented with pairs of items (Fig. 1b) and must pick the "larger" item according to an implicit hierarchy ($A > B > \cdots > G$, Fig. 1a). Importantly, subjects are not informed of this task structure [cf. 33, 38] and, further, only receive feedback about the correct response on adjacent pairs ($AB, BC, \ldots, FG$). Hence, they must infer the underlying relation and use transitive inference to determine the correct response to non-adjacent pairs ($B > E$, Fig. 1c).

Notably, humans [39] and a variety of animals (ranging from monkeys [40, 41] and rodents [42] to wasps [43] and fish [44]) perform TI successfully. Moreover, they show consistent behavioral patterns (in their reaction time and accuracy, Fig. 1d), which may be informative of the underlying neural implementation. First, subjects' performance improves with increasing separation in the hierarchy [**"symbolic distance effect,"** e.g. 39, 45–49]. Second, subjects' performance tends to be better for trials involving more terminal (e.g. A and G) rather than more intermediate (e.g. D and E) items [**"terminal item effect,"** e.g. 32, 50–52]. Third, some studies indicate better performance on training trials (which have a symbolic distance of 1) than on test trials with a symbolic distance of 2 [33, 49, 53–55, Fig. 1e], a limited violation of the symbolic distance effect which we here refer to as the **memorization effect**. Finally, subjects' performance is often better for item pairs towards the beginning of the hierarchy than item pairs towards the end (the **"asymmetry effect,"** which we do not address but return to in the discussion) [e.g. 33].

Many simple learning models have been shown to perform transitive inference. However, the inductive biases that give rise to this ability are not well understood. Various learning models associate a numerical

"rank" with each presented item and choose whichever item has a higher rank [32, 33, 56–58]. Such models are pre-configured to generalize transitively and leave unclear how transitive generalization could arise from more basic learning principles. They also leave unclear how the brain, which is not constrained in this way [59], could implement TI. Intriguingly, several studies have found that generic neural networks can generalize transitively, suggesting that statistical learning principles can sometimes give rise to a suitable relational inductive bias [38, 60–62]. However, TI in neural networks has largely been studied through simulations, rather than analytically [cf. 63], raising the question of when and how statistical learning models can implement TI.

Here we show, via both analytical approaches and simulations, that a broad range of biologically relevant learning models generalize transitively and recapitulate the symbolic distance, terminal item, and memorization effect. We first consider models with "additive" representations that represent the two presented items independently, and show that they are constrained to implementing a transitive relation. If models additionally represent nonlinear conjunctions between items (as is important for many other tasks) but use norm minimization to determine their readout weights, they also generalize transitively. Remarkably, the same learning principle that underpins many instances of successful near transfer also enables this instance of successful far transfer. We further show that, for TI, the effect of a particular choice of internal representation can be characterized by a single scalar "conjunctivity factor." Finally, we consider models which adapt their internal representation to a given task, an ability thought critical to human and animal cognition. Surprisingly, we find that this impairs performance on TI and leads to behavioral patterns that deviate from those in living subjects.

At first glance, TI appears to be a complex task involving relational cognition and application of the transitive rule. Nevertheless, we can characterize a broad range of learning models performing this task in exact analytical terms, explaining how they give rise to the rich behavioral patterns observed in living subjects. In doing so, our investigation clarifies systematically how a learning principle that has largely been considered in the context of near transfer, also implements an important instance of far transfer and compositional generalization.

2 Model setup

We represent individual items as high-dimensional vectors. A trial input is a concatenation of the two vectors X, Y corresponding to the two presented items. We generally consider a learning model f which represents this input as a numerical vector $g(X, Y)$ (in the simplest case, this could be the input vector itself) and computes a linear readout from that representation: $f(X, Y) = w \circ g(\mathrm{X}, \mathrm{Y})$. A positive model output ($f(X, Y) > 0$) corresponds to $X > Y$, whereas a negative output ($f(X, Y) < 0$) corresponds to $X < Y$.

Inputs where $f(X, Y) = 0$ lie on the model's decision boundary. Accordingly, a higher magnitude of the margin ($f(X, Y)$ if $X > Y$ or $-f(X, Y)$ if $X < Y$) corresponds to a larger distance from the decision boundary. We take this to indicate better performance, corresponding to higher choice accuracy and lower reaction time. This is a standard assumption in models of decision making [38, 64–66]. For example, in a drift diffusion model [67], the output $f(X, Y)$ determines the model's drift rate. The higher the drift rate, the less susceptible the model is to noise (improving accuracy) and the faster it will cross its threshold (improving reaction time) [68].

3 Results

3.1 An additive representation yields transitive generalization

To perform TI, or indeed any kind of relational inference, a learning model's representation of items X and Y, $g(X, Y)$, should reflect the fact that X and Y are separate items. In the most extreme case, this would amount to an *additive* representation, where $g(X, Y)$ is a sum of two separate representations $g_1(X)$ and $g_2(Y)$: $g(X, Y) = g_1(X) + g_2(Y)$. A simple instance of this is a model architecture where nodes respond exclusively either to X or Y (Fig. 2a).

A change in one of the two items will leave half of the additive representation unchanged (Fig. 2b), implementing a kind of compositionality [69]. A linear readout from an additive representation, $f(X, Y) =$

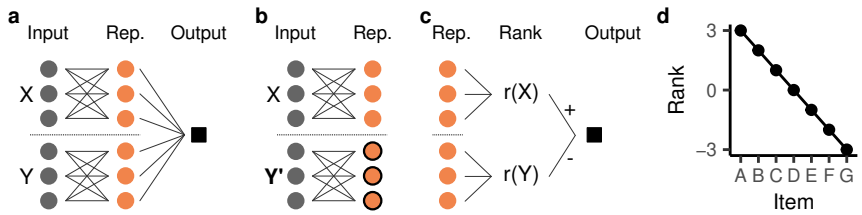

Figure 2: An additive representation yields transitive generalization and the symbolic distance effect. **a**, Schematic illustration of an additive representation (rep.). The first three orange nodes represent the first item (X) and the latter three represent the second item (Y). **b**, Replacing the second item only results in changes to the highlighted half of the units. **c**, The readout weights of the model can be grouped into those pertaining to item X and those pertaining to item Y. The model's output can be understood as assigning a "rank" $r(X)$ and $r(Y)$ to each item and then computing $r(X) - r(Y)$. **d**, Example of a model's rank assignment.

$w \circ g(X, Y)$, is a sum of responses to each individual item, $w \circ g_1(X) + w \circ g_2(Y)$, and therefore also additive. The consequences of this are especially clear in the case of a model without a choice bias, i.e. if $f(X, X) = w \circ g_1(X) + w \circ g_2(X) = 0$ (Appendix S1.1 considers the biased case). In this case, we know that $-w \circ g_1(X) = w \circ g_2(X)$, and, as a result, the model's decision function can be expressed as

$$f(X, Y) = w \circ g_1(X) - w \circ g_1(Y).$$

This means that the model necessarily learns to assign a scalar "rank" $r(X) = w \circ g_1(X) = -w \circ g_2(X)$ to each item, and computes its decision by comparing the two ranks (Fig. 2c):

$$f(X, Y) = r(X) - r(Y).$$

Note that $f(X, Y) > 0$, i.e. the decision to choose the left item, is equivalent to $r(X) > r(Y)$.

As shown in previous work [e.g. 32, 33, 50], learning systems that are pre-configured to have such a ranking system yield both transitive generalization and the symbolic distance effect. This is because, to learn the training set, the model's ranks must be monotonically decreasing:

$$r(A) > r(B) > \cdots > r(G).$$

As an example, if the model output has a margin of one for all inputs composed of adjacent items, the rank must decrease by one for each successive item (Fig. 2d). A monotonically decreasing rank directly implies that non-adjacent items are also ordered correctly and, consequently, that the model will generalize transitively. Further, item pairs with larger symbolic distance will have a larger difference in their ranks, giving rise to a symbolic distance effect.

Critically, and in distinction to previous work, the above model structure (additive representation) does not explicitly pre-configure or assume a ranking system. Rather, the above analysis shows that any learning model that implements an additive input representation necessarily implements a ranking system and thus obeys transitivity. Further, the above analysis makes no assumptions about how the model learns. As long as the model correctly classifies the training cases, it will correctly classify all test cases, i.e. transitively generalize, and it will exhibit a symbolic distance effect.[1]

3.2 A single scalar fully characterizes a broad range of relational representations

The above analysis indicates that an additive representation would enable the brain to perform TI. However, neural representations are not thought to be fully additive [70]. Indeed, non-additive representations (Fig. 3a) are important for learning relevant tasks across a broad range of domains and, with some differences in

[1] A related paper [62] has come to our attention, in which the authors are concerned with a linear item representation. This is a special case of the additive representations here. Further, the authors derive an analytic solution for a specific learning algorithm, whereas we demonstrate a transitive constraint for a broader class of models.

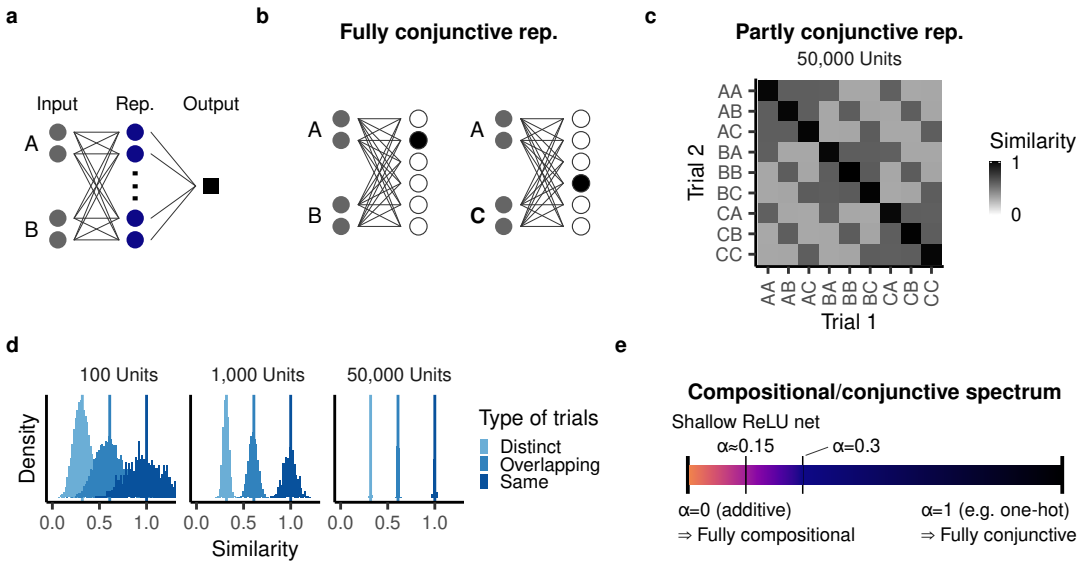

Figure 3: Non-additive representations of TI cases. **a**, A non-additive representation encodes nonlinear interactions between items. **b**, A one-hot representation represents each combination of items by a distinct node. **c**, Representational similarity (cross-correlation) between a subset of trials in a ReLU network with 50,000 units. **d**, Representational similarity between all possible trials in networks with different numbers of hidden units, organized according to the type of trials. **e**, The conjunctivity factor characterizes a given representation according to how similarly it represents overlapping trials. Additive representations lie at one end of the spectrum, whereas one-hot representations lie at the other end.

implementation, are known under a correspondingly broad range of names, including conjunctive [71] or configural [72] representations, nonlinear features, and representations with mixed selectivity [70]. We next asked whether and how non-additive representations can support transitive generalization.

To begin, consider the most extreme conjunctive case: a one-hot representation in which each composition of items is represented by a different hidden unit (Fig. 3b). In this case, a change in one of the two items yields a completely different representation. Such a model is able to memorize the training cases, but cannot generalize transitively.

Many representations, whether in the brain or in other learning systems, are neither fully conjunctive nor fully additive, but rather lie in between these two extremes [70, 73–76]. To characterize this spectrum formally, we considered the representational similarity between two trials (X, Y) and (X', Y'), as measured by their dot product, $\langle g(X, Y), g(X', Y') \rangle$. We assumed that the representational similarity between two trials only depends on whether these trials are distinct ((X, Y) and (X', Y')), overlapping ((X, Y) and (X', Y) or (X, Y')), or identical ((X, Y) and (X, Y)) (Fig. 3c). We call this the "exchangeability assumption."

The exchangeability assumption captures the fact that model behavior should not depend on the particular (i.e. arbitrary) hierarchy in which the set of items is arranged [77]. To promote exchangeability, we assumed that all input items are equally correlated with each other. In this case, most commonly used nonlinear representations of that input satisfy exchangeability as well. As a paradigmatic learning model, we considered a neural network with a ReLU nonlinearity and random weights. By determining the expected value of the representational similarity analytically (Appendix S1.3), we found that the network's hidden layer, in expectation, satisfies the exchangeability assumption (Fig. 3c). This is because even though the hidden layer computes a nonlinear transformation, it partially inherits the input's similarity structure [78]. In network simulations, the empirical representational similarity exhibits some variance around the expected value due to the random weight initialization. However, this variance vanishes as the network's hidden layer becomes wider (Fig. 3d).

Under the exchangeability assumption, we found that for learning models that only modify their readout weights, a single scalar, which we call the **"conjunctivity factor"** α, fully determines their TI task behavior, i.e. all models that have representations with the same α exhibit the same behavior on TI (see Appendix S1.2).

The conjunctivity factor is given by:

$$\alpha := 1 - 2\frac{\kappa_o - \kappa_d}{\kappa_s - \kappa_d}, \quad \kappa_s = \langle g(X,Y), g(X,Y)\rangle, \quad \kappa_o = \langle g(X,Y), g(X,Y')\rangle, \quad \kappa_d = \langle g(X,Y), g(X',Y')\rangle. \quad (1)$$

Here κ_s, κ_o, and κ_d are the similarity between identical, overlapping, and distinct pairs, respectively. Notably, we can extend the definition of the conjunctivity factor to non-exchangeable representations by taking the average over all identical, overlapping, or distinct pairs.

For an additive representation, half the units share their activation between different overlapping pairs. As a result, the similarity between overlapping pairs is halfway between that of distinct and that of identical pairs, corresponding to $\alpha = 0$. At the other extreme, $\alpha = 1$ indicates that overlapping pairs are encoded in the model with equal similarity to each other as completely distinct pairs (as is the case for a one-hot representation). Consequently, α systematically characterizes the spectrum from fully additive to fully conjunctive representations (Fig. 3e).

In most commonly used representations, overlapping trials are more similar to each other than distinct trials and therefore have an intermediate value for α ("partly conjunctive" representations). This is because the input space represents overlapping trials as more similar than distinct trials and, as noted above, the hidden layer partly inherits the input's similarity structure. Through analytical computation, we found that a random ReLU network with one hidden layer, for example, has $\alpha \approx 0.15$ (Eq. (S24); Fig. 3c,d).

Importantly, the network giving rise to the representation could have an arbitrary number of layers; all we need to know is the conjunctivity factor of the network's final layer. In fact, in Appendix S1.3, we compute the conjunctivity factor of neural networks with arbitrary depth, finding that α increases with depth. Indeed, as the networks become deeper, α eventually approaches one. However, if the connectivity structure is modified so a part of each layer is additive (but is, in turn, connected to the next layer's conjunctive representation as well), α can be reduced and, in fact, will eventually approach zero. This illustrates the crucial impact that network connectivity can have on representational structure.

For transitive inference, the conjunctivity factor raises two questions. First, how does $\alpha > 0$ affect TI behavior? Second, if the learning model's internal representation is modifiable rather than fixed, how does this affect the conjunctivity factor and subsequent TI behavior?

3.3 Norm minimization and partly conjunctive representations yield transitive generalization

Unlike an additive representation, a representation with $\alpha > 0$ is not constrained to implementing a transitive relation. To understand how models with partly conjunctive representations perform on TI, we need to consider additional constraints. In particular, we analyzed models in which the learning of readout weights implements norm minimization, a paradigmatic statistical inductive bias (Fig. 4a). Under the exchangeability assumption, we were able to characterize model behavior on TI through exact analytical solutions. As noted above, to perform this analysis, we do not need to know the particular representation implemented, but only its conjunctivity factor α.

On the training cases, the minimal norm model necessarily assigns a margin of ± 1, as dictated by the desired output. On the test cases, however, our analysis revealed an intriguing emergent behavior: the model's response to item pair (i, j) invariably reflects a ranking system:

$$f(i,j) = r_i(\alpha) - r_j(\alpha), \text{ with } r_i(\alpha) = \frac{\sinh((\frac{n+1}{2} - i)\lambda(\alpha))}{\sinh(\frac{n+1}{2}\lambda(\alpha)) - \sinh(\frac{n-1}{2}\lambda(\alpha))} \text{ and } \lambda(\alpha) := \operatorname{arccosh}\left(\frac{1}{1-\alpha}\right), \quad (2)$$

where n is the total number of items (see Appendix S1.4.4 for the derivation of this result). This is remarkable as the model architecture is not constrained or pre-configured to implement a ranking system. Rather, the behavior is a consequence of the principle of norm minimization, operating on a partly conjunctive representation. Importantly, as implied by its ranking system, the model generalizes transitively (as long as overlapping trials are represented as more similar than distinct ones, i.e. $\alpha < 1$), and it exhibits a symbolic distance effect.

For an intuition as to why a ranking system emerges, consider a particular class of representations having one-hot representations of the first and second item individually as well as their conjunction (Fig. 4b).

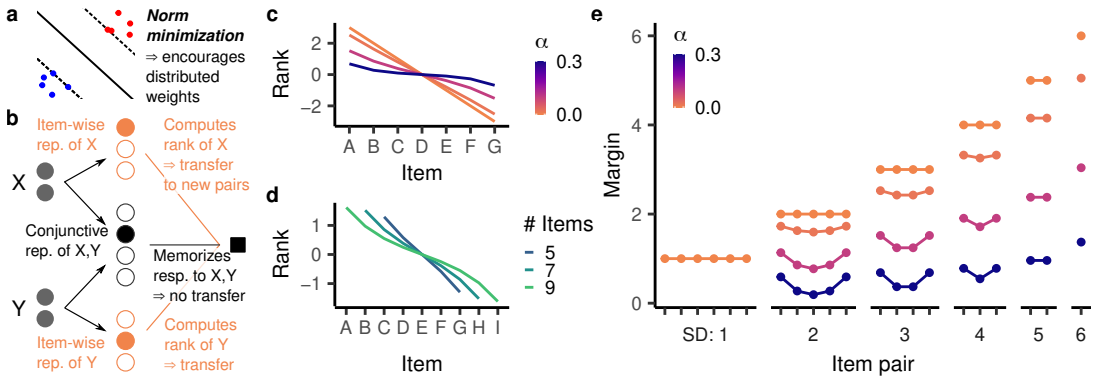

Figure 4: A scalar parameter ("conjunctivity factor") parameterizing compositional vs conjunctive representations determines TI behavior in models trained with gradient descent. **a**, Norm minimization implements a useful inductive bias for generalization to nearby data points. On categorization tasks, it determines the hyperplane separating the two categories with the maximal margin. **b**, Intuitively, a partly conjunctive representation is given by an item-wise representation of X and Y concatenated with a fully conjunctive representation of X and Y. The readout from the item-wise representations computes a rank for each item that transfers to overlapping pairs. The readout from the fully conjunctive representation memorizes a response to a given pair and does not transfer to overlapping pairs. Because norm minimization encourages distributed weights, it finds a solution that partly uses the item-wise representation and hence computes a rank. This leads to transitive generalization. **c**,**d**, The emergent rank representation at the end of training (Eq. (2)) for (c) seven items and different values for α; and (d) $\alpha = 0.1$ and different numbers of items. **e**, For seven items and different values for α, the corresponding margin for all trials. Item pairs are arranged by their position in the hierarchy, as in Fig. 1e.

This representation will have a conjunctivity factor of α if the item-wise units are weighted by $1 - \alpha$ and the conjunctive units are weighted by α. Changes in the item-wise unit weights correspond to changes in the model's rank, as they generalize to overlapping trials. In contrast, changes in the conjunctive unit weights correspond to memorization, as they have no effect on overlapping trials (akin to a fully conjunctive representation). In principle, the model could learn the training set through pure memorization, i.e. changes in the conjunctive unit weights alone. However, because more distributed weights tend to have a smaller weight norm, norm minimization causes the model to learn the training premises by changing both the conjunctive unit weights (resulting in memorization) and the item-wise unit weights (resulting in transitive generalization). Thus, the partial conjunctivity is necessary for the existence of an item-wise population; the principle of norm minimization ascertains that this population is implicated in the learning process.

Beyond transitive generalization, the principle of norm minimization gives rise to several empirically observed behavioral patterns. The hyperbolic sine making up the rank expression compresses more intermediate items more strongly than more terminal items (Fig. 4c), thus giving rise to a terminal item effect (Fig. 4e). This effect becomes stronger for higher values of α. A higher α also compresses the ranking more strongly overall, leading to lower margins on the test set. As α approaches one (the fully conjunctive case), the ranking becomes entirely flat and therefore no longer supports transitive generalization.

The form of the ranking also depends on the total number of items. Specifically, intermediate items are compressed more strongly when there are more items in total, an effect that is moreover dependent on α. For $\alpha = 0$, the ranking grows linearly with the number of items, whereas at higher values of α, the model ranking's overall range (between the first and last item) is almost invariant to the number of items (Fig. 4d).

Finally, when $\alpha > 0$, the model assigns a larger margin to the training cases than specified by the ranking. Intuitively, this is because the conjunctive unit weights contribute to model behavior on the training cases, but do not transfer to the test cases. Since a higher α compresses the ranks further, it leads to a higher discrepancy between training and test behavior. At a sufficiently high α, the margin of the training cases is larger than that of the test cases with a small symbolic distance (Fig. 4e), giving rise to a memorization effect. While Ciranka *et al.* [33] explained this effect by fitting an explicit memorization parameter, our analysis reveals that it is an emergent consequence of having a nonlinear representation

with sufficiently high conjunctivity factor. Notably, the fact that the memorization effect, according to our model, only arises in a subset of relational representations may explain why it is only occasionally observed in living subjects. In contrast, the symbolic distance and terminal item effect arise across the full spectrum of relational representations (except the fully conjunctive case) and indeed are also observed much more consistently in living subjects.

The above analytical solutions depend on the exchangeability assumption. In practice, the trials might be represented in a manner that violates this assumption. Indeed, we already saw that though randomly sampled features satisfy exchangeability in expectation, a model with insufficiently many of those features will have a non-exchangeable representation due to finite samples (Fig. 3d). In simulations (see Appendix S2.2), we found that a slight violation of exchangeability (e.g. resulting from a representation with many random features) does not change model behavior substantially. For larger violations (e.g. resulting from a representation with fewer random features), model behavior deviates from our theoretical account, but is still well approximated by it. Further, our account captured the average behavior across many models with a small number of random features almost exactly. This suggests that models with non-exchangeable representations behave differently from those with exchangeable representations but our analytical solutions can still be useful for understanding their behavior.

Our analysis highlights a potential tradeoff that learning multiple kinds of relations imposes on α. As noted above, a model with larger α has a smaller margin on the test dataset. It is not surprising that a small α is desirable on TI, a task which can be solved by a fully additive representation. However, for a small α, learning non-transitive relations such as transverse patterning [79, "rock, paper, scissors" with more than three items] is difficult or, for $\alpha = 0$, impossible. Thus, on a broader class of tasks, a higher value of α will enable learning non-transitive relations and nonlinear tasks more generally.

Our analysis demonstrates that norm minimization can explain not only transitive generalization, but also the symbolic distance effect, the terminal item effect (though only on test cases), and the memorization effect. We next characterize two popular statistical learning models implementing norm minimization: ridge regression and gradient flow, as applied to learning of network readout weights.

3.4 Learning through gradient flow or with weight regularization smoothly approaches the minimal norm solution

To see how the principle of norm minimization governs TI behavior across learning, we analyzed ridge regression [23] and gradient flow, using mean squared error as the loss function $L(w)$ (with the target response on the training cases being ± 1). Ridge regression minimizes the sum of this loss and the squared L_2-norm of the model weights, i.e.

$$L(w) + \tfrac{1}{c}\|w\|_2^2.$$

Here, the regularization coefficient c balances the weight penalty $\|w\|_2^2$ with the minimization of the loss function $L(w)$. Biologically, such weight regularization could, for instance, be implemented by mechanisms of synaptic homeostasis [80]. Gradient flow, on the other hand, assumes that a model minimizes $L(w)$ by following the pointwise gradient exactly. This approximates gradient descent with a small learning rate, and is more amenable to formal analysis [81]. In this case, the learning duration t determines how well the model has learned to minimize the loss function in the allotted time. At initialization, the model should be agnostic to all choices (i.e. output zero) and we therefore assumed that the weights are initialized at zero.

The regularization coefficient c and the learning duration t play a similar role in the two learning models. With $c = 0$, the weight penalization is infinitely more important than the task-based component of the loss function and therefore the model weights are all zero. Similarly, at $t = 0$, the weights are initialized to zero. On the other hand, in the limit of infinite training ($t \to \infty$), the model converges to the minimal norm solution determined in the section above [82]. This is also the case for the limit of models trained with increasingly small weight penalization ($c \to \infty$) [83]. Ridge regression and gradient flow are therefore two instances of common learning models which can perform TI by implementing the principle of norm minimization.

Going beyond these limits, we obtained exact solutions to model behavior for arbitrary t or c. For ridge regression (see Appendix S1.4.4), we found that the test behavior of a model with weight regularization and a given conjunctivity factor α is equivalent to that of a model without weight regularization but a different

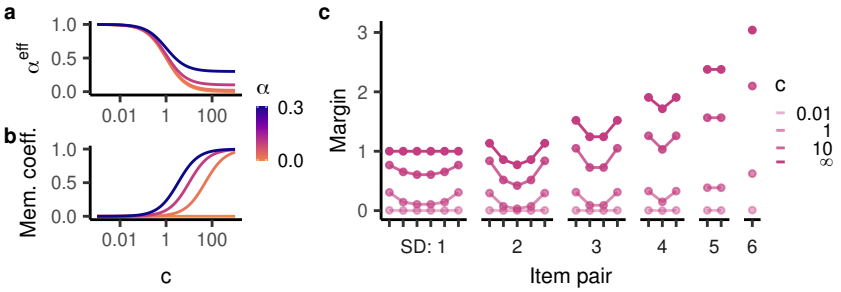

Figure 5: TI behavior for regularized regression. **a**,**b**, The effective conjunctivity factor (a) and the memorization coefficient (b) as a function of the inverse regularization coefficient c. **c** Generalization behavior for $\alpha = 0.1$ and different values of c. The margins overall become larger as c increases.

conjunctivity factor. This "effective conjunctivity factor" α^{eff} depends on both α and c (see Lemma S1.7):

$$\alpha^{\text{eff}} = \frac{\alpha + \frac{1}{c}}{1 + \frac{1}{c}}. \tag{3}$$

α^{eff} is generally higher (i.e. more conjunctive) than α and, as c becomes larger, gradually approaches α from above (Fig. 5a). Thus, a model with smaller c has a more compressed rank and a more pronounced terminal item effect.

On the training cases, the model's performance is boosted compared to the rank difference, just like in the case of norm minimization. Specifically, its behavior is given by

$$f(i, i \pm 1) = m \cdot \pm 1 + (1 - m) \cdot (r_i(\alpha^{\text{eff}}) - r_{i\pm 1}(\alpha^{\text{eff}})), \quad m := \frac{\alpha}{\alpha + \frac{1}{c}}, \tag{4}$$

i.e. a mixture of memorization (assigning a constant margin of ± 1) and reliance on the ranking system. The balance between the two behaviors is specified by the "memorization coefficient" $m \in [0, 1]$. m starts out at zero, indicating no memorization and full reliance on the ranking system. As c grows larger, m increases as well (Fig. 5b). The smaller m, the more strongly model behavior relies on the ranking. In particular, this partial reliance leads to a terminal item effect on the training cases in addition to the test cases (Fig. 5c). For $c \to \infty$, m converges to 1, indicating full memorization as observed for the minimal norm solution.

For gradient flow, we found qualitatively similar solutions for model behavior (Appendix S1.4.3). In particular, the model's behavior on test cases can be described by a ranking system throughout learning. Further, the model ranks gradually approach the minimal norm solution and, at earlier stages of training, have a more pronounced terminal item effect. Finally, the learning dynamics give rise to a transient terminal item effect on the training cases that vanishes as $t \to \infty$.

Humans and animals generally exhibit a terminal item effect on the training set. Our analysis suggests that this could be caused by either weight regularization or incomplete training (or both), arising from a mechanism that is related to but distinct from the mechanism giving rise to the terminal item effect on the test cases.

3.5 Neural networks with adaptive representations deviate from living subjects' TI behavior

Deep neural networks, which learn by updating their internal weights, have become increasingly relevant both as artificial learning systems [84] and models of cognitive processes [85, 86]. Importantly, recent work has revealed that their generalization behavior fundamentally depends on the magnitude of their initial weights [87, 88]. For large initial weights, learning dynamics can be approximated by gradient flow on a particular fixed-feature model called the neural tangent kernel [NTK; 89, 90]. Thus, even though the network updates its internal weights, it effectively still relies on a fixed representation; accordingly, this regime is often called "lazy" [87]. In contrast, neural networks initialized from sufficiently small values learn truly task-specific

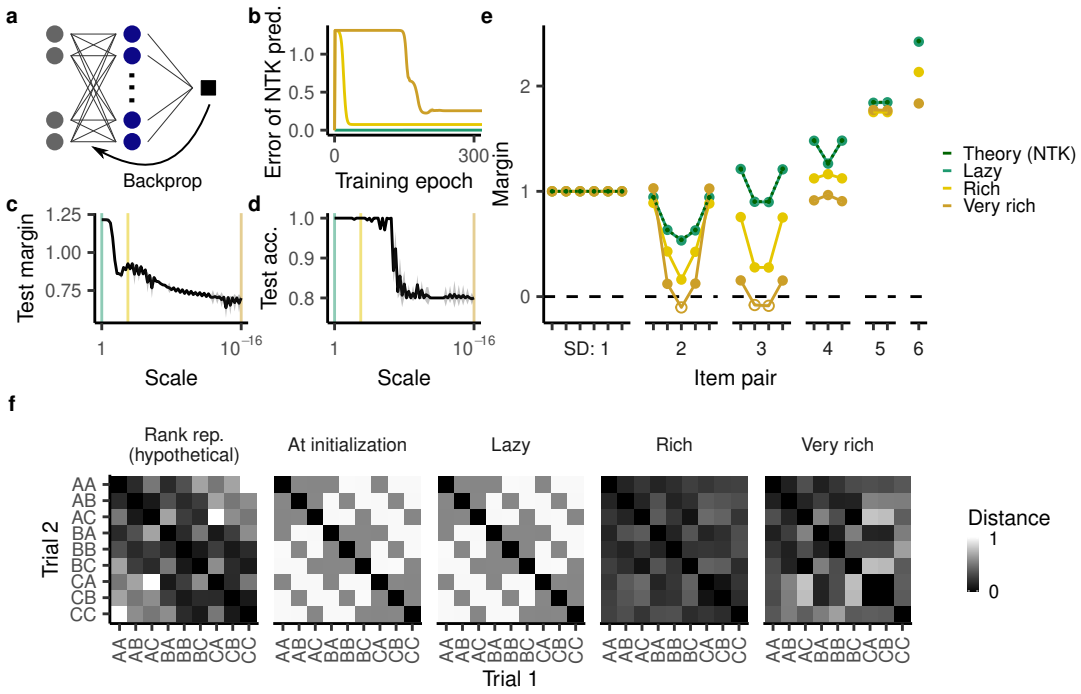

Figure 6: TI behavior in deep neural networks with modifiable representations. We considered twenty instances of a network with a ReLU nonlinearity and one hidden layer with 50,000 units, trained using mean squared error. Appendix S2.5 details other architectures, nonlinearities, and objective functions. Shaded regions indicate mean ± one standard deviation. (Note that some of the behaviors have sufficiently small variance that this shaded region may be invisible.) **a**, Illustration of network training. In contrast to the previous setups, the hidden layer weights were also trained using backpropagation. **b**, The mean squared error of the prediction made by the neural tangent kernel (NTK) at three different initialization scales (lazy: 1; rich: 10^{-3}; very rich: 10^{-16}). **c**,**d**, The (c) average test margin and (d) test accuracy as a function of initialization scale. The colored lines highlight the three representative values analyzed in more detail in panels b and e. **e**, TI performance according to our NTK-based prediction as well as of networks trained with backpropagation at the three representative scales. **f**, Representational dissimilarity matrix (RDM) of a subset of trials for a hypothetical network that represents item rank in its hidden layer as well as for the neural network at initialization and after training in different regimes.

representations ("rich regime"). Broadly considered, adapting a model's internal representation to a given task could address the competing demands imposed by the wide range of tasks subjects need to perform.

Indeed, the rich regime is seen as essential to the remarkable generalization capabilities of deep neural networks [87, 91, 92]. There is also widespread evidence that subjects adapt their representation to a given task [93, 94] and may do so similarly to deep neural networks trained in the rich regime [95]. In relational tasks, in particular, neural representations change to reflect how different items are related to each other [38, 96–100] and this may be essential for successful generalization on certain tasks [38, 101]. In light of this lazy vs. rich distinction and its potential relevance to biological learning, we investigated through simulations whether deep neural networks are suitable as a model of relational representation learning (Fig. 6a). Specifically, we trained neural networks from different scales of initialization using gradient descent (for details on training, see Appendix S2.3). In light of the important role played by the initialization scale, we covered a broad range of potential values, focusing on three representative values: 1 (resulting in lazy behavior), 10^{-3} (resulting in rich behavior), 10^{-16} (to characterize model behavior in the limit of small initialization).

In the lazy regime, we computed the NTK's conjunctivity factor analytically (Appendix S1.5). For large initialization, the learning trajectory of neural networks trained with full weight updating is approximated by gradient flow on the fixed NTK representation. Hence we were able to use the gradient flow solutions

determined in the previous section to predict the network behavior over the course of training, finding a virtually perfect match with simulations in the case of wide neural networks (Fig. 6b,e, green line). Our account may therefore be able to explain why previous studies [60, 61] found empirically that feedforward neural networks generalize transitively.

Surprisingly, we found that networks trained in the rich regime performed worse at TI. They had a smaller test margin (Fig. 6c) and also made systematic errors on the cases CE, BE, and CF for a sufficiently small initialization scale (Fig. 6d,e, hollow points). Further, and in contrast to all learning models considered thus far, the behavior of neural networks in the rich regime was not consistent with a ranking system. This is apparent from the fact that the networks' margin at a symbolic distance of three was smaller than at a symbolic distance of two (Fig. 6e). Importantly, in contrast to the previous limited violation of the symbolic distance effect, this cannot be explained by memorization as none of these cases are in the training set. Finally, the networks in the rich regime exhibited a violation of the terminal item effect at a symbolic distance of 4, a somewhat surprising pattern that no models examined thus far have exhibited. The networks' unconvential behavior was not due to the specific setup considered here: we observed similar behavior for alternative activation functions (Fig. S5), different numbers of items (Fig. S6), alternative loss functions (Fig. S7a), and deeper networks (Fig. S7b). In particular, the networks exhibited even more overtly idiosyncratic behavior for larger numbers of items (Fig. S6d).

This indicates that, for TI, standard neural network training does not confer the benefits of representation learning suggested in prior work. To further investigate the representation learned by the network, we considered the representational similarity of the hidden-layer activity (Fig. 6f). As expected, the lazy-regime networks expressed an exchangeable representation after training (Fig. 6f, middle). In contrast, the internal representation of networks trained in the rich regime violated the exchangeability assumption (Fig. 6f, right), confirming that the networks specialized their representation to the input. What kind of specialization did the networks learn? Most existing accounts of relational learning would suggest that an explicit representation of each item's rank, as observed by Nelli *et al.* [38] (e.g. Fig. 6f, left), would be particularly useful. However, the rich-regime network's representation was not well-characterized by such a rank representation (Fig. S2c). This suggests that the network's representation, though consistent between random initializations, did not encode the kinds of variables that would be useful for transitive inference.

The deviation from rank-based behavior provides a clue as to why neural networks in the rich regime are worse at transitive inference. Specifically, if the behavior of a learning model can be described by a rank computation, this indicates that the model "encodes" the transitive relation in a one-dimensional space (Fig. S3a). In contrast, rich-regime neural networks appear to learn a higher-dimensional representation. Notably, the behavior of humans and animals in TI is more similar to that of neural networks in the lazy regime than in the rich regime, contrary to another task in which more robust, human-like performance was best predicted by neural networks trained in the rich regime [95].

Our findings in this section may appear contrary to those of Nelli *et al.* [38], who also studied TI performance of neural networks, but found that the rich regime did not impair transitive generalization and further yielded an explicit rank representation in the hidden layer. In Appendix S2.6, we show that this is because the networks in their study had symmetric weights for the two items in the input layer. This weight symmetry constrains these networks to an additive internal representation (i.e. with $\alpha = 0$) and further is necessary and sufficient for the emergence of an explicit rank representation in the rich regime (Fig. S8). This highlights an important divergence between fixed-feature models and deep neural networks: removing the additive constraint in models with fixed representations (to make them suitable for a wider range of tasks) resulted in gradually altered yet realistic TI behavior. In contrast, even though deep neural networks with an additive constraint learn to represent item rank in their hidden layer, removing this constraint resulted in radically different, and indeed non-realistic, behavior. It remains an important open challenge to identify models that can learn sensible relational representations for TI without an additive constraint.

4 Discussion

We found that standard statistical learning models can perform transitive inference (TI) and recapitulate three empirically observed behavioral patterns (symbolic distance effect, terminal item effect, and memorization effect). The behavior of a given model is sufficiently captured by a single scalar "conjunctivity factor"

α, which characterizes the model's internal representation (of task items) on a spectrum from fully compositional ($\alpha = 0$) to fully conjunctive ($\alpha = 1$). For $\alpha = 0$, the model is constrained to encoding a transitive relation. For partly conjunctive representations ($0 < \alpha < 1$), the model is not constrained in this way, but the principle of norm minimization nevertheless yields transitive generalization. For fully conjunctive representations ($\alpha = 1$), the model cannot generalize transitively. Finally, we found that when representation learning is enabled in hidden layers, networks perform worse on TI and exhibit different behavioral patterns than living subjects.

Models of relational cognition often represent relations explicitly and are pre-configured to have a particular relational inductive bias [102–104]. In particular, alternative accounts of TI are either pre-configured to associate a rank (or value) with each item [32, 33, 56–58] or suggest that humans rely on abstract knowledge of transitivity [e.g. 32, 41, 105] (as implemented, for instance, in a cognitive map [106–108]). In contrast to these accounts, we took a "minimal principles" approach, representing nothing but the input itself (i.e. the two presented items). This perspective casts higher-level behavioral capacities (in this case transitive generalization) as emergent from minimally structured learning systems [109, 110]. Studies within this paradigm usually rely on simulations, which can leave the mechanisms for emergent behavior, such as generalization, unclear [cf. 111]. In contrast, our analytical account identifies the specific model components responsible for transitive generalization. This clarifies how the brain could implement relational inference and compositional generalization without pre-configured representations or compositional constraints.

Our account could be seen as an endpoint to a series of investigations of TI behavior: expanding upon previous studies [38, 60, 61], we show comprehensively that the principle of norm minimization enables any model with partly conjunctive representations to generalize transitively and further gives rise to naturalistic behavior on TI. Importantly, norm minimization is implicated not only in gradient flow and ridge regression (the examples we consider), but also a much broader range of learning models [112], including reinforcement learning [113]. Accordingly, the consistent behaviors many different animals exhibit on the task could be due to this shared, underlying learning principle. This is an alternative to the view that the ubiquity of TI stems from its ecological role in social cognition, in particular in animals which form hierarchies [114, 115], and the view that TI entails explicit reasoning.

Our results, while providing an alternative explanatory account of TI, should nevertheless be interpreted cautiously with regards to the basis of TI in living subjects, as our model is limited in a number of ways. In terms of behavioral predictions, it cannot account for the asymmetry effect, the observation that performance in living subjects is often better for items towards the start of the hierarchy (e.g. AB) than items towards of the end (e.g. FG) [33]. Further, our model does not generate predictions for what behavior we might expect from subjects that, after being trained on item pairs, are presented with three items at a time and expected to pick the largest item [40, 47, 116, 117].

These limitations, and extensions of TI, can inspire refining or expanding the statistical modeling approach. For example, Nelli *et al.* [38] found that conventional statistical learning models could not account for human behavior on a particular variant of TI ["list linking," 118], but that a modification in which uncertainty is encoded could. More broadly, TI, with its rich and well-established set of empirical phenomena [e.g. 119–121], may serve as a useful model task to compare statistical learning models (and their failure modes) to human and animal behavior. For example, the lack of an asymmetry effect and the inability to perform the three-item task indicate what is lost by treating TI as a binary categorization task. A model that understands its decision as tied to one of the presented items (rather than as a choice between two arbitrary categories) may be able to better account for these behaviors [33]. Notably, better accounts of behavior within the TI framework may not only elucidate how subjects perform relational learning, but even inform more general models of statistical learning in living subjects.

Despite these potential limitations, our account still makes a set of falsifiable predictions that can produce evidence for or against the biological relevance of our insights. In particular, we identified the conjunctivity factor α as a broadly important parameter for TI task behavior. Our account would predict that changes in α should lead to a set of coordinated changes in behavior. For example, a stronger memorization effect should be associated with a weaker terminal item effect on the training cases (as both are caused by an increase in α). Observing that these different behavioral patterns indeed change in a coordinated manner (for example between different subjects performing the same task) may indicate that our theory can accurately describe behavior.

Changes in the conjunctivity factor may arise from inter-individual differences (different subjects may

have representations that are best described by different values of α), but could also be introduced through experimental interventions. In particular, certain task variations may promote a representation of the input as either two separate items or a single conjunctive stimulus [122]; for example, presenting the two items in a common scene rather than as distinct stimuli may result in a higher α. More broadly, this suggests that the conjunctivity factor may be a useful parameter for comparing subjects and TI task variants.

Beyond behavioral predictions, the conjunctivity factor also suggests a new approach for clarifying the neural basis of TI. On the one hand, perturbations in relevant neural areas, for example through lesions [123], may affect the conjunctivity factor. On the other hand, using neural recordings to estimate the empirical representational similarity [124] between distinct, overlapping, and identical trials could enable the estimation of an effective conjunctivity factor in a particular neural area. If differences in this estimated conjunctivity factor (either due to inter-individual differences or as a result of lesions) are associated with the corresponding behavioral differences predicted by our model, this may suggest a role for statistical learning as described here, and, in addition, that the recorded neural area is indeed involved in TI.

Finally, our findings on rich-regime neural networks performing worse on transitive inference may have important implications for machine learning, as deep neural networks have been observed to struggle with tasks involving compositional generalization more broadly [125–131]. A common strategy in attempting to address these shortcomings consists in training ever larger models on ever larger datasets [132]. Such models have shown impressive results on tasks such as natural language production [133, 134], but the overall scale and complexity of the training data, model, and learning algorithm make it difficult to attribute successful or unsuccessful generalization behaviors to specific components of the model [though see 135–138]. In contrast, TI is a particularly simple relational task on which deep networks exhibit non-naturalistic behavior. Its simple structure may allow us to understand the exact inductive bias of current deep learning methods [139–141] and, based on this understanding, design principled modifications to address their limitations [142–145]. This, in turn, may help us understand how relational representation learning is implemented in the brain.

In summary, we here investigated transitive inference, a task that tests for a fundamental logical capacity and has fascinated researchers across neuroscience and psychology for many decades. We derived exact mathematical equations describing the TI behavior of a large class of statistical learning models. This allowed us to understand systematically how these models can generalize compositionally, using TI as an example. Our theory provides a basis for using TI to investigate the neural structures implementing relational cognition. At the same time, our findings also suggest that TI, in its standard form, may not be sufficient to clarify relational inference abilities that require more than statistical learning. For this purpose, other relational tasks should be considered.

Acknowledgments

We are grateful to the members of the Center for Theoretical Neuroscience at Columbia University, the SueYeon Chung lab at Flatiron/NYU, and attendees at Cosyne 2023 and the Gatsby Tri-Centre Meeting 2023 for helpful comments. We thank Taiga Abe, Veronica Bossio, Tala Fakhoury, Ching Fang, Jeff Johnston, Erica Shook, Sharon Su, and Denis Turcu for detailed feedback. The work was supported by NSF 1707398 (Neuronex), Gatsby Charitable Foundation GAT3708, and NIH-R01MH111703.

References

1. Craik, K. J. W. The Nature of Explanation (1943).
2. Chomsky, N. Aspects of the Theory of Syntax (1965).
3. Gentner, D. Structure-mapping: A theoretical framework for analogy. en. *Cognitive Science* **7,** 155–170. ISSN: 0364-0213. https://www.sciencedirect.com/science/article/pii/S0364021383800093 (2022) (Apr. 1983).
4. Brousse, O. Generativity and Systematicity in Neural Network Combinatorial Learning; CU-CS-676-93 (1993).
5. Frankland, S. M. & Greene, J. D. Concepts and Compositionality: In Search of the Brain's Language of Thought. *Annual Review of Psychology* **71.** _eprint: https://doi.org/10.1146/annurev-psych-122216-011829, 273–303. https://doi.org/10.1146/annurev-psych-122216-011829 (2023) (2020).
6. Dekker, R. B., Otto, F. & Summerfield, C. Curriculum learning for human compositional generalization. *Proceedings of the National Academy of Sciences* **119.** Publisher: Proceedings of the National Academy of Sciences, e2205582119. https://www.pnas.org/doi/full/10.1073/pnas.2205582119 (2023) (Oct. 2022).

7. Halford, G. S., Wilson, W. H. & Phillips, S. Relational knowledge: the foundation of higher cognition. en. *Trends in Cognitive Sciences* **14,** 497–505. ISSN: 1364-6613. https://www.sciencedirect.com/science/article/pii/S1364661310002020 (2023) (Nov. 2010).
8. Cheney, D. L., Seyfarth, R. M. & Silk, J. B. The responses of female baboons (Papio cynocephalus ursinus) to anomalous social interactions: Evidence for causal reasoning? *Journal of Comparative Psychology* **109.** Place: US Publisher: American Psychological Association, 134–141. ISSN: 1939-2087(Electronic),0735-7036(Print) (1995).
9. Peake, T. M., Terry, A. M. R., McGregor, P. K. & Dabelsteen, T. Do great tits assess rivals by combining direct experience with information gathered by eavesdropping? *Proceedings of the Royal Society of London. Series B: Biological Sciences* **269.** Publisher: Royal Society, 1925–1929. https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2002.2112 (2022) (Sept. 2002).
10. Paz-y-Miño C, G., Bond, A. B., Kamil, A. C. & Balda, R. P. Pinyon jays use transitive inference to predict social dominance. en. *Nature* **430.** Number: 7001 Publisher: Nature Publishing Group, 778–781. ISSN: 1476-4687. https://www.nature.com/articles/nature02723 (2022) (Aug. 2004).
11. Etienne, A. S. & Jeffery, K. J. Path integration in mammals. en. *Hippocampus* **14.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/ 180–192. ISSN: 1098-1063. https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.10173 (2022) (2004).
12. Seed, A. & Byrne, R. Animal Tool-Use. en. *Current Biology* **20,** R1032–R1039. ISSN: 0960-9822. https://www.sciencedirect.com/science/article/pii/S0960982210011607 (2022) (Dec. 2010).
13. Baldwin, M. W. Relational schemas and the processing of social information. *Psychological bulletin* **112.** Publisher: American Psychological Association, 461 (1992).
14. Tolman, E. C. Cognitive maps in rats and men. *Psychological Review* **55.** Place: US Publisher: American Psychological Association, 189–208. ISSN: 1939-1471 (1948).
15. Penn, D. C. & Povinelli, D. J. Causal Cognition in Human and Nonhuman Animals: A Comparative, Critical. *Annu. Rev. Psychol* **58,** 97–118 (2007).
16. Mitchell, T. M. *The need for biases in learning generalizations* (Department of Computer Science, Laboratory for Computer Science Research ..., 1980).
17. Krogh, A. & Hertz, J. A. Generalization in a linear perceptron in the presence of noise. en. *Journal of Physics A: Mathematical and General* **25,** 1135. ISSN: 0305-4470. https://dx.doi.org/10.1088/0305-4470/25/5/020 (2023) (Mar. 1992).
18. Watkin, T. L. H., Rau, A. & Biehl, M. The statistical mechanics of learning a rule. *Reviews of Modern Physics* **65.** Publisher: American Physical Society, 499–556. https://link.aps.org/doi/10.1103/RevModPhys.65.499 (2023) (Apr. 1993).
19. Sollich, P. *Learning Curves for Gaussian Processes* in *Advances in Neural Information Processing Systems* **11** (MIT Press, 1998). https://proceedings.neurips.cc/paper_files/paper/1998/hash/5cbdfd0dfa22a3fca7266376887f549b-Abstract.html (2023).
20. Liang, T. & Rakhlin, A. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics* **48.** Publisher: Institute of Mathematical Statistics, 1329–1347. ISSN: 0090-5364, 2168-8966. https://projecteuclid.org/journals/annals-of-statistics/volume-48/issue-3/Just-interpolate-Kernel-Ridgeless-regression-can-generalize/10.1214/19-AOS1849.full (2023) (June 2020).
21. Jacot, A., Simsek, B., Spadaro, F., Hongler, C. & Gabriel, F. *Kernel Alignment Risk Estimator: Risk Prediction from Training Data* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 15568–15578. https://proceedings.neurips.cc/paper/2020/hash/b367e525a7e574817c19ad24b7b35607-Abstract.html (2023).
22. Canatar, A., Bordelon, B. & Pehlevan, C. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. en. *Nature Communications* **12.** Number: 1 Publisher: Nature Publishing Group, 2914. ISSN: 2041-1723. https://www.nature.com/articles/s41467-021-23103-1 (2023) (May 2021).
23. Hoerl, A. E. & Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12.** Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/00401706.1970.10488634, 55–67. ISSN: 0040-1706. https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634 (2023) (Feb. 1970).
24. Cortes, C. & Vapnik, V. Support-vector networks. en. *Machine Learning* **20,** 273–297. ISSN: 1573-0565. https://doi.org/10.1007/BF00994018 (2022) (Sept. 1995).
25. Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. & Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* **19.** Publisher: JMLR. org, 2822–2878 (2018).
26. Krogh, A. & Hertz, J. *A Simple Weight Decay Can Improve Generalization* in *Advances in Neural Information Processing Systems* **4** (Morgan-Kaufmann, 1991). https://proceedings.neurips.cc/paper/1991/hash/8eefcfdf5990e441f0fb6f3fad709e21-Abstract.html (2023).

27. Moody, J. *The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems* in *Advances in Neural Information Processing Systems* **4** (Morgan-Kaufmann, 1991). https://proceedings.neurips.cc/paper/1991/hash/d64a340bcb633f536d56e51874281454-Abstract.html (2023).
28. Barnett, S. M. & Ceci, S. J. When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin* **128.** Place: US Publisher: American Psychological Association, 612–637. ISSN: 1939-1455 (2002).
29. Ye, H. *et al. Towards a Theoretical Framework of Out-of-Distribution Generalization* in *Advances in Neural Information Processing Systems* **34** (Curran Associates, Inc., 2021), 23519–23531. https://proceedings.neurips.cc/paper/2021/hash/c5c1cb0bebd56ae38817b251ad72bedb-Abstract.html (2023).
30. Abbe, E., Bengio, S., Lotfi, A. & Rizk, K. *Generalization on the Unseen, Logic Reasoning and Degree Curriculum* arXiv:2301.13105 [cs, stat]. June 2023. http://arxiv.org/abs/2301.13105 (2023).
31. Battaglia, P. W. *et al. Relational inductive biases, deep learning, and graph networks* arXiv:1806.01261 [cs, stat]. Oct. 2018. http://arxiv.org/abs/1806.01261 (2022).
32. Jensen, G., Muñoz, F., Alkan, Y., Ferrera, V. P. & Terrace, H. S. Implicit Value Updating Explains Transitive Inference Performance: The Betasort Model. en. *PLOS Computational Biology* **11.** Publisher: Public Library of Science, e1004523. ISSN: 1553-7358. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004523 (2022) (Sept. 2015).
33. Ciranka, S. *et al.* Asymmetric reinforcement learning facilitates human inference of transitive relations. en. *Nature Human Behaviour* **6.** Number: 4 Publisher: Nature Publishing Group, 555–564. ISSN: 2397-3374. https://www.nature.com/articles/s41562-021-01263-w (2022) (Apr. 2022).
34. Burt, C. in *Experimental psychology and child study* 53–63 (Sir Isaac Pitman & Sons, London, Great Britain, 1922).
35. Piaget, J. *Judgment and reasoning in the child* Pages: viii, 260 (Harcourt, Brace, Oxford, England, 1928).
36. Vasconcelos, M. Transitive inference in non-human animals: An empirical and theoretical analysis. en. *Behavioural Processes* **78,** 313–334. ISSN: 0376-6357. https://www.sciencedirect.com/science/article/pii/S0376635708000818 (2022) (July 2008).
37. Jensen, G. in *APA handbook of comparative psychology: Perception, learning, and cognition, Vol. 2* 385–409 (American Psychological Association, Washington, DC, US, 2017). ISBN: 978-1-4338-2352-7 978-1-4338-2353-4.
38. Nelli, S., Braun, L., Dumbalska, T., Saxe, A. & Summerfield, C. Neural knowledge assembly in humans and neural networks. *Neuron.* Publisher: Elsevier (2023).
39. Bryant, P. E. & Trabasso, T. Transitive Inferences and Memory in Young Children. en. *Nature* **232.** Number: 5311 Publisher: Nature Publishing Group, 456–458. ISSN: 1476-4687. https://www.nature.com/articles/232456a0 (2022) (Aug. 1971).
40. McGonigle, B. O. & Chalmers, M. Are monkeys logical? *Nature* **267.** Place: United Kingdom Publisher: Nature Publishing Group, 694–696. ISSN: 1476-4687 (1977).
41. Gillan, D. J. Reasoning in the chimpanzee: II. Transitive inference. *Journal of Experimental Psychology: Animal Behavior Processes* **7.** Place: US Publisher: American Psychological Association, 150–164. ISSN: 1939-2184 (1981).
42. Davis, H. Transitive inference in rats (Rattus norvegicus). *Journal of Comparative Psychology* **106.** Place: US Publisher: American Psychological Association, 342–349. ISSN: 1939-2087 (1992).
43. Tibbetts, E. A., Agudelo, J., Pandit, S. & Riojas, J. Transitive inference in Polistes paper wasps. *Biology Letters* **15.** Publisher: Royal Society, 20190015. https://royalsocietypublishing.org/doi/10.1098/rsbl.2019.0015 (2022) (May 2019).
44. Grosenick, L., Clement, T. S. & Fernald, R. D. Fish can infer social rank by observation alone. en. *Nature* **445.** Number: 7126 Publisher: Nature Publishing Group, 429–432. ISSN: 1476-4687. https://www.nature.com/articles/nature05511 (2022) (Jan. 2007).
45. Potts, G. R. Storing and retrieving information about ordered relationships. *Journal of Experimental Psychology* **103.** Place: US Publisher: American Psychological Association, 431–439. ISSN: 0022-1015 (1974).
46. D'Amato, M. R. & Colombo, M. The symbolic distance effect in monkeys (Cebus apella). en. *Animal Learning & Behavior* **18,** 133–140. ISSN: 1532-5830. https://doi.org/10.3758/BF03205250 (2022) (June 1990).
47. McGonigle, B. & Chalmers, M. Monkeys are rational! *The Quarterly Journal of Experimental Psychology Section B* **45.** Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1080/14640749208401017, 189–228. ISSN: 0272-4995. https://www.tandfonline.com/doi/abs/10.1080/14640749208401017 (2022) (Oct. 1992).
48. Merritt, D., MacLean, E. L., Jaffe, S. & Brannon, E. M. A comparative analysis of serial ordering in ring-tailed lemurs (Lemur catta). *Journal of Comparative Psychology* **121.** Publisher: American Psychological Association, 363 (2007).
49. Scarf, D. & Colombo, M. Representation of serial order: A comparative analysis of humans, monkeys, and pigeons. en. *Brain Research Bulletin. Special Issue:Brain Mechanisms, Cognition and Behaviour in Birds* **76,**

307–312. ISSN: 0361-9230. https://www.sciencedirect.com/science/article/pii/S0361923008000373 (2022) (June 2008).
50. Von Fersen, L., Wynne, C. D., Delius, J. D. & Staddon, J. E. Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* **17.** Publisher: American Psychological Association, 334 (1991).
51. Wynne, C. D. L. Pigeon transitive inference: Tests of simple accounts of a complex performance. en. *Behavioural Processes* **39,** 95–112. ISSN: 0376-6357. https://www.sciencedirect.com/science/article/pii/S0376635796000484 (2022) (Jan. 1997).
52. Merritt, D. J. & Terrace, H. S. Mechanisms of inferential order judgments in humans (Homo sapiens) and rhesus monkeys (Macaca mulatta). *Journal of Comparative Psychology* **125.** Publisher: American Psychological Association, 227 (2011).
53. Acuna, B. D., Sanes, J. N. & Donoghue, J. P. Cognitive mechanisms of transitive inference. en. *Experimental Brain Research* **146,** 1–10. ISSN: 1432-1106. https://doi.org/10.1007/s00221-002-1092-y (2022) (Sept. 2002).
54. Van Elzakker, M., O'Reilly, R. C. & Rudy, J. W. Transitivity, flexibility, conjunctive representations, and the hippocampus. I. An empirical analysis. en. *Hippocampus* **13.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hipo.10083 334–340. ISSN: 1098-1063. https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.10083 (2022) (2003).
55. Wynne, C. D. L. in *Models of Action* Num Pages: 40 (Psychology Press, 1998). ISBN: 978-0-203-77386-4.
56. Wynne, C. D. L. Reinforcement accounts for transitive inference performance. en. *Animal Learning & Behavior* **23,** 207–217. ISSN: 1532-5830. https://doi.org/10.3758/BF03199936 (2022) (June 1995).
57. Couvillon, P. A. & Bitterman, M. E. A conventional conditioning analysis of "transitive inference" in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* **18.** Place: US Publisher: American Psychological Association, 308–310. ISSN: 1939-2184 (1992).
58. Jensen, G., Terrace, H. S. & Ferrera, V. P. Discovering Implied Serial Order Through Model-Free and Model-Based Learning. *Frontiers in Neuroscience* **13.** ISSN: 1662-453X. https://www.frontiersin.org/articles/10.3389/fnins.2019.00878 (2022) (2019).
59. Wynne, C. D. L. Transverse patterning in pigeons. en. *Behavioural Processes* **38,** 119–130. ISSN: 0376-6357. https://www.sciencedirect.com/science/article/pii/S0376635796000320 (2022) (Nov. 1996).
60. De Lillo, C., Floreano, D. & Antinucci, F. Transitive choices by a simple, fully connected, backpropagation neural network: implications for the comparative study of transitive inference. en. *Animal Cognition* **4,** 61–68. ISSN: 1435-9448. https://doi.org/10.1007/s100710100092 (2022) (July 2001).
61. Kay, K. *et al. Neural dynamics and geometry for transitive inference* en. Pages: 2022.10.10.511448 Section: New Results. Oct. 2022. https://www.biorxiv.org/content/10.1101/2022.10.10.511448v1 (2022).
62. Antonio, G. D., Raglio, S. & Mattia, M. *Ranking and serial thinking: A geometric solution* en. Pages: 2023.08.03.551859 Section: New Results. Aug. 2023. https://www.biorxiv.org/content/10.1101/2023.08.03.551859v1 (2023).
63. Carmesin, H. .-. & Schwegler, H. Parallel versus sequential processing of relational stimulus structures. en. *Biological Cybernetics* **71,** 523–529. ISSN: 1432-0770. https://doi.org/10.1007/BF00198470 (2022) (Oct. 1994).
64. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. eng. *Psychological Review* **113,** 700–765. ISSN: 0033-295X (Oct. 2006).
65. Gold, J. I. & Shadlen, M. N. The Neural Basis of Decision Making. *Annual Review of Neuroscience* **30.** _eprint: https://doi.org/10.1146/annurev.neuro.29.051605.113038, 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038 (2023) (2007).
66. Churchland, A. K. *et al.* Variance as a Signature of Neural Computations during Decision Making. English. *Neuron* **69.** Publisher: Elsevier, 818–831. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(10)01087-1 (2023) (Feb. 2011).
67. Ratcliff, R. A theory of memory retrieval. *Psychological Review* **85.** Place: US Publisher: American Psychological Association, 59–108. ISSN: 1939-1471 (1978).
68. Ratcliff, R. & McKoon, G. The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation* **20,** 873–922. ISSN: 0899-7667. https://doi.org/10.1162/neco.2008.12-06-420 (2023) (Apr. 2008).
69. Frankland, S. M. & Greene, J. D. Two Ways to Build a Thought: Distinct Forms of Compositional Semantic Representation across Brain Regions. *Cerebral Cortex* **30,** 3838–3855. ISSN: 1047-3211. https://doi.org/10.1093/cercor/bhaa001 (2023) (May 2020).
70. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. en. *Nature* **497.** Number: 7451 Publisher: Nature Publishing Group, 585–590. ISSN: 1476-4687. https://www.nature.com/articles/nature12160 (2022) (May 2013).

71. Rudy, J. W. & O'Reilly, R. C. Conjunctive representations, the hippocampus, and contextual fear conditioning. en. *Cognitive, Affective, & Behavioral Neuroscience* **1,** 66–82. ISSN: 1531-135X. https://doi.org/10.3758/CABN.1.1.66 (2023) (Mar. 2001).
72. Rescorla, R. A. "Configural" conditioning in discrete-trial bar pressing. eng. *Journal of Comparative and Physiological Psychology* **79,** 307–317. ISSN: 0021-9940 (May 1972).
73. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. en. *Nature* **503.** Number: 7474 Publisher: Nature Publishing Group, 78–84. ISSN: 1476-4687. https://www.nature.com/articles/nature12742 (2023) (Nov. 2013).
74. Komorowski, R. W., Manns, J. R. & Eichenbaum, H. Robust Conjunctive Item–Place Coding by Hippocampal Neurons Parallels Learning What Happens Where. en. *Journal of Neuroscience* **29.** Publisher: Society for Neuroscience Section: Articles, 9918–9929. ISSN: 0270-6474, 1529-2401. https://www.jneurosci.org/content/29/31/9918 (2023) (Aug. 2009).
75. McKenzie, S. *et al.* Hippocampal Representation of Related and Opposing Memories Develop within Distinct, Hierarchically Organized Neural Schemas. English. *Neuron* **83.** Publisher: Elsevier, 202–215. ISSN: 0896-6273. https://www.cell.com/neuron/abstract/S0896-6273(14)00405-X (2023) (July 2014).
76. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. en. *Current Opinion in Neurobiology. Neurobiology of cognitive behavior* **37,** 66–74. ISSN: 0959-4388. https://www.sciencedirect.com/science/article/pii/S0959438816000118 (2023) (Apr. 2016).
77. Davis, H. in *Cognitive Aspects of Stimulus Control* Num Pages: 25 (Psychology Press, 1992). ISBN: 978-1-315-78910-1.
78. Cho, Y. & Saul, L. *Kernel Methods for Deep Learning* in *Advances in Neural Information Processing Systems* **22** (Curran Associates, Inc., 2009). https://papers.nips.cc/paper/2009/hash/5751ec3e9a4feab575962e78e006250d-Abstract.html (2022).
79. Spence, K. W. The nature of the response in discrimination learning. *Psychological Review* **59.** Place: US Publisher: American Psychological Association, 89–93. ISSN: 1939-1471 (1952).
80. Turrigiano, G. G. The Self-Tuning Neuron: Synaptic Scaling of Excitatory Synapses. English. *Cell* **135.** Publisher: Elsevier, 422–435. ISSN: 0092-8674, 1097-4172. https://www.cell.com/cell/abstract/S0092-8674(08)01298-1 (2023) (Oct. 2008).
81. Saxe, A. M., McClelland, J. L. & Ganguli, S. *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks* arXiv:1312.6120 [cond-mat, q-bio, stat]. Feb. 2014. http://arxiv.org/abs/1312.6120 (2023).
82. Gunasekar, S., Lee, J., Soudry, D. & Srebro, N. *Characterizing implicit bias in terms of optimization geometry* in *International Conference on Machine Learning* (PMLR, 2018), 1832–1841.
83. Ji, Z., Dudík, M., Schapire, R. E. & Telgarsky, M. *Gradient descent follows the regularization path for general losses* in *Proceedings of Thirty Third Conference on Learning Theory* (eds Abernethy, J. & Agarwal, S.) **125** (PMLR, July 2020), 2109–2136. https://proceedings.mlr.press/v125/ji20a.html.
84. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. en. *Nature* **521.** Number: 7553 Publisher: Nature Publishing Group, 436–444. ISSN: 1476-4687. https://www.nature.com/articles/nature14539 (2023) (May 2015).
85. Richards, B. A. *et al.* A deep learning framework for neuroscience. en. *Nature Neuroscience* **22.** Number: 11 Publisher: Nature Publishing Group, 1761–1770. ISSN: 1546-1726. https://www.nature.com/articles/s41593-019-0520-2 (2023) (Nov. 2019).
86. Saxe, A., Nelli, S. & Summerfield, C. If deep learning is the answer, what is the question? en. *Nature Reviews Neuroscience* **22.** Number: 1 Publisher: Nature Publishing Group, 55–67. ISSN: 1471-0048. https://www.nature.com/articles/s41583-020-00395-8 (2023) (Jan. 2021).
87. Chizat, L., Oyallon, E. & Bach, F. *On Lazy Training in Differentiable Programming* in *Advances in Neural Information Processing Systems* **32** (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html (2023).
88. Woodworth, B. *et al. Kernel and Rich Regimes in Overparametrized Models* en. in *Proceedings of Thirty Third Conference on Learning Theory* ISSN: 2640-3498 (PMLR, July 2020), 3635–3673. https://proceedings.mlr.press/v125/woodworth20a.html (2023).
89. Jacot, A., Gabriel, F. & Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems* **31** (2018).
90. Lee, J. *et al.* Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems* **32** (2019).
91. Fort, S. *et al. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 5850–5861. https://proceedings.neurips.cc/paper/2020/hash/405075699f065e43581f27d67bb68478-Abstract.html (2023).

92. Vyas, N., Bansal, Y. & Nakkiran, P. *Limitations of the NTK for Understanding Generalization in Deep Learning* arXiv:2206.10012 [cs]. June 2022. http://arxiv.org/abs/2206.10012 (2023).
93. Weinberger, N. Dynamic regulation of receptive fields and maps in the adult sensory cortex. *Annual review of neuroscience* **18,** 129–158 (1995).
94. Poort, J. *et al.* Learning Enhances Sensory and Multiple Non-sensory Representations in Primary Visual Cortex. en. *Neuron* **86,** 1478–1490. ISSN: 0896-6273. https://www.sciencedirect.com/science/article/pii/S0896627315004766 (2023) (June 2015).
95. Flesch, T., Juechems, K., Dumbalska, T., Saxe, A. & Summerfield, C. Orthogonal representations for robust context-dependent task performance in brains and neural networks. en. *Neuron* **110,** 1258–1270.e11. ISSN: 0896-6273. https://www.sciencedirect.com/science/article/pii/S0896627322000058 (2023) (Apr. 2022).
96. Van Opstal, F., Fias, W., Peigneux, P. & Verguts, T. The neural representation of extensively trained ordered sequences. en. *NeuroImage* **47,** 367–375. ISSN: 1053-8119. https://www.sciencedirect.com/science/article/pii/S1053811909003838 (2023) (Aug. 2009).
97. Hinton, E. C., Dymond, S., von Hecker, U. & Evans, C. J. Neural correlates of relational reasoning and the symbolic distance effect: involvement of parietal cortex. en. *Neuroscience* **168,** 138–148. ISSN: 0306-4522. https://www.sciencedirect.com/science/article/pii/S0306452210004471 (2023) (June 2010).
98. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352.** Publisher: American Association for the Advancement of Science, 1464–1468. https://www.science.org/doi/full/10.1126/science.aaf0941 (2023) (June 2016).
99. Munoz, F. *et al.* Learned representation of implied serial order in posterior parietal cortex. *Scientific Reports* **10.** Publisher: Springer, 1–14 (2020).
100. Ramawat, S. *et al.* Different Contribution of the Monkey Prefrontal and Premotor Dorsal Cortex in Decision Making During a Transitive Inference Task. en. *Neuroscience* **485,** 147–162. ISSN: 0306-4522. https://www.sciencedirect.com/science/article/pii/S0306452222000276 (2023) (Mar. 2022).
101. Whittington, J. C. R. *et al.* The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. en. *Cell* **183,** 1249–1263.e23. ISSN: 0092-8674. https://www.sciencedirect.com/science/article/pii/S009286742031388X (2023) (Nov. 2020).
102. Glymour, C. Learning, prediction and causal Bayes nets. en. *Trends in Cognitive Sciences* **7,** 43–48. ISSN: 1364-6613. https://www.sciencedirect.com/science/article/pii/S1364661302000098 (2022) (Jan. 2003).
103. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proceedings of the National Academy of Sciences* **105.** Publisher: Proceedings of the National Academy of Sciences, 10687–10692. https://www.pnas.org/doi/10.1073/pnas.0802631105 (2022) (Aug. 2008).
104. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to Grow a Mind: Statistics, Structure, and Abstraction. *Science* **331.** Publisher: American Association for the Advancement of Science, 1279–1285. https://www.science.org/doi/full/10.1126/science.1192788 (2022) (Mar. 2011).
105. Jacobs, L. F. From movement to transitivity: the role of hippocampal parallel maps in configural learning. eng. *Reviews in the Neurosciences* **17,** 99–109. ISSN: 0334-1763 (2006).
106. De Soto, C. B., London, M. & Handel, S. Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology* **2.** Place: US Publisher: American Psychological Association, 513–521. ISSN: 1939-1315 (1965).
107. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal Cortex as a Cognitive Map of Task Space. en. *Neuron* **81,** 267–279. ISSN: 0896-6273. https://www.sciencedirect.com/science/article/pii/S0896627313010398 (2022) (Jan. 2014).
108. Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. en. *Neuron* **100,** 490–509. ISSN: 0896-6273. https://www.sciencedirect.com/science/article/pii/S0896627318308560 (2022) (Oct. 2018).
109. McClelland, J. L. *et al.* Letting structure emerge: connectionist and dynamical systems approaches to cognition. en. *Trends in Cognitive Sciences* **14,** 348–356. ISSN: 1364-6613. https://www.sciencedirect.com/science/article/pii/S1364661310001245 (2022) (Aug. 2010).
110. Griffiths, T. L., Chater, N., Kemp, C., Perfors, A. & Tenenbaum, J. B. Probabilistic models of cognition: exploring representations and inductive biases. English. *Trends in Cognitive Sciences* **14.** Publisher: Elsevier, 357–364. ISSN: 1364-6613, 1879-307X. https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(10)00112-9 (2022) (Aug. 2010).
111. Saxe, A. M., McClelland, J. L. & Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences* **116.** Publisher: National Acad Sciences, 11537–11546 (2019).
112. Telgarsky, M. & Singer, Y. A Primal-Dual Convergence Analysis of Boosting. *Journal of Machine Learning Research* **13** (2012).
113. Kumar, A. *et al. DR3: Value-Based Deep Reinforcement Learning Requires Explicit Regularization* arXiv:2112.04716 [cs]. Dec. 2021. http://arxiv.org/abs/2112.04716 (2023).

114. Bond, A. B., Kamil, A. C. & Balda, R. P. Social complexity and transitive inference in corvids. en. *Animal Behaviour* **65,** 479–487. ISSN: 0003-3472. https://www.sciencedirect.com/science/article/pii/S0003347203921014 (2022) (Mar. 2003).
115. MacLean, E. L., Merritt, D. J. & Brannon, E. M. Social complexity predicts transitive reasoning in prosimian primates. en. *Animal Behaviour* **76,** 479–486. ISSN: 0003-3472. https://www.sciencedirect.com/science/article/pii/S0003347208002030 (2023) (Aug. 2008).
116. Harris, M. & McGonigle, B. A model of transitive choice. *The Quarterly Journal of Experimental Psychology Section B* **47.** Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1080/14640749408401362, 319–348. ISSN: 0272-4995. https://www.tandfonline.com/doi/abs/10.1080/14640749408401362 (2023) (Aug. 1994).
117. Bryson, J. J. & Leong, J. C. S. Primate errors in transitive 'inference': a two-tier learning model. en. *Animal Cognition* **10,** 1–15. ISSN: 1435-9456. https://doi.org/10.1007/s10071-006-0024-9 (2023) (Jan. 2007).
118. Treichler, F. R. & Van Tilburg, D. Concurrent conditional discrimination tests of transitive inference by macaque monkeys: list linking. *Journal of Experimental Psychology: Animal Behavior Processes* **22.** Publisher: American Psychological Association, 105 (1996).
119. Gazes, R. P., Chee, N. W. & Hampton, R. R. Cognitive mechanisms for transitive inference performance in rhesus monkeys: Measuring the influence of associative strength and inferred order. *Journal of Experimental Psychology: Animal Behavior Processes* **38.** Place: US Publisher: American Psychological Association, 331–345. ISSN: 1939-2184(Electronic),0097-7403(Print) (2012).
120. Lazareva, O. F. & Wasserman, E. A. Transitive inference in pigeons: Measuring the associative values of Stimuli B and D. en. *Behavioural Processes* **89,** 244–255. ISSN: 0376-6357. https://www.sciencedirect.com/science/article/pii/S0376635711002403 (2023) (Mar. 2012).
121. Jensen, G., Alkan, Y., Muñoz, F., Ferrera, V. P. & Terrace, H. S. Transitive inference in humans (Homo sapiens) and rhesus macaques (Macaca mulatta) after massed training of the last two list items. *Journal of Comparative Psychology* **131.** Place: US Publisher: American Psychological Association, 231–245. ISSN: 1939-2087(Electronic),0735-7036(Print) (2017).
122. Eichenbaum, H. *The cognitive neuroscience of memory: An introduction* (Oxford University Press, 2011).
123. Dusek, J. A. & Eichenbaum, H. The hippocampus and transverse patterning guided by olfactory cues. *Behavioral Neuroscience* **112.** Place: US Publisher: American Psychological Association, 762–771. ISSN: 1939-0084 (1998).
124. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience.* Publisher: Frontiers, 4 (2008).
125. Fodor, J. A. & Pylyshyn, Z. W. Connectionism and cognitive architecture: A critical analysis. *Cognition* **28.** Publisher: Elsevier, 3–71 (1988).
126. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350.** Publisher: American Association for the Advancement of Science, 1332–1338. https://www.science.org/doi/10.1126/science.aab3050 (2022) (Dec. 2015).
127. Shalev-Shwartz, S. & Shashua, A. *On the Sample Complexity of End-to-end Training vs. Semantic Abstraction Training* arXiv:1604.06915 [cs]. Apr. 2016. http://arxiv.org/abs/1604.06915 (2022).
128. Bender, E. M. & Koller, A. *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data* in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Online, July 2020), 5185–5198. https://aclanthology.org/2020.acl-main.463 (2022).
129. Klinger, T. *et al.* A study of compositional generalization in neural models. *arXiv preprint arXiv:2006.09437* (2020).
130. Summerfield, C., Luyckx, F. & Sheahan, H. Structure learning and the posterior parietal cortex. en. *Progress in Neurobiology* **184,** 101717. ISSN: 0301-0082. https://www.sciencedirect.com/science/article/pii/S0301008219303351 (2023) (Jan. 2020).
131. Yuille, A. L. & Liu, C. Deep Nets: What have They Ever Done for Vision? en. *International Journal of Computer Vision* **129,** 781–802. ISSN: 1573-1405. https://doi.org/10.1007/s11263-020-01405-z (2022) (Mar. 2021).
132. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
133. Brown, T. *et al. Language Models are Few-Shot Learners* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 1877–1901. https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html (2022).
134. Mahowald, K. *et al. Dissociating language and thought in large language models: a cognitive perspective* arXiv:2301.06627 [cs]. Jan. 2023. http://arxiv.org/abs/2301.06627 (2023).
135. Saphra, N. & Lopez, A. Understanding learning dynamics of language models with SVCCA. *arXiv preprint arXiv:1811.00225* (2018).

136. Tirumala, K., Markosyan, A. H., Zettlemoyer, L. & Aghajanyan, A. *Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models* arXiv:2205.10770 [cs]. Nov. 2022. http://arxiv.org/abs/2205.10770 (2023).
137. Kandpal, N., Deng, H., Roberts, A., Wallace, E. & Raffel, C. *Large Language Models Struggle to Learn Long-Tail Knowledge* arXiv:2211.08411 [cs]. Nov. 2022. http://arxiv.org/abs/2211.08411 (2023).
138. Grosse, R. *et al. Studying Large Language Model Generalization with Influence Functions* arXiv:2308.03296 [cs, stat]. Aug. 2023. http://arxiv.org/abs/2308.03296 (2023).
139. Lyu, K. & Li, J. *Gradient Descent Maximizes the Margin of Homogeneous Neural Networks* in *International Conference on Learning Representations* (2020). https://openreview.net/forum?id=SJeLIgBKPS.
140. Chizat, L. & Bach, F. *Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss* en. in *Proceedings of Thirty Third Conference on Learning Theory* ISSN: 2640-3498 (PMLR, July 2020), 1305–1338. https://proceedings.mlr.press/v125/chizat20a.html (2023).
141. Saxe, A., Sodhani, S. & Lewallen, S. J. *The Neural Race Reduction: Dynamics of Abstraction in Gated Networks* en. in *Proceedings of the 39th International Conference on Machine Learning* ISSN: 2640-3498 (PMLR, June 2022), 19287–19309. https://proceedings.mlr.press/v162/saxe22a.html (2023).
142. Smolensky, P. On the proper treatment of connectionism. en. *Behavioral and Brain Sciences* **11.** Publisher: Cambridge University Press, 1–23. ISSN: 1469-1825, 0140-525X. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/on-the-proper-treatment-of-connectionism/4B8871A82A932DB96D183AAC9 (2023) (Mar. 1988).
143. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences* **40.** Publisher: Cambridge University Press (2017).
144. Kipf, T., Fetaya, E., Wang, K.-C., Welling, M. & Zemel, R. *Neural Relational Inference for Interacting Systems* en. in *Proceedings of the 35th International Conference on Machine Learning* ISSN: 2640-3498 (PMLR, July 2018), 2688–2697. https://proceedings.mlr.press/v80/kipf18a.html (2022).
145. Hill, F., Santoro, A., Barrett, D., Morcos, A. & Lillicrap, T. *Learning to Make Analogies by Contrasting Abstract Relational Structure* in *International Conference on Learning Representations* (2019).
146. Fukushima, K. Visual Feature Extraction by a Multilayered Network of Analog Threshold Elements. *IEEE Transactions on Systems Science and Cybernetics* **5.** Conference Name: IEEE Transactions on Systems Science and Cybernetics, 322–333. ISSN: 2168-2887 (Oct. 1969).
147. Maas, A. L., Hannun, A. Y., Ng, A. Y., *et al. Rectifier nonlinearities improve neural network acoustic models* in *Proc. icml* **30.** Issue: 1 (Atlanta, Georgia, USA, 2013), 3.
148. Tsuchida, R., Roosta-Khorasani, F. & Gallagher, M. *Invariance of Weight Distributions in Rectified MLPs* arXiv:1711.09090 [cs, stat]. May 2018. http://arxiv.org/abs/1711.09090 (2022).
149. Tsuchida, R., Roosta, F. & Gallagher, M. *Richer priors for infinitely wide multi-layer perceptrons* arXiv:1911.12927 [cs, stat]. Nov. 2019. http://arxiv.org/abs/1911.12927 (2022).
150. Han, I. *et al. Fast Neural Kernel Embeddings for General Activations* arXiv:2209.04121 [cs, stat]. Sept. 2022. http://arxiv.org/abs/2209.04121 (2022).
151. He, K., Zhang, X., Ren, S. & Sun, J. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification* in (2015), 1026–1034. https://openaccess.thecvf.com/content_iccv_2015/html/He_Delving_Deep_into_ICCV_2015_paper.html (2023).
152. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org/ (R Foundation for Statistical Computing, Vienna, Austria, 2022).
153. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org (Springer-Verlag New York, 2016).
154. Pedersen, T. L. *patchwork: The Composer of Plots* https://CRAN.R-project.org/package=patchwork (2022).
155. Paszke, A. *et al.* in *Advances in Neural Information Processing Systems 32* 8024–8035 (Curran Associates, Inc., 2019). http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
156. Hu, G. Y. & O'Connell, R. F. Analytical inversion of symmetric tridiagonal matrices. *Journal of Physics A: Mathematical and General* **29,** 1511. https://dx.doi.org/10.1088/0305-4470/29/7/020 (Apr. 1996).
157. Böttcher, A. & Grudsky, S. M. *Spectral properties of banded Toeplitz matrices* (SIAM, 2005).
158. Gill, J. The use of the sequence Fn(z)=fnf1(z) in computing fixed points of continued fractions, products, and series. en. *Applied Numerical Mathematics* **8,** 469–476. ISSN: 0168-9274. https://www.sciencedirect.com/science/article/pii/016892749190109D (2023) (Dec. 1991).

S1 Mathematical companion

S1.1 Models with additive representations and choice bias can make mistakes in TI but generalize correctly on average

Section 3.1 considered a model with no choice bias (i.e. $f(X, X) = 0$). Here we discuss the possible ramifications of a choice bias in additive representations. As a reminder, we consider a model

$$f(X, Y) = w \circ g_1(X) + w \circ g_2(Y).$$

We can understand the model's decision making in terms of two "implicit" ranks

$$r_1(X) := w \circ g_1(X), \quad r_2(Y) := -w \circ g_2(Y),$$

so that

$$f(X, Y) = r_1(X) - r_2(Y).$$

f will classify adjacent items i and $i + 1$, $i \in \{1, \ldots, n - 1\}$, correctly if and only if those ranks satisfy

$$r_1(i) > r_2(i + 1), \quad r_2(i) > r_1(i + 1).$$

We can characterize this condition more easily by defining two sequences: one starting at $i = 1$ with r_1 (we call this one s_1) and one starting at $i = 1$ with r_2 (we call this one s_2). At each item those sequences flip whether they specify the rank of r_1 or r_2. That is: s_1 specifies r_1 at odd items and r_2 at even items and s_2 behaves in the exact opposite way:

$$r_1(i) = \begin{cases} s_1(i) & \text{if } s \text{ odd}, \\ s_2(i) & \text{if } s \text{ even}, \end{cases} \quad r_2(i) = \begin{cases} s_2(i) & \text{if } s \text{ odd}, \\ s_1(i) & \text{if } s \text{ even}. \end{cases} \tag{S1}$$

By construction, f will classify adjacent items if and only if both s_1 and s_2 are monotonically decreasing. For items i, j with an odd symbolic distance, this immediately implies that f must also correctly order those, as $r_1(i)$ and $r_2(j)$ as well as $r_2(i)$ and $r_1(j)$ will be assigned to the same sequence and therefore be monotonic. For items with an even symbolic distance, however, the two ranks will be mapped to different sequences. The models do not have to generalize transitively and can generalize pairs of items incorrectly for arbitrarily large symbolic distances.

We leave with two notes. First, the average margin

$$\overline{f}(i, j) := \frac{1}{2} \left(f(i, j) - f(j, i) \right), \tag{S2}$$

has no choice bias even when f itself does (i.e. $\overline{f}(i, i) = 0$). As a result, this average margin still generalizes transitively even when f itself does not in generalize. Second, if the model classifies pairs (i, j) correctly for some even symbolic distance, it must also do so for any higher symbolic distance.

S1.2 The conjunctivity factor

We consider a general representation

$$g(i, j) \in \mathbb{R}^h, \quad h \in \mathbb{N},$$

where i and j denotes the numerical rank of the two presented items X, Y. We then consider the representational similarity (measured by the dot product) between two trials (i, j) and (i', j'):

$$K(i, j; i', j') := \langle g(i, j), g(i', j') \rangle. \tag{S3}$$

We assume that there are three values K can take, κ_s if $i = i'$ and $j = j'$, κ_o if either $i = i'$ or $j = j'$, and κ_d if $i \neq i'$ and $j \neq j'$. α is then defined as

$$\alpha := 1 - 2 \frac{\kappa_o - \kappa_d}{\kappa_s - \kappa_d}. \tag{S4}$$

We assume $\kappa_s \neq \kappa_d$. If $\kappa_s = \kappa_d$, then $\kappa_s = \kappa_o$, meaning that $g(i, j)$ is constant and unable to learn anything.

Lemma S1.1. *α is restricted to the range $[0,3]$. For arbitrary n, α can take on values in the range $[0,1]$.*

Proof. For compactness, we denote κ_o by o, κ_d by d, and κ_s by s. We then consider the Gram Matrix of all possible combinations of the first three items, i.e.

$$z := ((1,2),(2,1),(1,3),(3,1),(2,3),(3,2)), \tag{S5}$$

$$G := (K(z_i;z_j))_{ij} = \begin{pmatrix} s & d & o & d & d & o \\ d & s & d & o & o & d \\ o & d & s & d & o & d \\ d & o & d & s & d & o \\ d & o & o & d & s & d \\ o & d & d & o & d & s \end{pmatrix}. \tag{S6}$$

G is positive definite (as it is a Gram matrix), implying that all its eigenvalues are positive.

Using Mathematica, we determine two of these eigenvalues as

$$\lambda_1 = s + d - 2o, \quad \lambda_2 = s + o - 2d. \tag{S7}$$

We can transform

$$\alpha = \frac{s+d-2o}{s-d}.$$

As $s - d \geq 0$ (by basic properties of a kernel), $\lambda_1 \geq 0$ implies that $\alpha \geq 0$. Further, $\lambda_2 \geq 0$ implies that

$$0 \leq s + o - 2d = (s-d) + (o-d),$$

and therefore

$$0 \leq 1 + \frac{o-d}{s-d},$$

implying that $\alpha \leq 3$. Finally, α can take on arbitrary values in $[0,1]$ as $\alpha = 0$ corresponds to an itemic encoding and $\alpha = 1$ corresponds to a onehot encoding. By linear combinations of those encodings, we can achieve arbitrary values in-between. □

Remark S1.2. *The lemma leaves open which values in the range $(1,3]$ are attainable. In Appendix S3.2, we provide a construction that can attain $\alpha > 1$ for all n. Notably, this indicates a feature basis that encodes overlapping item pairs as less similar than distinct item pairs. For $n = 3$ attains $\alpha = 3$. As n increases, however, α decreases, gradually approaching 1. It remains an open question whether there are features that, at larger n attain higher values for α than our construction.*

In subsequent sections, we will constrain ourselves to $\alpha \in [0,1]$.

S1.3 Impact of nonlinearity, depth, and connectivity structure on a deep neural network's conjunctivity factor

We here compute the conjunctivity factor for a broad range of hidden representations, to better understand how it is impacted by different choices in architecture.

S1.3.1 Densely connected neural networks

We consider features that are determined in the following iterative manner:

$$h^{(0)} = z, \quad h^{(l+1)} = \tfrac{1}{\sqrt{H_{l+1}}}\phi(Wh^{(l)}), \quad W \in \mathbb{R}^{H_{l+1}\times H_l}. \tag{S8}$$

where ϕ is some nonlinearity. In particular, we focus on

$$\phi(x) := A\min(0,x) + B\max(0,x), \tag{S9}$$

which has as special cases ReLU [$A = 0$, $B = 1$, 146], leaky ReLU [$A < 1$, $B = 1$, 147], and the absolute value ($A = -1$, $B = 1$). As shown below, the conjunctivity factor of $h^{(l)}$ is invariant to ϕ's "magnitude"

$$m := A^2 + B^2, \tag{S10}$$

but depends on its "relative slope"

$$\rho := \frac{(B-A)^2}{A^2+B^2}. \tag{S11}$$

In particular, $\rho = 1$ for ReLU. (Note that similar analytical expressions could be derived for a broad range of other nonlinearities.)

We consider weights W that are sampled from a rotationally invariant distribution and denote their variance by σ_l^2. The expected similarity (which the empirical similarity will converge to in the large-width limit), is

$$\begin{aligned} K^{(l+1)}(z,z') := \mathbb{E}\left[\langle h^{(l+1)}(z), h^{(l+1)}(z')\rangle\right] &= \mathbb{E}_w\left[\phi(w^T h^{(l)}(z))\phi(w^T h^{(l)}(z'))\right] \\ &= \sigma_l^2 \|h^{(l)}(z)\|_2 \|h^{(l)}(z')\|_2 k\left(\hat{K}^{(l)}(z,z')\right), \end{aligned} \tag{S12}$$

where

$$\|h^{(l)}(z)\|_2^2 = K^{(l)}(z,z), \quad \hat{K}^{(l)}(z,z') = \frac{K^{(l)}(z,z')}{\|h^{(l)}(z)\|_2 \|h^{(l)}(z')\|_2}, \tag{S13}$$

and k is derived in previous work [78, 148–150] to be

$$k(u) = \frac{(B-A)^2}{2\pi}\left(\sqrt{1-u^2} + (\pi - \cos^{-1}(u))u\right) + ABu. \tag{S14}$$

Defining

$$\begin{aligned} \kappa_s^{(l)} &:= K^{(l)}((x,y),(x,y)), \\ \kappa_o^{(l)} &:= K^{(l)}((x,y),(x',y)) = K^{(l)}((x,y),(x,y')), \\ \kappa_d^{(l)} &:= K^{(l)}((x,y),(x',y')), \end{aligned} \tag{S15}$$

we infer

$$\kappa_s^{(l+1)} = \sigma_{l+1}^2 \kappa_s^{(l)} k(1) = \sigma_{l+1}^2 \kappa_s^{(l)}\left(\frac{(B-A)^2}{2} + AB\right) = \sigma_{l+1}^2 \kappa_s^{(l)} \frac{m}{2}. \tag{S16}$$

Further,

$$\kappa_o^{(l+1)} = \sigma_{l+1}^2 \kappa_s^{(l)} k\left(\frac{\kappa_o^{(l)}}{\kappa_s^{(l)}}\right), \quad \kappa_d^{(l+1)} = \sigma_{l+1}^2 \kappa_s^{(l)} k\left(\frac{\kappa_d^{(l)}}{\kappa_s^{(l)}}\right). \tag{S17}$$

Defining

$$\begin{aligned} \tilde{k}(u) := \frac{2}{m} k(u) &= \rho\left(\frac{\sqrt{1-u^2}}{\pi} + \left(1 - \frac{\cos^{-1}(u)}{\pi}\right)u\right) + (1-\rho)u = \\ &= u + \rho\frac{\sqrt{1-u^2} - \cos^{-1}(u)u}{\pi}, \end{aligned} \tag{S18}$$

and

$$\hat{\kappa}_\Delta^{(l)} := \frac{\kappa_\Delta^{(l)}}{\kappa_s^{(l)}} \tag{S19}$$

we infer

$$\hat{\kappa}_o^{(l+1)} = \tilde{k}\left(\hat{\kappa}_o^{(l)}\right) = \tilde{k}^{(l+1)}\left(\hat{\kappa}_o^{(0)}\right), \quad \hat{\kappa}_d^{(l+1)} = \tilde{k}\left(\hat{\kappa}_d^{(l)}\right) = \tilde{k}^{(l+1)}\left(\hat{\kappa}_d^{(0)}\right), \tag{S20}$$

where $\tilde{k}^{(l+1)}$ denotes $l+1$ applications of $\tilde{k}$. This allows us to recursively compute

$$\alpha = 1 - 2\frac{\hat{\kappa}_o^{(L)} - \hat{\kappa}_d^{(L)}}{1 - \hat{\kappa}_d^{(L)}}. \tag{S21}$$

As a special case, we consider input items x, which are all orthogonal to each other, meaning that

$$\kappa_s^{(0)} = 1, \quad \kappa_o^{(0)} = \tfrac{1}{2}, \quad \kappa_d^{(0)} = 0. \tag{S22}$$

and compute the similarities of the first hidden layer:

$$\hat{\kappa}_o^{(1)} = \tilde{k}(1/2) = \frac{1}{2} + \rho\left(\frac{\sqrt{3}}{2\pi} - \frac{1}{6}\right), \quad \hat{\kappa}_d^{(1)} = \tilde{k}(0) = \rho\frac{1}{\pi}. \tag{S23}$$

The conjunctivity factor α in this case is given by

$$\alpha = 1 - 2\frac{\hat{\kappa}_o^{(1)} - \hat{\kappa}_d^{(1)}}{1 - \hat{\kappa}_d^{(1)}} = \frac{1 + \hat{\kappa}_d^{(1)} - 2\hat{\kappa}_o^{(1)}}{1 - \hat{\kappa}_d^{(1)}} = \rho\frac{\pi/3 + 1 - \sqrt{3}}{\pi - \rho} \approx 0.15. \tag{S24}$$

S1.3.2 Neural networks with additive nodes

We can now extend these formulas to networks with partly additive nodes. In particular, we assume that the fraction $(1-\lambda)$ is fully connected and the fraction λ is only connected to one of the two items, though followed by the same nonlinearity. This part of the representation has the similarities

$$\hat{\mu}_d^{(l)} = \hat{\kappa}_d^{(l)}, \quad \hat{\mu}_o^{(l)} = \frac{1 + \hat{\kappa}_d^{(l)}}{2}, \quad \hat{\mu}_s^{(l)} = 1, \tag{S25}$$

and so the new normalized dissimilarities are given by

$$\tilde{\kappa}_o^{(l)} = (1-\lambda)\hat{\kappa}_o^{(l)} + \frac{\lambda}{2}(\hat{\kappa}_d^{(l)} + 1), \quad \tilde{\kappa}_d^{(l)} = \hat{\kappa}_d^{(l)}. \tag{S26}$$

This, in turn, means that

$$\tilde{\alpha}^{(l)} := 1 - 2\frac{\tilde{\kappa}_o^{(l)} - \tilde{\kappa}_d^{(l)}}{1 - \tilde{\kappa}_d^{(l)}} = \frac{1 + \tilde{\kappa}_d^{(1)} - 2\tilde{\kappa}_o^{(1)}}{1 - \tilde{\kappa}_d^{(1)}} = (1-\lambda)\frac{(1 + \hat{\kappa}_d^{(l)} - 2\hat{\kappa}_o^{(l)})}{1 - \hat{\kappa_d}^{(l)}}. \tag{S27}$$

Predictably, for a single layer, the additive units therefore simply reduce the conjunctivity factor by a proportional factor. For multiple layers, this reduction has ramifications on the densely connected units as well, as these units draw upon both the additive and densely connected units. In fact, in the limit of infinite depth, these ramifications are particularly strong:

Lemma S1.3. *For a nonlinearity ϕ with relative slope ρ and a proportion of additive units λ, the infinite-depth conjunctivity factor*

$$\alpha^{(\infty)} := \lim_{l\to\infty} \alpha^{(l)}. \tag{S28}$$

is given by

$$\alpha^{(\infty)} = \begin{cases} 0 & \text{if } \lambda > 0 \vee \rho = 0, \\ 1 & \text{else.} \end{cases} \tag{S29}$$

Proof. We prove the lemma in Appendix S3.5. □

S1.4 Standard statistical learning models have a transitive inductive bias

S1.4.1 Setup

We consider $n \in \mathbb{N}$ items $(x^{(i)})_{i=1,\dots,n}$, $x^{(i)} \in \mathbb{R}^d$. The task presents two of those items together and we denote this by

$$x^{(i,j)} = \begin{pmatrix} x^{(i)} \\ x^{(j)} \end{pmatrix} \in \mathbb{R}^{2d}. \tag{S30}$$

The corresponding label is given by

$$y^{(i,j)} = \begin{cases} 1 & \text{if } i < j, \\ -1 & \text{else,} \end{cases} \tag{S31}$$

and the training dataset consists of all adjacent items, i.e.

$$\mathcal{D} = \left\{ (x^{(i,i+1)}, 1) | i = 1, \dots, n-1 \right\} \cup \left\{ (x^{(i+1,i)}, -1) | i = 1, \dots, n-1 \right\}. \tag{S32}$$

We consider a set of features

$$\phi : \mathbb{R}^{2d} \to \mathbb{R}^h, \quad h \in \mathbb{N}, \tag{S33}$$

and a linear model of those latent features that predicts the output, i.e.

$$\hat{y}^{(i,j)} = \langle w, \phi(x^{(i,j)}) \rangle. \tag{S34}$$

We train these weights using some convex, monotonically decreasing loss function $\ell(y\hat{y})$, for example the quadratic loss

$$\ell(u) = (1-u)^2, \tag{S35}$$

or the crossentropy

$$\ell(u) = \log(1 + \exp(-u)). \tag{S36}$$

To simplify our notation, we will denote

$$\phi^{(i,j)} := \phi(x^{(i,j)}). \tag{S37}$$

Our loss function is then given by

$$L(w) = \sum_{(x,y)\in\mathcal{D}} \ell(y\langle w, \phi(x)) = \sum_{i=1}^{n-1} \ell(\langle w, \phi^{(i,i+1)}) + \ell(-\langle w, \phi^{(i+1,i)}). \tag{S38}$$

The following sections consider different special cases. Appendix S1.4.2 considers gradient flow on $L(w)$ and Appendix S1.4.3 considers gradient flow for the mean squared error in particular. Finally, Appendix S1.4.4 discusses ridge regression.

S1.4.2 General gradient flow dynamics

Setting

$$g(u) := -\ell'(u), \tag{S39}$$

we write

$$\dot{w} = -\nabla L(w) = \sum_{i=1}^{n-1} g(\langle w, \phi^{(i,i+1)} \rangle)\phi^{(i,i+1)} - g(-\langle w, \phi^{(i+1,i)} \rangle)\phi^{(i+1,i)}, \tag{S40}$$

and so writing

$$g^{(i,j)}(t) := g(y^{(i,j)} \langle w(t), \phi^{(i,j)} \rangle), \tag{S41}$$

we can express

$$\dot{g}^{(i,j)} = y^{(i,j)} g'(y^{(i,j)} \langle w(t), \phi^{(i,j)} \rangle) \langle \dot{w}, \phi^{(i,j)} \rangle. \tag{S42}$$

Depending on the pair under consideration, this results in a characteristic differential equation depending on other g's. In general we consider the intercept

$$b = \kappa_d \sum_{i=1}^{n-1} (g^{(i,i+1)} - g^{(i+1,i)}) \tag{S43}$$

Moreover, we define

$$h = g' \circ g^{-1}. \tag{S44}$$

Then, for $j = i + 1$,

$$\dot{g}^{(i,i+1)} = \delta_s h(g^{(i,i+1)}) \left(g^{(i,i+1)} - \tfrac{1-\alpha}{2}(g^{(i,i-1)} + g^{(i+2,i+1)}) + \frac{\kappa_o}{\delta_s} b \right), \tag{S45}$$

where we define as boundary conditions $g^{(0,1)} \equiv g^{(1,0)} \equiv 0$ and $g^{(n,n+1)} \equiv g^{(n+1,n)} \equiv 0$. Similarly, for $j = i - 1$,

$$\dot{g}^{(i+1,i)} = \delta_s h(g^{(i+1,i)}) \left(g^{(i+1,i)} - \tfrac{1-\alpha}{2}(g^{(i-1,i)} + g^{(i+1,i+2)}) + \frac{\kappa_o}{\delta_s} b \right). \tag{S46}$$

At initialization, $w = 0$ and therefore,

$$g^{(i,i+1)}(0) = g^{(i+1,i)}(0) = -\ell'(0) > 0. \tag{S47}$$

The differential equations above imply that the equality $g^{(i,i+1)}(t) = g^{(i+1,i)}(t)$ keeps holding and we define

$$g^{(i)} \equiv g^{(i,i+1)} \equiv g^{(i+1,i)}, \tag{S48}$$

and can express

$$\dot{g}^{(i)} = \delta_s h(g^{(i)}) \left(g^{(i)} - \tfrac{1-\alpha}{2}(g^{(i-1)} + g^{(i+1)}) \right), \tag{S49}$$

with the boundary conditions $g^{(0)} \equiv g^{(n)} \equiv 0$. At initialization, $g^{(i)}(0) = -\ell'(0) > 0$, as ℓ is monotonically decreasing. As ℓ is convex, h is strictly negative. If $\alpha = 0$, at initialization, for $i = 2, \ldots, n-2$,

$$g^{(i)} - \tfrac{1-\alpha}{2}(g^{(i-1)} + g^{(i+1)}) = 0,$$

As a result, the gradient of all intermediate items initially remains unchanged. The boundary conditions break this symmetry for the terminal items $g^{(1)}$ and $g^{(n-1)}$, which immediately decrease. Once they decrease, the gradient of their neighboring items, $g^{(2)}$ and $g^{(n-2)}$ will also become negative and in this way, the changes in g percolate from outside in. This causes a terminal item effect.

For $\alpha > 0$,

$$g^{(i)} - \tfrac{1-\alpha}{2}(g^{(i-1)} + g^{(i+1)}) > 0,$$

all $g^{(i)}$ decrease from the start. However, the terminal items $g^{(1)}$ and $g^{(n-1)}$ decrease faster, and the faster decreases gradually percolate in as well. As a result, for higher α, the terminal item effect is more attenuated, but nevertheless still obtains.

On the other hand, for $|i - j| > 1$,

$$\dot{\hat{y}}^{(i,j)} = \langle \dot{w}, \phi^{(i,j)} \rangle = \delta_s \tfrac{1-\alpha}{2}(g^{(i)} + g^{(j-1)} - g^{(i-1)} - g^{(j)})) = \dot{r}^{(i)} - \dot{r}^{(j)}, \tag{S50}$$

where we define the rank representation as

$$r^{(i)}(0) := 0, \quad \dot{r}^{(i)} = \delta_s \tfrac{1-\alpha}{2}(g^{(i)} - g^{(i-1)}). \tag{S51}$$

The terminal item effect implies that this rank representation will be increasing for the $i < \frac{n}{2} + 1$ and decreasing for $i > \frac{n}{2} + 1$. Further, the rank representation will be monotonic in i if $g^{(i-1)} - g^{(i)}$ has higher magnitude for more terminal i. Remarkably, this means that a sufficient condition for the emergence of a rank representation from gradient descent is given by the observation of a convex terminal item effect during training.

S1.4.3 Gradient descent dynamics on mean squared error

For the mean squared error,

$$\ell(u) = \tfrac{1}{2}(1-u)^2,$$

the dynamics are simplified as

$$h \equiv -1.$$

As a result,

$$\dot{g}^{(i)} = \delta_s(g^{(i)} - \tfrac{1-\alpha}{2}(g^{(i-1)} + g^{(i+1)})). \tag{S52}$$

Defining the vector

$$g \in \mathbb{R}^{n-1}, \quad g_i := g^{(i)}, \tag{S53}$$

these dynamics can be written as

$$\dot{g} = -Kg, \quad K_{ij} = \begin{cases} 1 & \text{if } i = j, \\ -\frac{1-\alpha}{2} & \text{if } |i-j| = 1, \\ 0 & \text{else.} \end{cases} \tag{S54}$$

Further, as $g(u) = 1 - u$,

$$\hat{y}^{(i,i+1)}(t) = 1 - g^{(i)}(t). \tag{S55}$$

We determine $g(t)$ by using the eigenvalue decomposition of K, as provided by Theorem S3.2. After some algebraic manipulation, this yields the following formulas for g_j and r_j:

Lemma S1.4. *Given the mean squared error as a loss function, g_j and r_j amount to*

$$g_j = \sum_{\substack{k=1 \\ k \text{ odd}}} \frac{2\sin\left(\frac{jk\pi}{n}\right)\exp\left(-\delta_s\left(1 - 2\alpha\cos\left(\frac{k\pi}{n}\right)\right)t\right)}{n\tan\left(\frac{k\pi}{2n}\right)}. \tag{S56}$$

and

$$r_j(t) = (1-\alpha)\sum_{\substack{k=1 \\ k \text{ odd}}} \frac{\left(\sin\left(\frac{jk\pi}{n}\right) - \sin\left(\frac{(j-1)k\pi}{n}\right)\right)\left(1 - \exp\left(-\delta_s\left(1 - (1-\alpha)\cos\left(\frac{k\pi}{n}\right)\right)t\right)\right)}{n\tan\left(\frac{k\pi}{2n}\right)\left(1 - (1-\alpha)\cos\left(\frac{k\pi}{n}\right)\right)}. \tag{S57}$$

Proof. See Appendix S3.3. □

We can further simplify this solution for $t \to \infty$:

Lemma S1.5. *Given the mean squared error as a loss function,*

$$\lim_{t\to\infty} \hat{y}^{(j,j+1)}(t) = 1, \tag{S58}$$

$$\lim_{t\to\infty} r_j(t) = \frac{\sinh((\frac{n+1}{2} - j)\lambda)}{\sinh(\frac{n+1}{2}\lambda) - \sinh(\frac{n-1}{2}\lambda)}, \tag{S59}$$

$$\lambda := \operatorname{arccosh}\left(\frac{1}{1-\alpha}\right). \tag{S60}$$

Proof. The solutions found by gradient descent and ridge regression (see below) converge to the same limit as $t, c \to \infty$. The lemma thus follows directly from Lemma S1.7 below. □

Remark S1.6. *Models are often trained on crossentropy for classification problems. While a closed-form solution to the learning dynamics cannot be determined, we can characterize its infinite-time solution. First, we note that crossentropy cannot be minimized by any finite value, but rather attains its minimum as $u \to \infty$. As a consequence, the weights w diverge: $\|w\|_2^2 \to \infty$. However, we can normalize by the minimal margin. This normalized model converges to the support vector machine solution [25]. As all $2(n-1)$ training cases are support vectors for both the regression and classification problem, support vector regression and classification are equivalent and the normalized model is characterized by the lemma above.*

S1.4.4 Ridge regression

Ridge regression minimizes

$$L(w) + \frac{1}{c}\|w\|_2^2, \tag{S61}$$

where ℓ is the mean squared error. The corresponding dual problem is given by

$$a := (\tilde{K} + \tfrac{1}{c}I)^{-1}y, \tag{S62}$$

where $\tilde{K}$ is the cross-sample similarity,

$$\tilde{K} \in \mathbb{R}^{(2(n-1))^2}, \quad \tilde{K}_{ij} = \begin{cases} \kappa_s & \text{if } i = j, \\ \kappa_o & \text{if } |i-j| \in \{n-2, n\}, \\ \kappa_d & \text{else.} \end{cases} \tag{S63}$$

Writing a as

$$a = \begin{pmatrix} a^{(1)} \\ a^{(2)} \end{pmatrix}, \quad a^{(1)} \in \mathbb{R}^{n-1}, \quad a^{(2)} \in \mathbb{R}^{n-1}, \tag{S64}$$

this amounts to the equation

$$\begin{aligned} (\delta_s + c^{-1})a_i^{(1)} + \delta_o(a_{i-1}^{(2)} + a_{i+1}^{(2)}) + \kappa_d\langle a\rangle &= 1, \\ (\delta_s + c^{-1})a_i^{(2)} + \delta_o(a_{i-1}^{(1)} + a_{i+1}^{(1)}) + \kappa_d\langle a\rangle &= -1, \end{aligned} \tag{S65}$$

where $\langle a\rangle = \sum_{i=1}^{2(n-1)} a_i$ and, for notational simplicity, $a_0^{(j)} = a_n^{(j)} = 0$. This means that if $a^{(1)}$, $a^{(2)}$ are a solution to the equation, then so are

$$\tilde{a}^{(1)} := -a^{(2)}, \quad \tilde{a}^{(2)} := -a^{(1)}. \tag{S66}$$

Since (S62) is well-defined, the solution a must be unique and we can infer

$$a^{(1)} = -a^{(2)}. \tag{S67}$$

Setting

$$\tilde{c} := \delta_s c, \tag{S68}$$

(S65) amounts to

$$(1 + \tilde{c}^{-1})\overline{a}_i - \tfrac{1-\alpha}{2}(\overline{a}_{i-1} + \overline{a}_{i+1}) = \frac{1}{\delta_s}, \tag{S69}$$

where

$$a^{(1)} = \overline{a}, \quad a^{(2)} = -\overline{a}. \tag{S70}$$

In matrix form, this results in

$$\frac{1}{\delta_s}(K + \tilde{c}^{-1}I)^{-1}1 = \overline{a}, \tag{S71}$$

where K is defined as in (S54). Setting $B := K + \tilde{c}^{-1}I$, we note that B is a tridiagonal matrix with diagonal $\frac{1+\tilde{c}^{-1}}{\delta_s}$ and off-diagonal $-\frac{1-\alpha}{2\delta_s}$. The ratio between diagonal and off-diagonal is therefore given by

$$-2\tfrac{1+\tilde{c}^{-1}}{1-\alpha} \leq -2. \tag{S72}$$

As this ratio is smaller than -2, we can apply Theorem S3.1 to infer

$$B_{ij}^{-1} = \frac{\cosh\left((n - |j-i|)\lambda\right) - \cosh\left((n-i-j)\lambda\right)}{(1-\alpha)\sinh(\lambda)\sinh(n\lambda)}, \quad \lambda := \operatorname{arccosh}\left(\frac{1+\tilde{c}^{-1}}{1-\alpha}\right). \tag{S73}$$

The predictions made by $\overline{a}$ are expressed as

$$f(x_j, x_k) = \begin{cases} \delta_s(\overline{a}_j - \frac{1-\alpha}{2}(\overline{a}_{j-1} + \overline{a}_{j+1})) & \text{if } k = j+1, \\ -\delta_s(\overline{a}_k - \frac{1-\alpha}{2}(\overline{a}_{k-1} + \overline{a}_{k+1})) & \text{if } k = j-1, \\ \delta_s \frac{1-\alpha}{2}(\overline{a}_j + \overline{a}_{k-1} - \overline{a}_k - \overline{a}_{j-1}) & \text{else,} \end{cases} \tag{S74}$$

where we define the boundary conditions $\overline{a}_0 = \overline{a}_n = 0$. In particular, this makes apparent the emergent rank representation

$$r_j = \delta_s \tfrac{1-\alpha}{2}(\overline{a}_j - \overline{a}_{j-1}). \tag{S75}$$

Using hyperbolic trigonometric identities, we can simplify these expression further, resulting in the following lemma:

Lemma S1.7. *For non-adjacent items,*

$$f(x_j, x_k) = r_j - r_k, \quad r_j = \frac{\sinh((\frac{n+1}{2} - j)\lambda)}{\sinh(\frac{n+1}{2}\lambda) - \sinh(\frac{n-1}{2}\lambda)}, \quad \lambda := \operatorname{arccosh}\left(\frac{1+\tilde{c}^{-1}}{1-\alpha}\right). \tag{S76}$$

For adjacent items,

$$f(x_j, x_{j+1}) = m \cdot 1 + (1-m)(r_j - r_{j+1}), \quad m := \frac{\alpha}{\alpha + \tilde{c}^{-1}}. \tag{S77}$$

Proof. See Appendix S3.4. □

S1.5 Exact learning dynamics of neural networks trained with backpropagation in the lazy regime

S1.5.1 Neural tangent kernel

We consider the deep network

$$f_w^{(0)}(x) := W^{(0)}x, \quad f_w^{(l+1)}(x) := W^{(l+1)}\phi(f_{w^{(:l)}}^{(l)}(x)), \quad W^{(l)} \sim \mathcal{N}(0, \sigma_l^2/H_l), \tag{S78}$$

where

$$w^{(:l)} = \left(W^{(0)}, \ldots, W^{(l)}\right), \tag{S79}$$

is the sequence of weights up to the l-th layer. We denote

$$\theta = \operatorname{vec}\left(w^{(:L)}\right) \tag{S80}$$

as the vectorized total parameters of the network, writing $f_\theta(x) = f_{w^{(:L)}}^{(L)}(x)$. The *neural tangent features* are given by the linearization of the network around its initial parameters θ_0,

$$\tilde{f}(x) = f_{\theta_0}(x) + \langle \nabla_\theta f_{\theta_0}(x), (\theta - \theta_0) \rangle, \tag{S81}$$

and the *neural tangent kernel* is given by the similarity according to those features:

$$\overline{K}(x, x') := \langle \nabla_\theta f_{\theta_0}(x), \nabla_\theta f_{\theta_0}(x') \rangle. \tag{S82}$$

If the distance $\|\theta - \theta_0\|_2^2$ remains infinitesimally small throughout training, the linear approximation $\tilde{f}$ remains valid, and backpropagation can be approximated by gradient descent on the linear readout model given above. By computing the conjunctivity factor associated with $\overline{K}$, we can therefore predict the behavior of neural networks in the regime where the approximation is valid (the "lazy" regime). We make this computation in the next section, but previously remark on our choice of initialization.

Remark S1.8. *We here scale the weight variance by the input dimensionality, H_l, rather than the output dimensionality, H_{l+1}. This initialization is standard in practice as it conserves the variance of the network output throughout the layers with adequately chosen σ_l^2 [151]. In investigations of the neural tangent kernel, weights are frequently scaled by the output dimensionality instead. Part of the reason is that, as we will see below, standard scaling leads to less interesting behavior in wide neural networks. We here use the neural tangent kernel to show that our theory captures the behavior of neural networks in the lazy regime. For alternative initialization scales, this would still be the case, though the conjunctivity factor may be different.*

S1.5.2 The NTK's conjunctivity factor

We consider a neural network with one hidden layer. Note that

$$\nabla_{W_{i\cdot}^{(0)}} f_\theta(x) = W_i^{(1)} \phi'(W_{i\cdot}^{(0)} x), \quad \nabla_{W_i^{(1)}} f_\theta(x) = \phi(W_{i\cdot}^{(0)} x). \tag{S83}$$

As a result,

$$\begin{aligned}\mathbb{E}[\overline{K}(x, x')] &= \sum_{i=1}^{H_1} \left(\mathbb{E}\left[\left\langle \nabla_{W_{i\cdot}^{(0)}} f_\theta(x), \nabla_{W_{i\cdot}^{(0)}} f_\theta(x') \right\rangle \right] + \mathbb{E}\left[\left\langle \nabla_{W_i^{(1)}} f_\theta(x), \nabla_{W_i^{(1)}} f_\theta(x) \right\rangle \right] \right) \\ &= \sigma_1^2 \mathbb{E}_{w\sim\mathcal{N}(0,1)} \left[\phi'(w^T x)\phi'(w^T x') \right] + H_1 \frac{\sigma_0^2}{d} \mathbb{E}_{w\sim\mathcal{N}(0,1)} \left[\phi(w^T x)\phi(w^T x') \right].\end{aligned}$$

This means that as H_1 grows large, this similarity diverges. Normalizing the similarity by H_1, in turn, highlights that it is dominated by the similarity of the hidden layer. As a result, under standard initialization, the conjunctivity factor of the neural network with one hidden layer trained with backpropagation is equal to the conjunctivity factor of the hidden layer of the neural network.

In a sense, backpropagation is "trivial" in this regime as it is equivalent to training a linear readout. This is the origin of the different initialization paradigm highlighted in Remark S1.8. If the variance of $W^{(0)}$ is also scaled with H_1, the first term and the second term have comparable magnitude. While we refrain from computing the conjunctivity factor in that case, we note that similar analytical solutions can be derived and the resulting conjunctivity factor obeys the same qualitative trends observed in the previous section, though it is generally slightly higher.

S2 Experimental details

All experiments and figures can be reproduced from the public Github repository: https://github.com/sflippl/relational-generalization-in-ti.

S2.1 Data analysis and visualization

We analysed the data in R [152] and provide an R package containing all datasets listed below. In addition, we provide infrastructure to load these datasets into Python. The visualizations were created using `ggplot2` [153] and `patchwork` [154].

S2.2 TI behavior in models with non-exchangeable representations

To explore whether our theory can capture the behavior of a non-exchangeable representation, we explored the behavior of models with fixed, randomly sampled ReLU features. For a sufficiently small number of units, this stochasticity results in a drop in accuracy (Fig. S1a), though the margin remains the same (Fig. S1b). Further, because the representations violate exchangeability, the individual models' behavior is not fully captured by the theoretical account. For 1000 units, for example, the models exhibited considerable spread around the theoretical prediction, though they were still well approximated (Fig. S1c,e). Notably, our account captured the average behavior across many models much better, even for a small number of units (Fig. S1d,e).

S2.3 Deep neural network training

We trained deep neural networks with standard He initialization [151], transitioning between the lazy and rich regime by scaling the output of the networks. An output scaling of σ is equivalent to scaling the initialization by $\sigma^{1/(l+1)}$, where l is the depth of the network. (Indeed, the main text describes the transition between the lazy and rich regime in terms of scaling the initialization.) In the lazy regime, the network adheres to a linear readout model with an additional random intercept given by its initial values $f_{\theta_0}(x)$ (see

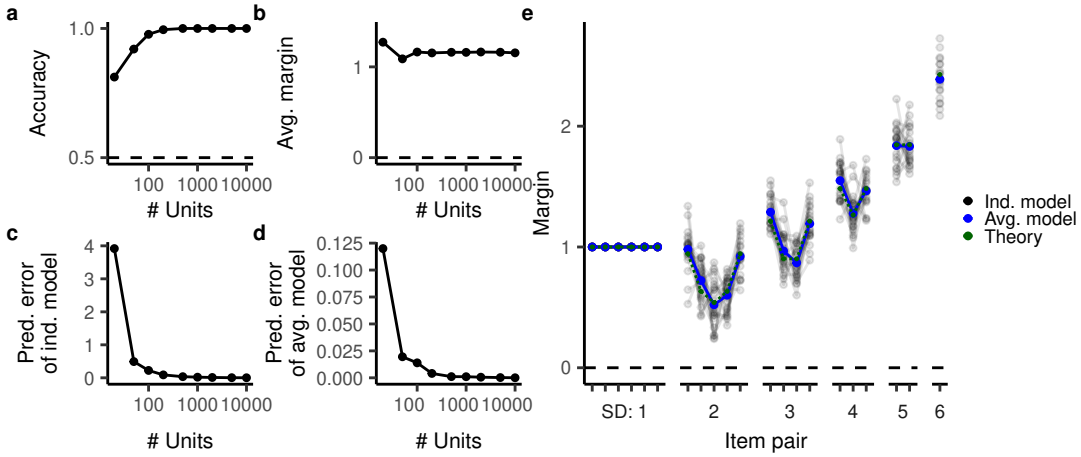

Figure S1: Behavior of models with fixed, randomly sampled ReLU representations with different numbers of units. **a-b**, Accuracy (a) and average margin (b). **c**, Error of the theoretical prediction compared to the individual model prediction. **d**, Error of the theoretical prediction compared to the average model prediction. **e** Margins for individual ReLU networks with 1000 units (black translucent lines), their average (blue line), and the theoretical prediction (green line).

Appendix S1.5.1). As this complicates analysis, but does not qualitatively change behavior, we subtract this initial intercept. The complete network was therefore given by

$$\tilde{f}_\theta(x) := \sigma(f_\theta(x) - f_{\theta_0}(x)). \tag{S84}$$

We implemented network training in Pytorch [155], using stochastic gradient descent without momentum. The learning rate was chosen automatically by the following process. We started at an extremely high learning rate of 10^{37}. (This is because we considered extremely big ranges of scalings, some of which required high learning rates.) If network training diverged, we divided the learning rate by 10 and restarted training. This is not an efficient way to determine the optimal learning rate, but suffices for our purposes. Further, we trained the networks until they 1) had a loss under 10^{-5} and 2) did not improve their loss in the last 50 epochs.

S2.4 Analysis of the hidden-layer representation

Fig. 6 presented the representational dissimilarity matrix (RDM) for a subset of item pairs. Fig. S2 depicts the full trial-by-trial RDM. Candidates for a useful hidden-layer representation are an exchangeable representation (Fig. 6a, top row) or a rank representation; such a rank representation may represent both items' ranks individually (Fig. 6a, middle row) or it may represent the difference between the items' ranks (Fig. 6a, bottom row). We compare the hypothetical representations to those empirically found in the network (Fig. S2b). While Fig. 6 presented those of the postactivation, $\phi(W^{(0)}x)$, we here additionally analyse the preactivation, $W^{(0)}x$ and the neural tangent kernel (NTK; Appendix S1.5). All of them have an exchangeable representation in the lazy regime. In the rich regime, the representation exhibits some similarity to an exchangeable representation (in particular for the NTK) and perhaps a rank representation. There is also some structure, however, that appear not to be captured by any of the hypothetical RDMs.

To make this more rigorous, we try to predict the RDM by fitting an exchangeable representation, a representation of each items' rank, a representation of rank difference, and a mixture of all three representations. The free parameters of the exchangeable representation are the three distinct similarities. The free parameters of the rank representation are distinct ranks for each item, where we allow for different ranks depending on whether an item was presented first or second. We fit these parameters using gradient descent and measure the proportion of variance explained (Fig. S2c). This confirms our previous interpretation: the lazy regime is perfectly captured by a ternary representation, whereas the rich regime is partly captured by a ternary representation (in particular for the NTK representation). In addition, the rich regime is

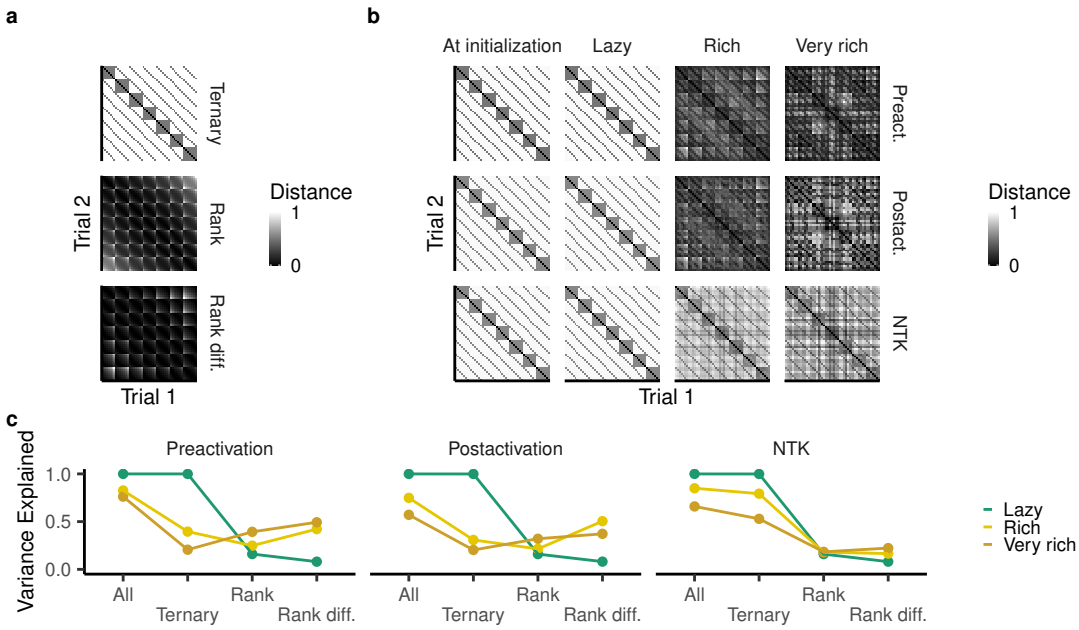

Figure S2: Representational dissimilarity of the hidden-layer representation, computed as the squared Euclidean distance between the representations. **a**, Trial-by-trial representational dissimilarity matrix (RDM) of a hypothetical hidden layer with either an exchangeable representation, a representation of both items' rank or a representation of the rank difference. The trials are arranged in lexical ordering, i.e. AA, AB, ..., AG, BA, ..., BG, ..., GG. **b**, RDM of the networks' preactivations, postactivations, and neural tangent features at initialization and after lazy or rich training. **c**, Variance explained by a model which predicts the RDM in terms of a ternary representation, a representation of rank, a representation of rank difference, or a mixture of all three ("All").

partly captured by a representation of rank difference. However, even a mixture of all three hypothetical representations cannot fully explain the representational structure.

S2.5 Detailed analysis of deep neural network behavior on transitive inference

This section presents a more detailed analysis of the deep network behavior. The results below demonstrate, in particular, that changing the activation function or considering deeper networks does not straightforwardly improve TI behavior. Indeed, these extended results further demonstrate the fundamental strangeness of the deep networks' behavior on TI.

S2.5.1 Rank representability

The "rank representability" measures how well TI behavior can be captured by a rank representation, a value of 0 indicating that it can be perfectly captured. Specifically, a linear model is fit to approximate

$$\hat{f}(i,j) = r_1(i) + r_2(j), \tag{S85}$$

where r_1 and r_2 are individually estimated for each i and j. The rank representability is then given by the mean squared error of this approximation. As expected, in the lazy regime, the rank representability is zero. As the scaling decreases (and training enters the rich regime), the error increases, indicating that network behavior cannot be described by a rank representation anymore (Fig. S3a). This confirms the picture observed in Fig. 6e.

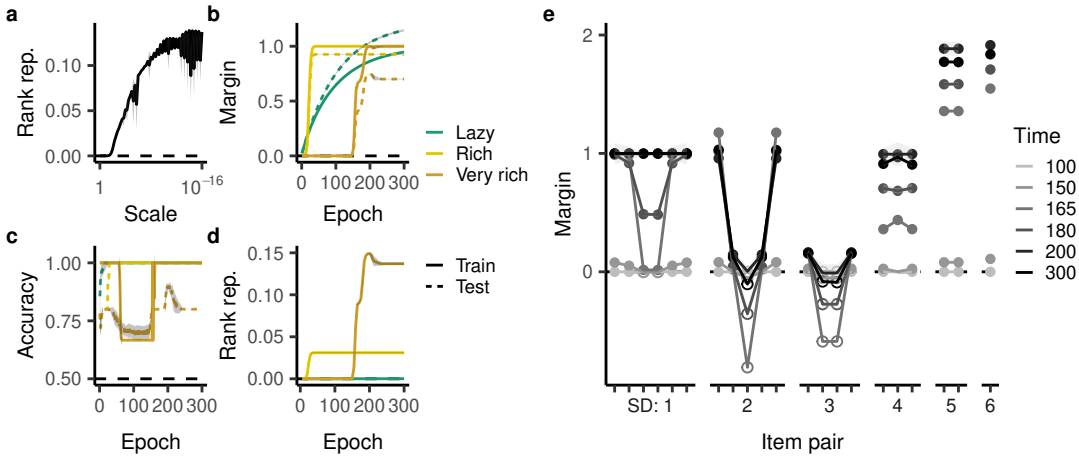

Figure S3: TI behavior of deep networks as a function of initialization scale and training duration. **a**, Rank representability at the end of training as a function of initialization scale. **b-d**, Margin, accuracy, and rank representability over the first 300 epochs. **e**, Margin at the different item pairs as a function of training duration (very rich regime).

S2.5.2 Time trajectory

Fig. S3b-e depict the behavior of the neural network with one hidden layer over training. Notably, learning in the rich regime occurs much more suddenly (Fig. S3b) than in the lazy regime. Further, the TI behavior over training exhibits several curious intermediate stages (Fig. S3e). Specifically, the terminal items are learned essentially completely before the intermediate item pairs even improve slightly. In fact, the margin of the item pairs CE, BE, and CF decreases considerable before increasing again.

S2.5.3 Effect of the number of hidden units

At a fixed initialization, neural networks approach the lazy regime as they become wide [89, 90]. Neural networks with few hidden units therefore tend to be in the rich regime at standard initialization. Further, the random initialization leads to deviations from the NTK predictions (which are generated from the expected neural tangent kernel) even in the lazy regime; these deviations are larger for fewer hidden units. We therefore decided to explore how the number of hidden units affects the behavior of the network at different initialization scales. We find that at 100 and 1000 hidden units, the network behavior deviates substantially from our theoretical prediction even at standard initialization (Fig. S4a). Part of this is due to the random noise of individual initializations. As a result, the prediction of twenty models averaged together are much more accurate and, in fact, close to zero for 1,000 hidden units (Fig. S4b). At 100 units, in contrast, we do observe a systematic bias (Fig. S4d). Further, at 100 units, model behavior cannot be entirely described by a rank representation, whereas even individual models with 1,000 units do implement a rank representation (Fig. S4c). Taken together, this suggests that at 100 hidden units, the neural network is in the rich regime even at standard initialization, but the effect of the rich regime can be exacerbated by further decreasing the scale of initialization.

S2.5.4 Effect of the nonlinearity

Here we explore the effect of different nonlinearities: both piecewise linear functions and a tanh function as an instance of a saturating nonlinearity. Generally, as the difference in slopes ρ of the piecewise linear function increases (see Appendix S1.3 for a definition of ρ), test margin and test accuracy decrease (Fig. S5a,b) and their rank representation error increases (Fig. S5c). In fact, for $\rho \geq 1.2$, the network performs *below chance-level*. While TI behavior is better at $\rho = 0.5$, it is still non-naturalistic, as an inverse terminal item effect is consistently observed (Fig. S5g). For the tanh nonlinearity, it is not as clear how the rich regime can be approached, as a small output scaling can only be overcome by making the readout large due to

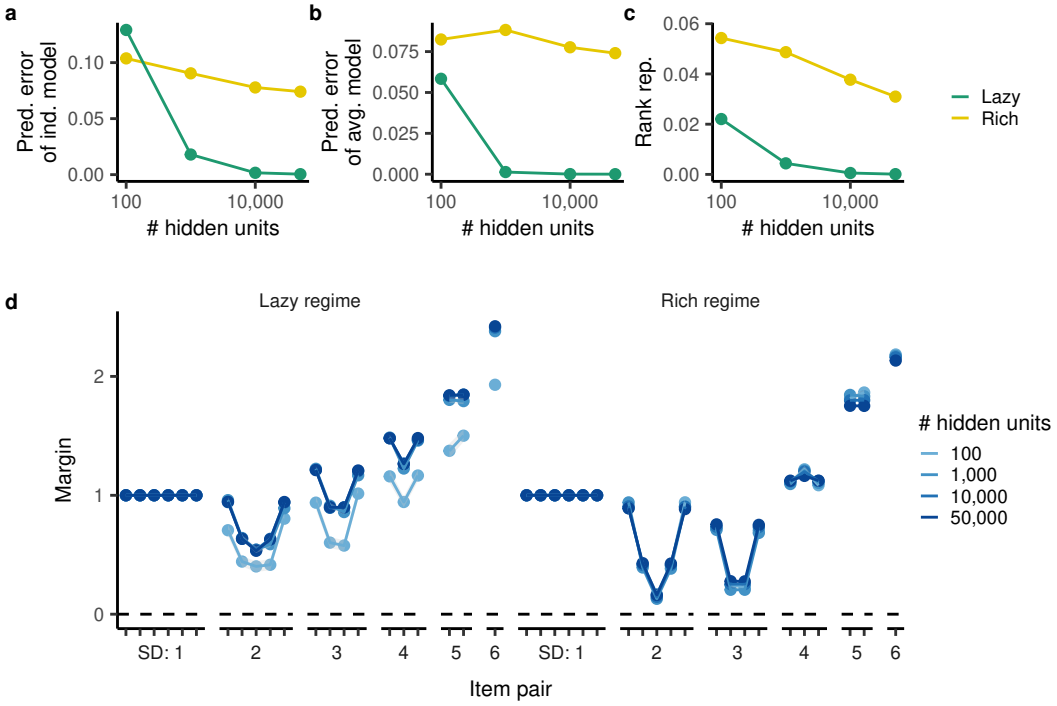

Figure S4: TI behavior for different numbers of hidden units and standard initialization ("lazy") as well as smaller initialization (scale 10^{-3}, "rich"). **a**, Error of the theoretical prediction compared to the individual model prediction. **b**, Error of the theoretical prediction compared to the average model prediction. **c**, Rank representability as a function of the number of hidden units. **d**, Predictions of the networks in different regimes and for different numbers of hidden units.

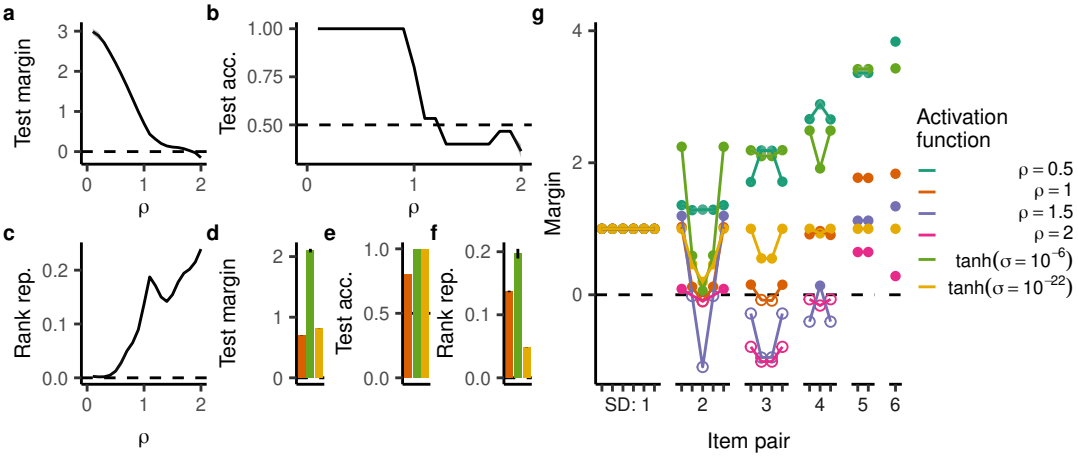

Figure S5: TI behavior for different activation functions. **a-c**, Margin (a), accuracy (b), and rank representability (c) as a function of ρ (see (S11)). **d-f**, Margin (d), accuracy (e) and rank representability (f) for the tanh activation function at two different output scalings σ (as well as, for comparison, the ReLU activation function ($\rho = 1$)). **g**, Margin of the different item pairs for selected activation functions.

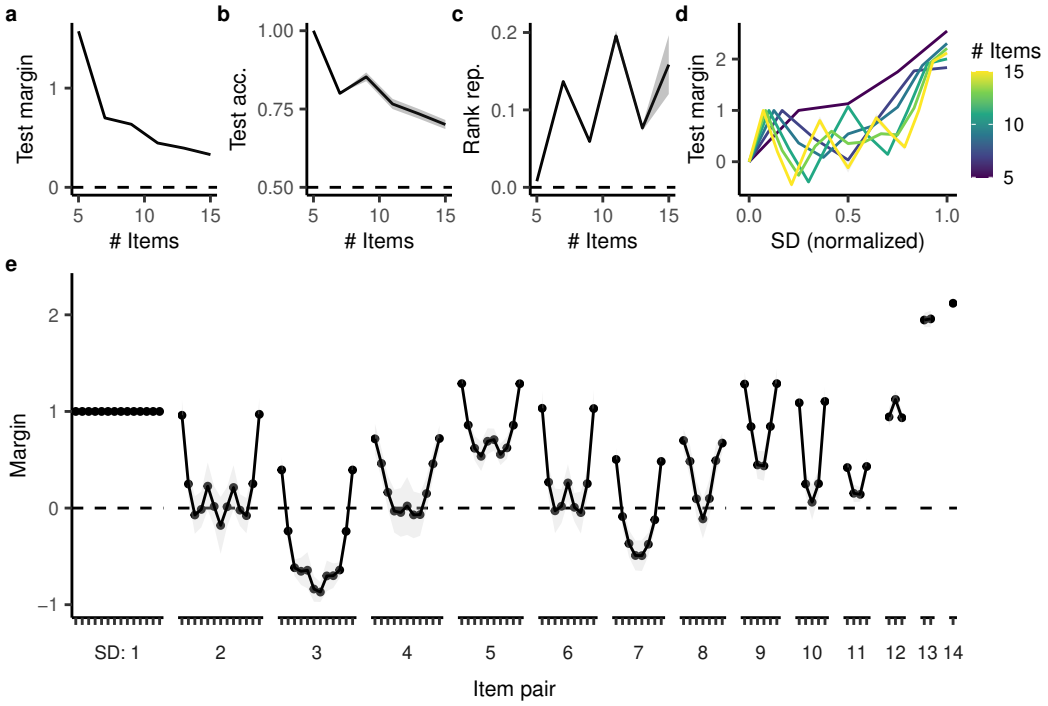

Figure S6: TI behavior as a function of the number of items. **a-c**, Margin, accuracy, and rank representability as a function of the number of items. **d**, The average test margin for a given symbolic distance (normalized by $n-1$). **e**, Margin at the different item pairs for 15-item TI.

the saturating nonlinearity. Here we explore two different scalings: 10^{-6} and 10^{-22}. We find that the tanh nonlinearity has better performance, though it is still characterized by an extremely sharp terminal item effect at a symbolic distance of two (Fig. S5g).

S2.5.5 Effect of the number of items

We next determined the behavior of ReLU networks with one hidden layer in the very rich regime for different numbers of items. Generally, both test margin and accuracy decreased as the number of items increased (Fig. S6a,b). The rank representability exhibited a more non-monotonic pattern. In particular, none of the networks except for those trained on 5-item TI exhibited a symbolic distance effect (Fig. S6d). Instead their margin oscillated at intermediate symbolic distances before eventually increasing at large symbolic distances.

S2.5.6 Networks trained on crossentropy

Neural networks are often trained on crossentropy rather than mean squared error. We therefore study the inductive bias of neural networks trained on TI with crossentropy. Notably, the learning trajectory of neural networks trained on crossentropy can generally not be approximated by their neural tangent kernel as the weights must diverge to minimize the loss function. This is because crossentropy is minimized as the margins diverge towards infinity (see Remark S1.6). Nevertheless, we consider networks with standard (scaling 1) or small (scaling 10^{-3}) initialization and find that they behave differently, presumably because we train for a finite number of epochs. Further, we consider both learning by backpropagation and only adjusting the linear readout.

While the margins grow over the duration of learning (Fig. S7a), we find that the normalized margins of the linear readout model are well approximated by our theory. This confirms our theoretical results (Remark S1.6). More surprisingly, the neural network trained by backpropagation with standard initialization is also well approximated by our theory (Fig. S7b). Only those networks trained with backpropagation and initialized with small weights exhibit behavior similar to that of neural network trained on MSE in the rich

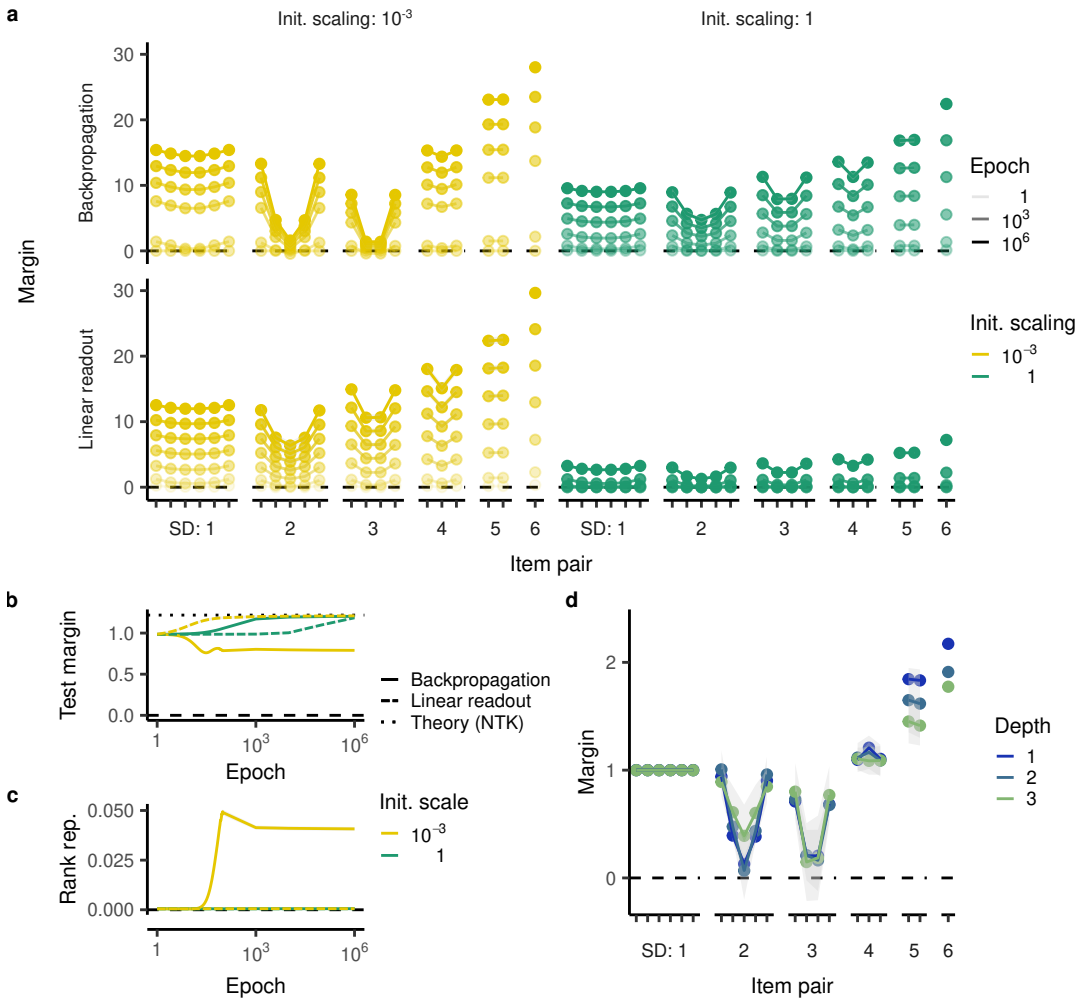

Figure S7: **a**, TI behavior of networks trained on crossentropy. The top row depicts a network that has been trained by backpropagation; the bottom row a network whose linear readout has been trained. Further, the left column is initialized with smaller weights than the right column. **b**, Test margin of the networks over training. **c**, Rank representability of the networks over training. **d**, TI behavior of networks with different depths.

regime (Fig. S7a-c). There are two potential reasons for this: it is possible that after longer amounts of training, the neural network with standard initialization may exhibit similar behavior to that with smaller initialization [140]. It is also possible that different initializations genuinely affect the inductive bias of neural networks trained with crossentropy. This is an interesting possibility especially in light of the fact that known theoretical results are largely silent on such an effect [139, 140].

S2.5.7 Deeper networks

While neural networks with one hidden layer already exhibit representation learning, the superior generalization of deep neural networks is often tied to their depth. We therefore explore how increasing depth changes the networks' inductive bias on TI. We use an initialization scale of 10^{-3} and train networks with one to three hidden layers. (Deeper networks do not converge to a minimal loss within six hours and we leave their exploration to future work.) We find little difference between the behavior of networks with different depth (Fig. S7d), suggesting that simply increasing the depth of the networks is not sufficient to improve their TI performance.

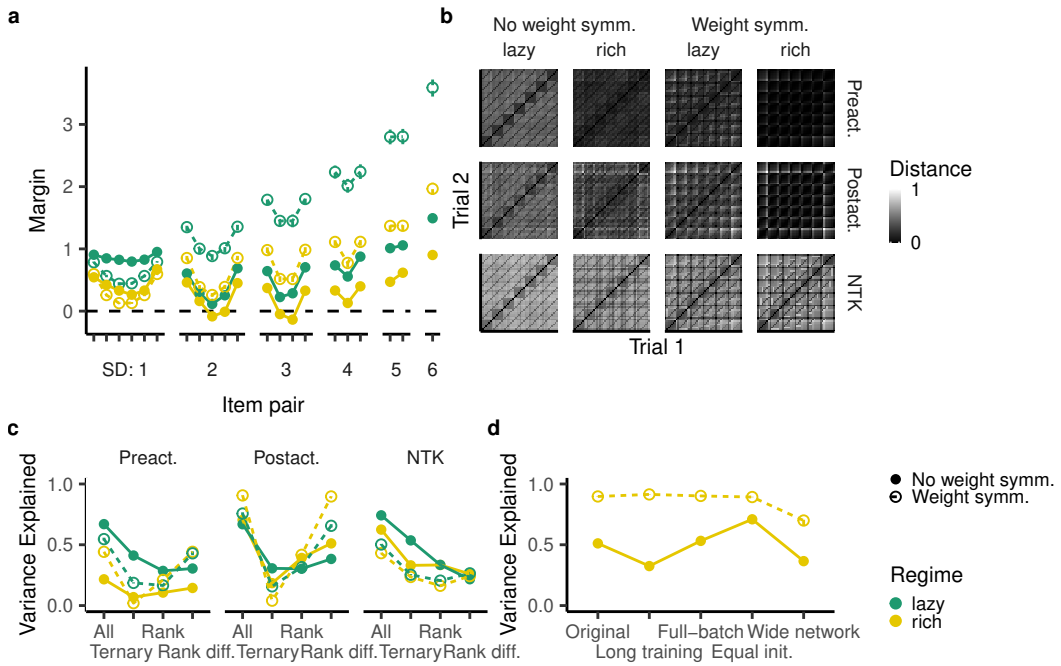

Figure S8: Effect of weight symmetry as implemented in [38]. **a**, Generalization behavior in the lazy and rich regime with and without weight symmetry. **b**, Trial-by-trial RDM of the preactivation, postactivation and NTK in the lazy and rich regime with and without weight symmetry. **c**, Variance of those RDMs explained by a ternary representation, a rank representation, a representation of rank difference, and a mixture of all representations. **d**, Variance explained of the RDMs that introduce one of the changes introduced by our training protocol. Long training indicates training until convergence and equal init. indicates that both the hidden layer and the readout layer were initialized with a scaling of 0.025. Finally, the wide network indicates a network with 1000 units.

S2.6 Symmetric weights in the input layer

In related work, Nelli *et al.* [38] also compare the TI behavior of deep neural networks in the lazy and rich regime, finding that neural networks trained in the rich regime have an emergent rank representation in their hidden layer. Notably, their networks implement a weight symmetry between the two input items. Specifically, given two items $x_1, x_2 \in \mathbb{R}^d$, their network computes the output as

$$f(x_1, x_2) = w_2^T \phi(W_1 x_1 - W_1 x_2), \quad \phi(u) = \max(u, 0), \quad w_2 \in \mathbb{R}^h, \quad W_1 \in \mathbb{R}^{h \times d}. \tag{S86}$$

At first glance, this network is not additive – after all, ϕ mixes the two items nonlinearly. However, only the directional bias can be nonlinear. The average choice of the network between either order of presentation,

$$\overline{f}(x_1, x_2) := f(x_1, x_2) - f(x_2, x_1) = w_2^T (W_1 x_1 - W_1 x_2), \tag{S87}$$

is additive, as $\phi(u) - \phi(-u) = u$.

Indeed, weight symmetry strongly improves test performance in the rich and lazy regime (Fig. S8a). In line with the parameters used in [38], the rich regime corresponds to an initialization scaling of 0.025 in the hidden layer and a scaling of 1 in the readout layer, whereas the lazy regime corresponds to a scaling of 1 in both layers. Note that training performance is suboptimal because the network is only trained for 40 epochs. The RDM of the postactivation (Fig. S8b) corresponds to a rank difference representation (Fig. S8c), whereas such a representation leaves a substantial proportion of variance unexplained in the lazy regime as well as without weight symmetry. Notably, neither the preactivation nor the NTK can be explained as a rank difference representation.

In addition to weight symmetry, Nelli *et al.* implement training with some differences to our implementation: they run stochastic gradient descent with a batch size of one, use networks with only twenty units, run

training for 40 epochs rather than until convergence, and initialize the hidden-layer weights with a smaller variance than the output weights (specifically multiplying them by a factor of 0.025). To investigate whether these differences are responsible for the emergence of a rank difference representation, we changed each of these hyperparameters independently. We found that weight symmetry is necessary and sufficient for the emergence of a rank difference representation. Note that in the wide network, the rank difference representation is weaker because the network is leaving the rich regime due to its increasing width. To re-enter the rich regime, we would have to initialize the weights by a smaller factor.

S3 Detailed mathematical analysis

S3.1 Prerequisites

S3.1.1 Results on Tridiagonal Matrices

Theorem S3.1 (Hu & O'Connell [156]). *Let $a, b \in \mathbb{R}$ such that $\frac{a}{b} \leq -2$ and*

$$B \in \mathbb{R}^{m\times m}, \quad B_{ij} = \begin{cases} a & \text{if } i = j, \\ b & \text{if } |i-j| = 1, \\ 0 & \text{else.} \end{cases} \tag{S88}$$

Then

$$B^{-1}_{ij} = -\frac{\cosh\left((m+1-|j-i|)\lambda\right) - \cosh\left((m+1-i-j)\lambda\right)}{2b\sinh(\lambda)\sinh\left((m+1)\lambda\right)}, \tag{S89}$$

where

$$\lambda := \operatorname{arccosh}\left(-\frac{a}{2b}\right). \tag{S90}$$

Proof. We define $\tilde{B} := \frac{1}{b}B$. Using [156],

$$\tilde{B}^{-1}_{ij} = -\frac{\cosh\left((m+1-|j-i|)\lambda\right) - \cosh\left((m+1-i-j)\lambda\right)}{2\sinh(\lambda)\sinh\left((m+1)\lambda\right)}, \tag{S91}$$

and thus

$$B^{-1} = \tfrac{1}{b}\tilde{B}^{-1}. \tag{S92}$$

□

Theorem S3.2 (Böttcher & Grudsky [157]). *A tridiagonal matrix*

$$B \in \mathbb{R}^{m\times m}, \quad B_{ij} = \begin{cases} a & \text{if } i = j, \\ b & \text{if } |i-j| = 1, \\ 0 & \text{else,} \end{cases} \tag{S93}$$

has the eigenvalue decomposition

$$\begin{aligned} A &= U\Lambda U, \\ \Lambda_{jj} &= a + 2|b|\cos\left(\frac{\pi j}{m+1}\right), \\ U_{jk} &= \sqrt{\frac{2}{m+1}}\sin\left(\frac{jk\pi}{m+1}\right). \end{aligned} \tag{S94}$$

Proof. This is a special case of [157, Theorem 2.4]. All that remains is proving the scaling factor. We use

$$\sin^2(x) = \frac{1-\cos(2x)}{2}, \tag{S95}$$

to infer

$$\begin{aligned}\sum_{k=1}^{m}\sin^2\left(\frac{jk\pi}{m+1}\right) &= \frac{m}{2}-\frac{1}{2}\sum_{k=1}^{m}\cos\left(2\frac{jk\pi}{m+1}\right)\\ &= \frac{1}{2}\left(m-\frac{\sin\left((2m+1)\frac{j\pi}{m+1}\right)-\sin\left(\frac{j\pi}{m+1}\right)}{2\sin\left(\frac{j\pi}{m+1}\right)}\right)\\ &= \frac{1}{2}\left(m+\frac{2\sin\left(\frac{j\pi}{m+1}\right)}{2\sin\left(\frac{j\pi}{m+1}\right)}\right)=\frac{m+1}{2},\end{aligned} \tag{S96}$$

where the last equality holds because

$$\sin\left((2m+1)\tfrac{j\pi}{m+1}\right)=\sin\left((2m+1-2(m+1))\tfrac{j\pi}{m+1}\right)=\sin\left(-\tfrac{j\pi}{m+1}\right). \tag{S97}$$

Note that we know that the eigenvectors are orthogonal because the matrix is symmetric and thus normal. □

S3.1.2 Hyperbolic identities

Lemma S3.3. *For any x, and $m,n\in\mathbb{N}$,*

$$\sum_{k=m}^{n}\cosh(kx+b)=\frac{\sinh\left(\frac{2n+1}{2}x+b\right)-\sinh\left(\frac{2m-1}{2}x+b\right)}{2\sinh(\frac{1}{2}x)}, \tag{S98}$$

$$\sum_{k=m}^{n}\sinh(kx)=\frac{\cosh\left(\frac{2n+1}{2}x\right)-\cosh\left(\frac{2m-1}{2}x\right)}{2\cosh\left(\frac{1}{2}x\right)}, \tag{S99}$$

Proof. We prove (S98) (as in this StackExchange response) by

$$\begin{aligned}\sum_{k=0}^{n}\cosh(kx+b) &= \frac{1}{2}\sum_{k=0}^{n}\left(e^b e^{kx}+e^{-b}e^{-kx}\right)=\frac{1}{2}\left(e^b\frac{1-e^{(n+1)x}}{1-e^x}+e^{-b}\frac{1-e^{-(n+1)x}}{1-e^{-x}}\right)\\ &= \frac{1}{2}\left(\frac{e^{\frac{n+1}{2}x+b}\left(e^{-\frac{n+1}{2}x}-e^{\frac{n+1}{2}x}+\right)}{e^{x/2}(e^{-x/2}-e^{x/2})}+\frac{e^{-(\frac{n+1}{2}x+b)}\left(e^{\frac{n+1}{2}x}-e^{-\frac{n+1}{2}x}\right)}{e^{-x/2}(e^{x/2}-e^{-x/2})}\right)\\ &= \frac{1}{2}\left(e^{nx/2+b}+e^{-(nx/2+b)}\right)\frac{\sinh\left(\frac{n+1}{2}x\right)}{\sinh(x/2)}=\cosh\left(\tfrac{n}{2}x+b\right)\frac{\sinh\left(\frac{n+1}{2}x\right)}{\sinh(x/2)}\\ &= \frac{\sinh\left(\frac{2n+1}{2}x+b\right)+\sinh\left(\frac{x}{2}-b\right)}{2\sinh(x/2)}.\end{aligned}$$

Analogously, we prove (S99) by

$$\sum_{k=0}^{n}\sinh(kx)=\sinh\left(\tfrac{n}{2}x\right)\frac{\sinh\left(\frac{n+1}{2}x\right)}{\sinh(x/2)}=\frac{\cosh\left(\frac{2n+1}{2}x\right)-\cosh\left(\frac{1}{2}x\right)}{2\cosh\left(\frac{1}{2}x\right)}.$$

□

S3.1.3 Fixed-point theorem

In Appendix S3.5, we use a fixed-point theorem that works with a sequence of functions:

Theorem S3.4 (Gill [158])**.** *Let $(f_n)_n$ be a sequence of functions that are analytic on a simply connected region S and continuous on the closure S' of S. Define the outer compositional sequence as*

$$F_n(z):=f_n\circ\cdots\circ f_1(z). \tag{S100}$$

Suppose there exists a compact set Ω contained in S such that $\Omega \supset f_n(S)$ for all n. Furthermore, suppose the sequence of fixed points $(\alpha_n)_n$ of $(f_n)_n$ (in Ω) converges to a number α. Then $F_n(z) \to \alpha$ uniformly on S'.

S3.2 Upper bounds on α

We here construct a set of features that encodes overlapping item pairs as less similar than distinct item pairs. To do so, we modify a onehot encoding to contain negative loadings at a trial's overlapping trial loadings. Specifically, our features

$$\phi(i,j) \in \mathbb{R}^{n\times n},$$

are defined as

$$\phi_{kl}(i,j) = \begin{cases} 1 & \text{if } k = i \wedge l = j, \\ -a & \text{if } k = i \text{ xor } l = j, \\ 0 & \text{else.} \end{cases} \tag{S101}$$

a is a hyperparameter that will depend on n.

We compute the similarities as

$$s = 1 + 2(n-1)a^2, \tag{S102}$$

$$o = -2a + (n-2), \tag{S103}$$

$$d = 2a^2, \tag{S104}$$

and therefore

$$\alpha = 1 - 2\frac{n-2-2a-2a^2}{1+2(n-2)a^2}. \tag{S105}$$

As the determinant of the polynomial in the numerator, $4 + 8(n-2)$, is positive for all n, this polynomial can take on negative values for all n. As a result $\max_a \alpha > 1$ for all n, demonstrating that this construction encodes overlapping item pairs as less similar than distinct item pairs. Determining its value numerically, we find that it exponentially approaches 1 as n grows large. While we provide a lower bound on the upper bound on α, it remains unclear whether this is a tight lower bound or whether there is a construction that can attain higher values for α.

S3.3 Gradient descent dynamics for mean squared error

We recall Lemma S1.4:

Lemma S1.4. *Given the mean squared error as a loss function, g_j and r_j amount to*

$$g_j = \sum_{\substack{k=1 \\ k \text{ odd}}} \frac{2\sin\left(\frac{jk\pi}{n}\right)\exp\left(-\delta_s\left(1-2\alpha\cos\left(\frac{k\pi}{n}\right)\right)t\right)}{n\tan\left(\frac{k\pi}{2n}\right)}. \tag{S56}$$

and

$$r_j(t) = (1-\alpha)\sum_{\substack{k=1 \\ k \text{ odd}}} \frac{\left(\sin\left(\frac{jk\pi}{n}\right) - \sin\left(\frac{(j-1)k\pi}{n}\right)\right)\left(1-\exp\left(-\delta_s\left(1-(1-\alpha)\cos\left(\frac{k\pi}{n}\right)\right)t\right)\right)}{n\tan\left(\frac{k\pi}{2n}\right)\left(1-(1-\alpha)\cos\left(\frac{k\pi}{n}\right)\right)}. \tag{S57}$$

Proof. We here solve the differential equation

$$\dot{g} = -Kg, \quad K_{ij} = \begin{cases} 1 & \text{if } i = j, \\ -\frac{1-\alpha}{2} & \text{if } |i-j| = 1, \\ 0 & \text{else.} \end{cases} \tag{S54 revisited}$$

Setting U as in Theorem S3.2 (meaning that $K = U\Lambda U$), we define

$$\tilde{g} := Ug, \tag{S106}$$

which results in the diagonalized equation

$$\dot{\tilde{g}}_j = -\delta_s \left(1 - (1-\alpha)\cos\left(\tfrac{j\pi}{m+1}\right)\right)\tilde{g}, \tag{S107}$$

We know that $g(0) = 1$ and so

$$\begin{aligned}\tilde{g}_j(0) = U_j 1 = \sqrt{\frac{2}{n}}\sum_{k=1}^{n-1}\sin\left(\frac{jk\pi}{n}\right) &= \frac{\cos\left(\frac{j\pi}{2n}\right) - \cos\left((n-\frac{1}{2})\frac{j\pi}{n}\right)}{\sqrt{2n}\sin\left(\frac{j\pi}{2n}\right)} \\ &= \frac{\cos\left(\frac{j\pi}{2n}\right) - \cos\left(j\pi - \frac{j\pi}{2n}\right)}{\sqrt{2n}\sin\left(\frac{j\pi}{2n}\right)}.\end{aligned} \tag{S108}$$

For odd j,

$$\cos\left(j\pi - \tfrac{j\pi}{2n}\right) = -\cos\left(\tfrac{j\pi}{2n}\right), \tag{S109}$$

and therefore

$$\tilde{g}_j(0) = \frac{\sqrt{2}}{\sqrt{n}\tan(\frac{j\pi}{2n})}, \tag{S110}$$

meaning that

$$\tilde{g}_j(t) = \frac{\sqrt{2}\exp\left(-\delta_s\left(1-(1-\alpha)\cos\left(\frac{j\pi}{n}\right)\right)t\right)}{\sqrt{n}\tan\left(\frac{j\pi}{2n}\right)}. \tag{S111}$$

For even j,

$$\cos\left(j\pi - \tfrac{j\pi}{2n}\right) = \cos\left(\tfrac{j\pi}{2n}\right), \tag{S112}$$

and therefore

$$\tilde{g}_j(0) = 0, \tag{S113}$$

meaning that

$$\tilde{g}_j(t) = 0. \tag{S114}$$

Therefore,

$$g_j(t) = \sum_{\substack{k=1\\ k \text{ odd}}} \frac{2\sin\left(\frac{jk\pi}{n}\right)\exp\left(-\delta_s\left(1-(1-\alpha)\cos\left(\frac{k\pi}{n}\right)\right)t\right)}{n\tan\left(\frac{k\pi}{2n}\right)}. \tag{S115}$$

This proves (S56).

As a reminder, the rank representation was defined as

$$r_i(0) := 0, \quad \dot{r}_i = \delta_s \tfrac{1-\alpha}{2}(g_i - g_{i-1}). \tag{S51 revisited}$$

As a result,

$$\begin{aligned} r_i(t) &= \delta_s \frac{1-\alpha}{2}\int_0^t g^{(i)}(s) - g^{(i-1)}(s)\mathrm{d}s \\ &= (1-\alpha)\sum_{\substack{j=1\\ j \text{ odd}}} \frac{\left(\sin\left(\frac{ij\pi}{n}\right) - \sin\left(\frac{(i-1)j\pi}{n}\right)\right)\left(1-\exp\left(-\delta_s\left(1-(1-\alpha)\cos\left(\frac{j\pi}{n}\right)\right)t\right)\right)}{n\tan\left(\frac{j\pi}{2n}\right)\left(1-(1-\alpha)\cos\left(\frac{j\pi}{n}\right)\right)}. \end{aligned} \tag{S116}$$

This proves (S57). □

S3.4 Ridge regression

We recall Lemma S1.7:

Lemma S1.7. *For non-adjacent items,*

$$f(x_j, x_k) = r_j - r_k, \quad r_j = \frac{\sinh((\frac{n+1}{2} - j)\lambda)}{\sinh(\frac{n+1}{2}\lambda) - \sinh(\frac{n-1}{2}\lambda)}, \quad \lambda := \operatorname{arccosh}\left(\frac{1+\tilde{c}^{-1}}{1-\alpha}\right). \tag{S76}$$

For adjacent items,

$$f(x_j, x_{j+1}) = m \cdot 1 + (1-m)(r_j - r_{j+1}), \quad m := \frac{\alpha}{\alpha + \tilde{c}^{-1}}. \tag{S77}$$

Proof. We had established that

$$\overline{a} = \frac{1}{\delta_s} B^{-1}, \tag{S117}$$

where

$$B^{-1}_{ij} = \frac{\cosh\left((n - |j-i|)\lambda\right) - \cosh\left((n-i-j)\lambda\right)}{(1-\alpha)\sinh(\lambda)\sinh(n\lambda)}, \quad \lambda := \operatorname{arccosh}\left(\frac{1+\tilde{c}^{-1}}{1-\alpha}\right). \tag{S73 revisited}$$

Thus

$$\overline{a}_i = \sum_{j=1}^{n-1} \frac{\cosh\left((n - |j-i|)\lambda\right) - \cosh\left((n-i-j)\lambda\right)}{(1-\alpha)\delta_s \sinh(\lambda)\sinh(n\lambda)}. \tag{S118}$$

Setting

$$b_i^{(1)} = \sum_{j=1}^{n-1} \cosh\left((n - |j-i|)\lambda\right), \quad b_i^{(2)} = \sum_{j=1}^{n-1} \cosh\left((n - (i+j))\lambda\right), \tag{S119}$$

we can express this as

$$\overline{a}_i = \frac{b_i^{(1)} - b_i^{(2)}}{2\delta_s \alpha \sinh(\lambda)\sinh(\lambda n)}. \tag{S120}$$

We simplify

$$\begin{aligned}
b_i^{(1)} &= \sum_{l=n-i+1}^{n} \cosh(l\lambda) + \sum_{l=i+1}^{n-1} \cosh(l\lambda) \\
&= \frac{\sinh((n+\frac{1}{2})\lambda) - \sinh((n-i+\frac{1}{2})\lambda) + \sinh((n-\frac{1}{2})\lambda) - \sinh((i+\frac{1}{2})\lambda)}{\sinh(\lambda/2)}, \\
b_i^{(2)} &= \sum_{l=0}^{i-1} \cosh(l\lambda) + \sum_{l=1}^{n-i-1} \cosh(l\lambda) \\
&= \frac{\sinh(\lambda/2) + \sinh((i-\frac{1}{2})\lambda) + \sinh((n-i-\frac{1}{2})\lambda) - \sinh(\lambda/2)}{\sinh(\lambda/2)} \\
&= \frac{\sinh((i-\frac{1}{2})\lambda) + \sinh((n-i-\frac{1}{2})\lambda)}{\sinh(\lambda/2)}.
\end{aligned} \tag{S121}$$

Moreover, note that

$$\frac{\sinh((m+\frac{1}{2})\lambda) + \sinh((m-\frac{1}{2})\lambda)}{2\sinh(\lambda/2)} = \frac{2\sinh(m\lambda)\cosh(\lambda/2)}{2\sinh(\lambda/2)} = \frac{\sinh(m\lambda)}{\tanh(\lambda/2)}. \tag{S122}$$

Applying this to $m \in \{n, n-i, i\}$ results in

$$\begin{aligned}
\overline{a}_i &= \frac{\sinh(n\lambda) - (\sinh((n-i)\lambda) + \sinh(i\lambda))}{\delta_s(1-\alpha)\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)} \\
&= \frac{\sinh(n\lambda) - 2\sinh(\frac{n}{2}\lambda)\cosh((\frac{n}{2} - i)\lambda)}{\delta_s(1-\alpha)\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)}.
\end{aligned} \tag{S123}$$

Note that the resulting boundary conditions are $\overline{a}_0 = \overline{a}_n = 0$, consistent with our definition.

Recalling that $r_j = \delta_s \frac{1-\alpha}{2}(\overline{a}_j - \overline{a}_{j-1})$, (S75), we find that

$$\begin{aligned} r_j &= -\frac{\sinh(n\lambda) - 2\sinh(\frac{n}{2}\lambda)\cosh((\frac{n}{2}-j)\lambda)}{2\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)} \\ &= \frac{\sinh(\frac{n}{2}\lambda)(\cosh((\frac{n}{2}+1-j)\lambda) - \cosh((\frac{n}{2}-j)\lambda))}{\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)} \\ &= \frac{2\sinh(\frac{n}{2}\lambda)\sinh((\frac{n+1}{2}-j)\lambda)\sinh(\lambda/2)}{\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)} \\ &= \frac{2\cosh(\lambda/2)\sinh(\frac{n}{2}\lambda)\sinh((\frac{n+1}{2}-j)\lambda)}{\sinh(\lambda)\sinh(\lambda n)} = C^{-1}\sinh((\tfrac{n+1}{2}-j)\lambda), \end{aligned} \tag{S124}$$

where

$$C := \frac{\sinh(\lambda)\sinh(\lambda n)}{2\cosh(\lambda/2)\sinh(\frac{n}{2}\lambda)}. \tag{S125}$$

Noting that

$$\begin{aligned} &\sinh(\tfrac{n+1}{2}\lambda) - \sinh(\tfrac{n-1}{2}\lambda) = 2\cosh(\tfrac{n}{2}\lambda)\sinh(\lambda/2) = \\ &\frac{2\cosh(\frac{n}{2}\lambda)\sinh(\lambda)}{\sqrt{2\cosh(\lambda)+1}} = \frac{\cosh(\frac{n}{2}\lambda)\sinh(\lambda)}{\cosh(\lambda/2)} = \\ &\frac{\sinh(\lambda)\sinh(\frac{n}{2}\lambda)\cosh(\frac{n}{2}\lambda)}{\cosh(\lambda/2)\sinh(\frac{n}{2}\lambda)} = C, \end{aligned} \tag{S126}$$

proves (S76).

Analysing the adjacent items, we observe that by definition,

$$\delta_s((1+\tilde{c}^{-1})\overline{a}_j - \tfrac{1-\alpha}{2}(\overline{a}_{j-1} + \overline{a}_{j+1})) = 1, \tag{S127}$$

and thus

$$f(x_j, x_{j+1}) = \delta_s(\overline{a}_j - \tfrac{1-\alpha}{2}(\overline{a}_{j-1} + \overline{a}_{j+1})) = 1 - \delta_s\tilde{c}^{-1}\overline{a}_j = 1 - \frac{\sinh(n\lambda) - 2\sinh(\frac{n}{2}\lambda)\cosh((\frac{n}{2}-j)\lambda)}{(1-\alpha)\tilde{c}\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)}. \tag{S128}$$

We split this up into two components: the memorization coefficient

$$m := 1 - \frac{\sinh(n\lambda)}{(1-\alpha)\tilde{c}\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)}, \tag{S129}$$

and the generalization coefficient g, which we pick such that

$$f(x_j, x_{j+1}) = m + g(r_j - r_{j+1}). \tag{S130}$$

We first simplify

$$\begin{aligned} m &= 1 - \frac{\tilde{c}^{-1} + 1 - 1}{(1-\alpha)\tanh(\lambda/2)\sinh(\lambda)} = 1 - \frac{(1-\alpha)\cosh(\lambda) - 1}{(1-\alpha)\tanh(\lambda/2)\sinh(\lambda)} \\ &= 1 - \frac{(1-\alpha)\cosh(\lambda) - 1}{(1-\alpha)(\cosh(\lambda)-1)} = \frac{\alpha}{(1-\alpha)(\cosh(\lambda)-1)} = \frac{\alpha}{\alpha + \tilde{c}^{-1}}. \end{aligned} \tag{S131}$$

To determine g, we solve the equation

$$g(r_j - r_{j+1}) = \frac{2\sinh(\frac{n}{2}\lambda)\cosh((\frac{n}{2}-j)\lambda)}{(1-\alpha)\tilde{c}\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)}. \tag{S132}$$

We observe

$$r_{j+1} - r_j = \frac{\sinh((j+1-\frac{n+1}{2})\lambda) - \sinh((j-\frac{n+1}{2})\lambda)}{\sinh((\frac{n+1}{2})\lambda) - \sinh((j+1-\frac{n-1}{2})\lambda)} = \frac{\cosh((j-\frac{n}{2})\lambda)}{\cosh(\frac{n}{2}\lambda)}. \tag{S133}$$

Thus,

$$g = \frac{2\sinh(\frac{n}{2}\lambda)\cosh(\frac{n}{2}\lambda)}{(1-\alpha)\tilde{c}\tanh(\lambda/2)\sinh(\lambda)\sinh(\lambda n)} = \frac{1}{(1-\alpha)\tilde{c}\tanh(\lambda/2)\sinh(\lambda)} = 1-m, \tag{S134}$$

and we can write

$$f(x_j, x_{j+1}) = m + (1-m)(r_j - r_{j+1}), \tag{S135}$$

proving (S77). □

S3.5 Deep neural networks' conjunctivity factor in the limit of infinite depth

We recall Lemma S1.3:

Lemma S1.3. *For a nonlinearity ϕ with relative slope ρ and a proportion of additive units λ, the infinite-depth conjunctivity factor*

$$\alpha^{(\infty)} := \lim_{l\to\infty} \alpha^{(l)}. \tag{S28}$$

is given by

$$\alpha^{(\infty)} = \begin{cases} 0 & \text{if } \lambda > 0 \vee \rho = 0, \\ 1 & \text{else.} \end{cases} \tag{S29}$$

Proof. Note that for $\rho = 0$, ϕ is the identity function and as a results $\alpha^{(l)} = 0$. We thus assume $\rho \neq 0$. As a reminder,

$$\hat{\kappa}_d^{(l+1)} = \tilde{k}(\hat{\kappa}_d^{(l)}), \quad \hat{\kappa}_o^{(l+1)} = (1-\lambda)\tilde{k}(\hat{\kappa}_o^{(l)}) + \frac{\lambda}{2}(\hat{\kappa}_d^{(l+1)} + 1), \tag{S136}$$

where

$$\tilde{k}(u) = u + \frac{\rho}{\pi}\left(\sqrt{1-u^2} - \cos^{-1}(u)u\right). \tag{S137}$$

First, $\tilde{k}([-1,1]) = [0,1]$. Further, $\tilde{k}$ has precisely one fixed point at $u = 1$. As a result,

$$\hat{\kappa}_d^{(l)} \to 1, \quad \hat{\kappa}_o^{(l)} \to 1. \tag{S138}$$

Note that

$$\alpha^{(l+1)} = (1-\lambda)\frac{1 + \hat{\kappa}_d^{(l+1)} - 2\hat{\kappa}_o^{(l+1)}}{1 - \hat{\kappa}_d^{(l+1)}} = (1-\lambda)\frac{1 + \hat{\kappa}_d^{(l)} - 2\hat{\kappa}_o^{(l)} + h(\hat{\kappa}_d^{(l)}) - 2h(\hat{\kappa}_o^{(l)})}{1 - \hat{\kappa}_d^{(l)} - h(\hat{\kappa}_d^{(l)})}, \tag{S139}$$

where

$$h(u) = \frac{\rho}{\pi}\left(\sqrt{1-u^2} - \cos^{-1}(u)u\right). \tag{S140}$$

We further simplify

$$\alpha^{(l+1)} = (1-\lambda)\frac{(1 - \hat{\kappa}_d^{(l)})\alpha^{(l)} + h(\hat{\kappa}_d^{(l)}) - 2h(\hat{\kappa}_o^{(l)})}{1 - \hat{\kappa}_d^{(l)} - h(\hat{\kappa}_d^{(l)})}. \tag{S141}$$

Noting that

$$\hat{\kappa}_o^{(l)} = \frac{1+\alpha}{2}\hat{\kappa}_d^{(l)} + \frac{1-\alpha}{2}, \tag{S142}$$

we can define updates to $\alpha^{(l)}$ by the iterated map

$$\alpha^{(l+1)} = f_l(\alpha^{(l)}) := f(\alpha^{(l)}, \hat{\kappa}_d^{(l)}), \tag{S143}$$

where

$$f(\alpha, d) := (1-\lambda)\frac{(1-d)\alpha + h(d) - 2h\left(\frac{1+\alpha}{2}d + \frac{1-\alpha}{2}\right)}{1 - d - h(d)}. \tag{S144}$$

Case $\lambda = 0$. For $\lambda = 0$, the fixed point equation amounts to

$$\alpha_d^*(1 - d - h(d)) = (1 - d)\alpha_d^* + h(d) - 2h\left(\frac{1+\alpha}{2}d + \frac{1-\alpha}{2}\right), \tag{S145}$$

which is equivalent to

$$-h(d)\alpha_d^* = h(d) - 2h\left(\frac{1+\alpha_d^*}{2}d + \frac{1-\alpha_d^*}{2}\right), \tag{S146}$$

and thus

$$2h\left(\frac{1+\alpha_d^*}{2}d + \frac{1-\alpha_d^*}{2}\right) - h(d)(1+\alpha_d^*) =: g_d(\alpha_d^*) = 0. \tag{S147}$$

Clearly, $g_d(1) = 0$. Note that

$$h'(u) = \frac{\rho}{\pi}\left(-\frac{u}{\sqrt{1-u^2}} - \cos^{-1}(u) + \frac{u}{\sqrt{1-u^2}}\right) = -\frac{\rho}{\pi}\cos^{-1}(u). \tag{S148}$$

Therefore,

$$g_d'(\alpha) = \frac{\rho}{\pi}\cos^{-1}\left(\frac{1+\alpha}{2}d + \frac{1-\alpha}{2}\right)(1-d) - h(d) \tag{S149}$$

As $\cos^{-1}$ is monotonically decreasing,

$$g_d'(\alpha) \geq g_d'(0) = \frac{\rho}{\pi}\cos^{-1}\left(\frac{1}{2}d + \frac{1}{2}\right)(1-d) - h(d). \tag{S150}$$

Visual inspection of this plot reveals $g_d'(0) \geq 0$ for all d. As a result, g_d is monotonically increasing on α and can therefore only have one zero. Put differently, the fixed point $\alpha_d^* = 1$ is unique. This means that each f_l has a unique fixed point at 1, and, by Theorem S3.4, $\alpha^{(l)} \to 1$.

Case $\lambda > 0$. Because $f(\cdot, d)$ maps $[0,1]$ onto $[0,1]$, it has a fixed point. If

$$\lim_{d\to 1}\frac{1-d}{1-d-h(d)} = 1, \quad \forall_{\alpha\in[0,1]}\lim_{d\to 1}\frac{h(d) - 2h(\frac{1+\alpha}{2}d + \frac{1-\alpha}{2})}{1-d-h(d)} = 0, \tag{S151}$$

we can write

$$f(\alpha, d) = (1-\lambda)(1+\epsilon_d^{(1)})\alpha + \epsilon_d^{(2)}(\alpha), \quad \lim_{d\to 1}\epsilon_d^{(1)} = 0, \quad \lim_{d\to 1}\epsilon_d^{(2)} \equiv 0. \tag{S152}$$

For any fixed point α_d^*, we thus observe that

$$(\lambda - (1-\lambda)\epsilon_d^{(1)})\alpha_d^* = \epsilon_d^{(2)}(\alpha). \tag{S153}$$

As $(\lambda - (1-\lambda)\epsilon_d^{(1)}) \to \lambda > 0$, $\alpha_d^* \to 0$. This means that any sequence of fixed points given by f_l as $l \to \infty$ must converge to zero, and, as a result, $\alpha^{(l)} \to 0$.

All that remains to show is (S151). We first note that $0 \leq h(\frac{1+\alpha}{2}d + \frac{1-\alpha}{2}) \leq h(d)$ and thus

$$-\frac{h(d)}{1-d-h(d)} \leq \frac{h(d) - 2h(\frac{1+\alpha}{2}d + \frac{1-\alpha}{2})}{1-d-h(d)} \leq \frac{h(d)}{1-d-h(d)}. \tag{S154}$$

As the left limit in (S151) implies

$$\lim_{d\to 1}\frac{h(d)}{1-d-h(d)} = 0, \tag{S155}$$

it immediately implies the right half of the equation. All that remains to show is the left half. Note that it is equivalent to

$$\lim_{d\to 1}\frac{1-d-h(d)}{1-d} = 1, \tag{S156}$$

and thus

$$\lim_{d \to 1} -\frac{h(d)}{1-d} = 0. \tag{S157}$$

We observe that

$$-\frac{h(d)}{1-d} = \frac{h(1) - h(d)}{1-d}, \tag{S158}$$

and thus

$$\lim_{d \to 1} -\frac{h(d)}{1-d} = h'(1) = 0, \tag{S159}$$

which concludes the proof. □