

Interactions on structured networks

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades einer Doktorin der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Claudia Merger
aus Heidelberg

Berichter: Prof. Dr. Moritz Helias
Prof. Dr. Carsten Honerkamp

Tag der mündlichen Prüfung: 10.01.2024

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Contents

1	Introduction	5
2	Order and spreading processes on structured networks: accounting for self-feedback	9
2.1	The emergence of the self-feedback effect	9
2.2	Spurious self-feedback and the emergence of order	11
2.2.1	Introduction to the Barabási-Albert Ising model	12
2.2.2	The conundrum: mean-field theory	15
2.2.3	Self-feedback and TAP equations	18
2.2.4	Effective magnetic transition on finite networks and Monte-Carlo observables	25
2.2.5	Transition temperature	26
2.3	Spurious self-feedback in the spread of disease	29
2.3.1	Models for the spread of disease	29
2.3.2	Fluctuation correction	31
2.3.3	Further fluctuation corrections to infection models	37
2.3.4	The epidemic threshold	38

3	Inference of higher-order interactions	43
3.1	Introduction to inference problems	43
3.1.1	Example: Pairwise interactions	44
3.1.2	Polynomial actions for continuous variables	45
3.2	Learning polynomial actions with invertible neural networks	46
3.2.1	Training	46
3.2.2	Architecture & layer transforms	48
3.2.3	Example: From Gaussian to fourth order action	54
3.2.4	Truncation in the interaction order	56
3.2.5	Learning rules in coefficient space	56
3.3	Experiments	58
3.3.1	Teacher Student scenario	58
3.3.2	Out of class distributions	60
3.3.3	Interactions on a lattice with external coupling	63
3.3.4	Three-point interactions in pictures of handwritten digits	64
3.4	Other approaches to inference problems	67
4	Discussion	69

Appendix	75
A	Parallel-tempering Monte-Carlo 75
B	Plefka expansion at equilibrium 75
C	Self-averaging on BA networks 76
D	Fluctuation-dissipation theorem 77
E	Epidemic dynamics in a Spiking Simulator code 78
F	Dynamical fluctuation expansion for infection models 78

F.1	Noninteracting system	81
F.2	Mean-field	82
F.3	Second order correction	83
G	Extensions to SIS and SIRS model	86
G.1	Noninteracting case	86
G.2	Mean-field equations	87
G.3	Second order correction	88
H	Decomposed tensors	88
I	Random generation of multi-modal actions	91
I.1	Coefficient distributions for random actions	91
I.2	Multimodality of random actions	92
J	Sampling actions with MCMC	93
K	Lattice model in low dimensions	94
L	Training on MNIST digits	96

Bibliography

Abstract

Structured systems appear ubiquitously in nature. Indubitably, the structure of a system determines its characteristic behavior. However, predicting the behavior of a system given its structure, or vice versa, is not straightforward. We here demonstrate that the mapping from structure to behavior can be tackled using a systematic fluctuation expansion, and develop a new method to infer structure given observations of the system. Often, structure can be represented as a network of nodes, where the nodes represent the agents, the elementary degrees of freedom of the system, and the connections define their interactions. One common feature of structured systems are hubs: nodes with significantly more connections than average, which are expected to be key to the observed overall system behavior. To understand the influence of hubs, we investigate to which extent the hubs of a scale-free network can drive a system of binary agents into an ordered or disordered state. We find that a typical mean-field approach to these systems introduces a nonphysical process: the signal sent by a node to its neighbors may travel back and influence the same node, leading to a self-feedback loop. The phenomenon is most prominent in the presence of hubs; their accumulated self-feedback grows with the number of connections. We show that a second-order fluctuation correction eliminates this spurious self-feedback. These insights are then translated to a model of disease spreading: We investigate the SIR model, where each agent can be in one of three states (susceptible, infected, or recovered), and transitions between these states follow a stochastic process. A typical approach in literature to predict average infection curves is to assume that all agents are statistically independent, introducing self-feedback artificially into the system, which yields inflated infection curves. We use a dynamical Plefka expansion to calculate a fluctuation correction, which eliminates the self-feedback effect, leading to more accurate predictions on the spread of disease. We then approach the reverse direction: inferring pairwise and higher-order interactions from data, these interactions constitute the structure of the underlying system. In principle, inference problems require an optimization over the space of all possible interactions, whose number increases exponentially with the system size. Nevertheless, machine learning models can infer structures efficiently from data. Typically, however, the inferred structure is hidden in the parameters of the trained model. We here show how to extract the learned structure, formulated in terms of interactions up to the fourth order. This process uncovers how the model hierarchically constructs interactions via nonlinear transformations of pairwise relations. This yields a fully understandable AI-powered tool for inference. Thus, we close the loop, demonstrating how collective behavior can emerge from structure and vice versa.

Zusammenfassung

Strukturierte Systeme sind in der Natur allgegenwärtig. Zweifelsohne bestimmt die Struktur eines Systems sein charakteristisches Verhalten, aber das Verhältnis zwischen Struktur und Verhalten eines Systems ist komplex. In dieser Arbeit nutzen wir eine systematische Entwicklung in den Fluktuationen des Systems um das Verhalten bei gegebener Struktur zu bestimmen und entwickeln eine neue Methode, um aus Beobachtungen des Systems auf die Struktur zu schließen. Häufig kann solch eine Struktur als ein Netzwerk von Knoten dargestellt werden, wobei die Knoten die elementaren Freiheitsgrade des Systems, hier Agenten genannt, darstellen und die Verbindungen die Wechselwirkungen derselben. Ein gemeinsames Merkmal strukturierter Systeme sind „hubs“: Knoten mit überdurchschnittlich vielen Verbindungen, die als entscheidend für das beobachtete Gesamtsystemverhalten gelten. Um den Einfluss derselben zu verstehen, untersuchen wir, inwiefern die Anwesenheit von hubs ein System aus binär interagierenden Agenten in einen geordneten oder ungeordneten Zustand treiben kann. Dabei stellt sich heraus, dass ein typischer Mean-Field-Ansatz für diese Systeme einen nicht-physikalischen Prozess einführt: Das von einem Agenten an seine Nachbarn gesendete Signal kann zurückkehren und denselben Agenten beeinflussen, was zu einer Rückkopplungsschleife führt. Das Phänomen ist besonders ausgeprägt, wenn hubs vorhanden sind; denn die kumulierte Rückkopplung wächst mit der Anzahl der Verbindungen. Wir zeigen, dass eine Fluktuationskorrektur zweiter Ordnung diese ungewollte Rückkopplung eliminiert und somit zu besseren Vorhersagen für das Verhalten führt, in denen der Einfluss von hubs nicht mehr überschätzt wird. Wir übertragen diese Erkenntnisse auf ein Modell der Krankheitsausbreitung: Wir untersuchen das SIR-Modell, bei dem sich jeder Agent in einem von drei Zuständen befinden kann (anfällig, infiziert oder genesen), und die Übergänge zwischen diesen Zuständen einem stochastischen Prozess folgen. Ein typischer Ansatz zur Vorhersage durchschnittlicher Infektionskurven ist die Annahme, dass alle Agenten statistisch unabhängig sind, wodurch wiederum eine künstliche Rückkopplung in das System eingeführt wird, die zu überhöhten Infektionskurven führt. Mithilfe einer dynamischen Plefka-Entwicklung berechnen wir eine Fluktuationskorrektur, die die Korrelationen von Agenten berücksichtigt. Diese Korrektur eliminiert den Effekt der Rückkopplung, und führt zu genaueren Vorhersagen über die Verbreitung von Krankheiten. Anschließend gehen wir den umgekehrten Weg: Wir schließen von den Daten auf paarweise Wechselwirkungen und Wechselwirkungen höherer Ordnung zwischen den Freiheitsgraden. Im Prinzip erfordert die Lösung solcher Inferenzprobleme eine Optimierung über den Raum aller möglichen Interaktionen, deren Anzahl exponentiell mit der Systemgröße zunimmt. Dennoch

können maschinelle Lernmodelle effizient Strukturen aus Daten ableiten, typischerweise ist die abgeleitete Struktur jedoch in den Parametern der trainierten Modelle versteckt. Wir zeigen hier, wie man die gelernte Struktur, die in Form von Wechselwirkungen bis zur vierten Ordnung formuliert ist, mit Hilfe von neuronalen Netzen extrahieren kann. Durch diesen Prozess wird klar, wie das Netz die Wechselwirkungen durch nichtlineare Transformationen paarweiser Beziehungen hierarchisch aufbaut. Das Ergebnis ist ein vollständig verständliches KI-gestütztes Werkzeug für Inferenzprobleme. Auf diese Weise schließen wir den Kreis und zeigen, wie kollektives Verhalten aus der Struktur hervorgehen kann und umgekehrt.

Introduction

Consider an interaction network originating from a) the world wide web, with web pages as the nodes and hyperlinks between them as the links [1], b) the members of a Karate club, whose interactions outside the karate club constitute the edges [2], c) proteins in yeast, whose physical interactions constitute the edges [3]. In all three cases, the structure of the network is believed to be key to the system behavior, for example, it can be used to predict which faction the members of the karate club belong to after the club split into two due to a conflict [2]. We can rank the agents in these three systems by their degree k ; the number of edges connecting them to other agents. It is remarkable that all three systems show a strongly hierarchical interaction structure: all three networks in the examples a)-c) feature a fat-tailed degree distribution $p(k)$, meaning that some agents have far more connections than average, a property they share with a plethora of other systems [4, 5]. This then characterizes hierarchy of the agents, with the hubs of the system (agents with far more connections than average) as the key system constituents. Structured systems hence appear ubiquitously across scientific disciplines. In this work, we study structured systems using interacting theories: a type of models central to physics.

In general, we consider a structured system to be constituted of degrees of freedom, or *agents*, whose interactions give rise to the properties of the system. Further examples of such degrees of freedom include Ising spins, firing rates of neurons, social agents, field points, or pixels of an image. Their interactions can be formulated in terms of a network, with the nodes symbolizing the degrees of freedom, and the edges their interactions. A structured system therefore, is one where the nature and organization of the interactions is non-trivial and characteristic of the system. In general, predicting a system's global behavior from the known interacting structure is a hard computational problem; the size of the system typically prohibits an exact treatment. However, statistical physics provides a range of tools with which powerful effective models as approximations of the system can be derived. We will use a subset of these tools to investigate quantitatively how a heterogeneous network structure can give rise to collective behavior.

In the context of our initial set of examples, it is a priori not clear, to which extent the hubs of these systems drive the behavior. To understand the role of the network

topology, including hubs, better, we begin by investigating ordering and spreading processes on structured networks of agents which interact in a pairwise fashion in Chap. 2. We characterize the behavior of a system via its statistics, and aim to compute moments of the systems observables, averaged over all possible states. Since the number of possible states typically increases exponentially with the size of the system, and these systems are typically large, such averages cannot be computed exactly. However, they can be approximated to high accuracy using systematic expansion techniques. In Chap. 2, we will hence use such a systematic fluctuation expansion to second order. Although the two systems under study originate from completely different fields, one being a system of Ising spins, the other a model for disease propagation, we find that they can be treated with the same type of expansion method, where the fluctuation corrections of both systems remove a spurious self-feedback effect which dominates the behavior otherwise.

Rather than solving the forward problem, namely which specific behavior follows from a specific problem, we can then ask the reverse question: can we infer the structure of a system from observations of the same system? This is known as an inference or inverse problem, the known quantities being observations of the system from which (empirical) averages can be computed directly. The inferred quantities are then the interactions of the agents. However, inference problems are typically hard for two reasons: First, the space of candidate solutions is typically prohibitively large. Consider, for example, a network of N agents. In this network, there can be at most $N(N-1)$ edges, each of which can be present or not. Hence there are $2^{N(N-1)}$ possible networks, counting pairwise interactions alone. Furthermore, for each edge, the strength of the interaction may be different and must therefore be determined as well. The second reason is that direct optimization, (e.g. via gradient descent) of the likelihood of observing the data given the proposed structure requires computing averages over the prohibitively large state space at every optimization step, thus the forward problem must be solved at every optimization step. Despite this difficulty, significant progress has been made on pairwise phenomena, including the inverse problem to the Ising model we study in Chap. 2.

Recently, however, a series of works has uncovered that higher-order interacting systems [6] can produce wealth of complex phenomena [7, 8], for example explosive synchronization [9, 10]. Higher-order interactions are also used to characterize brain activity [11] or for dimensionality reduction [12]. A typical definition of a higher-order interaction is that it induces a type of coordination between degrees of freedom which cannot be reproduced by pairwise interactions between the same degrees of freedom. In a higher-order network, these interactions constitute hyper-edges, for example a three-point interaction is represented by a filled triangle between three nodes, a four-point interaction by a tetrahedron and so on [6, 7]. At a first glance, inference beyond pairwise interactions seems prohibitively hard as allowing third-order interactions (hyper-edges between triplets of agents) and higher-order interactions increases the size of the already very large space of possible solutions.

We here present a solution to the inference problem which uses generative neural networks. Generative neural networks such as Normalizing Flows [13–15], transformers [16], or diffusion models [17] are able to mimic the typical behavior of structured systems. In practice, they are used to generate new samples similar to the data they were trained on, such as text or images. In training these models, the essential properties of the underlying system must be inferred. Furthermore, many real-world systems, and in particular machine learning data sets, are believed to possess an underlying structure which neural networks are able to reproduce [18, 19], which is organized into different orders of complexity. The authors of [18, 19] showed that neural networks can pick up on higher-order statistics of data sets, beyond pairwise interactions. However, these higher-order statistics are stored implicitly in the parameters of the trained networks, the interpretation of which is typically difficult. For this reason, the neural networks are often referred to as "black boxes". Therefore, an important step towards understanding generative neural networks is to find a language in which the learned structure can be expressed.

We here demonstrate that it is possible to extract the learned structure in the form of pairwise and higher-order interactions between the system constituents. In doing so, we translate the learned structure from an implicit form in terms of the network parameters into an explicit form, which is central to physics. This process can hence not only inform us about the functional organization of the systems underlying the data, but also help us understand how it is learned in neural networks.

We here focus on a special type of generative neural network, namely Normalizing Flows [13–15]. Normalizing Flows are likelihood-based generative neural networks, meaning that they are optimized to approximate the probability distribution of the observed data. This is achieved via an implicit, rather than an explicit parameterization of the probability density: these models constitute a mapping from data space to a completely unstructured latent space, which is optimized such that the latent variables follow a normal distribution. Hence the name *Normalizing Flow*: the highly structured distribution in input space is transformed by the mapping to assimilate a Gaussian distribution in latent space. The mapping is invertible, hence its inverse generates the structure; it transforms the unstructured latent distribution back into the structured data distribution. In Chap. 3, we show how to extract this structure, in the form of interactions between the systems constituents, from trained Normalizing Flows. Extracting the learned structure from the trained models then allows us both to characterize how these models perform inference, and obtain the structure of the underlying system in a computationally feasible way.

The overarching structure of the thesis is this: in Chap. 2, we solve the forward problem, predicting a systems behavior from a given interaction structure. In Chap. 3, we then tackle the reverse problem. Finally in Chap. 4, we discuss our contributions towards a theory of interacting system and provide an outlook to possible future research directions.

Order and spreading processes on structured networks: accounting for self-feedback

2.1 The emergence of the self-feedback effect

The emergence of long-range order and the speed at which signals travel are important global properties of any system. The former property tells us when we can expect a system to be in a globally homogeneous state. The latter informs us about the speed at which perturbations of the system travel in space, and at which they can be expected to affect the system at large.

We will first study the transition to an ordered state in a system at thermodynamic equilibrium in Sec. 2.2. If the system is ergodic, then the equilibrium distribution will tell us about a system's properties at infinitely large time scales. This approach is therefore suitable for systems in which the dynamics play a minor role and we rather want to characterize typical states in which we can find the system. Concretely, we investigate when a set of binary variables, namely Ising spins, whose interactions are given by a Barabási-Albert network [20], undergo a transition to a globally ordered state. Both the network topology and the interactions of the agents are archetypal: the Ising model [21] acts as a minimal model to study ordering

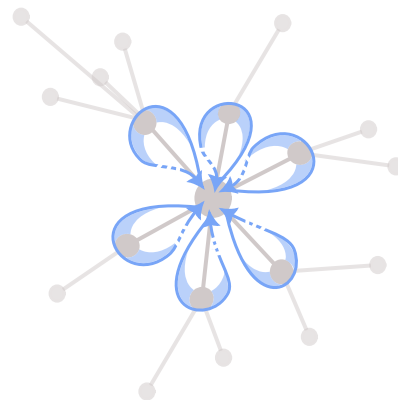


FIG. 2.1: Illustration of self-feedback of a hub (central large node) on a tree network. The self-feedback loops of the hub via its direct neighbors is illustrated as blue arrows.

processes. In the ferromagnetic Ising model, an alignment of all spins (perfect ordering) is energetically favored, while there are far more disordered than ordered states, thus there is an entropic drive towards disorder. The temperature T weights these two effects against each other, such that it is possible to find an intermediate regime, where the system transitions from disorder to order. The network topology is chosen such that there are few, highly connected agents while the average agent has a relatively small number of connections: Barabási-Albert networks have a scale-free degree distribution [20]. The degree is the number of connections of a node of a network. The scale-free property of the network implies that the probability that a node of the network has k connections, follows a power-law $p(k) \propto k^{-\gamma}$. At $\gamma \leq 3$, the variance of the degree distribution $\langle k^2 \rangle - \langle k \rangle^2$ diverges with the system size, meaning there is no typical scale of the degree. While the full universality of scale-free or power-law behavior in the underlying connectivity structure is still under debate [4, 5], many real-world networks have hubs, whose role we will characterize in this study: we will find that hubs play a role in driving local order, but long range order is not driven by a single hub explicitly, rather, it emerges as a global phenomenon on the whole network.

We then move to a non-equilibrium system, namely the susceptible-infected-recovered (SIR) model [22], which is used to predict the spread of disease. In this model, each agent can be in one of three states: susceptible, infected, or recovered. Agents on adjacent sites of the interaction network can infect each other with probability β at each time step. Once an agent is infected, in each time step he may recover with probability μ . Hence, at each time point, an agent can move from the susceptible to the infected, or from the infected to the recovered state, but never backwards. The allowed transitions are hence $S \rightarrow I \rightarrow R$. Once all infected agents have recovered, the dynamics stop. Starting from a few infected agents, the dynamic evolution hence depends on β, μ and the network topology. In the SIR model, in contrast to the Ising system, the non-equilibrium dynamics are hence the quantities of interest: how fast does the number of infected agents grow, how high is the peak of the infection curve, and how many agents are in the recovered state once the dynamics have converged, i.e. an endemic state has been reached.

It turns out that in both systems under study, a phenomenon we call *self-feedback* plays an important role. We call a signal, which travels from one node to its nearest neighbors and then travels back to the original node, a self-feedback signal. This process is illustrated in Fig. 2.1: a central node receives a self-feedback signal from all of its nearest neighbors. In both systems we study, self-feedback is not allowed by construction:

- In the thermodynamic limit and at thermodynamic equilibrium, the global behavior of a system does not depend on a single Ising spin. This is the idea underlying cavity theory [23].
- In the SIR model, an infected agent cannot be re-infected, it can only recover. Hence an agent can never experience a self-infection via a self-feedback loop.

How then, does self-feedback appear in these systems? The answer is that in both systems, typical first order approximations introduce self-feedback into the system artificially. This is so because self-feedback is a second order effect, quadratic in the interaction strength: for a self-feedback loop to take place, a signal must traverse an edge in the system twice, hence the relevant term must contain the interaction strength to the power of two. Typical first order approximations only account for terms up to the first order in the interaction strength, hence they introduce this spurious self-feedback. In both cases, we will show that the presence of self-feedback leads to inaccurate predictions, and that it is corrected by a second order fluctuation expansion.

In the following, we will work out the fluctuation corrections to both systems and demonstrate their effect. These fluctuation corrections, in principle, are derived from the systems definitions, and do not include any prior knowledge on whether self-feedback is present or not. Rather, the cancellation of self-feedback here emerges as a second order correction in a systematic series, such that, in principle, higher-order corrections can be computed. However, we find that the second order approximation typically agrees well with simulations of the system, such that we expect the effect of higher-order terms to be less relevant in this case. The cancellation of self-feedback then acts as an intuitive interpretation of the result of the fluctuation expansion.

2.2 Spurious self-feedback and the emergence of order

This section, parts of Chap. 4 and appendices A, B, C and D are based on the following publication:

Merger, C., Reinartz, T., Wessel, S., Honerkamp, C., Schuppert, A., Helias, M., 2021. Global hierarchy vs local structure: Spurious self-feedback in scale-free networks. *Phys. Rev. Res.* 3, 033272. <https://doi.org/10.1103/PhysRevResearch.3.033272>

Author contributions

Under the supervision of Moritz Helias and Carsten Honerkamp, the author worked on all parts of the above publication presented, except the Monte-Carlo simulations of the Ising system, which were performed by Timo Reinartz under the supervision of Stefan Wessel. The author contributed to the general formalism, performed the corresponding numerical experiments on TAP-and mean-field theory and wrote the original draft of the manuscript. All authors jointly developed the ideas of the publication and contributed to finalizing the manuscript. The idea of employing the TAP theory to compute corrections to the mean-field result is also present in the author's master thesis, however, the link between degree-based mean-field-theory and the TAP correction via self-feedback was developed during the PhD thesis.

2.2.1 Introduction to the Barabási-Albert Ising model

In this section, we set the stage for our investigation into ordering processes in the presence of hubs by introducing the Barabási-Albert (BA) network model and the Ising model. We then give a brief overview to the most important results gained on the Barabási-Albert Ising (BAI) model in the literature.

BA networks [20] are defined via a construction algorithm, based on *growth* and *preferential attachment*. Starting with a fully connected network of m_0 nodes, one iteratively adds nodes, until the desired system size N is reached. Once a node is added, it is connected to randomly chosen existing nodes, where the probability to connect to a node i is proportional to the degree k_i of the same node.

The construction algorithm invokes a hierarchy on the nodes of the systems: early nodes have the highest chance of accumulating many connections, they will later constitute the hubs of the system. This is characterized by the degree distribution, which is derived in [20] using an averaged version of the dynamical attachment algorithm. It states that the probability that a randomly chosen node has a specific degree k is given by

$$p(k) = \frac{2m_0^2}{k^3}. \quad (2.1)$$

For a finite system, the largest degree of the system is typically

$$k_{\max} = m_0 \sqrt{N}, \quad (2.2)$$

while the average degree is just $\langle k \rangle = 2m_0$, by construction: for each node added to the system, we obtain m_0 further links. The factor two in $\langle k \rangle$ arises since each edge increases the degrees of the nodes it connects by one.

The final network of size N is then specified by its $N \times N$ adjacency matrix \mathbf{A} , where for nodes $i, j \in \{1, \dots, N\}$

$$A_{ij} = A_{ji} = \begin{cases} 1 & \text{if } i \neq j \text{ and } i, j \text{ are connected} \\ 0 & \text{else.} \end{cases} \quad (2.3)$$

As a simple model of how agents interact on such a network structure, we consider an Ising model [21]. We assign a local binary degree of freedom (spin) $x_i \in \{-1, 1\}$ to each node $i \in \{1, \dots, N\}$. The probability for a given configuration $x = (x_1, \dots, x_N)$ is determined by the Boltzmann-factor $p(\mathbf{x}) \propto \exp(-\beta H(\mathbf{x}))$, where $H(\mathbf{x}) = -\frac{J}{2} \mathbf{x} \cdot \mathbf{A} \mathbf{x}$ is the energy of each configuration. The coupling constant $J > 0$ in H favors the alignment of connected spins into either direction, hence we have a *ferromagnetic* coupling. Since the factor $\beta = T^{-1}$, which constitutes the inverse temperature of the system (we set $k_B = 1$), is always measured in relation to the elementary energy scale J , we are free to fix $J = 1$ and to vary only β in the following. In addition, an external

magnetic field \mathbf{h} is included in the Hamiltonian H , which will be set to zero for our further analysis. However, the inclusion of such a term is convenient for the derivation of the fluctuation correction, which we detail in App. (B). There, \mathbf{h} plays the role of an infinitesimal bias. The response of the system to the inclusion of such a source term will be treated in Sec. 2.2.4 and Sec. 2.2.5. The full Hamiltonian of the system hence reads

$$H(\mathbf{x}, \mathbf{h}) = -\frac{1}{2} \mathbf{x} \cdot \mathbf{A} \mathbf{x} - \mathbf{h} \cdot \mathbf{x}. \quad (2.4)$$

The quantities of interest here are then defined as statistical averages over all possible spin configurations, $\langle O \rangle = \sum_{\mathbf{x}} O(\mathbf{x}) p(\mathbf{x})$, of appropriate observables O , specified further below. The inherent difficulty in computing these averages lies in the fact that the number of possible configurations increases as 2^N with the system size N . For large systems, performing such an average exactly is hence computationally infeasible, and we must rely on approximations. We use parallel-tempering Monte Carlo simulations to calculate these statistical averages (see. App. (A) for details regarding the employed simulation scheme). The Monte-Carlo simulation here serves as a ground truth to which we will compare several field-theoretic approximation schemes further specified in Sec. 2.2.2 and Sec. 2.2.3.

The degree resolved view on BA networks. BA networks exhibit linear degree-degree correlations [24], where the probability that two nodes with degrees k_i and k_j are connected is given by

$$p_c(k_i, k_j) = \frac{k_i k_j}{2m_0 N}. \quad (2.5)$$

A number of works have hence used equations (2.1) and (2.5) to define a degree-resolved statistic of the system, where all nodes of equal degree are expected to be statistically equivalent. They then analyze the emergence of order using mean-field theory [24], recursion methods for tree-like networks [25] or the replica trick [26]. These studies demonstrate that, for finite network sizes, a strong alignment of the Ising spins emerges below a specific temperature T_T . This transition to a globally ordered state resembles the onset of ferromagnetic order in conventional lattice Ising models. There, the ordered phase emerges below a finite transition temperature out of the paramagnetic high-temperature phase in the thermodynamic (large- N) limit. For the BAI model, however, Monte Carlo simulations [27, 28] as well as the degree-resolved theories [24–26] indicate that the effective transition temperature T_T instead grows logarithmically with the network size, $T_T \propto \log(N)$. They further found that close to T_T in the ordered phase, the average magnetization m_i of a given node i increases proportional to its degree

$$m_i \propto k_i. \quad (2.6)$$

This implies a rather simple structure, in which the magnetization of a node is determined solely by its degree and not by its local environment. However, we will demonstrate using Monte-Carlo simulations that the average behavior of a node is far from being determined by its degree alone. Rather, it depends strongly on the local environment of the node: Two nodes of the same degree will have different statistics depending on whether they are connected to a hub or not.

This calls for an investigation which retains full information on the network connectivity, namely the adjacency matrix defined in Eq. (2.3). We will consider both a variant of the mean-field theory in [24], which is informed about the full connectivity structure, and a self-consistent Thouless-Anderson-Palmer (TAP) approach [29–31], which takes second order interaction effects into account.

We find that a mean-field calculation on the level of individual Ising variables, rather than their degree, yields vastly different results than those reported in [24]. To distinguish the two mean-field theories henceforth, we will use terminology established in studies for the spread of disease [32]. We will call the approach of Ref. [24] degree-based mean-field theory and the approach using the full connectivity matrix individual-based mean-field theory. We find that individual-based mean-field theory severely overestimates the role of hubs. This overestimation arises due to a spurious self-feedback effect: A node's local alignment field h_u induces a strong magnetization on its nearest neighbor nodes. The magnetization of the nearest neighbor nodes in turn raises the magnetization of the original node itself. The node thus effectively feels its own field. The more nearest neighbors a node has, the stronger this self-feedback effect becomes, making it most severe in hubs of the system. Due to this self-feedback, the individual-based mean-field approach predicts a much higher transition temperature than actually observed, scaling proportionally to $N^{\frac{1}{4}} \gg \log(N)$, as we will find in Sec. 2.2.2. Moreover, these ordered states are strongly localized around the hubs, in contradiction to the simple proportionality relation (2.6). The ordered states are therefore also highly sensitive to the presence of individual hubs in the network. This makes them fragile in terms of the stability to small perturbations in the network topology, such as the random removal of nodes. This sensitivity also appears to be in conflict with the collective nature of the onset of global order.

We hence face a conundrum, in which a theory of microscopic interactions performs worse upon inclusion of full information of the same interactions into the system. The contrast between the individual-based and the degree-based mean-field theory ultimately exposes an inherent inconsistency of the degree-resolved approach to heterogeneous systems.

To reconcile this inconsistency, we will show that the self-feedback effect outlined above is non-physical, as one can also anticipate from the cavity argument [23], which states that the local alignment field h_i at a given node of the network must be calculated in the absence of this node, hence the self-feedback must be eliminated. In the degree-based mean-field theory, this is achieved through the elimination of

local information, since all nodes are coupled via Eq. (2.5), but this approach cannot describe the local magnetisation adequately. As we report below, the cancellation of self-feedback is obtained in an individual-based approach upon expanding beyond mean-field theory and analyzing the BAI using a the TAP approach, which takes second order interaction effects into account.

The remainder of this section is organized as follows: in Sec. 2.2.2 we expose the inherent inconsistency of the mean-field theory. We introduce the second order fluctuation correction in Sec. 2.2.3 and show that the self-feedback effect is canceled by fluctuations. We then demonstrate how an effective coupling to a global ordering field approximates the network state. Finite size effects and the appropriate observables of Monte-Carlo simulations are treated in Sec. 2.2.4. We then discuss our results for the transition temperature in Sec. 2.2.5.

2.2.2 The conundrum: mean-field theory

For a set of Ising spins coupled according to a network topology with adjacency matrix A , the exact equation for the magnetization $m_i = \langle x_i \rangle$ of a node i reads

$$m_i = \left\langle \tanh \left(\beta \sum_j A_{ij} x_j \right) \right\rangle. \quad (2.7)$$

To evaluate this average, we must however have access to the full statistics of the system, which is computationally infeasible. Hence we must resort to approximations, one typical approximation being a mean-field theory, which approximates the interactions between spins to first order in β . We will outline the derivation of this mean-field theory in the context of a systematic expansion Sec. 2.2.3. It yields a self-consistency equation for the magnetizations, hence in mean-field theory Eq. (2.7) is replaced by

$$m_i = \tanh \left(\beta \sum_j A_{ij} m_j \right). \quad (2.8)$$

The trivial solution to this equation, $m_i = 0 \forall i$, exists for all inverse temperatures β . We are interested in the emergence of order, characterized by $m_i \neq 0$. Since in the absence of an external field, Eq. (2.4) is symmetric under a global sign change of x , solutions always emerge in pairs which differ only by the signs of all m_i . We hence restrict ourselves to the case $m_i > 0 \forall i$ without loss of generality.

We will now consider the smallest value of β , or equivalently the highest temperature T , at which a non-trivial solution exists. We call this temperature *transition temperature*¹. Since we expect these solutions to emerge continuously from zero at

¹We discuss the growth of this quantity with the system size in Sec. 2.2.4.

the transition temperature $\beta_T = T_T^{-1}$, we expand Eq. (2.8) up to linear order in the m_i , $i = 1, \dots, N$:

$$m_i = \beta_T \sum_j A_{ij} m_j. \quad (2.9)$$

Degree-based and individual-based mean-field theory now solve this equation in two different ways. We will first follow Ref. [24] to derive the degree-based mean-field theory result for T_T .

2.2.2.1 Degree-based mean-field theory

We first assume that all nodes of equal degree are statistically equivalent, replacing m_i by $m(k_i)$ in Eq. (2.9). Next, we substitute $\sum_j A_{ij} m(k_j)$ by $N \sum_k p(k) p_c(k_i, k) m(k)$. This yields the following degree-resolved linear equation, where the local structure of the coupling has been eliminated

$$m(k_i) = \frac{k_i \beta_T}{2m_0} \sum_k p(k) k m(k). \quad (2.10)$$

Observe that the k_i dependence of $m(k_i)$ here enters only through the linear prefactor to a global order parameter

$$S := \frac{1}{2m_0} \sum_k p(k) k m(k), \quad (2.11)$$

such that we find a linear scaling $m(k) = \beta_T k S$. Inserting this into the definition of the order parameter (2.11), this global order parameter obeys $S = \frac{\langle k^2 \rangle}{\langle k \rangle} \beta_T S$, from which the transition temperature of the degree-based mean-field theory approach follows as

$$T_{\text{MF}}^{\text{DB}} = \frac{\langle k^2 \rangle}{\langle k \rangle} \approx \frac{m_0}{2} \log(N). \quad (2.12)$$

This prediction indeed matches the transition temperature found in Monte-Carlo simulations here and in [28, 33]. However, in the steps leading to Eq. (2.10), we have discarded all local structure except the degree hierarchy of nodes. We will now demonstrate that a different result is obtained for T_T if one retains local structure.

2.2.2.2 Individual-based mean-field theory

Observe that Eq. (2.9) has the shape of an eigenvalue equation. In individual-based mean-field theory, we hence find the transition temperature as the smallest value of β for which the eigenvalue equation admits a solution, $T_{\text{MF}}^{\text{IB}} = \lambda_{\mathbf{A}, \text{max}}$. In other words, the transition temperature is the largest eigenvalue of A . The authors of [34] found that for BA networks the leading eigenvalue of the adjacency matrix scales

as $\lambda_{\mathbf{A},\max} \propto \sqrt{m_0} N^{1/4}$, with the corresponding eigenvector strongly localized at the node of highest degree, i.e. the largest hub.

We illustrate this result by considering how a vector localized at the largest hub u of the system transforms upon multiplication by the adjacency matrix. We define a vector \mathbf{v} of unit length that takes on the value $v_u = 1/\sqrt{2}$ at the hub and $v_i = A_{iu}/\sqrt{2k_u}$ for all other nodes, $i \neq u$. Thus v is only nonzero for nearest neighbors of u . Multiplication with the adjacency matrix yields

$$\sum_i A_{ui} v_i = \sqrt{\frac{k_u}{2}},$$

for the entry corresponding to the hub u . Since \mathbf{A} is real and symmetric, we find

$$\lambda_{\mathbf{A},\max} = \|\mathbf{A}\| \geq \|\mathbf{A} \mathbf{v}\| \geq \sqrt{\frac{k_u}{2}},$$

where $\|\mathbf{A}\|$ is the matrix norm. Finally, using Eq. (2.2), we find that the largest eigenvalue scales at least as $\lambda_{\mathbf{A},\max} \propto N^{1/4}$. In conclusion, we find that individual-based mean-field theory predicts a different transition temperature than degree-based mean-field theory,

$$T_{\text{MF}}^{\text{IB}} \propto N^{1/4} \gg \log(N) \quad (2.13)$$

which is larger than $T_{\text{MF}}^{\text{DB}}$ for sufficiently large N , at variance with Monte-Carlo simulations. This result is remarkable, since the inclusion of full information on the network structure should improve the prediction, not vice versa. We will explain why degree-based theory performs better in Sec. 2.2.3 and present our results on the transition temperature in Sec. 2.2.5.

We now compare further predictions of degree-based mean-field theory to simulations. Degree-based mean-field theory predicts a linear scaling of the magnetization with k , at the transition temperature. We verify this by computing the average degree-wise magnetization

$$m(k) = \frac{1}{\sum_i \delta_{k_i,k}} \sum_i \delta_{k_i,k} m_i, \quad (2.14)$$

at the transition temperature in panel a) of Fig. 2.2. We find that the prediction $m_i \propto k_i$ averaged over all vertices of the same degree holds for small degrees (see Fig. 2.2(a)), albeit at a slightly downward shifted effective transition temperature (cf. Sec. 2.2.4 for a discussion on the exact position of this crossover). For large degrees, however, the average behavior in the Monte-Carlo data seems to deviate from this rule. This here results from the finite size of the simulated system: by virtue of having a high degree, these nodes tend to be few. For finite systems, the modulus of the magnetization is never exactly zero, even in the disordered case, but rather converges to a finite value

(we will revisit this point in Sec. 2.2.4). One can then understand the deviation from the $m_i \propto k_i$ scaling for large k_i by this finite magnetization converging to a plateau as the magnetization eventually has to saturate to unity.

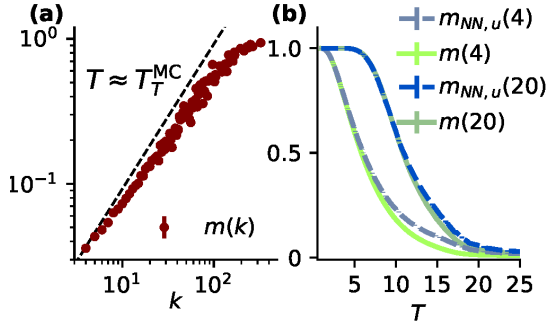


FIG. 2.2: (a) **Degree-wise magnetization** $m(k)$ as a function of the degree k obtained by Monte Carlo simulation for $N = 10^4$ and $m_0 = 4$ at the effective transition temperature $T_T^{\text{MC}} = 16.12 \pm 0.05$. The dashed line is parallel to the identity, to show the approximate linear relationship. (b) **Local structure.** Green curves: $m(k)$, for $k \in \{4, 20\}$ as a function of temperature. Blue dashed curves: Degree-wise averages over the nearest neighbors of the hub u Eq. (2.15).

From the linear scaling of the magnetization with the degree and the position of the transition temperature, one might conclude that the degree-based description more adequately models the underlying stochastic process. However, the Monte-Carlo simulation, performed on the full connectivity matrix, also reveals that local information is not lost. In Fig. 2.2(b) we compare the average degree-wise magnetization (2.14) of all nodes which have a given degree k to the average magnetization of the nearest neighbors of a hub u , given by

$$m_{\text{NN}u}(k) = \frac{1}{\sum_i A_{iu} \delta_{k_i, k}} \sum_i A_{iu} \delta_{k_i, k} m_i. \quad (2.15)$$

This shows that hubs elicit a stronger alignment of their nearest neighbors i than predicted by their degree alone, $m_{\text{NN}u}(k_i) > m(k_i)$.

All in all, we conclude that neither individual-based nor degree-based mean-field theory can adequately describe the model, as they either overestimate the role of hubs or fail to account for local structure. We will resolve this contradiction in the following by showing that the transition temperature Eq. (2.12) emerges naturally from the microscopic TAP equations, which simultaneously also predict the local structure of magnetization, thus solving the conundrum.

2.2.3 Self-feedback and TAP equations

The TAP approach [29, 31, 35] accounts for second order interaction effects. It can be derived from a second order Plefka expansion [36]. To obtain the correction, one usually starts with the free F energy of the system, obtained from the sum over all spin configurations as

$$e^{-\beta F(\mathbf{h}, \beta)} = \sum_{\mathbf{x}} e^{-\beta H(\mathbf{x}; \mathbf{h})}. \quad (2.16)$$

One can immediately observe, that F has the form of a generating function: to obtain the local magnetizations for example, we must simply take the derivative $m_i = -\partial_{h_i} F$.

We now define the Legendre-Fenchel transform $G(\mathbf{m}, \beta)$ of the free energy

$$G(\mathbf{m}, \beta) = \sup_{\mathbf{h}} F(\mathbf{h}, \beta) + \mathbf{h} \cdot \mathbf{m}, \quad (2.17)$$

which depends on the mean local magnetizations m_i rather than the external fields h_i . It fulfills the equation of state

$$\frac{dG}{dm_i} = h_i = 0. \quad (2.18)$$

To systematically include fluctuations in the model, G is then expanded around the non-interacting case $J = 0$:

$$G \approx G \Big|_{J=0} + J \partial_J G \Big|_{J=0} + \frac{J^2}{2} \partial_J^2 G \Big|_{J=0} \quad (2.19)$$

where we again set $J = 1$ later. The detailed calculation can be found in Ref. [36] or App. (B). In essence, the expansion around the non-interacting case allows the evaluation of the averages in a simple fashion, as the non-interacting model can be solved exactly. The expansion yields the following expression for G :

$$\begin{aligned} \beta G(\mathbf{m}, \beta) &= \frac{1}{2} \sum_i (1 + m_i) \ln \frac{1 + m_i}{2} + (1 - m_i) \ln \frac{1 - m_i}{2} \\ &\quad - \frac{\beta}{2} \sum_{i \neq j} A_{ij} m_i m_j \\ &\quad - \frac{\beta^2}{4} \sum_{i \neq j} A_{ij} (1 - m_i^2)(1 - m_j^2) + \mathcal{O}(\beta^3), \end{aligned} \quad (2.20)$$

This sum is organized as follows: the first line is the Shannon entropy of a set of independent binary variables. The second line takes the form of the inner energy in mean-field approximation multiplied by β . The third line, proportional to β^2 , is known as the TAP or Onsager correction term [31]. To obtain the TAP self-consistency equations, we insert Eq. (2.20) into the equation of state, which yields

$$m_i = \tanh \left[\underbrace{\beta \sum_j A_{ij} m_j}_{\text{mean-field}} - \underbrace{\beta^2 m_i \sum_j A_{ij}^2 (1 - m_j^2)}_{\text{TAP}} \right], \quad (2.21)$$

In comparison to Eq. (2.8), we obtain a correction to second order in β , which we must anticipate from the fact that we expanded to second order in the interaction strength, which always appears together with a factor of β . We hence view Eq. (2.19) as a high-temperature expansion, as we must expect it to be increasingly accurate at small β , or equivalently high T . Since we know that the transition temperature diverges with the system size, and we are interested in the behavior at the transition,

it is reasonable to expect $\beta \ll 1$. In practice, however, we find that TAP and mean-field theory lead to similar results in the low temperature regime, while they differ close to the transition temperature. The similarity at low temperature stems from the factor $1 - m_j^2$ in Eq. (2.21): for a binary variable x_j , the expression $1 - m_j^2$ is just its variance, hence the second term in Eq. (2.21) arises due to the fluctuations of the variables x around their mean value. As the temperature decreases, the magnetization of each node approaches one and the fluctuations $1 - m_j^2$ approach zero.

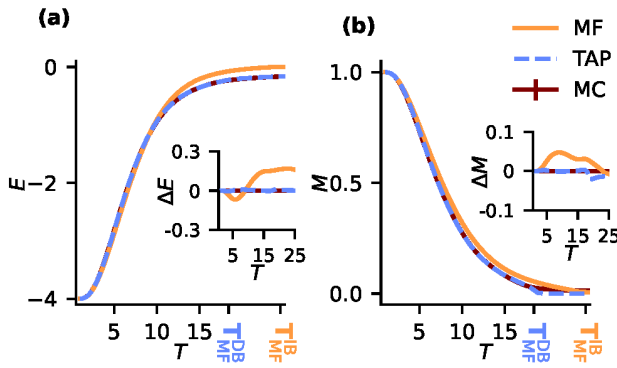


FIG. 2.3: Energy E (a) and magnetization M (b) as functions of temperature for a BAI network with $N = 10^4$ and $m_0 = 4$. Results from individual-based mean-field theory (MF) and TAP are compared to Monte Carlo simulations (MC). All data are obtained for the same adjacency matrix \mathbf{A} . The insets in both panels show the differences of MF and TAP results to MC.

Figure 2.3(b) demonstrates that non-zero solutions of Eq. (2.8) indeed emerge at higher temperatures than Eq. (2.12), at odds with Monte-Carlo simulations. The TAP approach is able to describe the high temperature behavior better than individual-based mean-field theory. Indeed, the onset of non-zero solutions for Eq. (2.21) approximately coincides with the transition temperature predicted by degree-based mean-field theory. In the next section, we confirm that the transition temperatures predicted by TAP and degree-based mean-field theory coincide.

2.2.3.1 TAP transition temperature

We will proceed analogously to Sec. 2.2.2.2 and solve the linearized Eq. (2.21) to find the transition temperature from TAP theory. The linearization yields another eigenvalue problem:

$$m_i = \beta \sum_j (A_{ij} - \beta \delta_{ij} k_i) m_j, \quad (2.23)$$

We now solve Eqs. (2.8) and (2.21) numerically and compare the resulting averages to the Monte-Carlo simulation in Fig. 2.3. Given the expectation values m_i , we can find the expectation value of the energy via

$$E = \langle H \rangle = -N^{-1} \partial_\beta \beta G(m, \beta), \quad (2.22)$$

where we omit the term proportional to β^2 to compute the mean-field result. We find that the predictions for the average magnetization M and energy E , while showing a similar overall behavior at low temperatures, exhibit different high-temperature behavior. In partic-

where $k_i = \sum_j A_{ij}$ is the degree of node i . We will hence define a β -dependent matrix

$$B_{ij} = A_{ij} - \beta \delta_{ij} k_i,$$

and seek its largest eigenvalue. In analogy to the arguments presented in Sec. 2.2.2.2, the transition temperature is the largest value of β for which such a solution to $\lambda_{\mathbf{B}(\beta), \max} \beta = 1$ exists. To do so, we compute the spectrum of B for different values of β and then look for the intersection point where

$$T_{\text{TAP}} = \lambda_{\mathbf{B}(\beta), \max}. \quad (2.24)$$

We show the comparison between $\lambda_{\mathbf{B}(\beta), \max}$ and T for different network sizes in Fig. 2.4(a). We find that this procedure reproduces the degree-based mean-field transition temperature Eq. (2.12). We also check whether the linearization (2.23) is consistent with the solution of the full TAP equations (2.21). To do so, we compute the projection of the magnetization \mathbf{m} onto the leading eigenvector $\mathbf{v}_{\mathbf{B}(\beta), \max}$ of $\mathbf{B}(\beta)$,

$$p_{\mathbf{B}(\beta)} = \frac{\mathbf{v}_{\mathbf{B}(\beta), \max} \cdot \mathbf{m}}{|\mathbf{m}|} \Theta(|\mathbf{m}|) \quad (2.25)$$

Indeed, as we show in Fig. 2.4(b), the projection approaches unity as $T \rightarrow T_{\text{TAP}}$ from below, so the linearized equation yields the same result as Eq. (2.21) there.

2.2.3.2 Cancellation of self-feedback

We have seen that the TAP approach can accurately predict the transition temperature while retaining the full local connectivity structure. We will now argue that this is so, because the TAP correction term eliminates the self-feedback effect which leads to the localization of states around the hubs.

For given values of m_i , the presence of node i affects the network like a heterogeneous external field. This field enters into the self-consistency equations of the nearest neighbors j of i . We illustrate this in the linearized equation Eq. (2.21), valid at the transition. We expand m_j for $j \neq i$ around $m_i = 0$:

$$m_j = m_j|_{m_i=0} + \beta A_{ji} m_i + \mathcal{O}(\beta^2) \quad (2.26)$$

keeping only terms up to linear order in m_i , m_j , and β . We then insert this expression into (2.23) to obtain

$$m_i = \beta \sum_j \left[A_{ij} \left(m_j|_{m_i=0} + \beta A_{ji} m_i \right) - \beta \delta_{ij} k_i m_i \right], \quad (2.27)$$

up to quadratic order in β , consistent with the high-temperature expansion employed

here. Using $\sum_j A_{ji} = k_i$, Eq. (2.27) simplifies to

$$m_i = \beta \underbrace{\sum_j A_{ij} m_j}_{\text{field in the absence of } i} \Big|_{m_i=0}, \quad (2.28)$$

which is similar to the linearized mean-field equation Eq. (2.9), but consistent with the notion of global order: the field at a node i has no local contribution from the value of m_i , as it is calculated in the absence of i . This is the mechanism which prevents the localization of states around hubs. It shows that order emerges as a global effect, on the whole network simultaneously, and cannot be driven by a single node.

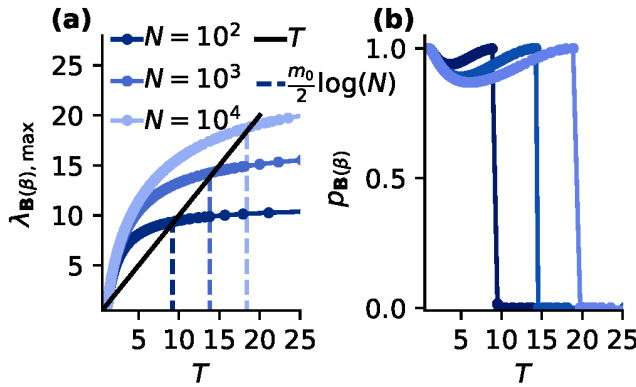


FIG. 2.4: (a) Leading eigenvalue $\lambda_{\mathbf{B}(\beta), \max}$ of $\mathbf{B}(\beta)$ as a function of temperature T for different system sizes N and $m_0 = 4$. The black solid line shows the identity and the vertical dashed lines mark temperatures $T = \frac{m_0}{2} \log N$ (b) Projection $p_{\mathbf{B}(\beta)}$ (2.25) of the magnetization obtained from solving the TAP equations onto the leading eigenvector of $\mathbf{B}(\beta)$ for different system sizes N .

The cancellation of the self-feedback, Eq. (2.28), also extends to lower temperatures beyond the regime of validity of the linearized Eq. (2.21). Using the same expansion method, one finds that up to quadratic terms in the inverse temperature β , the self-consistency equation reads

$$m_i \approx \tanh \left(\beta \sum_j A_{ij} m_j \Big|_{m_i=0} \right). \quad (2.29)$$

Thus, the inconsistency between degree-based mean-field theory and individual-based mean-field theory is resolved: it is not necessary to eliminate local structure, and thus self-feedback, by looking only at a degree-resolved statistic.

Rather, we must account for the spurious self-feedback introduced by the mean-field approximation.

2.2.3.3 Local order

Almost immediately below T_{TAP} , we observe that hubs of the system are "frozen", meaning that their mean value approaches one very fast, $m_u \approx 1$ for $k_u \gg \langle k \rangle$. Due to their large degree, they connect to a representative collection of nodes on the network and therefore do not deviate much from the expectation value $m(k)$, as

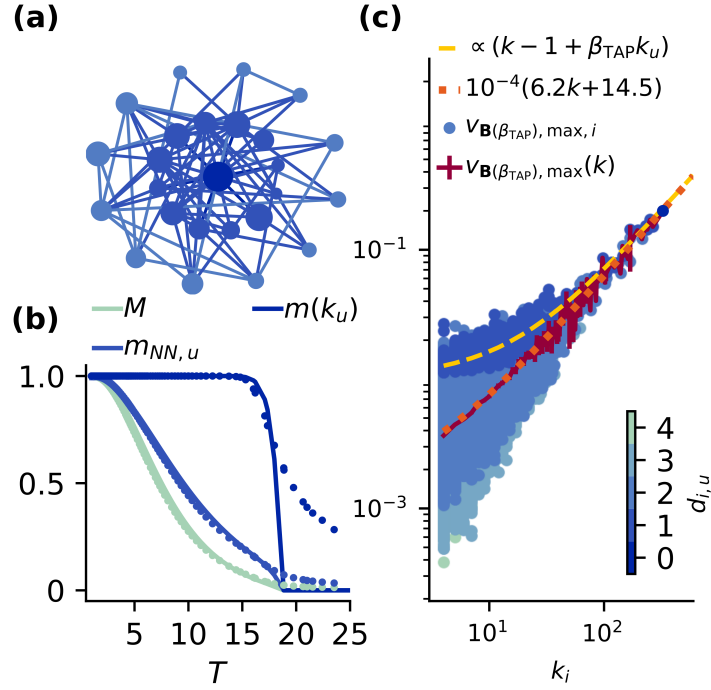


FIG. 2.5: (a) Hub u , the node with the largest degree, shown as central dark dot. Subset of the first 25 nodes i of the network, colored according to the minimal number $d_{i,u}$ of edges between i and hub u . (b) Average magnetization M (2.32), magnetization $m_{NN,u}$ (2.34) averaged over the nearest neighbors of u , and magnetization $m(k_u)$ (2.30) of the hub as functions of the temperature T within the TAP approach for a BAI model. Symbols denote the corresponding Monte Carlo data. (c) Eigenvector of the largest eigenvalue of $\mathbf{B}(\beta)$ at $T = \frac{m_0}{2} \log(N)$. Dots: entries of the eigenvector, sorted by their degree k_i , and colored according to the distance $d_{i,u}$ to the central hub u . Red curve: Average of the entries with nodes of equal degree. Dotted orange curve: linear fit of the averaged entry per degree. Dashed yellow curve: Scaling (2.33) of entry of a nearest neighbor of the central hub u . The common prefactor is calculated as $\beta_{\text{TAP}} S = k_u^{-1} v_{\mathbf{B}(\beta_{\text{TAP}}), u}$. All three panels show data from the same BAI model with $N = 10^4$, $m_0 = 4$.

seen in Fig. 2.5(c). We will hence refine the degree-based picture by computing the influence a hub has on its nearest neighbors. Our starting point for the analysis is the approximate result for the degree-resolved average magnetization from the degree-based mean-field theory [24],

$$m(k) \approx \tanh(\beta k S). \quad (2.30)$$

For nodes with small degrees k_i , this is not a good approximation, since their magnetization depends on their local environment, as we demonstrated in Fig. 2.2. Nodes of larger degree however contribute more strongly to the global order parameter S , for which $m_i \approx m(k_i)$ holds quite precisely, since their input fields are averaged over a larger local environment. The resulting equation for S is

$$S = \frac{1}{\langle k \rangle} \sum_k p(k) k \tanh(\beta k S). \quad (2.31)$$

This self-consistency equation can be solved numerically for arbitrary network sizes. From this, we compute the total magnetization via

$$M = \sum_k p(k) \tanh(\beta k S), \quad (2.32)$$

which can predict the magnetization measured in Monte-Carlo simulations accurately (Fig. 2.5(b)). We now make use of these results, obtained in the degree-resolved picture, to account for local structure. To this end we calculate the average magnetization of nearest neighbors of a hub u . Suppose we choose a node of degree k which is adjacent to u . Its effective field is composed of one contribution from $m(k_u)$ and $k-1$ connections coupling to the global order parameter S

$$h_{\text{NN},u}(k) = \beta(k-1)S + \beta m(k_u). \quad (2.33)$$

In the linear regime, where $m(k_u) = \beta_{\text{TAP}} k_u S$ and $m_{\text{NN},u}(k) = h_{\text{NN},u}(k)$ we find that this approximates the entries for the nearest neighbors of the hub u in the leading eigenvector of $B(\beta_{\text{TAP}})$ in Fig. 2.5(c). Making use of p_c and averaging over all k_u neighbors of u yields

$$m_{\text{NN},u} = \frac{1}{\langle k \rangle} \sum_k p(k) k \tanh[\beta(k-1)S + \beta m(k_u)], \quad (2.34)$$

which predicts an elevated magnetization for the neighbors of u , compared to the average value (Fig. 2.4(b)). We thus find that the degree-based view can be extended to account for the local differences in the ordered state of the system in a consistent manner. Equation (2.34) shows that hubs have a local effect: they elevate the magnetization of their nearest neighbors above the network average, Eq. (2.32).

2.2.4 Effective magnetic transition on finite networks and Monte-Carlo observables

Typically, one studies the onset of ferromagnetic order in the thermodynamic limit, i.e. for infinite system sizes. Only in this limit the spontaneous symmetry breaking associated with the onset of global order can emerge (see, e.g. the discussion in [37], Sec. 2). However, here the transition temperature diverges with the system size, such that the transition we want to study is never observed in the thermodynamic limit, as the system is always in the ordered state. Instead, we take a more pragmatic view and describe the properties of large but finite systems. Hence, we wish to determine the temperature below which we will typically find the system in an aligned state, which approaches the true ordering transition temperature in the thermodynamic limit. We here define the transition temperature as the peak position of the magnetic susceptibility (a detailed discussion of this approach for the Ising model on a regular lattice geometry can be found, e.g., in Ref. [38]). Since criticality is defined only in the thermodynamic limit, we choose to call this temperature the transition temperature rather than critical temperature.

At finite size, a global sign flip of all spins becomes feasible, as the energy difference is finite. This means that the average $\langle x_i \rangle$ vanishes exactly for a sufficiently long run, due to the Z_2 -symmetry of the Ising model Hamiltonian. For finite systems, this means that the magnetization in the absence of an external field must vanish

$$M_0 = \frac{1}{N} \left\langle \sum_i x_i \right\rangle. \quad (2.35)$$

This quantity also enters in the explicit form of the zero-field susceptibility

$$\chi_0 = \left. \frac{dM(h)}{dh} \right|_{h=0} = \beta \left(\frac{1}{N} \sum_{i,j} \langle x_i x_j \rangle - N (M_0)^2 \right). \quad (2.36)$$

For Ising models with a finite critical temperature, such as the Ising model on regular lattices, this quantity converges to the zero-field susceptibility in the thermodynamic limit only for temperatures $T > T_T$ [38], hence, in the paramagnetic regime. In the ferromagnetic regime, i.e. for temperatures below the transition temperature, the mean value of the absolute magnetization, which we denote by

$$M_{\text{MC}} = \frac{1}{N} \left\langle \left| \sum_i x_i \right| \right\rangle, \quad (2.37)$$

can instead be used to probe a unimodal symmetry broken state. We compare this quantity to the magnetic solutions of mean-field and TAP theory. Within the symmetry broken regime, we measure the susceptibility using M_{MC} , explicitly

$$\chi_{\text{MC}} = \beta \left(\frac{1}{N} \sum_{i,j} \langle x_i x_j \rangle - N (M_{\text{MC}})^2 \right), \quad (2.38)$$

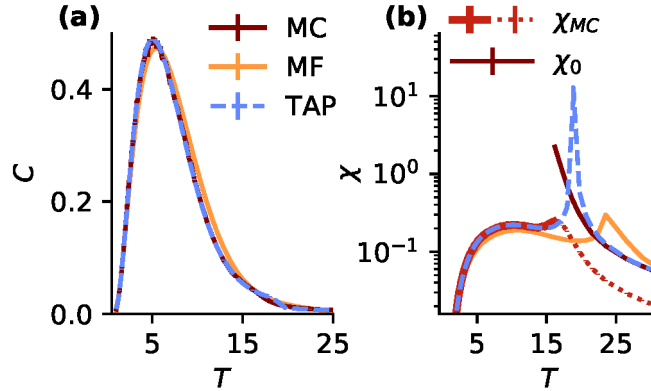


FIG. 2.6: Specific heat C (a) and susceptibility χ (b) as functions of temperature T for a BAI with $N = 10^4$ and $m_0 = 4$. χ_{MC} Eq. (2.38) is shown as a full line below (overlaid by TAP) and as a dotted line above the Monte Carlo transition temperature. Monte Carlo data for χ_0 Eq. (2.36) above the effective transition temperature, which is the appropriate estimator for the susceptibility in the paramagnetic regime. All data obtained for the same realization of the adjacency matrix \mathbf{A} of the network, where $N = 10^4$ and $m_0 = 4$.

which is the appropriate estimator here [38].

On the single spin level, we use a similar estimator to compute the magnetization, namely

$$m_{i,MC} = \left\langle x_i \operatorname{sign} \left(\sum_j x_j \right) \right\rangle, \quad (2.39)$$

which quantifies the alignment of individual spins with the system.

With the prescription for the measurement of the susceptibility and the specific heat in the Monte-Carlo simulations in hand, we can now proceed to compare these quantities to their counterparts originating from individual-based mean-field and TAP theory.

2.2.5 Transition temperature

In this section we compare the susceptibility and specific heat obtained from all three methods. We then estimate the transition temperature from the peak of the susceptibility. We use numerical differentiation to compute the specific heat $C = \partial_T E$ from E . For mean-field theory and TAP, we make use of the fact that the susceptibility, namely the response of spin i to an external field at site j is

$$\chi_{ij} = \partial_{h_j} m_i = -\partial_{h_j} \partial_{h_i} F(\mathbf{h}, \beta), \quad (2.40)$$

meaning that we can compute it from the Hessian of F . We now use that the Hessian $G^{(2)}$ is the negative inverse of the Hessian of the free energy, $F^{(2)}$, provided that Eq. (2.18) holds. We then compute the average response to a globally homogeneous field $h_i = h$, $\chi = \frac{1}{N} \sum_{ij} \chi_{ij}$. Figure 2.6 shows a comparison of the different predictions. The three methods yield only marginally different results for the specific heat, with the agreement between TAP and the Monte-Carlo simulations slightly higher, consistent with their better agreement for E in Fig. 2.3.

For all three methods, we now determine the position of the peak of χ , which defines the transition temperature. For low temperatures, both individual-based mean-field theory and TAP agree well with the Monte-Carlo simulations. However, as we approach the transition temperature, both theories show a pronounced peak in susceptibility, which is absent in the Monte-Carlo simulations. This pronounced peak is a result of the approximations made to derive these theories: For a finite system, the susceptibility must necessarily always remain finite, and may never diverge. In the high temperature regime, we again find a good agreement between TAP prediction and the estimator χ_0 in the paramagnetic regime. Finally, we find that both TAP and individual-based mean-field theory overestimate the transition temperature, but the TAP peak is considerably closer to the peak in the Monte-Carlo susceptibility. As we argued in the previous section, this is so because TAP theory eliminates non-physical self-feedback in the self-consistency equations for m_i .

A comparison for the transition temperatures obtained from all three methods, shown in Fig. 2.7, reveals that the scaling of the TAP transition temperature with the system size, identical to the prediction from degree-based mean-field theory, indeed matches the scaling obtained from simulations up to a constant shift. This is consistent with previous work on the BAI model [27, 28]. Further, we also find that individual-based mean-field theory, while approximately equal to TAP and degree-based mean-field theory for small systems, eventually scales with $N^{\frac{1}{4}}$, as we anticipated for the largest eigenvalue of the adjacency matrix, Eq. (2.13).

In conclusion, we find that the degree-based theory yields accurate results for the transition temperature only due to its ignorance of local structure, which would

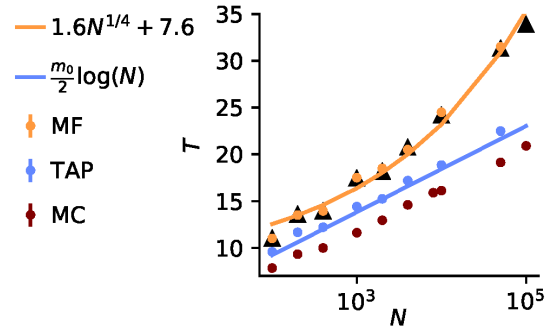


FIG. 2.7: Effective transition temperature as extracted from the maximum of the susceptibility as a function of system size N for $m_0 = 4$. Red dots: Estimated from Monte Carlo simulations using Eq. (2.38). Yellow dots: Solutions of the local mean-field equation Eq. (2.8). Violet dots: Solutions of the TAP equation Eq. (2.21). Violet curve: Global mean-field (2.12) or, equivalently, TAP prediction. Black triangles: Leading eigenvalue of the adjacency matrix $\lambda_{A,\max}$. Orange curve: Fit of $\lambda_{A,\max}$ to $aN^{1/4} + b$ to illustrate the $N^{1/4}$ scaling.

otherwise cause a spurious self-feedback effect in individual-based mean-field theory. Nevertheless, accurate predictions of both local magnetization and degree-based global averages can be obtained via TAP theory.

2.3 Spurious self-feedback in the spread of disease

This section, parts of Chap. 4 and Appendices E, F and G are based on the following publication:

Merger, C., Albers, J., Honerkamp, C., Helias, M., 2023. Spurious self-feedback of mean-field predictions inflates infection curves, *in preparation*

Author contributions

Under the supervision of Moritz Helias and Carsten Honerkamp, the author worked on all parts of the above publication. The author contributed to the general formalism and wrote the original draft of the manuscript. The numerical experiments and implementation of the SIR, SIRS and SIS dynamics were completed jointly with Jasper Albers. All authors contributed to finalizing the manuscript.

In the previous section, we considered a system at equilibrium. Provided that the system is ergodic, this is equivalent to computing long time averages of the system. Now, we will deal with a dynamic process, namely the spread of disease, where we are particularly interested in the dynamics.

2.3.1 Models for the spread of disease

In modeling the spread of disease, we are interested in predicting the *rate* at which it spreads, the temporal position and height of the *peak* of the infection curve, and whether we can expect the wave of infection to die out, or if there exists a state in which a significant proportion of the population is infected in the so-called endemic phase.

These phenomena are modeled by the SIR model [22] and its variants. In these models, each agent can be in one of three states (susceptible (S), infected (I), or recovered (R)). Transitions between the states occur stochastically: at each time step, an infection travels from an agent j to its nearest neighbor i with probability β . The agents interact via

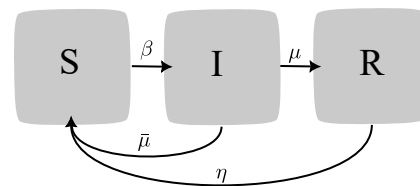


FIG. 2.8: Models of disease spreading. Susceptible agents with n infected neighbors transition to the infected state with probability βn . Infected agents return to the susceptible state with probability $\bar{\mu}$ or recover with probability μ and move to the R state, in which they are immune. Finally, agents in the R state lose their immunity and move back to the susceptible state with probability η .

a network with adjacency matrix A , which is constant in time². The probability that agent i is infected at time $t + 1$ is $\theta_i(t)$ with $\theta_i(t)$ defined by

$$\theta_i(t) = h_i(t) + \beta \sum_j A_{ij} I_j(t) \quad (2.41)$$

with an external field $h_i(t)$ (corresponding to influx of infections from outside the network) and A the adjacency matrix of a network. We will later set $h_i = 0$ to investigate the endogenous dynamics. At each time step, an infected agent recovers with probability μ , in the recovered state, the agent is immune to infection. Hence, in the SIR model, the number of infected agents must eventually decline to zero, when all infected agents are recovered.

One variant of the SIR model is the SIS model: Here, infected nodes transition back to the susceptible state with probability $\bar{\mu}$ meaning that they recover without attaining immunity, and may thus be infected again. Hence in the SIS model, it is possible to reach an endemic state where a finite fraction of agents is infected at all times. The minimal fraction $\bar{\lambda}_c = \frac{\beta_c}{\bar{\mu}}$, where this occurs, is called the epidemic threshold.

The SIRS model is a variant of the SIR model with waning immunity; here, recovered agents lose their immunity with probability η in each time step, but the transition from $I \rightarrow S$ is not allowed. Like the SIS model, the SIRS model may have an endemic state in which the finite fraction of agents is infected.

An overview of the transition probabilities in the different models is shown in Fig. 2.8.

We will model N agents i via a set of two binary variables, $S_i, I_i \in \{0, 1\}$ each. The susceptible state hence corresponds to $S_i = 1$ and $I_i = 0$, and the infected one to $S_i = 0$ and $I_i = 1$. Once the individual recovers, we set $S_i = I_i = 0$. We are now interested in the evolution of the probabilities ρ_i^α that individual i belongs to compartment $\alpha \in \{S, I, R\}$. Here, we will compute only the change in ρ_i^I and ρ_i^S , since ρ_i^R follows from $\rho_i^R = 1 - \rho_i^S - \rho_i^I$. We will treat the most general case, where in principle $\beta, \eta, \mu, \bar{\mu}$ are all finite. We will then set some of these probabilities to zero to treat the SIR ($\bar{\mu} = \eta = 0$), SIS ($\mu = \eta = 0$), or SIRS ($\bar{\mu} = 0$) case. Using Eq. (2.41), the difference between two time steps $\Delta\rho_i^\alpha(t+1) = \rho_i^\alpha(t+1) - \rho_i^\alpha(t)$ is

$$\begin{aligned} \Delta\rho_i^S(t+1) &= \eta\rho_i^R(t) + \bar{\mu}\rho_i^I(t) - \beta \langle S_i(t)\theta_i(t) \rangle \\ \Delta\rho_i^I(t+1) &= -(\mu + \bar{\mu})\rho_i^I(t) + \beta \langle S_i(t)\theta_i(t) \rangle. \end{aligned} \quad (2.42)$$

The terms in the first line of Eq. (2.42) are organised as follows: $\eta\rho_i^R(t) + \bar{\mu}\rho_i^I(t)$ is the influx from the R, I state to the S state, respectively. The term $\beta \langle S_i(t)\theta_i(t) \rangle$ constitutes the probability of infection, provided that the agent is currently in the S state ($S_i(t) = 1$). The same term appears with the opposite sign in the second line of Eq. (2.42), as it must increase the probability of infection. Finally, the term $-(\mu + \bar{\mu})\rho_i^I(t)$ denotes

²Time dependent connectivities $A(t)$ may however be treated within the same formalism.

the out-flux from the infected state to the S state with probability $\bar{\mu}$ and to the R state with probability μ , respectively.

Evaluating Eq. (2.42) exactly is infeasible for large system sizes because of the average $\beta \langle S_i(t)\theta_i(t) \rangle$, which couples the evolution equations for different agents to each other. To evaluate the average exactly, we must sum over all possible trajectories of all variables up to the time point t , weighted by their respective probabilities. However, the sum of these possible trajectories grows exponentially in time and with the number of agents, rendering an exact computation infeasible.

A typical approximation stipulates that agents are statistically independent, meaning that averages of the type $\langle S_i(t)\theta_i(t) \rangle$ factorize into $\langle S_i(t) \rangle \langle \theta_i(t) \rangle$, since θ_i neither depends on S_i nor I_i . This yields the following dynamics

$$\begin{aligned}\Delta\rho_i^S(t+1) &= \eta\rho_i^R(t) + \bar{\mu}\rho_i^I(t) - \beta\rho_i^S(t) \sum_j A_{ij}\rho_j^I(t) \\ \Delta\rho_i^I(t+1) &= -(\mu + \bar{\mu})\rho_i^I(t) + \beta\rho_i^S(t) \sum_j A_{ij}\rho_j^I(t),\end{aligned}\tag{2.43}$$

which is a closed recursion in ρ_i^α , and can hence be solved efficiently even for large systems. We will call this approximation (quenched) mean-field approximation³. The assumption of statistical independence however, is typically not well-justified. In particular, this formulation introduces a spurious self-feedback effect as we outlined in Sec. 2.1. We here improve upon mean-field theory by deriving a correction to Eq. (2.43) which takes fluctuations into account.

Previous approaches to the SIR model and its variants also take the co-dependence of different agents' activities into account, either by tracing the evolution of correlations (the pair approximation) [39] in addition to the mean, or via dynamic message passing [40–42]. These approaches introduce new dynamic variables which describe the correlations between pairs of agents. We discuss our results in the context of the pair approximation, individual-based mean-field approximation and the dynamic message passing approach in Sec. 2.3.3. In the next section, we introduce the dynamic fluctuation correction to Eq. (2.43), derived here via a systematic expansion in β .

2.3.2 Fluctuation correction

Formally, one can write the average over all possible realizations of the stochastic processes as a path integral. In the case of $\beta = 0$, all trajectories of all agents decouple, and the system can be solved exactly. We make use of this fact to compute a fluctuation correction up to second order in β , starting from the non-interacting case $\beta = 0$. The procedure is the same as in Sec. 2.2.3 (the explicit derivation is given in

³"quenched" here refers to the quenched connectivity of the graph, as opposed to an ensemble over realizations of the connectivity matrix

App. (B)), with the important distinction that we are now interested in an evolution of averages over time. The authors of [43] demonstrate how to calculate a dynamical fluctuation correction for the Ising model outside of equilibrium. The resulting update equations are of the type of the dynamical TAP equations for spin systems, a non-equilibrium version of the TAP correction term [31, 35], which we already used in Sec. 2.2.3. We here adapt this approach to the SIR model and its variants. Our derivation, detailed in App. (F) for the SIR model, and extended to the SIRS and SIS model in App. (G), follows the same steps as [43], with two important distinctions: first, we must track two binary variables per agent. Second, the update probabilities for each agent i depend on the state of the same agent in addition to their dependence on θ_i .

To first order in β , one obtains Eq. (2.43), meaning that the systematic expansion reproduces mean-field theory. To second order, each update equation acquires a correction term of order β^2 , which reads

$$\begin{aligned}\Delta\rho_i^S(t+1) &= \eta\rho_i^R(t) + \bar{\mu}\rho_i^I(t) - \rho_i^S(t)\beta\sum_j A_{ij}(\rho_j^I(t) - \rho_{j\leftarrow i}^I(t)) \\ \Delta\rho_i^I(t+1) &= -(\mu + \bar{\mu})\rho_i^I(t) + \rho_i^S(t)\beta\sum_j A_{ij}(\rho_j^I(t) - \rho_{j\leftarrow i}^I(t)),\end{aligned}\tag{2.44}$$

where $\rho_{j\leftarrow i}^I(t)$, to first order in β , is just the probability that agent j was infected by agent i at an earlier point in time

$$\rho_{j\leftarrow i}^I(t) := \beta A_{ji} \sum_{t' \leq t-1} \rho_j^S(t') \rho_i^I(t') (1 - \mu)^{t-t'-1}\tag{2.45}$$

Here, $\beta A_{ji} \rho_j^I(t') \rho_i^S(t')$ describes the probability that agent j becomes infected at $t' + 1$ due to agent i , and $(1 - \mu)^{t-t'-1}$ is the probability that node j does not recover between $t' + 1$ and t .

Thus, in the context of the SIR model, the second order correction exactly cancels the self-feedback loop of the node i via its nearest neighbors. In the following sections, we will illustrate the impact of this correction for the SIR and SIRS model.

We compare the predictions of Eq. (2.43) and Eq. (2.44) to simulations implemented in NEST [44]. NEST is a simulator for spiking neurons, we describe the implementation of the SIR dynamics in NEST in App. (E). Just as the Monte-Carlo simulations in the BAI model, the simulations here serve as a ground truth, by means of a comparison to them, we assess the validity of our theoretical predictions.

2.3.2.1 SIR model

We here study the effect which the average connectivity has on the spread of the infection. To do so, we generate Erdős-Rényi graphs [45] of different average degrees

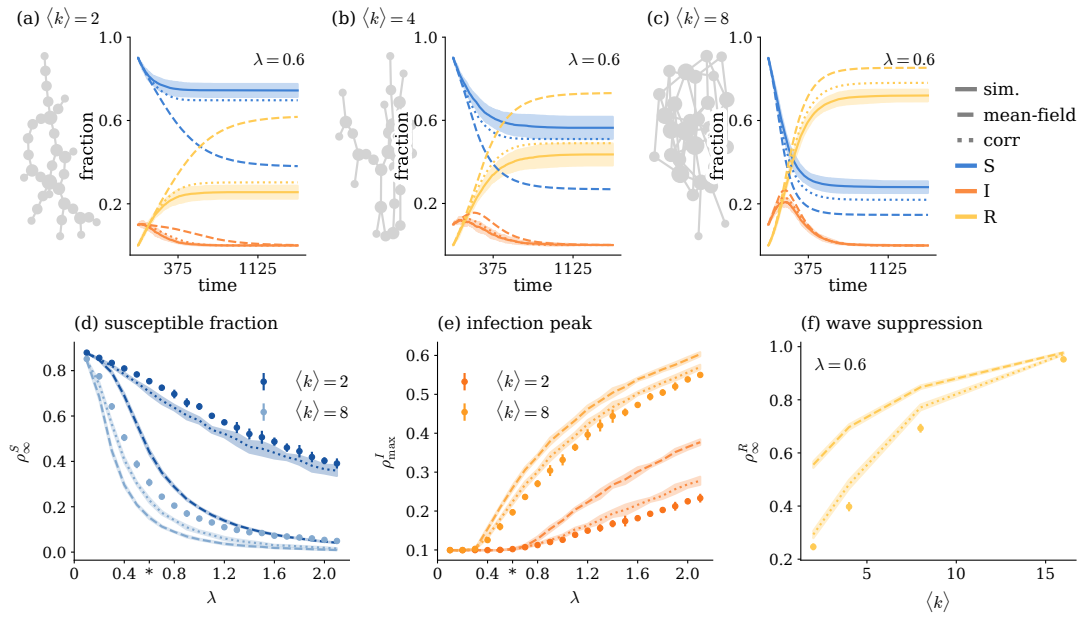


FIG. 2.9: Sparsity suppresses spread of infection. (a)-(c) Full lines are results of simulation for fixed $\mu = 10^{-2}$, $\lambda = 0.6$ and theory, shaded areas show one standard deviation, averaged over 10 different realizations of the stochastic process on the same network realization with the same initial infected agents. Dashed lines show mean-field result; dotted lines the second order correction. (d) Fraction of agents in the susceptible state after the dynamics have stopped. Dots are simulation results, curves show mean-field and second order corrections, shaded areas indicate one standard deviation around average result. All data are averaged over ten realizations of the adjacency matrix. (e) Same as (d), but for the peak of the infection curve. Stars along the axis mark the value of λ used in (a)-(c). (f) Fraction of agents in the recovered state after the dynamics have stopped vs average degree of the network. All results were obtained on networks with average size $N = 10^3$

$\langle k \rangle$, infect a small fraction of randomly chosen agents, and then let the dynamics evolve. Erdős-Rényi graphs possess no hierarchical structure; on an Erdős-Rényi graph of size N , each of the $N(N-1)$ possible edges exists with probability $\frac{\langle k \rangle}{N}$. Being completely random, they are suited to study the effect which the average connectivity of the graph has on the spread of the infection. After construction, we ensure that the graph is connected, i.e. there exists a path from any agent to any other agent on the graph, such that the infection can in principle reach any agent. This may reduce the size of the graph, we accept graphs which are in an interval of $\pm 20\%$ of the target system size. If the effective size of the graph is outside this range, the graph is redrawn with an adjusted initial size, and the procedure is repeated, until the connected component of the graph is in the desired size range.

We compare the predictions of the mean-field and second-order predictions to simulations of the stochastic process. In Fig. 2.9 (a)-(c) we show the infection curves, averaged over the whole population. For densely connected networks, the predictions of mean-field theory, the second order correction and the simulation all agree. As the networks become more sparse, however, the number of infections decreases much faster than mean-field theory predicts, but is still well approximated by the second order corrected theory. This shows that self-feedback is highly relevant for sparse networks.

The latter observation holds true over a broad range of fractions $\lambda = \frac{\beta}{\mu}$. After the infection wave has died down, all agents must be either susceptible or recovered. The number of susceptible agents in the final state, ρ_∞^S , decreases with increasing λ - but much slower than predicted by mean-field theory. We find that, on sparse graphs in particular, the difference between the mean-field prediction and the simulation is large, while 2.44 captures the dynamics well.

Finally, we compare the height of the peak for different network topologies and different values of λ in Fig. 2.9(e). We find that below a certain value of λ , which depends on the average degree $\langle k \rangle$, the number of agents individuals does not grow beyond the number of initially infected agents, thus the infection dies out immediately. Beyond this point, the height of the infection peak increases with λ , and again the mean-field theory predicts a higher value than observed in simulations. We find that the number of infected agents overall decreases rapidly with the average degree. After the dynamics have stopped, the number of recovered agents is equivalent to the overall number of agents that experienced an infection, it is the area under the "infection wave". In Fig. 2.9(f), we compare the fraction ρ_∞^R of recovered agents after the infection has died out to predictions from both theories, finding again that as the graph becomes sparser, far fewer agents become infected than predicted by mean-field theory.

Despite the good agreement between simulations and the second-order correction Eq. (2.44), we find that the corrected theory consistently overestimates the fraction of infected agents slightly, both at the height of the wave, as shown in Fig. 2.9(e),

and in the average over the time trajectory, the pandemic, see Fig. 2.9(d). This is so, because the correction Eq. (2.44), does not correct for higher-order feedback loops: the correction eliminates self-feedback via direct nearest neighbors, but self-feedback can also occur via over-next nearest neighbors or loops of three or more distinct nodes in the graph. As the connectivity increases, the networks become more clustered, and more potential self-feedback loops appear. These higher-order self-feedback effects are necessarily $\mathcal{O}(\beta^3)$ and hence not treated here. Nevertheless, in comparison to the difference between mean-field theory and the second order correction, these higher-order terms appear to play a minor role.

2.3.2.2 SIRS model

In the SIRS model, an infected agent can reach the susceptible stage again, hence, in principle, a self-infection loop is allowed, provided that the transition back to the susceptible state has taken place in the mean-time. Given this, the correction Eq. (2.44) is counter-intuitive: clearly, self-feedback may exist here. However, if $\eta \ll \mu, \beta$, then the loss of immunity is a slow process compared to infection and recovery. In this case, the positive self-feedback effect may be very small, as such a correction comes with a factor of $\eta\beta^2$. We apply Eq. (2.44) with $\mu = 10^{-2}$, $\eta = 10^{-3}$ and vary β between 10^{-3} and 2μ . We show the results of this procedure in Fig. 2.10.

We find that both for very sparse graphs, e.g. $\langle k \rangle = 2$ and highly connected graphs, e.g. $\langle k \rangle = 16$, the fluctuation correction reproduces the average dynamics well (see Fig. 2.10 (a), (c)). Furthermore, both for small and large connectivity, the final fraction of susceptible and infected agents is also captured quite accurately Fig. 2.10 (d), (f). In particular, in the sparse connectivity model, the infection wave dies out completely, which is accurately predicted by the fluctuation correction, whereas mean-field theory predicts $\rho_\infty^S < 1$ for almost all values of λ . For large connectivity, the difference between mean-field and the fluctuation correction is also small, both underestimate ρ_∞^S (and overestimate ρ_∞^I), slightly. The height of the infection peak is predicted with high accuracy by the fluctuation correction, at all levels of connectivity, while it is overestimated by the mean-field result, especially for sparse graphs. For intermediate connectivity, here $\langle k \rangle = 8$ however, the fluctuation correction overestimates the activity in the endemic state: It underestimates the final fraction of susceptible individuals, while it overestimates the number of infected agents in the final epidemic state, see Fig. 2.10 (d), (f).

If the mechanism which leads to the deviation between the fluctuation correction and the simulation were the omission of the positive self-feedback loop which we described at the outset of this section, then the fraction of infected agents in the endemic state should be underestimated by the fluctuation correction. Contrary to this, we rather find that it is overestimated by the fluctuation correction. Examining the trajectory of a single realization of a graph at intermediate connectivity Fig. 2.10

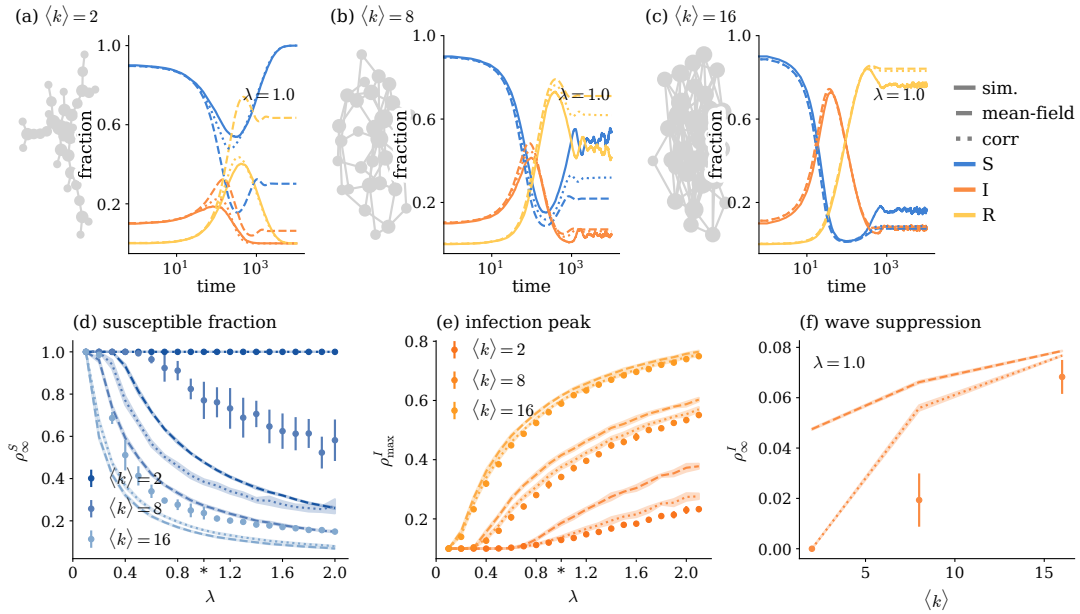


FIG. 2.10: Fluctuation correction for SIRS model with slowly waning immunity, $\eta = 10^{-3}$.
(a)-(c) Full lines are results of simulation for fixed $\mu = 10^{-2}$, $\lambda = 1.0$ and theory, shaded areas show one standard deviation, averaged over 10 different realizations of the stochastic process on the same network realization with the same initial infected agents. Dashed lines show mean-field result; dotted lines the second order correction. **(d)** Fraction of agents in the susceptible state after the dynamics have stopped. Dots are simulation results, curves show mean-field and second order corrections, shaded areas indicate one standard deviation around average result. All data are averaged over ten realizations of the adjacency matrix. **(e)** Same as (d), but for the peak of the infection curve. Stars along the axis mark the value of λ used in (a)-(c). **(f)** Fraction of agents in the infected state after the dynamics have stopped vs average degree of the network. All results were obtained on networks with average size $N = 10^3$

(b), we find that even when the same set of agents is infected at the outset of the simulation, the final number of recovered and susceptible agents shows large variations. Both the fluctuation correction and the mean-field prediction initially predict the average activities well - before they deviate from them, the latter at an earlier time point than the former. On the level of single realizations of the stochastic process, we observe oscillations in ρ^α , which are not present neither in the first nor the second order dynamical evolution.

2.3.3 Further fluctuation corrections to infection models

Apart from the mean-field method, there exist further methods to approximate the dynamics Eq. (2.42). They are dynamical message-passing (DMP) [40, 41] and the so-called pair approximation [39]. Both of these approaches yield a larger set of coupled differential equations.

Within the pair-approximation, one performs a so-called moment closure. Here, in addition to the first moments $\langle S_i \rangle, \langle I_i \rangle$, one also tracks higher second moments such as $\langle I_i S_j \rangle$ for different agents i, j , thus taking correlations between these variables into account. However, the evolution of first and second moments of the distribution is not closed, to compute the time evolution of $\langle I_i S_j \rangle$, one needs third moments such as $\langle I_i S_j I_l \rangle$, which are then approximated using second and first moments. While the pair approximation certainly improves upon the mean-field equations, it is less computationally efficient, requiring the tracking of $\mathcal{O}(N^2)$ coupled differential equations.

Within DMP, one tracks so-called messages, namely the probabilities that an infection is passed along a given edge of the graph. The DMP equations follow from a strict prohibition of all self-feedback. For the SIR model, DMP is exact on trees. DMP has also been used for recurrent infection models such as the SIRS model [42], even though self-feedback is known to be present in these systems. Despite this, the authors of [42] note that their recurrent DMP method yields an improvement of the results.

This observation is in line with our result Eq. (2.44), which shows that indeed fluctuations cancel self-feedback up to the second order in the interaction. However, while the DMP approach imposes the non-existence of self-feedback explicitly, here its cancellation emerges from a systematic expansion. Furthermore, while the DMP equations of [42] eliminate all self-feedback, our update equations in principle allow for self-feedback of higher orders.

Within recurrent infection models, the allowed self-feedback loop, in which an initially infected node recovers and becomes susceptible again between the original infection and the receipt of the self-feedback signal, comes with a factor $\beta^2 \mu \eta \ll \beta^2$ for the SIRS model, or a factor $\beta^2 \bar{\mu} \ll \beta^2$ for the SIS model. In the case in which the loss of immunity is a significantly slower effect than the spread of the infection

therefore, one could attempt to argue that this effect becomes negligible, thus justifying the use of DMP theory. However, depending on the network topology, the positive self-feedback effect exists, and can considerably change the predictions [46]. A systematic comparison between the pair approximation and the DMP approach to the SIRS model on scale-free graphs has been provided in [47]. The authors of the latter study find that neither theory predicts the endemic activity or the localization of the activity above it accurately.

For the SIR model therefore, our result eliminates the strongest self-feedback effect, while it does not eliminate self-feedback altogether. For recurrent infection models such as the SIRS model, this partial elimination has a different interpretation: there, it is only a correction on the level of fluctuations. From the form of this correction term, we can not conclude that self-feedback in all forms is absent in SIRS models. Nevertheless, it can explain the partial success of DMP theory also in recurrent infection models, since it illustrates that fluctuations can lead cancellation of self-feedback to a certain extent. We expect that higher than second order corrections to the dynamics will differ in their nature depending on which variant of the three models one chooses.

2.3.4 The epidemic threshold

In this section, we present our results on the endemic activity in the SIRS model.

In infection models, the epidemic threshold is defined as the configuration of parameters below which there exists an absorbing phase: a state where almost all individuals remain healthy [32]. In the SIR model, this means that an infected agent transmits the infection to no more than one other agent on average (corresponding to a reproduction number ≤ 1). In the SIRS or SIS model, this means that the wave of infection eventually dies out, such that the dynamics stop in the end. Figure 2.10 shows an example of this: in Fig. 2.10 a), we see that the dynamics eventually die out, $\rho^S \rightarrow 1 = \rho_\infty^S$ as time progresses. The system in Fig. 2.10 a) is therefore below the epidemic threshold. Indeed, note that for $\langle k \rangle = 2$, we here find $\rho_\infty^S = 1$ for all configurations of λ in Fig. 2.10 d), such that we do not observe any transition. For larger values of $\langle k \rangle$, seen in Fig. 2.10 b) and c), we find a finite fraction of perpetually infected individuals above a certain value of λ , the number of infected individuals in the endemic state, ρ_∞^S , drops below one there, see in Fig. 2.10 d).

We will now compute how the endemic state depends on the network topology and the rates of infection and recovery. The epidemic threshold bears a resemblance to a phase transition: at a certain parameter λ , the system properties fundamentally change: the absorbing state, where the number of infected agents is zero, becomes unstable. Rather, the number of infected agents increases. Before we derive the epidemic threshold from Eq. (2.44), we state two known estimates for the epidemic threshold originating from mean-field theory. We find that all three estimates are the same independent of whether one chooses the SIS or SIRS model.

Individual-based mean-field theory, also referred to as quenched mean-field theory, corresponds to the evolution equations (2.43) [32]. Linearizing these equations for small initial infection levels yields the condition

$$\lambda_c^{\text{IBMF}} = \lambda_{A,\text{max}}^{-1} \quad (2.46)$$

with $\lambda_{A,\text{max}}$ the largest eigenvalue of the adjacency matrix. If λ is larger than this value, the number of infected agents will grow according to Eq. (2.43).

Degree-based mean-field theory, which stipulates that all agents of equal degree are statistically equivalent, yields the following result [48]

$$\lambda_c^{\text{DBMF}} = \frac{\langle k \rangle}{\langle k^2 \rangle} \quad (2.47)$$

provided that the network has an uncorrelated degree distribution: the probability $p(k'|k)$ that a node of degree k is connected to a node of degree k' is $p(k')k'/\langle k \rangle$. This degree-based approach to infection models is equivalent to the degree-based approach to Ising spins on a scale-free lattice which we introduced in Sec. 2.2.2.1.

For $\gamma \leq \frac{5}{2}$, degree-based mean-field theory and individual-based mean-field theory are expected to yield equivalent results for large N [32]. For $\gamma > \frac{5}{2}$, individual-based mean-field theory is known to overestimate the epidemic threshold. This coincides with the point where the largest eigenvalue of the adjacency matrix localizes due to the self-feedback effect, see [49]. The authors of Ref. [49] suggest that to assess the relative importance of a node in a graph, the leading eigenvector of the following matrix $2N \times 2B$ should be considered

$$M = \begin{pmatrix} A & \mathbb{1} - K \\ \mathbb{1} & 0 \end{pmatrix}, \quad (2.48)$$

where $K_{ij} = \delta_{ij}k_i$ is a diagonal matrix of with the degree k_i of the node i on the diagonal. The underlying idea is that the self-feedback effect makes hubs appear more relevant than they are. This suggestion is remarkable, because it coincides with the estimator of the epidemic threshold for DMP. The estimator for the DMP epidemic threshold is then the inverse leading eigenvalue of M [42, 46]. We will see in the following, that the TAP estimation follows from a similar $2N \times 2N$ matrix, whose off-diagonal blocks however differ from those in M .

TAP prediction. We now compute an estimate for λ_c from Eq. (2.44). To simplify the analysis, we define another dynamical variable $\tau_i(t)$ which has the following

properties

$$\begin{aligned}\tau_i(0) &= 0 \\ \Delta\tau_i(t+1) &= -(\mu + \bar{\mu})\tau_i(t) + \rho_i^I(t)\beta \sum_j a_{ij}a_{ji}\rho_j^S(t).\end{aligned}$$

Using τ , we may write Eq. (2.44) as

$$\begin{aligned}\Delta\rho_i^I(t+1) &= -(\mu - \bar{\mu})\rho_i^I(t) + \rho_i^S(t)\beta \left(\sum_j A_{ij}\rho_j^I(t) - \tau_i(t) \right) \\ \Delta\rho_i^R(t+1) &= \mu\rho_i^I(t) - \eta\rho_i^R(t),\end{aligned}$$

and we further require that $\rho_i^S(t) = 1 - \rho_i^I(t)$ for the SIS and $\rho_i^S(t) = 1 - \rho_i^R(t) - \rho_i^I(t)$ for the SIRS model. We now examine the regime in which ρ^I is infinitesimally small, hence we keep only terms up to linear order in ρ^I . This can be seen as a small perturbation from the state of zero activity. If this state is unstable, the number of infected agents grows, $\Delta\rho^I > 0$, and we are above the epidemic threshold. Linearizing the update equations in ρ^I , we find that

$$\begin{pmatrix} \Delta\rho^I(t+1) \\ \Delta\tau(t+1) \end{pmatrix} = \left[-(\mu + \bar{\mu}) \begin{pmatrix} \mathbb{1} & 0 \\ 0 & \mathbb{1} \end{pmatrix} + \beta \begin{pmatrix} A & -\mathbb{1} \\ K & 0 \end{pmatrix} \right] \begin{pmatrix} \rho^I(t) \\ \tau(t) \end{pmatrix}, \quad (2.49)$$

with $K_{ij} = \delta_{ij}k_i$ a diagonal matrix with the degrees of the nodes along the diagonal. We now ask that the right hand side of this equation must be equal to zero: This is exactly the point where the activity neither decays nor increases and therefore marks the transition point. This gives us an estimate of the epidemic threshold:

$$\lambda^{-1} \begin{pmatrix} \rho^I(t) \\ \tau(t) \end{pmatrix} = \underbrace{\begin{pmatrix} A & -\mathbb{1} \\ K & 0 \end{pmatrix}}_{=:D} \begin{pmatrix} \rho^I(t) \\ \tau(t) \end{pmatrix}, \quad (2.50)$$

where we use $\lambda = \frac{\beta}{\mu}$ in case of the SIS and $\lambda = \frac{\beta}{\bar{\mu}}$ for the SIRS model. This is an eigenvalue equation, the inverse largest eigenvalue of the matrix on the right hand side yields the epidemic threshold. Since the matrix D is not symmetric, we can obtain complex eigenvalues, corresponding to oscillatory solutions to Eq. (2.49). To find the point where the activity decreases or increases, it is however sufficient to look at the real part of the eigenvalues. However, it may be fruitful to test whether the frequencies of the oscillations in the SIRS model at intermediate connectivities can be predicted by the linearized Ansatz.

We now compare these different approaches to each other on scale-free networks. We generate networks with scale-free distributions $p(k) \sim k^{-\gamma}, k \geq k_0$ using the configuration model [50]. To ensure that the networks are connected with high probability, we choose a minimal degree $k_0 = 3$. We then perform simulations of the SIRS model for

various values of λ . The epidemic threshold is determined by finding the maximum of the susceptibility

$$\chi = N \frac{\langle (\rho^I)^2 \rangle - \langle \rho^I \rangle^2}{\langle \rho^I \rangle}, \quad (2.51)$$

where the average is computed in the endemic state, see e.g. [32].

Figure 2.11 shows a comparison of the three different predictions to simulations. For small system sizes, all theories underestimate the epidemic threshold. As the system size increases, degree-based mean-field theory and the TAP prediction yield more and more similar predictions. For $\gamma = 2.4$, all three theories eventually agree with the simulation, see Fig. 2.11 c). For $\gamma = 2.9$, individual-based and degree-based mean-field theory do not agree, as we expect. However, both the TAP estimator and the degree-based mean-field result show a good agreement with the simulation.

This analysis also shows that the TAP prediction and dynamical message passing are not equivalent: dynamical message passing overestimates the epidemic threshold in the same parameter regime [46]. We hence see that although our theory eliminates self-feedback effects up to second order in β , it is not equivalent to DMP.

We note that the scaling of the different theories and the simulation should be tested also for larger system sizes, as the differences in scaling between theory and simulation are sometimes only resolved at $N \sim 10^6$ [32]. Investigating whether the good agreement between TAP and simulations also extends to different architectures and larger system sizes is an interesting direction of future research. Finally, understanding the working mechanism behind the similarity of the TAP and the degree-based mean-field result (see e.g. Fig. 2.11 d)) better may shed light on the nature of the degree-based approach to heterogeneous systems, in full analogy to our result on the Ising system.

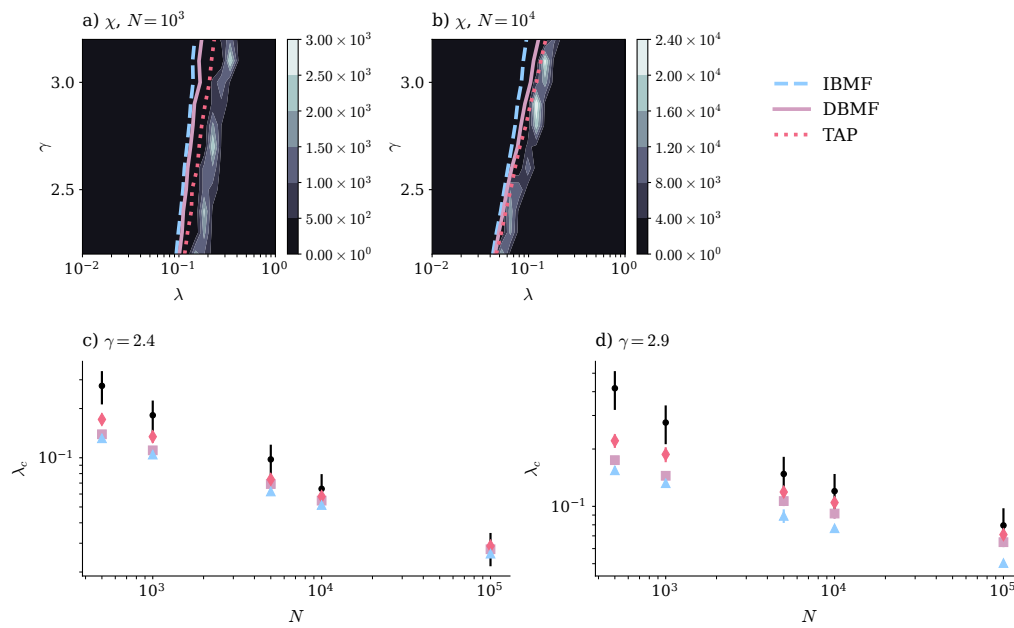


FIG. 2.11: Epidemic threshold for the SIRS model on scale free networks with $\gamma = 0.05$
 a) shows the susceptibility, Eq. (2.51), for scale free networks of size 10^3 , Lines indicate the predictions for λ_c . b) Same as a) but for larger system sizes $N = 10^4$. c) Black dots show λ_c from simulations over system size for $\gamma = 2.4$. Blue triangles, purple squares and pink diamonds show the predictions from individual-based mean-field theory, degree-based mean-field theory, and TAP respectively. d) Same as c) but for $\gamma = 2.9$. All data in panels a)-b) are averaged over ten different realizations of the network architecture and ten runs of the stochastic process per realization.

Inference of higher-order interactions

This chapter, parts of Chap. 4 and appendices H, I, J, K, and L are based on the following publication:

Merger, C., René, A., Fischer, K., Bouss, P., Nestler, S., Dahmen, D., Honerkamp, C., Helias, M., 2023. Learning Interacting Theories from Data. <https://doi.org/10.48550/arXiv.2304.00599>

Author contributions

Under the supervision of Moritz Helias and Carsten Honerkamp, the author worked on all parts of the above publication. The author contributed to the general formalism, performed the corresponding numerical experiments and wrote the original draft of the manuscript. All authors contributed to finalizing the manuscript. The code basis for the tensor computations was developed jointly with Alexandre René. The code for sampling from random actions was developed by Alexandre René. The code basis for the training of the networks was developed jointly with Alexandre René, Peter Bouss and Sandra Nestler.

3.1 Introduction to inference problems

In the previous chapter, we studied the effect a specific, given structure has on a system. Now, we ask the reverse question: Given (sufficiently many) observations of a system, can we infer the underlying structure?

Inference problems are typically difficult to study. To understand why the difficulty in inference problems arises, we will suppose that we are given a set \mathcal{D} of observations of random variables $x \in \mathbb{R}^d$, drawn i.i.d. from an unknown probability

distribution p_x , which defines our system. Here, the x_i are the degrees of freedom of the system which we want to describe. They could stem from any system, for example, they could encode neuron activity, local magnetizations, amplitudes of sound recordings, particles, or pixels of an image. We would now like to approximate p_x by a density p_θ , which depends on a (potentially very high-dimensional) parameter vector θ . Typically, we will use a density of the form

$$p_\theta(x) = \exp(S_\theta(x) - \ln Z_\theta). \quad (3.1)$$

with the partition function

$$Z_\theta = \int dx \exp(S_\theta(x)). \quad (3.2)$$

If x is discrete, we must replace the integral in Eq. (3.2) by a sum. We will call S_θ the action of the system, in the canonical ensemble it is related to the Hamiltonian via $-\beta H = S_\theta$. The terms in S_θ will constitute the interactions of the degrees of freedom of our system. Inferring these terms gives us insights on how the systems constituents interact with each other.

We would now like to find the parameter θ^* which maximizes the log-likelihood of observing the data. We can equivalently minimize the negative log-likelihood

$$\mathcal{L}(\theta) = - \sum_{x \in \mathcal{D}} \ln p_\theta(x). \quad (3.3)$$

We formulate the problem as a minimization of a cost function as it is usual practice in machine-learning literature. Supposing that θ is continuous and p_θ differentiable, we could try to perform gradient descent on the log-likelihood. In gradient descent (GD), one iteratively minimizes a function by computing the direction of steepest descent $-\nabla_\theta \mathcal{L}(\theta)$ locally, and then moving into this direction, changing the parameters by a small amount $\Delta\theta \sim -\nabla_\theta \mathcal{L}(\theta)$. However, the gradient of \mathcal{L} , which is

$$\nabla_\theta \mathcal{L}(\theta) = \sum_{x \in \mathcal{D}} \nabla_\theta \ln Z_\theta - \nabla_\theta S_\theta(x), \quad (3.4)$$

contains the term $\nabla_\theta \ln Z_\theta$. This term is typically very hard to compute, as it involves an integral or sum over the space of all possible states, see Eq. (3.2). In the next section, we will explicitly treat the special case of pairwise interactions as an example. We then formulate the general setting of our inference problem in Sec. 3.1.2, and show how it can be solved using invertible neural networks in Sec. 3.2.

3.1.1 Example: Pairwise interactions

We will now suppose that our density is a second order polynomial in x , parametrized by

$$\ln p_\theta = \left(A^{(1)}\right)^T x + x^T A^{(2)} x - \ln Z_\theta \quad (3.5)$$

This is a model of pairwise interactions because the degrees of freedom are only coupled to each other in a pairwise fashion via $A^{(2)}$, plus a coupling to a constant field $A^{(1)}$.

Minimizing the negative log-likelihood using gradient descent requires the computation of

$$\begin{aligned}\nabla_{A_i^{(1)}} \mathcal{L}(A^{(1)}, A^{(2)}) &\propto \langle x_i \rangle_{A^{(1)}, A^{(2)}} - \langle x_i \rangle_{\mathcal{D}}, \\ \nabla_{A_{ij}^{(2)}} \mathcal{L}(A^{(1)}, A^{(2)}) &\propto \langle x_i x_j \rangle_{A^{(1)}, A^{(2)}} - \langle x_i x_j \rangle_{\mathcal{D}},\end{aligned}\tag{3.6}$$

where $\langle \cdot \rangle_{\mathcal{D}}$ is the empirical average with respect to \mathcal{D} , and $\langle \cdot \rangle_{A^{(1)}, A^{(2)}}$ is the average with respect to p_{θ} .

For continuous variables $x_i \in \mathbb{R}$, Eq. (3.5) is simply a Gaussian theory, with covariance $\Sigma = -\frac{1}{2}(A^{(2)})^{-1}$ and mean $\mu = \frac{1}{2}(A^{(2)})^{-1}A^{(1)}$. Hence in this case, we know how to express the averages $\langle x \rangle_{A^{(1)}, A^{(2)}}$, $\langle x^2 \rangle_{A^{(1)}, A^{(2)}}$ exactly. Setting $\nabla_{\theta} \mathcal{L}$ to zero and solving for the parameters is therefore possible.

For binary variables $x_i \in \{-1, 1\}^d$, on the other hand, we have the same setting as in Sec. 2.2.1, with the notable difference that now the adjacency matrix A and the external field h are unknown. In this case, the inference problem is called the inverse Ising model. For the Ising model, however, averages of the type $\langle \cdot \rangle_{A^{(1)}, A^{(2)}}$ are typically hard to calculate (see Sec. 2.2.1) and many evaluations of these averages are required by gradient descent. Nevertheless, a multitude of techniques to solve the inverse Ising model have emerged, an overview of which is provided in Sec. 3.4.

In this work, however, we will not treat the inverse Ising model. Rather, we will focus on continuous random variables, but go beyond pairwise interactions. We specify our inference problem in the following section.

3.1.2 Polynomial actions for continuous variables

From here on, our degrees of freedom will be continuous random variables $x \in \mathbb{R}^d$, and we will infer a polynomial action S_{θ} . To express this action S_{θ} , we will first define a shorthand notation to express multivariate polynomials of arbitrary degrees.

Notation. In the following, we use the notation $u^{\otimes k}$ for the outer product of k instances of a tensor u

$$u^{\otimes k} = \underbrace{u \otimes u \otimes \cdots \otimes u}_{k \text{ times}},\tag{3.7}$$

and $A^{(k)} \cdot (u)^{\otimes l}$ for $l \leq k$ to denote the contraction of the first l indices of a rank k tensor $A^{(k)}$ with the first index of each tensor u :

$$\left(A^{(k)} \cdot (u)^{\otimes l} \right)_{\beta_1, \dots, \beta_l, i_{l+1}, \dots, i_k} = \sum_{i_1, \dots, i_l} A_{i_1, \dots, i_k}^{(k)} u_{i_1 \beta_1} \cdots u_{i_l \beta_l}, \quad (3.8)$$

where β_1, \dots, β_l are multi-indices whose rank depends on the rank of u . In the special case that u is a vector, the indices β_1, \dots, β_l vanish from the expression. If additionally $k = l$, the result is a scalar. We symmetrize a tensor by averaging over the set $\mathcal{P}(\alpha)$ of all permutations of the multiindex α :

$$\left(\text{sym } A^{(k)} \right)_\alpha = |\mathcal{P}(\alpha)|^{-1} \sum_{\pi \in \mathcal{P}(\alpha)} A_\pi^{(k)}; \quad (3.9)$$

this operation does not change the result of polynomial contractions: $A^{(k)} \cdot (x_l)^{\otimes k} = (\text{sym } A^{(k)}) \cdot (x_l)^{\otimes k}$.

Interaction coefficients. We now wish to infer an action of the type

$$\begin{aligned} S_\theta(x) = & A^{(0)} + A^{(1)} \cdot x \\ & + A^{(2)} \cdot (x)^{\otimes 2} + A^{(3)} \cdot (x)^{\otimes 3} + \dots, \end{aligned} \quad (3.10)$$

with $-\ln Z_\theta = A^{(0)}$. The action S_θ is fully defined via its coefficients, which are hence the central objects of the theory we want to learn. We will say that polynomial terms of type $A_{i_1, \dots, i_k}^{(k)} x_{i_1} \cdots x_{i_k}$ for i_1, \dots, i_k all unequal constitute k -point interactions between degrees of freedom, because they encode a coordination between the variables x_{i_1}, \dots, x_{i_k} which cannot be assimilated through interactions of lower order.

For actions of type Eq. (3.10), in general, no exact mapping between the coefficients $\{A^{(k)}\}_k$ and the averages $\{\langle x^{\otimes k} \rangle\}_k$ exists. Thus a direct optimization of Eq. (3.3) via gradient descent is infeasible. However, generative neural networks can be optimized to infer such actions. The underlying trick is to parametrize the action implicitly, not explicitly via the coefficients. In the next section, we explain both how the networks are optimized, and the mechanism by which the interaction coefficients $A^{(k)}$ can be extracted from the implicit parametrization.

3.2 Learning polynomial actions with invertible neural networks

3.2.1 Training

Invertible neural networks [13–15] parametrize an invertible function $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$. They are trained to map from a set of data points x to a latent space $f_\theta(x) = z$ such

that the variables z follow a Gaussian distribution, $z \sim \mathcal{N}(0, \mathbb{1})$, hence their density is

$$p_Z(z) = \frac{e^{-\frac{z^T z}{2}}}{\sqrt{2\pi}^d}. \quad (3.11)$$

The learned distribution in data space and p_Z are then related via the change of variables formula

$$p_\theta(x) = p_Z(f_\theta(x)) |\det J_{f_\theta}(x)|, \quad (3.12)$$

with J_{f_θ} the Jacobian of f_θ . In this way, the distribution on data space is parametrized implicitly via the mapping f_θ . The network is trained to match the p_θ to the unknown distribution p_x of the data as closely as possible. Hence, the training objective is to minimize the negative log-likelihood

$$\begin{aligned} \mathcal{L} &= - \sum_{x \in \mathcal{D}} \ln p_\theta(x) \\ &= - \sum_{x \in \mathcal{D}} (\ln p_Z(f_\theta(x)) + \ln |\det J_{f_\theta}(x)|). \end{aligned} \quad (3.13)$$

The gradients of this function can now be computed, provided that f_θ and J_θ are differentiable with respect to θ . The differentiation of these functions can be done by most machine learning libraries via back-propagation, i.e. an efficient implementation of the chain rule. Thus we avoid having to compute averages as in Eq. (3.6), at the cost of specifying the action implicitly, via f_θ . In practice, we do not use gradient descent, but a variant of gradient descent called stochastic gradient descent (SGD). Within stochastic gradient descent, for each update on θ , we do not perform pure gradient descent $\Delta\theta = -\eta \nabla_\theta \mathcal{L}$, rather we randomly choose a subset $\mathcal{D}_t \subset \mathcal{D}$, on which the update is performed:

$$\Delta\theta = -\eta \nabla_\theta \sum_{x \in \mathcal{D}_t} S_\theta(x), \quad (3.14)$$

where we used $\ln p_\theta(x) = S_\theta(x)$. Here $\eta \in \mathbb{R}, \eta > 0$ is called the learning rate, and is typically in the range 10^{-3} to 10^{-6} , depending on the dimensionality of the problem. We will discuss the consequences of training with SGD for the learning of coefficients in Sec. 3.2.5.

Conventionally, invertible neural networks are used to generate new samples resembling those in the data set [13–15], such as images. This is done by sampling from the simple distribution in latent space, p_Z , and then performing the inverse mapping f_θ^{-1} on these samples. The network thus transforms a vector of uncorrelated random variables, z , into a vector x whose entries are no longer statistically independent. The nature of the co-dependence of the entries in x is then encoded in f_θ^{-1} in a non-trivial way.

Here, we cast these co-dependencies of the degrees of freedom x_i into an interpretable form, using the language of interactions, which is central to physics. Specifically, we extract the action coefficients in Eq. (3.10) from the parameters θ of the

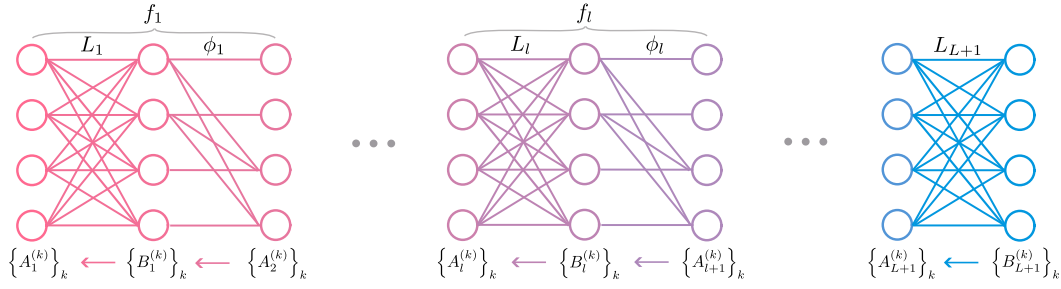


FIG. 3.1: Architecture of invertible neural networks, composed out of alternating affine linear and nonlinear transforms. The bottom row specifies the order by which the action coefficients are computed.

trained networks via an iteration over the network layers. Consequently, we obtain not only a method to generate new samples, but also cast the distribution p_θ of these samples into an interpretable form. In the following, we will define the network architecture in conjunction with the corresponding coefficient transforms.

3.2.2 Architecture & layer transforms

Invertible neural networks are more constrained than general feed-forward neural networks: first, they must be invertible, and this inverse must be computed in a numerically efficient way. Second, the determinant of their Jacobian J_{f_θ} must also be computed efficiently, since it must be evaluated many times during training. The authors of [13–15] hence devise an architecture composed out of alternating affine linear L_l and nonlinear mappings ϕ_l ,

$$f_l = \phi_l \circ L_l$$

which each fulfill these requirements. For a network of depth L we stack L of these layers, plus one additional affine linear transform on top of each other

$$f_\theta = L_{L+1} \circ \phi_L \circ L_L \cdots \phi_1 \circ L_1$$

thus, the composition of all linear and nonlinear mappings defines f_θ (see Fig. 3.1 for a visualization). Besides f_θ , $\ln J_{f_\theta}$ also appears in the loss Eq. (3.13). It is sufficient to ensure that $\ln J_{\phi_l}$ and $\ln J_{L_l}$ can be computed efficiently for all l ; these terms are then summed up to compute $\ln J_{f_\theta}$.

Our strategy to compute the action $S_\theta = \ln p_\theta$ is an iteration over layers from the output to the input space, thus backwards through the network. We define $p_{\theta,l}$ to be the probability density of the layer activations x_l . Starting from the known density in latent space, we iterate over the layers and compute the action $S_{X,l} = \ln p_{\theta,l}(x_l)$ of

the previous layer from the action of the later layer

$$S_{X,l}(x_l) = S_{X,l+1}(f_l(x_l)) + \ln |\det J_{f_l}(x_l)|. \quad (3.15)$$

We will design ϕ_l, L_l such that $S_{X,l}$ is always a polynomial, and can hence be written via a set of action coefficients $\{A_l^{(k)}\}_k$. Thus, the transform on the level of the actions

$$S_{X,l} \xleftarrow{f_l} S_{X,l+1}$$

can be equivalently expressed as a transform on a set of coefficients,

$$\{A_l^{(k)}\}_k \xleftarrow{f_l} \{A_{l+1}^{(k)}\}_k. \quad (3.16)$$

We will later choose f_l to be a polynomial with a Jacobian determinant constant in x , $\ln J_{f_l}(x) = C_\theta \forall x$ for all l . Thus, up to an additive constant, Eq. (3.15) corresponds to the composition of two polynomials, namely $S_{X,l+1}$ and f_l . To find the action coefficients $\{A_l^{(k)}\}_k$ of the pre-activations to layer l from $\{A_{l+1}^{(k)}\}_k$, all we must do is to multiply out these polynomials, which amounts to tensor contractions on the level of the coefficients of the polynomials. However, since $x_l, x_{l+1} \in \mathbb{R}^d$, the action coefficients $A^{(k)}$ are tensors, we must take care to contract along the correct indices. We find that this procedure can be simplified without any loss in expressivity by using only symmetric action coefficients $A^{(k)}$. Furthermore, we will use a diagrammatic language to express the contractions, which facilitates the computation of multiplicity factors, and acts as a visual guide in how higher-order interactions may arise out of lower order interactions, see Fig. 3.2.

The transform of the coefficients, Eq. (3.16) further decomposes into two steps, corresponding to the linear and nonlinear mappings in the layer

$$\{A_l^{(k)}\}_k \xleftarrow{L_l} \{B_l^{(k)}\}_k \xleftarrow{\phi_l} \{A_{l+1}^{(k)}\}_k. \quad (3.17)$$

An overview of the order in which the coefficients are computed is shown in Fig. 3.1. We will now define ϕ_l, L_l and compute the corresponding transforms on the action coefficients.

3.2.2.1 Affine linear mapping

We define the affine linear mapping

$$h_l = L_l(x_l) = W_l x_l + b_l$$

with $W_l, b_l \in \theta$. On the level of the distribution, we hence find that L_l encodes a stretch and rotation of the space, plus a shift via b_l . Suppose now that we have computed

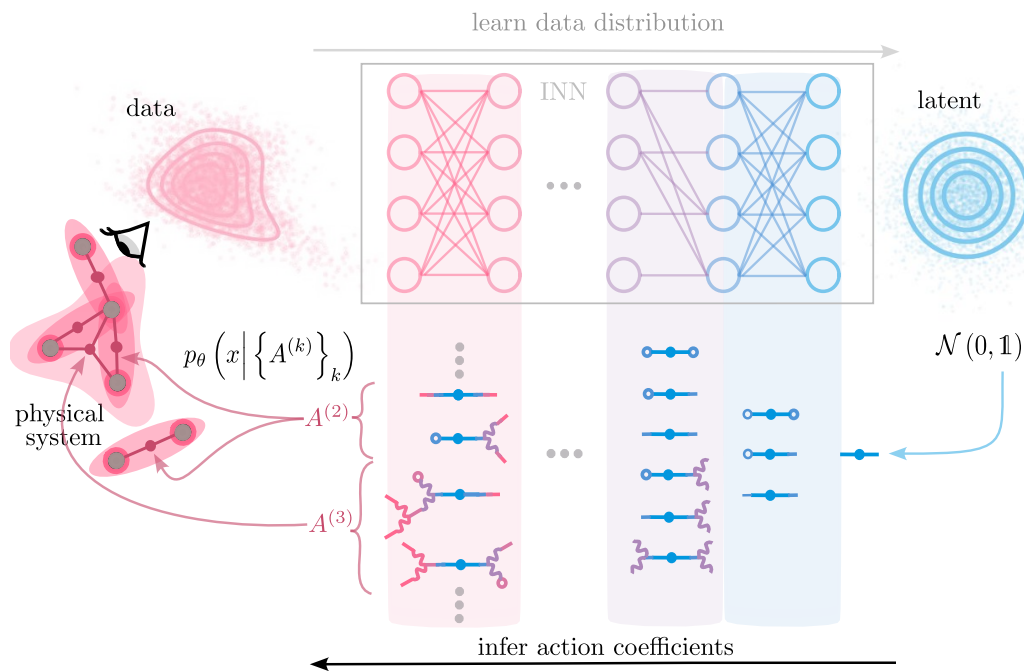


FIG. 3.2: Learning actions from data. We observe a physical system of interacting degrees of freedom (gray dots), whose precise interactions are unknown (shaded areas). We train a neural network on measurements of the system. The network learns in unsupervised fashion an estimate of the distribution of training data. We extract the action from the network parameters layer by layer, using a diagrammatic language. The final action coefficients $A^{(k)}$ represent the learned interactions (pink nodes).

the action coefficients $\{B_l^{(k)}\}_k$ of h_l , thus we have

$$S_{H,l}(h_l) = B_l^{(0)} + B_l^{(1)} \cdot h_l \\ + B_l^{(2)} \cdot (h_l)^{\otimes 2} + B_l^{(3)} \cdot (h_l)^{\otimes 3} + \dots$$

The action of the preactivations x_l then follows as

$$S_{X,l}(x_l) = S_{H,l}(L_l(x_l)) + \ln |\det W_l| \\ = B_l^{(0)} + \ln |\det W_l| + B_l^{(1)} \cdot (W_l x_l + b_l) \\ + B_l^{(2)} \cdot (W_l x_l + b_l)^{\otimes 2} + B_l^{(3)} \cdot (W_l x_l + b_l)^{\otimes 3} + \dots, \quad (3.18)$$

which we multiply out to obtain the new coefficients. Note that the degree K_l of $S_{X,l}$ is equal to the degree of $S_{H,l}$. We get the new normalization $A_l^{(0)}$ by adding the $\ln |\det W_l|$ to $B_l^{(0)}$, plus the contraction of all coefficients with the bias b_l along all legs,

$$A_l^{(0)} = B_l^{(0)} + \ln |\det W_l| + \sum_{k=1}^{K_l} B_l^{(k)} \cdot (b_l)^{\otimes k}$$

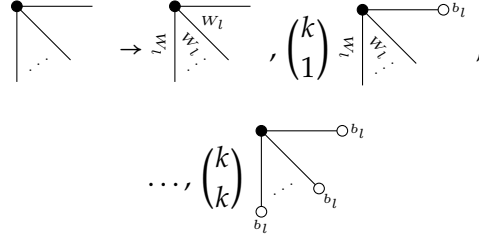
The coefficients $A_l^{(k)}$ for $k \geq 1$ are then computed from all coefficients $B_l^{(k')}$ of equal or higher order $k' \geq k$ by contracting with $k' - k$ factors of b_l , and subsequently contracting all remaining indices with the first index of W_l . Since the coefficients $\{B_l^{(k)}\}_k$ are symmetric tensors, it does not matter which indices we contract the biases b_l and matrices W_l with, but we must sum over the $\binom{k'}{k'-k}$ possibilities of doing so. This yields the coefficient transforms for $1 \leq k \leq K_l$

$$A_l^{(k)} = \left(\sum_{k' \geq k} \binom{k'}{k'-k} B_l^{(k')} \cdot (b_l)^{\otimes (k'-k)} \right) \cdot (W_l)^{\otimes k}. \quad (3.19)$$

We will now introduce the diagrammatic notation which allows us to do these calculations graphically. Each coefficient of order k will be represented by a vertex with k legs; the legs represent the indices over which the tensor contractions take place. A contraction with a vector b_l reduces the order of the coefficient by one, we will hence represent it by contracting the vertex with an empty circle. A contraction with a matrix W_l along one index yields a coefficient of the same order, thus we symbolize it by an elongation of the corresponding vertex leg, which we decorate with W_l to symbolize that the contraction has taken place.

The diagrammatic rule which encodes Eq. (3.19) is: attach empty circles to the legs of all vertices in all possible ways, then elongate the remaining legs with W_l . Then sum over all diagrams with k legs to get the coefficient of order k . A vertex with k

legs therefore produces the following diagrams:



The binomial factors are the same as in Eq. (3.19), diagrammatically they arise from the different ways in which the empty circles can be attached to the legs of the vertex. We show an example of this procedure in Sec. 3.2.3.

3.2.2.2 Nonlinear mapping

We now perform the same steps as in Sec. 3.2.2.1, but for the nonlinear mapping. We first state the method by which the authors of [13] introduce flexible nonlinear mappings which are nevertheless trivial to invert. We split the activation vectors and activation functions into two halves¹ denoting them by

$$h_l = \begin{pmatrix} h_l^1 \\ h_l^2 \end{pmatrix} \quad \text{and} \quad \phi_l = \begin{pmatrix} \phi_l^1 \\ \phi_l^2 \end{pmatrix}$$

We then perform a general nonlinear transform $\tilde{\phi}_l : \mathbb{R}^{\lfloor \frac{d}{2} \rfloor} \rightarrow \mathbb{R}^{\lfloor \frac{d}{2} \rfloor}$ only on the first half h_l^1 and add this onto the second half

$$x_{l+1} = \phi_l(h_l) = \begin{pmatrix} h_l^1 \\ h_l^2 + \tilde{\phi}_l(h_l^1) \end{pmatrix} \quad (3.20)$$

This function is trivially invertible,

$$\phi_l^{-1}(x_{l+1}) = \begin{pmatrix} x_{l+1}^1 \\ x_{l+1}^2 - \tilde{\phi}_l(x_{l+1}^1) \end{pmatrix}$$

no matter, what $\tilde{\phi}_l$ is. In particular, $\tilde{\phi}_l$ could be parameterized by another neural network. Furthermore, observe that $\det J_{\phi_l}(h_l) = 1 \forall h_l$. The latter property follows directly from the fact that J_{ϕ_l} is a triangular matrix with ones on the diagonal.

Equation (3.20), inserted into the change of variables formula, then gives us the following action transform

$$S_{H,l}(h_l) = S_{X,l+1}(\phi_l(x_l)) \quad (3.21)$$

¹If the dimension d is uneven, we take the first $\lfloor \frac{d}{2} \rfloor$ entries of h_l to be in h_l^1 .

Thus it becomes clear that $S_{H,l}$ is a polynomial if both $S_{X,l+1}$ and $\tilde{\phi}_l$ are polynomials. Since our objective is to obtain a polynomial action S_θ , we choose $\tilde{\phi}_l$ to be a polynomial. Specifically, $\tilde{\phi}_l$ here is a quadratic nonlinearity

$$\tilde{\phi}_l(h_l^1) = \tilde{\chi}_l \cdot (h_l^1)^{\otimes 2},$$

where $\tilde{\chi}_l \in \mathbb{R}^{\lfloor \frac{d}{2} \rfloor \times \lfloor \frac{d}{2} \rfloor \times \lfloor \frac{d}{2} \rfloor}$ is a third-order tensor whose coefficients are trained, $\tilde{\chi}_l \in \theta$. This choice of $\tilde{\phi}_l$ is the most elementary non-linearity, which, through the composition of multiple layers, gives rise to higher-order polynomial functions. Consequently, the network mapping f_θ is a polynomial.

To simplify the calculations, we now define a third order tensor χ_l , such that

$$\phi_l(h_l) = h_l + \chi_l \cdot (h_l)^{\otimes 2}.$$

The coefficient transforms then follow from multiplying out Eq. (3.21),

$$\begin{aligned} S_{H,l}(h_l) &= A_{l+1}^{(0)} + A_{l+1}^{(1)} \cdot (h_l + \chi_l \cdot (h_l)^{\otimes 2}) + A^{(2)} \cdot (h_l + \chi_l \cdot (h_l)^{\otimes 2})^{\otimes 2} \\ &\quad + A^{(3)} \cdot (h_l + \chi_l \cdot (h_l)^{\otimes 2})^{\otimes 3} + \dots \end{aligned} \quad (3.22)$$

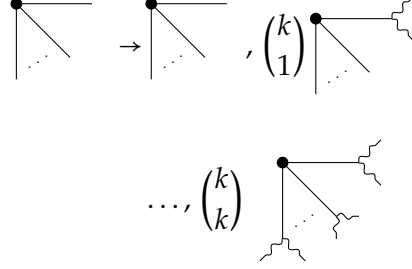
In this case, the degree of the polynomial increases, $K_l = 2K_{l+1}$. We now find that lower order coefficients, through the contraction with χ_l , contribute to higher-order coefficients. Explicitly, we have

$$\begin{aligned} B_l^{(k)} &\stackrel{k \leq 1}{=} A_{l+1}^{(k)} \\ B_l^{(k)} &\stackrel{k > 1}{=} \text{sym} \sum_{k'=0}^k \binom{k-k'}{k'} A_{l+1}^{(k-k')} \cdot (\chi_l)^{\otimes k'} \end{aligned} \quad (3.23)$$

with the binomial factor arising due to the $\binom{k-k'}{k'}$ ways of contracting the tensor $A_{l+1}^{(k-k')}$ with k' factors of χ_l . After the contraction, the tensors are not necessarily symmetric, since indices originating from χ_l are not equivalent to remaining indices originating from $A_{l+1}^{(k-k')}$. We perform a symmetrization operation in Eq. (3.23) to obtain symmetric coefficients $B_l^{(k)}$.

The contraction with χ_l increases the rank of a tensor. Therefore, the corresponding diagram must have more legs than the original diagram. This is why contractions with χ_l are here represented by a splitting of legs. To keep track of which legs originate from χ_l , and which from the original coefficients, we use wavy lines for the split legs. The number of legs of each tensor increases by one for each contraction with χ_l , in the same way as the number of indices of the underlying tensor increases by one. The diagrammatic rule to compute the coefficient transforms is thus: split the legs of all vertices in all possible ways, then sum over all diagrams with k legs to get the k -th action coefficient. A vertex with k legs therefore produces the following

diagrams:



Again, the combinatorial factors are the same as those in Eq. (3.23), they here arise as the number of possible choices of which legs to split. We will now exemplify these diagrammatic rules by computing the combined transform of the coefficients via a linear and a non-linear mapping. This calculation also demonstrates how higher-order interactions emerge, as here a Gaussian theory is transformed into a fourth-order interacting one.

3.2.3 Example: From Gaussian to fourth order action

Our starting point is the latent density p_Z , whose non-zero action coefficients are

$$\begin{aligned} B_{L+1}^{(0)} &= \frac{d}{2} \ln(2\pi) \\ B_{L+1}^{(2)} &= -\frac{\mathbb{1}}{2}. \end{aligned} \quad (3.24)$$

We now use that $z = W_{L+1}x_L + b_{L+1}$ and Eq. (3.19), (see 3.1 for the order of the layer mappings and coefficient transforms) to compute the coefficients of the activations x_L . We get

$$\begin{aligned} A_{L+1}^{(0)} &= B_{L+1}^{(0)} - \frac{b_{L+1}^T b_{L+1}}{2} + \ln |\det W_L| \\ A_{L+1}^{(1)} &= -W_{L+1}^T b_{L+1} \\ A_{L+1}^{(2)} &= -\frac{W_{L+1}^T W_{L+1}}{2} \end{aligned} \quad (3.25)$$

This is just another Gaussian theory, albeit with a different covariance and mean. Diagrammatically, the same equations read:

$$\begin{aligned} A_{L+1}^{(0)} &= \bullet + \overset{b_{L+1}}{\circ} \text{---} \bullet \text{---} \overset{b_{L+1}}{\circ} + \ln |\det W_{L+1}| \\ A_{L+1}^{(1)} &= 2 \overset{b_{L+1}}{\circ} \text{---} \bullet \text{---} \overset{W_{L+1}}{\bullet} \\ A_{L+1}^{(2)} &= \overset{W_{L+1}}{\bullet} \text{---} \bullet \text{---} \overset{W_{L+1}}{\bullet} \end{aligned}$$

Here, the two-point vertex is $B_{L+1}^{(2)}$, whose legs are either elongated by factors W_{L+1} or truncate on b_{L+1} . If our network was purely linear, $L = 0$, then the affine linear mapping which minimizes 3.13 is just the PCA where all components are retained, hence we would get a purely Gaussian approximation of the data distribution.

We now add a nonlinear mapping ϕ_L and compute the distribution of the activations h_L prior to this mapping. Inserting our result Eq. (3.25) into 3.23, we find that the change in the first two coefficients is

$$\begin{aligned} B_L^{(k)} &\stackrel{k \leq 1}{=} A_{L+1}^{(k)} \\ B_L^{(2)} &= A_{L+1}^{(2)} + A_{L+1}^{(1)} \cdot \chi_L \\ &= B_{L+1}^{(2)} \cdot (W_{L+1})^{\otimes 2} + 2 \operatorname{sym} \left(\left[B_{L+1}^{(2)} \cdot b_{L+1} \right] \cdot W_{L+1} \right) \cdot \chi_L \end{aligned} \quad (3.26)$$

the zeroth and first coefficients remain unchanged, while the second order coefficient obtains a contribution from the first order coefficient via the contraction with χ_L . Diagrammatically, the transform for $B_L^{(2)}$ is

$$B_L^{(2)} = \overline{\text{---} \bullet \text{---}}_{W_{L+1} \quad W_{L+1}} + 2 \text{---} \circ \text{---} \bullet \text{---} \begin{array}{l} \diagup \\ \diagdown \end{array},$$

where the contraction of $A_{L+1}^{(1)}$ with χ_L is symbolized in a leg split. Besides the change in the second coefficient, a third-order and fourth-order interaction coefficient appears from contractions of $A_{L+1}^{(2)}$ with χ_L

$$\begin{aligned} B_L^{(3)} &= \binom{2}{1} \operatorname{sym} \left[B_{L+1}^{(3)} \cdot (W_{L+1})^{\otimes 2} \right] \cdot \chi_L, \\ B_L^{(4)} &= \operatorname{sym} \left[B_{L+1}^{(3)} \cdot (W_{L+1})^{\otimes 2} \right] \cdot (\chi_L)^{\otimes 2}. \end{aligned} \quad (3.27)$$

Diagrammatically, the same equations for $B_L^{(3)}$ and $B_L^{(4)}$ read

$$\begin{aligned} B_L^{(3)} &= \binom{2}{1} \overline{\text{---} \bullet \text{---}}_{W_{L+1} \quad W_{L+1}} \begin{array}{l} \diagup \\ \diagdown \end{array}, \\ B_L^{(4)} &= \begin{array}{l} \diagup \\ \diagdown \end{array} \overline{\text{---} \bullet \text{---}}_{W_{L+1} \quad W_{L+1}} \begin{array}{l} \diagup \\ \diagdown \end{array}. \end{aligned}$$

In this way, we have obtained a fourth order interacting theory from a pairwise theory. This corresponds to a network with one linear and one nonlinear mapping. For deeper networks, we must iterate the procedure; the next linear mapping corresponds to attaching empty circles to the existing diagrams in all possible ways, then elongating all remaining legs and summing up the result. Another nonlinear mapping then may split all the legs of the vertices in all possible ways. Thus, it can happen, that a diagram with three legs contributes to the second order coefficient

after attaching another empty circle to one of its legs, but a splitting of one of the two remaining legs transforms it into a contribution to the third-order coefficient again. Thus through the combination of multiple layer transforms, the network can move contributions of higher-order coefficients to lower order coefficients and vice versa, always building on the existing coefficients. Hence we obtain a hierarchical mixing of contributions across layers. We illustrate this procedure of building up an interacting theory hierarchically in Fig. 3.2.

3.2.4 Truncation in the interaction order

Through the composition of the action with a quadratic polynomial, Eq. (3.22), the degree of the resulting action is increased. Explicitly, we have found that $K_l = 2K_{l+1}$, i.e. that the order of the interactions doubles at each layer. However, in practice, we will find that the entries in the tensors χ_l , which build up the interactions, are typically small. Thus the entries coefficients of order $k \geq 2$, which must necessarily contain coefficients $k - 2$ factors of χ_l (potentially stemming from different layers), are expected to decrease in magnitude.

We therefore truncate the computation of the coefficients at order $k = 4$, which is equivalent to a truncation at order $\mathcal{O}(\chi_l \chi_{l'})$ (the factors of potentially stemming from different layers l, l').

Diagrammatically, we can identify the order of a contribution by the number of leg splits; since every such split comes with a factor χ_l , a truncation to second order thus implies that we must only compute diagrams with at most two leg splits.

We further write the χ_l tensors in a decomposed form, namely as outer products between vectors. This format is particularly efficient for computations. Furthermore, since higher-order coefficients, namely $A_l^{(3)}$ and $A_l^{(4)}$, emerge only via contractions of χ_l tensors with lower-order tensors, this decomposition translates to the higher-order tensors as well, which allows for an efficient implementation of the coefficient transforms. We detail the decomposition of χ_l and its consequences for the computation of coefficient transforms in App. (H).

3.2.5 Learning rules in coefficient space

Before we move to a demonstration of the method, we examine the training procedure in coefficient space. In Sec. 3.1, we saw that learning the coefficients $\{A^{(k)}\}_k$ directly would require computing the moments of the distribution depending on the coefficients at every step. In contrast, the update step on the network parameters Eq. (3.14) does not require the expensive computation of these moments. On the other hand, since the action S_θ is fully parametrized by the coefficients, we are free to rewrite the derivative in Eq. (3.14)

$$\begin{aligned}\Delta\theta &= -\eta \sum_{k \geq 1} \sum_{\alpha_k} \frac{d \langle S_\theta(x) \rangle_{\mathcal{D}_t}}{d A_{\alpha_k}^{(k)}} \nabla_\theta A_{\alpha_k}^{(k)} \\ &= -\eta \sum_{k \geq 1} \sum_{\alpha_k} \left(\left\langle \left(x^{\otimes k} \right)_{\alpha_k} \right\rangle_{\mathcal{D}_t} - \left\langle \left(x^{\otimes k} \right)_{\alpha_k} \right\rangle_A \right) \nabla_\theta A_{\alpha_k}^{(k)}\end{aligned}$$

where $\langle \cdot \rangle_{\mathcal{D}_t}$ is the empirical average over all samples in the batch \mathcal{D}_t . Each update in the network parameters θ also changes the coefficients. We now compute the change in the coefficients to first order in $\Delta\theta$,

$$\begin{aligned}\Delta A_\alpha^{(k)} &= -\eta \left(\left\langle \left(x^{\otimes k} \right)_\alpha \right\rangle_{\mathcal{D}_t} - \left\langle \left(x^{\otimes k} \right)_\alpha \right\rangle_A \right) \left(\nabla_\theta A_\alpha^{(k)} \right)^\top \nabla_\theta A_\alpha^{(k)} \\ &\quad - \eta \sum_{l \geq 1} \sum_{\alpha_l \neq \alpha} \left(\left\langle \left(x^{\otimes l} \right)_{\alpha_l} \right\rangle_{\mathcal{D}_t} - \left\langle \left(x^{\otimes l} \right)_{\alpha_l} \right\rangle_A \right) \left(\nabla_\theta A_\alpha^{(k)} \right)^\top \nabla_\theta A_{\alpha_l}^{(l)} + \mathcal{O}(\Delta\theta^2).\end{aligned}\quad (3.28)$$

The first term is directly proportional to the gradient $\nabla_{A_\alpha^{(k)}} \mathcal{L}$. The second term couples the coefficients through their mutual dependence on the network parameters. Apart from the factors $\nabla_\theta A^{(k)}$, this equation resembles the direct optimization of the coefficients, see e.g. Eq. (3.6). We can hence understand the optimization procedure to perform an inference of the action coefficients, while constrained to a subspace which is allowed by the implicit parametrization. This implicit parametrization also ensures that the action stays normalized at each training stage.

To make the noise which stochastic gradient descent introduces in the training process visible, we further split the averages $\langle (x^{\otimes k}) \rangle_{\mathcal{D}_t}$ into a deterministic part and a noise term $\zeta^{(k)}$. The latter comes from the subsampling of training batches $\mathcal{D}_t \subset \mathcal{D}$

$$\left\langle \left(x^{\otimes k} \right) \right\rangle_{\mathcal{D}_t} = \zeta^{(k)} + \left\langle \left(x^{\otimes k} \right) \right\rangle_{\mathcal{D}} \quad (3.29)$$

This then yields a stochastic update equation for a coefficients.

Fixpoints. In the average over all training batches, the noise vanishes $\langle \zeta^{(k)} \rangle = 0$. If the right hand side of Eq. (3.28) vanishes in the same average, then the average update on the coefficients is zero, thus training has converged. This condition is met if either the moments of the learned distribution match the moments computed on the training set, or alternatively when the sums vanish due to the limited flexibility of the network, in which case the terms $\nabla_\theta A_{\alpha_l}^{(l)}$ either vanish or cancel each other with their respective prefactors.

In either case, the learning of the coefficients is biased if the averages over the training set deviate from the average over the true distribution

$$\left\langle \left(x^{\otimes k} \right) \right\rangle_{\mathcal{D}} \neq \left\langle \left(x^{\otimes k} \right) \right\rangle_T,$$

where $\langle \cdot \rangle_T$ is the average over the (unknown) true distribution of the data.

Sensitivity to noise. The variance of the noise term $\zeta^{(k)}$ is

$$\left\langle \left(\zeta^{(k)} \right)^{\otimes 2} \right\rangle - \left\langle \zeta^{(k)} \right\rangle^{\otimes 2} = |\mathcal{D}_t|^{-1} \left(\left\langle x^{\otimes 2k} \right\rangle_{\mathcal{D}} - \left\langle x^{\otimes k} \right\rangle_{\mathcal{D}}^{\otimes 2} \right).$$

For many distributions, the estimation of moments in the empirical averages on finite data sets grows more noisy with increasing k (see [51, 52]). Thus, not only do we expect the terms corresponding to higher coefficients in 3.28 to be more noisy, but also their estimation on the whole training set may be more biased.

In the next section, we will illustrate the efficacy of the method outlined here to learn coefficients from data. We will also find qualitatively the consequences of the learning rule which we have discussed here, namely both the bias in learning of higher-order coefficients due to limited data (Sec. 3.3.1) and limited network flexibility (Sec. 3.3.2).

3.3 Experiments

With the diagrammatic rules for the coefficient transforms in hand, we now proceed to test the efficacy of our inference method in four different settings. In all experiments, the models we train must reconstruct the data statistics from samples alone.

We start by a teacher-student setup, where both teacher and student are networks with the architecture outlined in Sec. 3.2.2. The student must learn to mimic the teacher statistic. We then test the procedure on a distribution which is outside the class which can be modeled by the architecture. Thirdly, we learn a fourth order theory of degrees of freedom sitting on lattice sites, which is inspired by a physical model, namely the Ising model. Finally, we infer pixel interactions in the MNIST data set of handwritten digits.

3.3.1 Teacher Student scenario

We construct a teacher network through random initialization of the parameters θ , from which we compute the teacher coefficients $\{T^{(k)}\}_k$ in the way outlined in Sec. 3.2.2. We then draw samples from the teacher model by generating i.i.d samples $\sim \mathcal{N}(0,1)$ and transforming them with the inverse teacher network. We use these samples to train a student network initialized at identity ($W_l = \mathbb{1}$ and $b_l = \chi_l = 0$ for all l). Subsequently, we compute the student coefficients $\{A^{(k)}\}_k$ and compare them to the teacher coefficients in Fig. 3.3. This procedure tests whether the coefficients can

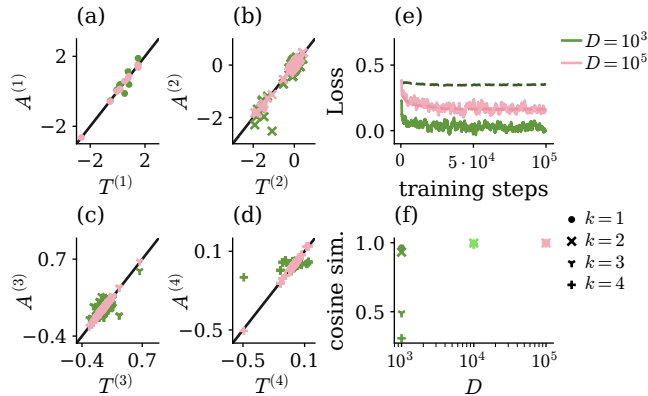


FIG. 3.3: Teacher-student coefficient comparison for varying training set sizes $D = |\mathcal{D}|$. Both teacher and student have depth $L = 1$. **(a–d)** Student coefficients $A^{(k)}$ over teacher coefficients $T^{(k)}$ up to fourth order for $D = 10^3$ in green and $D = 10^5$ in pink. **(e)** Training loss (full lines) and test loss (dashed lines) over training steps. **(f)** Cosine similarity of coefficients over number of training samples.

be reconstructed from samples alone by the student. Sampling from the teacher network ensures that the samples follow exactly the distribution defined by the teacher $\{T^{(k)}\}_k$. Thus, if the successful statistics are accurately learned by the student network, we find that $T^{(k)} = A^{(k)} \forall k$. We repeat the experiment, training several student models on data sets of different sizes D and compute the entries in their coefficients side by side in Fig. 3.3 (a) - (d). Indeed we find that given sufficient samples ($D = 10^4$ or more samples), the student recovers the teacher coefficients accurately. However, if the training data set is too small, the student overfits the the training set. We show that this is the case by computing the loss on a new set of data drawn from the same distribution, we call this quantity the test error. The test error is larger than the training error in Fig. 3.3 (e) for training a data set of size $D = 10^3$. This means that the network has adapted to the specific samples in the training set. Consequently, several entries in the student coefficients deviate from the teacher coefficients in Fig. 3.3 (a) - (d). Upon increasing the size of the training set, the student no longer overfits, where test and training error are approximately equal (see Fig. 3.3 (e)), and the teacher and student coefficients also match.

We quantify the agreement between the coefficients by computing the cosine similarity between the tensors

$$\cos \angle (T^{(k)}, S^{(k)}) = \frac{|\sum_{\alpha} T_{\alpha}^{(k)} S_{\alpha}^{(k)}|}{\sqrt{\sum_{\alpha} (S_{\alpha}^{(k)})^2 \sum_{\alpha} (T_{\alpha}^{(k)})^2}} \quad (3.30)$$

where the sum runs over all independent indices α , excluding duplicate tensors entries which are equal due to the symmetry of the coefficients. We find that the higher-order coefficients in particular show larger deviations when the student net-

work overfits the training data set. This corresponds to the case in which the training set is too small, introducing a bias in the moments, Eq. (3.29), which leads to a biased drift term in the coefficient updates Eq. (3.28). This bias is more severe for the higher-order coefficients, the cosine similarity between $T^{(k)}$, $A^{(k)}$ decreases with k in Fig. 3.3 (f). Learning higher-order coefficients hence requires more data.

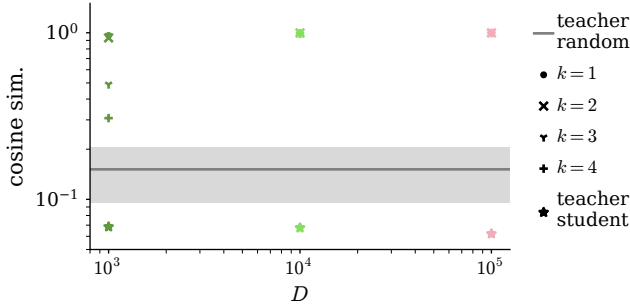


FIG. 3.4: Dissimilarity of parameters. Stars show the cosine similarities of the teacher and student network parameters trained on varying data set sizes D . The average cosine similarity between the teacher and 10^2 randomly generated random networks is marked by the grey line, the shaded area encompasses one standard deviation. The remaining markers display the cosine similarity between the teacher coefficients $T^{(k)}$ and student coefficients $S^{(k)}$.

network will be equal to the ones originating from the network without the additional rotation, but the parameters θ will change significantly. In conclusion, we find that the similarity of the interaction coefficients rather than the network parameters is the appropriate measure to determine whether two networks have learned the same statistics.

In this section, the target distribution was (by construction) inside the class of distributions which the network architecture can parametrize exactly. In the next section, we will test the method on a distribution outside this class.

3.3.2 Out of class distributions

The architecture outlined in Sec. 3.2.2 belongs to the class of *volume preserving flows* [13, 14]; since its Jacobian determinant is a constant in x . This means that the network mapping can bend and rotate the space, but it can only homogeneously (and not locally) stretch the space. Consequently, the number of maxima and minima in the learned and latent distribution will be equal

$$\nabla_x p_\theta(x) = 0 \iff \left| \det J_{f_\theta}(x) \right| = \text{const.} > 0 \iff \nabla_{f_\theta} p_Z(f_\theta(x)) = 0. \quad (3.31)$$

Note that $T^{(k)} = A^{(k)} \forall k$ does not imply that the parameters θ of the teacher and student network are equal. Indeed, we find that the parameters θ of teacher and student do not align in Fig. 3.4. To show that the parameters do not align, we compute the cosine similarity of the flattened parameter vectors θ of teacher and student networks, as well as the teacher and a randomly initialized network of the same architecture. To understand this, consider adding a linear mapping to the last layer, which rotates the latent space. Since the latent space is rotationally invariant, the interaction coefficients of the new

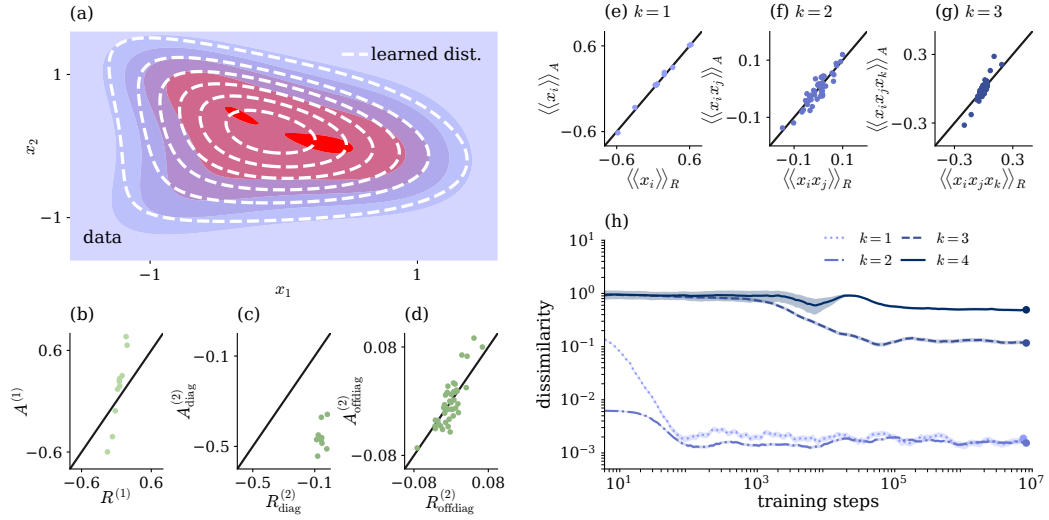


FIG. 3.5: Learning an effective monomodal theory. (a) Two-dimensional example of random density with multiple maxima. White lines are level lines of learned distribution for a five layer network. All other panels show results from a $d = 10$ -dimensional data set. (b–d) Learned over true coefficients for a three layer network on a $d = 10$. We distinguish diagonal tensor entries from off-diagonal ones, where at least two indices differ. (e–g) Learned over true cumulants, computed from samples. Error bars are typically smaller than marker size. (h) Dissimilarity of true and learned cumulants: $1 - \cos \angle (\langle \langle x^{\otimes k} \rangle \rangle_A, \langle \langle x^{\otimes k} \rangle \rangle_R)$ over training steps. We record the cumulants at logarithmically spaced intervals during training. The curves are then smoothed by averaging over ten adjacent recording steps. Shaded areas show the variation due to the estimation of the cumulants from samples. Dots indicate training stage of cumulants shown in (e–g).

This property arises through the additive nature of the non-linearity Eq. (3.20), and is not unique to polynomial activation functions. We now generate a target action S_R , by sampling coefficients $R^{(k)}$ for $1 \leq k \leq 4$, randomly (details on the sampling scheme can be found in App. (I)), and ensure that it has at least two maxima. We then train a volume preserving network with the unimodal latent distribution Eq. (3.11).

Although the action S_R is an unnormalized log-probability², we can then sample a training set \mathcal{D} using Markov chain Monte Carlo (MCMC); for this work we used a Hamiltonian Monte Carlo [53–55] sampler implemented in PyMC3 [56]. (For details see App. (J).)

Figure 3.5 (a) illustrates the learned density versus the data distribution in a two-dimensional example; the learned density approximates the triangular shape of the data distribution, but has only a single maximum, while the data distribution has two. We then repeat the experiment on a ten-dimensional random action and compare learned coefficients $A^{(k)}$ with the true coefficients $R^{(k)}$ in Fig. 3.5 (b)-(d), finding

²We do not compute the constant term in S_R which ensures $\int dx \exp(S_R(x)) = 1$ as it is not needed for the sampling method or the comparison of the coefficients.

that they do not match, as we expect from Eq. (3.31). In this comparison, we distinguish the offdiagonal entries in $A_{\text{offdiag}}^{(2)} = \{A_{ij}^{(2)} | i, j \in \{1, \dots, d\} \text{ s.t. } i \neq j\}$ from the diagonal ones $A_{\text{diag}}^{(2)} = \{A_{ii}^{(2)} | i \in \{1, \dots, d\}\}$. While the offdiagonal entries of $A^{(2)}$ and $R^{(2)}$ typically agree (see Fig. 3.5 (d)), the diagonal entries systematically disagree.

Despite the difference in the coefficients, we find that the learned distribution reproduces the cumulants of the data distribution up to the third order. To see this, we sample from the learned model, and then compute the cumulants $\langle\langle x^{\otimes k} \rangle\rangle$. Cumulants are better suited than moments to compare two different distributions because they contain only independent statistical information. The cumulants of a distribution can be computed from its moments $\langle x^{\otimes k} \rangle$ and vice versa. For example, the first three cumulants are equal to the mean and the centered second and third moment of a distribution. We show the agreement between the first three cumulants in Fig. 3.5 (e)-(g).

The agreement between the cumulants shows that the learned coefficients provide an *effective* non-Gaussian theory S_A . The theory is effective, since the learned coefficients are unequal to the true ones, but compensate each other to yield the correct statistics, here expressed via cumulants. S_A is non-Gaussian, since the third and fourth order coefficients are non-zero, and produce the correct third-order cumulant. In a Gaussian theory, the third cumulant would vanish, it is hence a higher-order statistic of the data set.

Finally, we examine the stage at which the different cumulants are learned. To this end, we compute the cosine similarity between the cumulants of the two distributions in the same way as in Eq. (3.30), replacing coefficient tensors by cumulant tensors and summing only over non-redundant entries. We show the cosine similarities of the cumulants during training in 3.5 (h). We find that the third order cumulant, and partially also the fourth order cumulant, are learned much later in the training process than the first and second cumulant. Hence, higher-order statistics are learned much later than lower order statistics. Similar observations have previously been made for classifiers [18], namely that they pick up on higher-order statistics of data later in training.

In physics, multimodality often appears as a result of symmetry breaking. One such symmetry is the global flipping of all spins in an Ising model [57] with a finite critical temperature (e.g. an Ising model on a square lattice, as opposed to the model we studied in Sec. 2.2). Below the critical temperature, this symmetry is broken. In the action, two modes appear, one for positive and one for negative net magnetization. However, in a physical system, only one mode will be observed, because the probability of a global sign flip approaches zero as the system size increases. Other factors, such as the coupling to the environment or a measurement device will also break the symmetry. We will consider such a setting, the network can nevertheless find an informative *effective* theory, which characterizes the observed monomodal distribution.

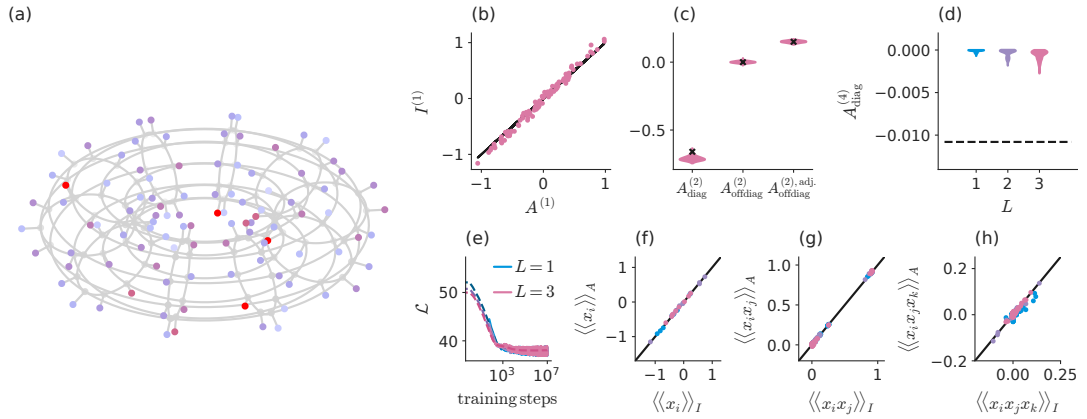


FIG. 3.6: Symmetry broken lattice model for networks of varying depth trained on a $d = 10^2$ dimensional data set with $D = 10^6$ samples. **(a)** Sites of square lattice with periodic boundary conditions distributed on a two-dimensional torus. Colored dots show strength of external field h at connected lattice sites. **(b)** Learned over true first order coefficients for network depth $L = 3$. **(c)** Distribution of learned coefficient entries $A^{(2)}$ compared to target values (black crosses) for network depth $L = 3$. We distinguish self-interaction terms $A_{\text{diag}}^{(2)}$ from off-diagonal entries $A_{\text{offdiag}}^{(2)}$. From the off-diagonal entries $A_{\text{offdiag}}^{(2)}$, we further separate those entries belonging to adjacent lattice sites $A_{\text{offdiag}}^{(2),\text{adj}}$. **(d)** Training loss (solid curves) and test loss (dashed curves). Colors distinguish different network depths L . **(e)** Distribution of learned fourth order self-interactions as function of network depth. The dashed line marks the target value. **(f–h)** Learned over true cumulants of up to third order. Cumulants were computed on a subset of 10 randomly chosen lattice sites. Colors distinguish different network depths L .

3.3.3 Interactions on a lattice with external coupling

We now apply our mechanism to the archetypical ϕ^4 theory. This model can be considered the lattice version of the effective long distance theory of an Ising model in two dimensions [57]. The action

$$\begin{aligned}
 S_I(x) &= -\beta \left[\frac{1}{2} \sum_{i,j} x_i (r_0 \delta_{ij} - \Lambda_{ij}) x_j + \sum_i (h_i x_i + u x_i^4) \right] \\
 &=: I^{(1)} \cdot x + I^{(2)} \cdot (x)^{\otimes 2} + I^{(4)} \cdot (x)^{\otimes 4}, \tag{3.32}
 \end{aligned}$$

defines our model. The coefficients $I^{(k)}$ are the coefficients of the true distribution and we have omitted the normalization. As in Sec. 2.2, β serves as an inverse temperature.

The degrees of freedom x are placed on a 10×10 square grid in two dimensions with periodic boundary conditions. Hence we have $d = 10^2$ degrees of freedom. Let a be the adjacency matrix of the corresponding graph, $a_{ij} = 1$ if i, j are nearest neighbors and $a_{ij} = 0$ otherwise. We use the matrix Laplacian $\Lambda_{ij} = -\delta_{ij} k_i + a_{ij}$ to define the

interaction with the $k_i = 4$ nearest neighbors on the lattice. This interaction favors an alignment of the degrees of freedom (it assigns higher probability to states in which the signs of nearest neighbors i, j are equal $\text{sign}(x_i) = \text{sign}(x_j)$).

We now suppose that these degrees of freedom interact with their environment, which we encode in a linear coupling to an external field h , which we sample randomly from a normal distribution. This external field breaks the symmetry of the system. We illustrate the network topology and external field in Fig. 3.6 (a).

Finally, the second-order term r_0 , and the fourth-order term $-\beta u x_i^4$ constitute self-interactions, which suppress large absolute values of x .

We sample from the action Eq. (3.32) using the same procedure as in Sec. 3.3.2, then train a network on these samples. We then compare the coefficients of the learned model to the true coefficients in Fig. 3.6 (b) - (d). We find that the procedure recovers the correct external biases, as well as the nearest neighbor couplings. The self-coupling terms however, show deviations from the true couplings. Since both $A_{\text{diag}}^{(2)}$ and $A_{\text{diag}}^{(4)}$ control the width of the distributions along the axes, a more negative $A_{\text{diag}}^{(2)}$ may compensate for a smaller absolute value of $A_{\text{diag}}^{(4)}$. In support of this, we find in Fig. 3.6 (f) - (h) that the cumulants of the data distribution and the learned distribution match up to the third order. Thus, the learned coefficients again constitute an *effective* model.

In Sec. 3.2.5 we found that the dynamics of learning may also converge if the network is not sufficiently flexible to fit the data distribution exactly. In 3.3.2, this was the case since the data distribution and the latent distribution did not have the same number of modes. Here, the mismatch may result from the high dimensionality of the problem: The number of independent entries in the action coefficients up to the fourth order is $\mathcal{O}(d^4)$. In contrast, the number of free parameters of a single layer is $\lceil \frac{d}{2} \rceil^3$ in $\tilde{\chi}_l$, d^2 in W_l , and d in b_l , so all in all $\lceil \frac{d}{2} \rceil^3 + d(d+1)$. Therefore we estimate that the depth L of the network must be $\mathcal{O}(d)$ to fit all coefficients up to the fourth order. In this case, at a depth of $L = 34$ of the network, the number of free parameters in the network and in the coefficients approximately coincide. In App. (K), we show that deeper networks reproduce the data distribution more faithfully, and that the coefficients $A_{\text{diag}}^{(4)}$ grow more and more in magnitude as depth increases. Shallow networks on the other hand provide effective theories, as indicated by the good agreement of the cumulants. This behavior is equivalent to that of renormalized theories [57]: they feature the same statistical correlations while modifying the interaction strengths in a consistent manner.

3.3.4 Three-point interactions in pictures of handwritten digits

We now apply our method to a data set where no underlying theory is known. The data are grey-scale images of handwritten digits, namely the MNIST data set [58].

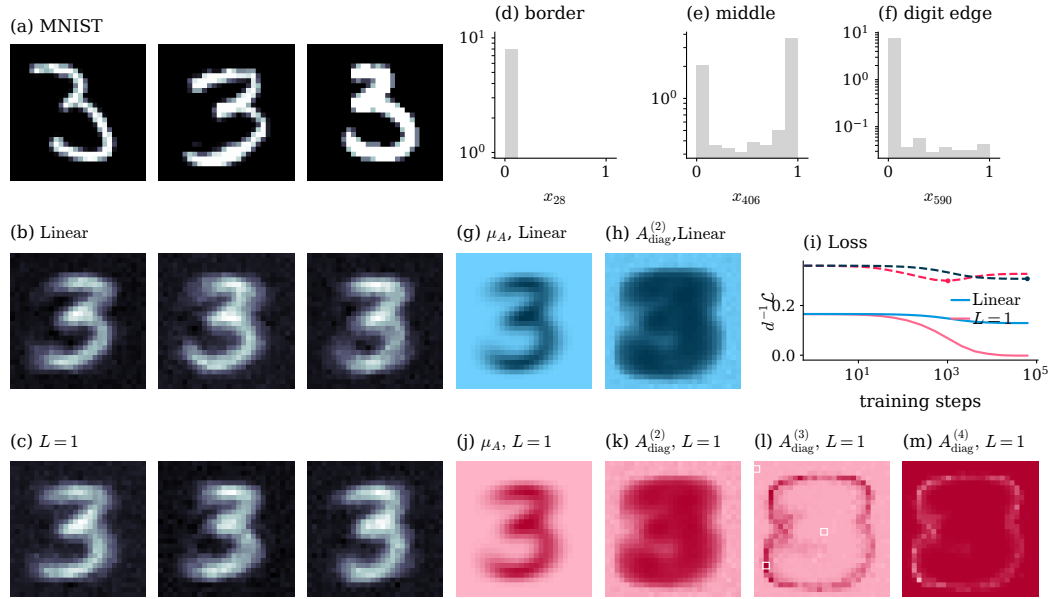


FIG. 3.7: Inference of interactions on MNIST for digit three. (a-c) Images from the data set, the linear model, and an $L = 1$ layer nonlinear model, respectively. (d-f) Single pixel activation statistics from three distinct locations in the image. (g) Entries of the mean μ_A of the Gaussian theory (linear model). (h) Entries on the diagonal of the second order coefficient $A_{\text{diag}}^{(2)}$ of the linear model. (i) Training loss (full lines) and test loss (dashed lines) over training steps. Dots mark the training stages from which the coefficients of both models were extracted. (j) Mean μ_A for the nonlinear model if $A^{(3)}, A^{(4)}$ were not present. (k-m) Entries on the diagonals of the remaining coefficients of the $L = 1$ layer nonlinear model. White squares in (l) mark the locations of the single pixel statistics shown in (d-f).

The degrees of freedom x_i hence encode the brightness of the individual pixels. The images have a size of 28×28 pixels, hence the dimensionality is $d = 784$.

We now train both a linear and a non-linear model on these data. Samples from both models, as well as the original data set are found in Fig. 3.7 (a-c). The linear model corresponds to a Gaussian approximation of these data distributions, whereas a non-linear model learns higher-order interactions. It has been demonstrated [18, 19] that feed-forward classifiers pick up on higher-order statistics of image data sets. In [18], this was done by training several generative models of varying complexity on the data sets which then reproduce the statistics of the data set more and more faithfully. Here, we make the higher-order statistics of the MNIST data set visible.

We take the training stage yielding the best test performance, (see Fig. 3.7 (i)) to compute the coefficients of the two models and compare them. We begin with $\mu_A = -\frac{1}{2}(A^{(2)})^{-1}A^{(1)}$, which, in the case of the linear model, is the mean of the distribution. In the nonlinear model, higher-order coefficients also appear in the

mean, nevertheless we here depict the same quantity for comparability. We then compute the diagonal entries of the interaction coefficients of the linear and nonlinear model. In Fig. 3.7 (g) - (h) and (j) - (m), we arrange the coefficient entries on the same 28×28 grid, such that the action coefficient entries lie in the same position as the pixels whose statistics they characterize. While for the first and second order coefficients, we find the same qualitative behavior, the entries of the third-order and fourth-order coefficients of the nonlinear model trace the edges of the typical digit locations, whereas the higher-order coefficients of the linear model are zero by construction. The distributions of the pixels at these locations are typically more skewed; we show the statistics of three representative pixels in Fig. 3.7 (d) - (f). Pixels located at the border are typically zero, whereas pixels in the center of the digit show a more bi-modal distribution, being either zero or one. By construction, neither the linear nor the nonlinear model can express the bimodality, rather the diagonal entry in $A^{(2)}$ in these locations is less negative, allowing for a broader distribution. Finally, pixels at the typical edges of the digits can be either zero or greyscale, making their distributions more skewed. This behavior is captured by the larger entries in $A_{\text{diag}}^{(3)}$.

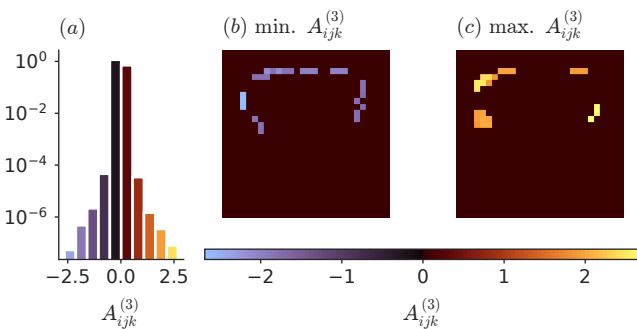


FIG. 3.8: Three-point interactions in MNIST for digit three. (a) Histogram of all entries of the third-order coefficient $A_{ijk}^{(3)}$ for $i \neq j, j \neq k$ and $i \neq k$, color coded according to their value. (b) - (c) Triplets corresponding to the ten most negative (b) or most positive (c) values of $A_{ijk}^{(3)}$. For each triplet, we color pixels i, j and k , according to the value of the interaction coefficient in $A_{ijk}^{(3)}$. Thus triplets of pixels corresponding to the same entry in $A^{(3)}$ have the same color.

many more entries in $A_{ijk}^{(3)}$ which would couple pixels that are distant from each other, and can therefore be assumed to be mostly independent. Rather, the strongest third-order interactions, shown in Fig. 3.8 (b) and (c) encode interactions between localized patches of pixels which trace the curved edges of the digit "three". Thus the third-order interactions encode the digit edges in a localized manner. We show

So far we have only looked at self-interactions of the degrees of freedom. To investigate the coordination between pixels, we now examine the third-order interactions between pixels. Positive offdiagonal entries in the third order coefficient $A_{ijk}^{(3)}$ can bias triplets of three distinct pixels (i, j, k) to be nonzero together, whereas negative entries favor configurations where at least one of the three pixels is zero. Hence the third order interactions can encode higher-order coordination between pixels.

Most entries in $A_{ijk}^{(3)}$ have small magnitude (see Fig. 3.8 (a)). This can be readily understood from the fact that there are

that the third order coefficients trained on images of the digit "two" show the same qualitative behavior in App. (L). Hence the results presented here do not depend on the choice of the digit.

An interesting extension of the analysis is to repeat the procedure on larger images. Here, the interactions on the pixel level may be of smaller relevance, since groups of three or four pixels constitute only small subsets of images. However, it has been demonstrated that wavelet transforms yield interpretable features which are at the same time efficient at encoding image data [59, 60]. Performing a wavelet transform on the data and feeding in these features as the degrees of freedom then yields an interacting theory on an interpretable set of features (namely wavelets) which can encode interactions along different length scales.

We herewith conclude the demonstration of the method. In the next section, we will put our results into the context of other approaches to inference.

3.4 Other approaches to inference problems

The general problem of inferring models from data which we address here, is a well-known challenge, to which a multitude of solution strategies have been developed. We now give a brief overview of the different strategies and settings and distinguish the approach outlined in Chap. 3 from them.

Dynamical systems. In this setting, the equations of motion of a dynamically evolving system are inferred. One way to do so is to use regression to infer the right hand side of a differential equation from a set of basis functions [61–64]. Other studies [65, 66] infer rules for the time-dependence of couplings (synaptic plasticity) using regression and genetic programming. Inference of parameters of stochastic processes [67–70] also relies prior knowledge on the specific form of the update equations. These approaches produce interpretable models, but require a predetermined set of basis functions or operations which are particularly suited to the problem at hand, such that through their combination system dynamics are approximated. Prior knowledge about likely terms in the dynamical equations or their exact functional form is therefore needed.

For a **quantum Many-Body system**, Zache et al. [71] first solve the forward problem, namely for the typically difficult step from interactions to correlation functions. They approximate a higher-order interacting theory to tree-level or one-loop-order in the effective action, and thereby obtain an invertible relation between interactions and correlations, the latter can be obtained from data directly. This approach relies on the accuracy of the approximations necessary to compute the forward step, which are not necessary to train INNs, as the correlation functions are implicitly generated by the network mapping.

Inverse problems for discrete random variables. A prominent example of a pairwise interaction model is the inverse Ising model. Here, a set of pairwise couplings between binary degrees of freedom is inferred. For the special case of a known tree-like network topology, or translationally invariant higher-order couplings along a linear chain, the strength of the couplings can be inferred exactly [72]. Aside from this special case, the difficulty in inferring pairwise couplings stems, as we illustrated in Sec. 3.1.1, from the necessity to maximize the likelihood over a very large parameter space and the difficulty of estimating gradients. Nevertheless, many algorithms now exist for inferring models with pairwise interactions, such as the inverse Ising or XY-model. Inference of patterns in Hopfield networks also fall into this broad class of inverse problems [73, 74]. Inference of pairwise Ising models is at the heart of training Boltzmann machines [75]. Other than Boltzmann machines, a range of techniques for the Ising or XY models first solve the forward problem, namely the statistics given the couplings – using variations of mean-field theory or the TAP equations [76–79] – and then invert these relations explicitly or iteratively [80–83]. Further works maximize the likelihood of the network model given the data, by using belief propagation to reconstruct the network structure from infection cascades [84] or Monte Carlo sampling to infer amino acid sequences in proteins [85]. Other approaches modify the objective function such that the optimization is numerically tractable. They derive optimal objective functions [86, 87], use the pseudo-likelihood [88, 89], or the interaction screening objective [90], the latter approach yielding a convex optimization problem. The method outlined in this work treats the case of continuous rather than discrete variables, is not restricted to pairwise interactions and does not require prior knowledge on the structure of interactions.

Neural Networks. Other approaches to inference using neural networks include training the networks themselves to infer the posterior probability of characteristic parameters given data [91, 92]. In [93], models are optimized to compute renormalized degrees of freedom that are maximally informative about the global state of a system. The authors of [94] use symbolic regression on trained graphical neural networks to derive interpretable interactions of particles. However, these approaches make no lucid connection between the learned model and the parameters of the neural network as we do in this study.

Discussion

In this work, we have treated both the forward and the inverse problem for interactions on structured systems.

Overall, on the level of structured networks with two-point interactions, we found that a spurious self-feedback effect can, when corrected, improve the predictions in the forward mapping significantly.

For the case of Ising spins on a scale-free network structure, this approach can reproduce observations made at the global scale, while also preserving local properties of the system. This allows for a more detailed description of the system than the degree-resolved level [24–26]. In contrast to a mean-field description of the system, it qualitatively reshapes the role of the hubs: while they can induce a local order in their nearest neighbors, a single hub cannot induce long range order in the system, rather global order emerges as a collective property. Nevertheless, it should be noted that the transition temperature grows with the system size N , such that the system is always in an ordered state in the thermodynamic limit. This defies the notion of regular phase transitions that would normally only acquire meaning in the thermodynamic limit. We here instead treat typical finite realizations of the system. The growth of the transition temperature with the network size is quite slow ($T_T \sim \log N$), hence for any real-world network, the transition occurs at a finite value.

We then showed that the same self-feedback effect can also confound predictions for the spread of disease, and when taken into account, improves predictions considerably. In the context of disease modeling, dynamic message passing has previously emerged as a means of taking the self-feedback effect into account [40–42]: dynamic message passing strictly prohibits all forms of self-feedback. Our calculation confirms that, indeed in the SIR model, self-feedback is eliminated.

However, whether self-feedback is present in the systems or not, depends on the disease model at hand: recurrent models of disease spreading such as the SIS or SIRS model permit temporal self-feedback loops [46], while the SIR model does not.

We found that even in systems where a positive self-feedback is in principle permitted, fluctuations can partially cancel the self-feedback effect. This observation

can explain why an elimination of all self-feedback effects, reported in [42] via dynamic message passing can yield improved results, even in systems where positive self-feedback loops are allowed. However, our results do not justify the complete cancellation of all self-feedback, which is known to lead to incorrect predictions [46]. Rather, our theory presents a middle ground between a complete elimination of the self-feedback effect and its over-representation in mean-field theory.

It is interesting to observe that both spin systems at equilibrium and disease models cannot only be linked via the same expansion method, but also that the expansion produces the cancellation of self-feedback in both models at second order. We have demonstrated that the expansion method can be applied also to other variants of the SIR model. This can be further extended, by incorporating for example the SEIR or SEIRS model. Moreover, the computation of higher-order corrections is straightforward. For example, it would be interesting to observe, if an expansion to higher order can reproduce the meta-stable states observed in recurrent disease spreading models, where hubs can self-sustain an infection for long times due to the same self-feedback effect [46, 95].

Overall, it is usually possible to argue whether self-feedback should be present or absent in a system without a systematic fluctuation expansion, and the presence or absence of the same effect can have dramatic effects on the system's behavior. For spin systems, this cancellation has been observed in previous studies [29], and can be anticipated from its relation to the fluctuation-dissipation theorem and cavity methods [23]. Here, we used an expansion of the effective action in the interaction strength to compute corrections to the mean-field case, in which the agents of the system are assumed to be independent. For the SIR model, it is also evident from the specification of the model. Understanding self-feedback effects can hence provide us with a valuable intuition on the properties of a system. We hence expect the results, derived here on several archetypal models, to translate to other systems with a pairwise interaction structure as well.

However, such an heuristic argument is not always sufficient to predict properties of the system as we saw in the SIS and SIRS models. In particular, the non-cancellation of self-feedback does not imply that a first order mean-field theory is accurate. In these cases, a systematic fluctuation expansion can yield improved predictions. Computing these corrections in different systems in which self-feedback is allowed is an interesting direction of further research.

The inverse problem, namely the inference of pairwise interactions from observations of the system, can be solved exactly for continuous random variables (see Sec. 3.1.1). For binary variables, considerable progress has been made (see [72–90] and Sec. 3.4 for a more detailed treatment). Beyond the pairwise regime, we showed that the reverse direction, namely inference of an interacting structure beyond the pairwise order can be efficiently realized by a special kind of generative neural network.

The method casts the structure learned by the neural network into an interpretable form. It hence yields two benefits: first, it can efficiently solve the inference problem for a broad class of higher-order interacting systems, and produces a nontrivial effective theory outside this class. Using this formalism, we can hence uncover higher-order interactions in data sets of high dimensionality. We provide four examples of learning such interactions. In three of these examples, the underlying data distribution is known, and we observe either an exact reproduction of this structure or the learning of an effective interaction structure. In the final example, we extract higher-order pixel interactions in an image, where we found that these interactions code for edges in the images. On the computational level, we can exploit the iterative nature of the coefficient transforms to parametrize the interaction coefficients, which are in principle, higher-order tensors, in a particularly efficient, decomposed form, which lends itself to systematic approximations. This decomposition cannot only be used to speed up the computations, but also to control the complexity of the learned structure, e.g. by directly parametrizing the weight of the quadratic mapping in a low-rank decomposed form during training. Alternatively, the network flexibility can be increased by raising the order of the polynomial activation function, for example, to a cubic activation function. Further, we can expand the class of learnable models to multi-modal ones, by the choice of a multi-modal latent distribution, such as e.g. a Gaussian Mixture. In this case the method yields one set of interaction coefficients per mixture component.

The second contribution of the method is that it uncovers the mechanism by which these networks learn, yielding a description on the level of the data, which is in particular insensitive to the concrete realization of the network parameters. The use of a diagrammatic formalism illustrates how complex data distributions emerge layer by layer, in a hierarchical fashion. The quadratic mapping constitutes the elementary building block of the mapping, higher-order interactions are thus decomposed into this simplest possible form of nonlinear interplay. As a result, the order of interaction in the data directly maps to the required depth of the network in an understandable manner, thus providing an explanation why deep networks are required to learn higher-order interactions. While deeper networks are required to offer sufficient flexibility to learn higher-order statistics, the larger number of trainable parameters at the same time requires more data to learn the statistics accurately. We demonstrated the dependence of the training outcome on the size of the data set in Sec. 3.3.1; confirming that higher-order interactions can only be learned accurately given sufficient data, which can be linked to the bias and noise in the training process. The underlying dynamical equation of motion, Eq. (3.28) relies on the network architecture only through the implicit parametrization of the coefficients via the parameters $\partial_\theta A^{(k)}$. Notably, effective noise and training bias enters on the level of the estimation of moments on subsets of the training data. Hence these phenomena can in principle be treated independent of the network architecture, given that the same architecture is sufficiently flexible. Investigating the dynamics of this learning process can yield insights into the effect of the learning rule on the learned structure on finite data sets. Overall, we have demonstrated that physics provides an efficient language, namely

interactions, in which to express data structure. We expect that the same language, when applied other contemporary generative neural network architectures, can yield valuable insights into their learning mechanisms.

From a physics point of view, the trained network provides a solution to an interacting classical field theory in a data-driven manner. The trained network maps each configuration of the interacting theory in data space to samples in latent space that follow a Gaussian theory, therefore a non-interacting one. The mapping allows one to compute arbitrary connected correlation functions of the interacting theory. This can be achieved via three routes. The traditional route is to use diagrammatic perturbation theory to obtain controlled approximations of connected correlation functions. The diagrams are constructed from propagators and interaction vertices of the inferred action. The second route directly constructs connected correlation functions hierarchically from the network mapping. Here, we exploit that we can express averages in data space via averages in latent space, which ultimately reduces to pairwise correlation functions on the level of the latent Gaussian. For example, the second order correlations read $\langle\langle x_i x_j \rangle\rangle_{p_\theta} = \langle\langle f_{\theta,i}^{-1}(z) f_{\theta,j}^{-1}(z) \rangle\rangle_{z \sim \mathcal{N}(0, \mathbb{1})}$. For the n -th order correlations $\langle\langle x_{i_1} \cdots x_{i_n} \rangle\rangle_{p_\theta}$, one can therefore work out the coefficients of the polynomial $f_{\theta,i_1}^{-1}(z) \cdots f_{\theta,i_n}^{-1}(z)$. The diagrammatic rules for these computations follow in a similar manner to the action transform. The average over p_Z can then be done efficiently using Wick's theorem. Finally, we may simply estimate the correlations by drawing samples from the generative network, this can be done also for more complex architectures which do not feature a tractable polynomial action.

On the whole, we observe that investigating the mapping between an interacting structure of a system on the microscopic level and its macroscopic properties is a highly complex yet fruitful task. Across the different systems we have studied, finding efficient ways in which to express the system's structure and global properties is paramount to finding an useful parametrization of a system, which is detailed enough to describe the observed phenomena, but at the same time yields interpretable, intuitive results and identifies the most relevant mechanisms. We here contribute to this effort by marking the self-feedback effect as an important mechanism in many structured systems, and by providing a new method for inference of higher-order interacting structures.

Appendix

A Parallel-tempering Monte-Carlo

In hierarchical networks such as BA networks featuring hubs, conventional Monte Carlo simulations of heterogeneous networks based on local update schemes are impeded by the freezing of the hub's magnetic states. To circumvent this, simulations of the Ising system were performed using parallel-tempering Monte Carlo instead [96]. This scheme samples over the full configuration space and hence renders an accurate picture of the behavior of hubs and their surrounding nodes throughout the entire relevant temperature range.

B Plefka expansion at equilibrium

The name TAP theory originates from [31], where the expressions are presented as a *fait accompli*. The TAP correction term Eq. (2.21) can be obtained as a second order correction in the interaction strength J . The Ansatz for the expansion leverages the fact that if $J = 0$, then all spins x_i are independent, and therefore, averages can be computed exactly. We here follow the derivation in [36]. The first derivation has been given in [35]; higher order corrections may be found in [97, 98].

In the main text, we set the interaction strength J to one. We here re-introduce J as our expansion parameter. The first order in the expansion is the non-interaction case, $G_{J=0}$ is consequently the entropy of independent binary variables with mean values m_i ,

$$\beta G_{J=0} = \sum_i \frac{1+m_i}{2} \log\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2} \log\left(\frac{1-m_i}{2}\right).$$

Observe that we may then write G as the logarithm of a modified action

$$G(\mathbf{m}) = -\beta^{-1} \ln \langle \exp \Omega_J(\mathbf{m}) \rangle$$

With the Ω_J defined by

$$\Omega_J(\mathbf{m}) = \frac{J\beta}{2} \mathbf{x} \cdot \mathbf{A} \mathbf{x} + \beta \mathbf{h} \cdot (\mathbf{x} - \mathbf{m}) \quad (\text{B.1})$$

Apart from being constrained to fulfilling the equation of state, G therefore has the form of a cumulant generating function. Derivatives of G with respect to J then follow as $-\beta^{-1}$ times the cumulants of the term proportional to J in Ω_J . The first derivative therefore produces the expectation value of the interaction term, for decoupled spins, it evaluates to

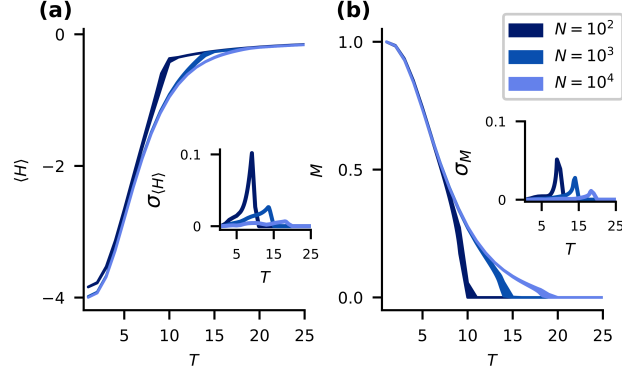


FIG. C.1: Average energy E (a) and magnetization M (b) from ten different realizations of the BAI model, as functions of temperature T for different system sizes N and $m_0 = 4$ obtained from TAP approximation. Dashed lines are drawn one standard deviation above and below the mean, calculated according to Eq. (C.1). Lines seem to align where the standard deviations are very small. Insets show the standard deviations alone.

$$\partial_J G|_{J=0} = -\frac{1}{2} \mathbf{m} \cdot \mathbf{A} \mathbf{m} \quad (\text{B.2})$$

Neglecting higher-order terms, and inserting this into the equation of state, this then yields the mean-field equations Eq. (2.8).

At the second order, we must now compute

$$\partial_J^2 G|_{J=0} = -\beta^{-1} \left(\langle \partial_J^2 \Omega_J \rangle + \langle (\partial_J \Omega_J - \langle \partial_J \Omega_J \rangle)^2 \rangle \right) |_{J=0} \quad (\text{B.3})$$

The first term vanishes. For the second term, namely for $\partial_J \Omega_J$, we must now take into account the fact that \mathbf{h} depends on J . To compute the derivative, we make use of the equation of state

$$\partial_J \mathbf{h} = \partial_J \nabla_{\mathbf{m}} G = -\mathbf{A} \mathbf{m}.$$

Inserted back into the expression for $\partial_J^2 G$, this yields

$$\partial_J^2 G|_{J=0} = -\frac{1}{4} \left\langle [(\mathbf{x} - \mathbf{m}) \cdot \mathbf{A} (\mathbf{x} - \mathbf{m})]^2 \right\rangle_{J=0}$$

evaluating this average then yields Eq. (2.20).

C Self-averaging on BA networks

In the main text, we reported results obtained from single realizations of BA networks. This approach is valid if global properties, such as the magnetization M

and the energy E do not depend significantly on the specific realization of the BA networks. By realization we here mean the specific outcome of the random process generating the network.

Let us denote by $\langle \cdot \rangle_R$ the average over independent realizations of the BA network. Then if the variance of observables O with respect to this average,

$$\sigma_O = \sqrt{\langle O^2 \rangle_R - \langle O \rangle_R^2}, \quad (\text{C.1})$$

are small, we can expect our results obtained from single realizations of the networks to generalize. In Fig. C.1, we compare mean and standard deviations of magnetization and energy for different realizations. As the system size increases, the curves concentrate more and more, thus we find that single realizations are typical of the system averaged over realizations of the connectivity. Therefore for large system sizes, it is sufficient to study single realizations.

D Fluctuation-dissipation theorem

In Sec. 2.2.3, we found that the fluctuations of an equilibrium system exactly cancels the self-feedback effect. We here argue that this is just another manifestation of the fluctuation-dissipation theorem [37]: to do so, we view the presence of the spin at node i as an in-homogeneous external field of strength $h_{\text{ext},j} = A_{ji}m_i$ at node j . The response to this field on the spin at node j is given by $\chi_j A_{ji}m_i$, where χ_j is the the susceptibility at node j , given by

$$\chi_j = \frac{\partial m_j}{\partial h_{\text{ext},j}} = \beta (1 - m_j^2). \quad (\text{D.1})$$

Up to factors of β , this is equal to the variance of the same spin

$$\langle x_j^2 \rangle - \langle x_j \rangle^2 = 1 - m_j^2. \quad (\text{D.2})$$

Thus we have demonstrated the fluctuation-dissipation theorem, which states the equivalence between the fluctuation Eq. (D.2) and the linear response Eq. (D.1) to a perturbation, the perturbation here being the presence of node i . In fact, here both the response and the fluctuation are calculated in exactly the same way, namely (up to factors of β) from the second derivative of the free energy F with respect to the external field h .

The self-feedback of the response field $\chi_j A_{ji}m_i$ at node j to the magnetization at node i is given by

$$A_{ij}^2 \beta^2 m_i (1 - m_j^2). \quad (\text{D.3})$$

Summing over all neighbor nodes j , this gives precisely the TAP term in (2.21), but with opposite sign.

For non-equilibrium systems (for example, the non-equilibrium kinetic Ising model or directed networks of binary units, where $A_{ji} \neq A_{ij}$), this equivalence between linear response and fluctuations is lost [43, 99]. There, the fluctuation correction does not produce the correct time arguments of the mean fields, neither the correct prefactors in A . This observation in particular also applies to the models treated in Sec. 2.3, where the cancellation of self-feedback arises through a different mechanism.

E Epidemic dynamics in a Spiking Simulator code

We implement the dynamics of the SIR, SIRS and SIS model in NEST [44], a simulator for spiking neurons. NEST has previously been used to model binary neurons [100]. We here follow the same approach as in [100], to adapt the simulator to the SIR, SIRS and SIS dynamics.

Note that each agent i need not know the exact state of its nearest neighbors to compute the update probability, rather, it suffices to know how many of these nearest neighbors are infected. This information is encoded in the input θ_i to the agent. We use spikes to transmit changes to these fields: When an agent i is infected, it sends out a spike. Upon receiving a spike, the nearest neighbors j of the agent hence increments their input fields θ_j by one. When agent i leaves the infected state, it sends to spikes at the same time. Two simultaneous spikes received by a nearest neighbor j then result in a deduction of θ_j by one.

F Dynamical fluctuation expansion for infection models

We here perform the fluctuation expansion for the SIR model following the same procedure as in [43]. We first write down a path integral which formally sums over all possible trajectories of the stochastic process. We then perform a self-consistent expansion analogous to App. (B), but for dynamic variables. We use a short-hand notation

$$X_i(t) = \begin{pmatrix} S_i(t) \\ I_i(t) \end{pmatrix} \in \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}$$

to describe the state of each individual, and $X(t)$ to indicate the state of the whole system of agents at time t . At each time step, we hence have a transition probability from $S(t), I(t)$ to $S(t+1), I(t+1)$

$$\begin{aligned} W_{t+1,t} = & \prod_{i=1}^N \{ S_i(t+1) [1 - \phi(\theta_i(t))] S_i(t) \\ & + I_i(t+1) [(1 - \mu)I_i(t) + \phi(\theta_i(t)) S_i(t)] \\ & + (1 - S_i(t+1) - I_i(t+1)) [1 - S_i(t) - (1 - \mu)I_i(t)] \} . \end{aligned} \quad (\text{F.1})$$

We here introduce a function $\phi : [0, \infty) \rightarrow [0, 1]$ to ensure that all probabilities remain in the range $[0, 1]$ since for agents of high degree k_i , the sum in Eq. (2.41) may exceed one. We further constrain ϕ to fulfill $\phi(0) = 0$ and $\phi'(0) = 1$. The first line in Eq. (F.1) corresponds to the agent remaining susceptible. The second corresponds to a new or sustained infection. The last line corresponds to the probability that the agent enters or remains in the recovered state.

Since we are interested averages of the X_i variables, we begin by defining a cumulant generating function

$$W(\psi, h) = \ln \left\langle \exp(\psi^T X) \right\rangle_h, \quad (\text{F.2})$$

the subscript h here emphasizes that the average depends on the input field h . In Eq. (F.2), a set of sources

$$\psi_i(t) = \begin{pmatrix} \psi_i^S(t) \\ \psi_i^I(t) \end{pmatrix}$$

are coupled to the field $X(t)$ s, where we used the shorthand notation

$$\psi^T X = \sum_{i,t} \psi_i^T(t) X_i(t).$$

We compute the cumulants of the vectors $X_i(t)$ by taking the derivatives of Eq. (F.2) by the sources ψ and subsequently setting $\psi = 0$. For example, we may obtain ρ_i^α with $\alpha \in \{S, I\}$ by computing

$$\rho_i^\alpha(t) = \langle X_i^\alpha(t) \rangle_h = \partial_{\psi_i^\alpha(t)} W(\psi, h) \Big|_{\psi=0}. \quad (\text{F.3})$$

We express the average in Eq. (F.2) by ordering the averages over individual time steps, from $t = 1$ to T . We find

$$\left\langle \exp(\psi^T X) \right\rangle_h = \prod_{i,t} \sum_{X_i(t)} \left\{ \exp(\psi_i^T(t) X_i) W_{t+1,t} [X(t+1) | \theta(t), X(t)] \right. \\ \left. \delta(\theta(t) - h(t) - \beta AI(t)) \right\} p(X(0))$$

where the starting density $p(X(0))$ must be given and we must enforce the condition on the external fields θ , Eq. (2.41) at each point. The latter is achieved via a Dirac delta distribution. We now introduce an auxiliary field $\hat{\theta}_i(t)$ to enforce the condition on $\theta_i(t)$ via the inverse Fourier transform $\delta(x) = \int \frac{d\hat{x}}{\sqrt{2\pi}} \exp(i\hat{x}x)$. This gives us the

following expression:

$$\begin{aligned} \langle \exp(\psi^T X) \rangle_h = & \prod_{i,t} \sum_{X_i(t)} \int \frac{d\theta_i(t) d\hat{\theta}_i(t)}{\sqrt{2\pi}} \exp(\psi_i^T(t) X_i(t)) \\ & \cdot W_{t+1,t} [X(t+1) | \theta(t), X(t)] p(X(0)) \\ & \cdot \exp\left(i\hat{\theta}_i(t) \left[\theta_i(t) - h_i(t) - \beta \sum_j a_{ij} I_j(t) \right]\right). \end{aligned} \quad (\text{F.4})$$

Writing Eq. (F.4) in this way, we can effectively split the role of the input fields θ from those of the binary variables $X(t)$. We can think of Eq. (F.4) as a reordering of the averages: starting with a known density $p(X(0))$, we compute the sum over the $X(0)$, which gives us a statistic of the $\theta(0)$ variables. We then integrate over $\theta(0), \hat{\theta}(0)$ to get the statistic of the $X(1)$, and then again sum over the $X(1)$ to get the statistic of the $\theta(1), \hat{\theta}(1)$, and so on. In practice however, these computations are not done exactly, as the average in each time step would require around 3^N terms to account for all possible states of each variable $X_i(t)$.

In the following, we will specify the expansion which allows us to compute averages $\rho^\alpha(t)$ perturbatively. The derivation closely resembles the Plefka expansion for the equilibrium statistics of the Ising model, which we outline in App. (B). In contrast to App. (B) however, here we must take one average per time point into account, as we have a dynamical system.

Observe that up to a prefactor $-i$, $\hat{\theta}_i(t)$ couples to $h_i(t)$ in the same manner as the source fields ψ couple to the physical observables X . We hence introduce another average

$$\rho_\theta = \langle \hat{\theta}_i(t) \rangle = -i \partial_{h_i(t)} W(\psi, h) \Big|_{\psi=0} = 0, \quad (\text{F.5})$$

which vanishes due to the normalization condition on $W(\psi, h)$. We then use Eq. (F.3) and Eq. (F.5) to define the Legendre-Fenchel transform of Eq. (F.2), which we will refer to as the effective action, via

$$\Gamma(\rho, \rho_\theta) = \sup_{\psi, h} \psi^T \rho - i h^T \rho_\theta - W(\psi, h). \quad (\text{F.6})$$

Together with Eq. (F.3) and Eq. (F.5), this yields the equations of state

$$\begin{aligned} \partial_{\rho_i^\alpha(t)} \Gamma(\rho, \rho_\theta) &= \psi_i^\alpha(t) \\ \partial_{\rho_{\theta_i}(t)} \Gamma(\rho, \rho_\theta) &= -i h_i(t). \end{aligned} \quad (\text{F.7})$$

To obtain the correction to the mean-field theory, we then proceed to expand Eq. (F.6) to second order in β around the non-interacting case $\beta = 0$,

$$\Gamma \approx \Gamma_{\beta=0} + \beta \partial_\beta \Gamma + \frac{1}{2} \beta^2 \partial_\beta^2 \Gamma. \quad (\text{F.8})$$

Reinserting this into Eq. (F.7), to first order in β , one obtains the mean-field approximation Eq. (2.43), and the dynamical TAP equation, Eq. (2.44), to second order.

We will now perform the expansion of Γ to second order in β term by term. To simplify the derivation, we first perform it only for the SIR case. We will generalize our result to the SIS and SIRS case in App. (G).

F.1 Noninteracting system

At $\beta = 0$, the cumulant generating function decomposes into a sum, since X_i, X_j are now independent for $i \neq j$,

$$W_{\beta=0}(\psi, h) = \sum_i \ln Z_i(\psi_i, h_i),$$

we can view the terms $Z_i(t)$ as single-agent partition functions.

We must find Eq. (F.6) under the conditions Eq. (F.5) and Eq. (F.3). To this end, we express the fields ψ, h as functions of the mean values $\rho_i^S, \rho_{\hat{\theta}}$. The Markov property of the random process ensures that the knowledge of the distribution of X at a specific time step is sufficient to compute the distribution at any future time step. Explicitly, this means that the distribution at the present time step depends only on the previous one and not on its history, neither on future time steps. Furthermore, in the non-interacting case, the full information about the distribution of each of the binary variables at any time point is contained in their mean. We use the latter two observations to write down Z_i for a single time step from t to $t+1$ with the averages at t given. We then have

$$\begin{aligned} Z_i(\psi_i, h_i) = \prod_t \left[e^{\psi_i^S(t+1)} \underbrace{\rho_i^S(t)(1 - \phi(h_i(t)))}_{S \rightarrow S} \right. \\ \left. + e^{\psi_i^I(t+1)} \left(\underbrace{\phi(h_i(t))\rho_i^S(t)}_{S \rightarrow I} + \underbrace{(1 - \mu)\rho_i^I(t)}_{I \rightarrow I} \right) \right. \\ \left. + \underbrace{\mu\rho_i^I(t)}_{I \rightarrow R} + \underbrace{1 - \rho_i^S(t) - \rho_i^I(t)}_{R \rightarrow R} \right] \end{aligned}$$

We have labeled the transition probabilities by the transitions in brackets, e.g. $S \rightarrow I$ for the transition from a susceptible to infected state. We must now find Eq. (F.6) under the conditions Eq. (F.5), Eq. (F.3). We first evaluate Eq. (F.5)

$$\rho_{\hat{\theta}_i(t)} = \frac{-i\partial_{h_i(t)} Z_i(\psi_i, h_i)}{Z_i(\psi_i, h_i)} = \frac{-i\phi'(h_i(t))\rho_i^S(t) \left(e^{\psi_i^I(t+1)} - e^{\psi_i^S(t+1)} \right)}{Z_i(\psi_i, h_i)} = 0,$$

hence we find for $\phi'(h_i(t))\rho_i^S(t) \neq 0$ that $\psi_i^I(t) = \psi_i^S(t) \forall i, t$. The derivatives by the sources yield

$$\begin{aligned}\rho_i^S(t+1) &= \frac{e^{\psi_i^S(t+1)}(1-\phi(h_i(t))\rho_i^S(t))}{Z_i(t)} \\ \rho_i^I(t+1) &= \frac{e^{\psi_i^S(t+1)}(\phi(h_i(t))\rho_i^S(t) + (1-\mu)\rho_i^I(t))}{Z_i(t)}\end{aligned}$$

we solve these equations for $\psi_i^I(t), h_i(t)$, and obtain

$$\begin{aligned}\psi_i^S(t) &= \ln \frac{\rho_i^S(t) + \rho_i^I(t)}{1 - \rho_i^S(t) - \rho_i^I(t)} - \ln \frac{\rho_i^S(t-1) + (1-\mu)\rho_i^I(t-1)}{1 - \rho_i^S(t-1) - (1-\mu)\rho_i^I(t-1)} \\ h_i(t) &= \phi^{-1} \left(\frac{\rho_i^I(t+1)\rho_i^S(t) - \rho_i^S(t+1)(1-\mu)\rho_i^I(t)}{\rho_i^S(t+1)\rho_i^S(t) + \rho_i^I(t+1)\rho_i^S(t)} \right).\end{aligned}$$

We insert this back into Z_i , to obtain for the single-agent partition function

$$Z_i(\psi_i, h_i) = \prod_t \frac{(1 - \rho_i^S(t) - (1-\mu)\rho_i^I(t))}{1 - \rho_i^S(t+1) - \rho_i^I(t+1)}.$$

This quantity must equal one, since the partition function is normalized. This equation then simply expresses that the probability $1 - \rho_i^S(t) - \rho_i^I(t)$ that node i is recovered, increases by $\mu\rho_i^I(t)$ in each time step.

F.2 Mean-field

We will now compute the first order correction to Γ . To do so, we write

$$\Gamma = \ln \prod_{i,t} \sum_{X_i(t)} \int \frac{d\theta_i(t)d\hat{\theta}_i(t)}{\sqrt{2\pi}} \exp(\Omega_\beta) W_t [X(t+1) | \theta(t), X(t)]$$

with Ω_β defined by

$$\begin{aligned}\Omega_\beta = \sum_{i,t} \left[\psi_i(t)^\top (\rho_i(t) - X_i(t)) + ih_i(t) (\hat{\theta}_i(t) - \rho_{\hat{\theta}_i}(t)) \right. \\ \left. - i\hat{\theta}_i(t) \left(\theta_i(t) - \beta \sum_j A_{ij} I_j(t) \right) \right].\end{aligned}\tag{E.9}$$

We may now use that, similar to the cumulant generating function, the derivative of Γ by β yields another average $\partial_\beta \Gamma = \langle \partial_\beta \Omega_\beta \rangle$. In the case that $\beta = 0$, all averages belonging to different indices factorize, because nodes i, j are independent for $i \neq j$.

We hence find that

$$\begin{aligned}\partial_\beta \Gamma|_{\beta=0} &= \langle \partial_\beta \Omega_\beta \rangle|_{\beta=0} \\ &= i \sum_{i,t} \rho_{\hat{\theta}_i}(t) \sum_j A_{ij} \rho_j^I(t),\end{aligned}$$

the analogous result to Eq. (B.2). To first order therefore, the effective action reads

$$\begin{aligned}\Gamma = \sum_{i,t} \left[\ln \left(\frac{1 - \rho_i^S(t+1) - \rho_i^I(t+1)}{(1 - \rho_i^S(t) - (1-\mu)\rho_i^I(t))} \right) \right. \\ + (\rho_i^S(t) + \rho_i^I(t)) \left(\ln \frac{\rho_i^S(t) + \rho_i^I(t)}{1 - \rho_i^S(t) - \rho_i^I(t)} + \ln \frac{1 - \rho_i^S(t-1) - (1-\mu)\rho_i^I(t-1)}{\rho_i^S(t-1) + (1-\mu)\rho_i^I(t-1)} \right) \\ - i \rho_{\hat{\theta}_i}(t) \phi^{-1} \left(\frac{\rho_i^I(t+1)\rho_i^S(t) - \rho_i^S(t+1)(1-\mu)\rho_i^I(t)}{\rho_i^S(t)(\rho_i^S(t+1) + \rho_i^I(t+1))} \right) \\ \left. + i \rho_{\hat{\theta}_i}(t) \beta \sum_j A_{ij} \rho_j^I(t) \right] + \mathcal{O}(\beta^2)\end{aligned}\quad (\text{F.10})$$

To obtain the mean-field equation, we must take the derivative after $\rho_i(t), \rho_{\hat{\theta}_i}(t)$ and set the right hand side to zero. We find that both identities in Eq. (F.7) are solved by

$$\rho_i^S(t) + \rho_i^I(t) = \rho_i^S(t-1) + (1-\mu)\rho_i^I(t-1). \quad (\text{F.11})$$

Taking the derivative by $\rho_{\hat{\theta}_i}(t)$ then yields the mean-field equations

$$\begin{aligned}\Delta \rho_i^S(t+1) &= -\rho_i^S(t) \phi \left(\beta \sum_j A_{ij} \rho_j^I(t) \right) \\ \Delta \rho_i^I(t+1) &= -\mu \rho_i^I(t) + \rho_i^S(t) \phi \left(\beta \sum_j A_{ij} \rho_j^I(t) \right).\end{aligned}$$

With ϕ the identity, we arrive at Eq. (2.43) for $\eta = \bar{\mu} = 0$. The extension for the SIRS and SIS models is shown in App. (G).

F.3 Second order correction

We compute the second derivative of Γ

$$\partial_\beta^2 \Gamma|_{\beta=0} = \langle \partial_\beta^2 \Omega_\beta \rangle|_{\beta=0} + \langle (\partial_\beta \Omega_\beta)^2 \rangle|_{\beta=0} - \langle \partial_\beta \Omega_\beta \rangle|_{\beta=0}^2. \quad (\text{F.12})$$

The first term vanishes. We are left with the variance of $\partial_\beta \Omega_\beta$ in the non-interacting case. Since we are at second order, the external fields ψ, h acquire a linear dependence

on β via their dependence on expectation values $\rho_i^\alpha, \rho_{\hat{\theta}}$. We must therefore consider derivatives of the type

$$\partial_\beta \psi_i^\alpha(t) = \partial_\beta \partial_{\rho_i^\alpha(t)} \Gamma = i \sum_j A_{ji} \rho_{\hat{\theta}_j}(t)$$

$$\partial_\beta h_i(t) = i \partial_\beta \partial_{\rho_{\hat{\theta}_i}(t)} \Gamma = - \sum_j A_{ij} \rho_j^I(t)$$

Combining these equations with Eq. (F.9) yields

$$\partial_\beta \Omega_\beta - \langle \partial_\beta \Omega_\beta \rangle_{\beta=0} = i \sum_{i,j,t} A_{ij} \delta \hat{\theta}_i(t) \delta I_j(t), \quad (\text{F.13})$$

where we use $\delta I_j(t) = I_j(t) - \rho_j^I(t)$ and $\delta \hat{\theta}_i(t) = \hat{\theta}_i(t) - \rho_{\hat{\theta}_i}(t)$. We must now take the average in the case $\beta = 0$. This again means that all agents are treated as independent. Until the end of this section, we will always evaluate expectation values in the non-interacting case and drop the subscript $\beta = 0$ for brevity. We hence arrive at

$$\partial_\beta^2 \Gamma_{\beta=0} = - \sum_{t,t'} \sum_{ijkl} A_{ij} A_{kl} \langle \delta \hat{\theta}_i(t) \delta \hat{\theta}_k(t') \delta I_j(t) \delta I_l(t') \rangle. \quad (\text{F.14})$$

Again, note the analogy of this result to Eq. (B.3). If all indices i, \dots, l are unequal, then the term under the sum vanishes, since all nodes decouple. We must therefore have at least two indices in Eq. (F.14) equal to get a meaningful contribution.

We now go through the different combinations of indices to determine those which yield a meaningful contribution. We immediately observe that due to the prefactor of $A_{ij} A_{kl}$, we must have that $i \neq j$ and $j \neq k$. The average in Eq. (F.14) must therefore always decompose into at least two factors. Furthermore, each index must be equal to at least one other index, otherwise the term is proportional to $\langle \delta \hat{\theta} \rangle = 0$ or $\langle \delta I \rangle = 0$. From the latter two observations it follows that exactly two independent indices are left. From

$$\langle \delta \hat{\theta}_i(t) \delta \hat{\theta}_i(t') \rangle = 0 \quad \forall i, t, t'$$

it follows that only the term where $i = l$ and $k = j$ remains. All in all, we find

$$\partial_\beta^2 \Gamma_{\beta=0} = - \sum_{t,t'} \sum_{ij} A_{ij} A_{ji} \langle \delta \hat{\theta}_i(t) \delta I_i(t') \rangle \langle \delta \hat{\theta}_j(t') \delta I_j(t) \rangle \quad (\text{F.15})$$

We must therefore compute correlations of the form $\langle \hat{\theta}_i(t) I_i(t') \rangle$. We will see in the following that $\langle \hat{\theta}_i(t) I_i(t') \rangle$ vanishes unless $t < t'$. This is so, because $\langle \hat{\theta}_i(t) I_i(t') \rangle$ has the role of a response function: it measures the effect of a change in the field $h_i(t)$ at time t on the random variable $I_i(t')$ at another time point. Since the stochastic process is causal (later changes in the external field can have no influence on earlier

time points), the response function is causal as well, and vanishes for $t \geq t'$. To evaluate the response function, we only need to consider the generating function from t to t' with the averages at $t - 1$ given. Again, this is because of the Markov property of the random process.

We now write down the single-agent partition function for an extended period of time: starting with a known distribution at time t (encoded in the expectation values $\rho_i^\alpha(t)$), we sum over all possible realizations of the stochastic process until time t'

$$\begin{aligned}
 Z_i(\psi_i, h_i) = & \rho_i^S(t) \left[\prod_{\tau=t}^{t'-1} (1 - \phi(h_i(\tau))) e^{\psi_i^S(\tau+1)} \right. \\
 & + \sum_{t_{\text{inf}}=t}^{t'-1} \phi(h_i(t_{\text{inf}})) e^{\psi_i^I(t_{\text{inf}}+1)} \left[\prod_{\tau=t_{\text{inf}}}^{t'-1} (1 - \phi(h_i(\tau))) e^{\psi_i^S(\tau+1)} \right] \\
 & \cdot \left\{ \prod_{v=t_{\text{inf}}+1}^{t'-1} [(1 - \mu) e^{\psi_i^I(v+1)}] + \sum_{t_{\text{rec}}=t_{\text{inf}}+1}^{t'-1} \mu \prod_{v=t_{\text{inf}}+1}^{t_{\text{rec}}} [(1 - \mu) e^{\psi_i^I(v+1)}] \right\} \\
 & \rho_i^I(t) \left\{ \prod_{v=t}^{t'-1} [(1 - \mu) e^{\psi_i^I(v+1)}] + \sum_{t_{\text{rec}}=t}^{t'-1} \mu \prod_{v=t}^{t_{\text{rec}}} [(1 - \mu) e^{\psi_i^I(v+1)}] \right\} \\
 & + 1 - \rho_i^S(t) - \rho_i^I(t) \Big] \tag{F.16}
 \end{aligned}$$

This sum is organized as follows: The first line corresponds to the node remaining susceptible. The second line counts all possible times t_{inf} of infection. The first term in the third line corresponds to a remaining infection until t' . The second term in the same line counts all possible recovery times respectively. The fourth line counts all trajectories with initial infection, and respectively the possibilities of recovery, analogous to the third line. The last line corresponds to the node beginning in the recovered state. To compute $\langle \hat{\theta}_i(t) I_i(t') \rangle$, we take the derivative

$$\begin{aligned}
 \langle \hat{\theta}_i(t) I_i(t') \rangle_{\beta=0} = & \frac{-i \partial_{h_i(t)} \partial_{\psi_i^I(t')} Z_i(\psi_i, h_i)}{Z_i(\psi_i, h_i)} \Big|_{\psi=h=0} \\
 \stackrel{t' \geq t}{=} & -i \rho_i^S(t) (1 - \mu)^{t'-t-1}. \tag{F.17}
 \end{aligned}$$

Where we used $\phi(0) = 0$, $\phi'(0) = 1$. Observe that the derivative in Eq. (F.17) picks out of all possible trajectories precisely the ones which correspond an infection of node i at time point t , which lasts until time point t' with probability $(1 - \mu)^{t'-t-1}$. For $t' < t$, no such trajectory exists, hence $\langle \hat{\theta}_i(t') I_i(t) \rangle = 0$. Therefore the product of the two averages always vanishes

$$\langle \hat{\theta}_k(t') I_k(t) \rangle \langle \hat{\theta}_i(t) I_i(t') \rangle = 0 \quad \forall i, k, t, t'. \tag{F.18}$$

Altogether, we have

$$\begin{aligned} \partial_{\beta}^2 \Gamma_{\beta=0} = & - \sum_{t,t'} \sum_{ij} A_{ij} A_{ji} \left(2i\Theta(t-t') \rho_j^S(t') (1-\mu)^{t-t'-1} \rho_{\hat{\theta}_i}(t) \rho_i^I(t') \right. \\ & \left. + \rho_{\hat{\theta}_i}(t) \rho_i^I(t') \rho_{\hat{\theta}_j}(t') \rho_j^I(t) \right) \end{aligned}$$

From which we can compute second order correction to Eq. (F.10). The correction only changes the equation of state originating from the derivative after $\rho_{\hat{\theta}}(t)$, because the factor $\rho_{\hat{\theta}}(t') = 0$ cancels all contributions of the correction term to the equation of state. The only relevant contribution to the equation of state is hence the term linear in $\rho_{\hat{\theta}}(t)$. All equations of state together finally yield Eq. (2.44) for $\eta = \bar{\mu} = 0$.

G Extensions to SIS and SIRS model

We now extend the calculation to for the SIS and SIRS model. Up to Eq. (F.4), the calculations remain unchanged, but now we must specify a different update probability

$$\begin{aligned} W_{t+1,t} = & \prod_{i=1}^N \left\{ S_i(t+1) \left([1 - \phi(\theta_i(t))] S_i(t) + \eta R_i(t) + \bar{\mu} I_i(t) \right) \right. \\ & + I_i(t+1) \left[(1 - \mu - \bar{\mu}) I_i(t) + \phi(\theta_i(t)) S_i(t) \right] \\ & \left. + R_i(t+1) \left[(1 - \eta) R_i(t) + \mu I_i(t) \right] \right\}. \end{aligned}$$

where now $\bar{\mu}$ is the probability for $I \rightarrow S$ and η for $R \rightarrow S$ and we used $R_i(t) = 1 - S_i(t) - I_i(t)$. We now follow the same steps as in App. (F) in an abbreviated form.

G.1 Noninteracting case

We first compute the Legendre-Fenchel transform for the non-interacting case. The derivatives of $W(\psi, h)$ by the sources yield

$$\begin{aligned} \rho_i^S(t+1) &= \frac{e^{\psi_i^S(t+1)} \left[(1 - \phi(h_i(t)) - \eta) \rho_i^S(t) + \eta + (\bar{\mu} - \eta) \rho_i^I(t) \right]}{Z_i(t)} \\ \rho_i^I(t+1) &= \frac{e^{\psi_i^I(t+1)} \left(\phi(h_i(t)) \rho_i^S(t) + (1 - \mu - \bar{\mu}) \rho_i^I(t) \right)}{Z_i(t)}. \end{aligned}$$

We solve these equations for $\psi_i^I(t), h_i(t)$, and obtain

$$\begin{aligned} \psi_i^S(t) &= \ln \frac{\rho_i^S(t+1) + \rho_i^I(t+1)}{\rho_i^S(t) + \rho_i^I(t) + \eta(1 - \rho_i^S(t) - \rho_i^I(t)) - \mu\rho_i^I(t)} \\ &\quad + \ln \frac{(1-\eta)(1 - \rho_i^S(t) - \rho_i^I(t)) + \mu\rho_i^I(t)}{1 - \rho_i^S(t+1) - \rho_i^I(t+1)} \\ h_i(t) &= \phi^{-1} \left(\frac{-(1-\mu-\bar{\mu})\rho_i^S(t+1)\rho_i^I(t) + \rho_i^I(t+1)(\eta + (1-\eta)\rho_i^S(t) + (\bar{\mu}-\eta)\rho_i^I(t))}{\rho_i^S(t)(\rho_i^S(t+1) + \rho_i^I(t+1))} \right). \end{aligned}$$

Inserted into Z_i , this yields the for the single-agent partition function

$$Z_i(\psi_i, h_i) = \prod_t \frac{(1-\eta)(1 - \rho_i^S(t) - \rho_i^I(t)) + \mu\rho_i^I(t)}{1 - \rho_i^S(t+1) - \rho_i^I(t+1)}.$$

This quantity must equal one, since the partition function is normalized. This condition then simply expresses that the probability $1 - \rho_i^S(t) - \rho_i^I(t)$ that node i is recovered, changes by $\mu\rho_i^I(t) - \eta(1 - \rho_i^S(t) - \rho_i^I(t))$ in each time step.

G.2 Mean-field equations

The mean-field equations follow analogously to App. (F.2). The derivative of Γ by β yields the same expression

$$\begin{aligned} \partial_\beta \Gamma|_{\beta=0} &= \langle \partial_\beta \Omega_\beta \rangle|_{\beta=0} \\ &= i \sum_{i,t} \rho_{\hat{\theta}_i}(t) \sum_j A_{ij} \rho_j^I(t). \end{aligned}$$

To first order therefore, the effective action reads

$$\begin{aligned} \Gamma &= \sum_{i,t} \left[\ln \left(\frac{1 - \rho_i^S(t+1) - \rho_i^I(t+1)}{(1-\eta)(1 - \rho_i^S(t) - \rho_i^I(t)) + \mu\rho_i^I(t)} \right) \right. \\ &\quad + (\rho_i^S(t) + \rho_i^I(t)) \ln \frac{\rho_i^S(t+1) + \rho_i^I(t+1)}{\rho_i^S(t) + \rho_i^I(t) + \eta(1 - \rho_i^S(t) - \rho_i^I(t)) - \mu\rho_i^I(t)} \\ &\quad + (\rho_i^S(t) + \rho_i^I(t)) \ln \frac{(1-\eta)(1 - \rho_i^S(t) - \rho_i^I(t)) + \mu\rho_i^I(t)}{1 - \rho_i^S(t+1) - \rho_i^I(t+1)} \\ &\quad - i\rho_{\hat{\theta}_i}(t) \phi^{-1} \left(\frac{-(1-\mu-\bar{\mu})\rho_i^S(t+1)\rho_i^I(t) + \rho_i^I(t+1)(\eta + (1-\eta)\rho_i^S(t) + (\bar{\mu}-\eta)\rho_i^I(t))}{\rho_i^S(t)(\rho_i^S(t+1) + \rho_i^I(t+1))} \right) \\ &\quad \left. + i\rho_{\hat{\theta}_i}(t) \beta \sum_j A_{ij} \rho_j^I(t) \right] + \mathcal{O}(\beta^2) \end{aligned} \tag{G.1}$$

To obtain the mean-field equation, we must take the derivative after $\rho_i(t), \rho_{\hat{\theta}_i}(t)$ and set the right hand side to zero. We find that both identities in Eq. (F.7) are solved by

$$\rho_i^S(t) + \rho_i^I(t) = \rho_i^S(t-1) + (1-\mu)\rho_i^I(t-1) + \eta(1-\rho_i^S(t-1) - \rho_i^I(t-1)). \quad (\text{G.2})$$

Taking the derivative of Γ by $\rho_{\hat{\theta}_i}(t)$ then yields the mean-field equations

$$\begin{aligned} \Delta\rho_i^S(t+1) &= \eta(1-\rho_i^S(t) - \rho_i^I(t)) + \bar{\mu}\rho_i^I(t) - \rho_i^S(t)\phi\left(\beta\sum_j A_{ij}\rho_j^I(t)\right) \\ \Delta\rho_i^I(t+1) &= -(\mu + \bar{\mu})\rho_i^I(t) + \rho_i^S(t)\phi\left(\beta\sum_j A_{ij}\rho_j^I(t)\right). \end{aligned}$$

which now contain the transitions $I \rightarrow S$ and $R \rightarrow S$ with the corresponding prefactors $\bar{\mu}$ and η .

G.3 Second order correction

We proceed in the same fashion as in App. (F.3), and find that we must compute the moments in Eq. (F.15). In principle, trajectories with multiple re-infections between time points t and t' must be taken into account to compute the average $\langle \hat{\theta}_i(t)I_i(t') \rangle$, since the transition $I \rightarrow S \rightarrow I$ and $I \rightarrow R \rightarrow S \rightarrow I$ are now allowed. But these terms drop out when we set $\beta = 0$. Therefore, the response function is

$$\langle \hat{\theta}_i(t)I_i(t') \rangle_{\beta=0} = -i \begin{cases} \rho_i^S(t)(1-\mu-\bar{\mu})^{t'-t-1} & t' > t \\ 0 & t' \geq t \end{cases}$$

which is just the same as Eq. (F.17), albeit that the probability to remain infected now decays with factors of $(1-\mu-\bar{\mu})$ rather than $(1-\mu)$. We therefore find that the second order correction yields the same terms as for the SIR model, when we replace $\mu \rightarrow \mu + \bar{\mu}$. This yields the final result Eq. (2.44).

H Decomposed tensors

The coefficient transforms outlined in 3.2 require computations with higher-order tensors. Computations with higher-order tensors $T^{(k)}$ of rank k can become numerically infeasible for large dimension d because the number of entries in $T^{(k)}$ grows as $\mathcal{O}(d^k)$. Specifically, two challenges arise: First, to store the entries of the tensors. Second, to compute contractions with matrices W such as

$$T^{(k)} \cdot (W)^{\otimes k}, \quad (\text{H.1})$$

which arise due to the linear coefficient transform Eq. (3.19), or other tensors, such as χ_l for the nonlinear transform.

These contractions require the computation of the sum over all entries

$$\left(T^{(k)} \cdot (W)^{\otimes k}\right)_{i_1 \dots i_k} = \sum_{j_1 \dots j_k=1}^d T_{j_1 \dots j_k}^{(k)} W_{j_1 i_1} \dots W_{j_k i_k}.$$

Without further simplification this entails the computation of d^k entries of $T^{(k)} \cdot (W)^{\otimes k}$ from d^k terms each. The total number of floating point operations hence scales as $\mathcal{O}(d^{2k})$. Although the number of steps required grows polynomially, thus sub-exponentially with d , for realistic data set sizes and $k = 4$ this still poses a problem.

Here, to facilitate the computation of coefficients with rank $k = 4$, we exploit that they are built from coefficients of lower rank: we express them in a decomposed form, which allows us to make the computations and storage of the tensors more efficient.

As a first step, we decompose the network parameters χ_l . In the following, we drop the layer index l for brevity, as the structure of the computation is the same for any layer. We may choose χ to be symmetric in its latter two indices $\chi_{\mu j k} = \chi_{\mu k j}$ without any loss of network expressivity. We then rearrange χ to be a list of d symmetric matrices $\bar{\beta}^\mu$, $\mu = 1, \dots, d$ such that $\chi_{\mu k j} = \bar{\beta}_{k j}^\mu$. Using the eigendecomposition of these symmetric matrices $\bar{\beta}^\mu$, we write

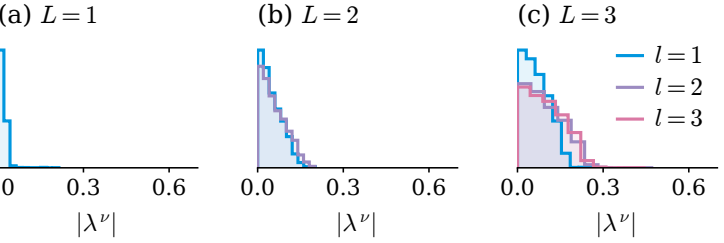


FIG. H.1: Eigenvalue distributions of decomposed χ_l for networks of different depths. We decompose trained network parameters χ_l from Sec. 3.3.3 to the form of Eq. (H.2) and distinguish eigenvalues from the decomposed form of different layers l .

$$\chi = \sum_{\mu, \nu=1}^d \gamma^{\mu, \nu} \otimes \beta^{\mu, \nu} \otimes \beta^{\mu, \nu}, \quad (\text{H.2})$$

where $\gamma^{\mu, \nu}, \beta^\nu$ are vectors. $\gamma_\tau^{\mu, \nu} = \delta_{\tau, \mu} \lambda_\nu^\mu$ has only one non-zero entry, namely the ν -th eigenvalue of the μ -th matrix $\bar{\beta}^\mu$. Storing (H.2) requires $2d^2$ vectors of length d , namely d^2 vectors $\beta^{\mu, \nu}$ and d^2 vectors $\gamma^{\mu, \nu}$. The magnitude of entries in χ is directly related to the magnitude of the eigenvalues λ_ν^μ , which is typically small for trained networks: we show distributions of eigenvalues from trained networks in Fig. H.1. The distributions broaden with increasing depth, however the peak of the distribution remains at $|\lambda_\nu^\mu| = 0$.

Since the eigenvalues are typically small, it is reasonable to reduce the space required to store χ and all tensors related to it by placing a cutoff $\bar{\lambda} \geq 0$ on the eigenvalues,

keeping only the $\bar{n} \leq d^2$ largest eigenvalues which have $|\lambda_\nu^\mu| \geq \bar{\lambda}$. Then the number of entries required to store χ then scales as $\mathcal{O}(2\bar{n}d)$, as again we need $2\bar{n}$ vectors of length d each. To further simplify the expression, we absorb the sum over μ, ν into a single index $\tau = 1, \dots, \bar{n}$.

An alternative way to achieve the decomposition of χ into a reduced number of components would be to use the decomposed form Eq. (H.2) directly during training and limit the number of independent vectors β^τ . This approach effectively trades the network expressivity for the tractability of the action coefficient transforms.

The decomposed form (H.2) also translates to all tensors computed via contraction with χ . Since all higher-order coefficients originate from contractions with χ (see Eq. (3.23)), we can hence write them as decomposed tensors as well. The contraction between a rank k symmetric tensor $T^{(k)}$ and χ is

$$T^{(k)} \cdot \chi = \sum_{\tau} \left(T^{(k)} \cdot \gamma^{\tau} \right) \otimes \beta^{\tau} \otimes \beta^{\tau}.$$

We now distinguish the cases $k = 1, 2, 3$. If $k = 1$, the result is a symmetric matrix. If $k = 2$, then $T^{(k)} \cdot \gamma^{\tau} =: \alpha^{\tau}$ is a vector, therefore $T^{(k)} \cdot \chi$ can be written as a sum of outer products between three vectors. If $k = 3$, the result is a sum of outer products between matrices $T^{(k)} \cdot \gamma^{\tau} =: \bar{\alpha}^{\tau}$ and two vectors;

$$T^{(3)} \cdot \chi = \sum_{\tau} \bar{\alpha}^{\tau} \otimes \beta^{\tau} \otimes \beta^{\tau}. \quad (\text{H.3})$$

The matrices $\bar{\alpha}^{\tau}$ are symmetric since $T^{(3)}$ is symmetric

$$\bar{\alpha}_{ab}^{\tau} = \sum_c T_{abc}^{(3)} \gamma_c^{\tau} = \sum_c T_{bac}^{(3)} \gamma_c^{\tau} = \bar{\alpha}_{ba}^{\tau}.$$

Storing the factors of Eq. (H.3) therefore requires \bar{n} matrices $\bar{\alpha}^{\tau}$ and \bar{n} vectors β^{τ} . The number of matrix and vector entries required to store this object is therefore $\mathcal{O}(\bar{n}d^2)$.

The case $k \geq 4$ does not arise, as any coefficient with degree $k \geq 4$ must already contain at least two factors χ , and we here truncate at second order in χ (see Sec. 3.2.4).

We now return to our original problem: the computation of contractions along all indices with a matrix. For the decomposed tensor $T^{(3)} \cdot \chi$ with matrices W , this is

$$\left(T^{(3)} \cdot \chi \right) \cdot (W)^{\otimes 4} = \sum_{\tau} \left(W^T \bar{\alpha}^{\tau} W \right) \otimes \left(W^T \beta^{\tau} \right) \otimes \left(W^T \beta^{\tau} \right).$$

This requires \bar{n} matrix-vector products $W^T \beta^\tau$ and $2\bar{n}$ matrix-matrix products for $\bar{\alpha}^\tau W$ and $W^T (\bar{\alpha}^\tau W)$. Each term in the matrix-matrix product is computed from d terms, therefore the number of terms required to compute $W^T \bar{\alpha}^\tau W$ is $\mathcal{O}(2d^\omega)$ with the matrix multiplication exponent ω , which depends on the concrete algorithm used for matrix multiplication, e.g. $\omega \approx 2.8$ for the Strassen algorithm [101]. To evaluate the contraction, we therefore need to compute $\mathcal{O}(2\bar{n}d^\omega)$ terms. Even in the case of no cutoff $\bar{\lambda} = 0 \Rightarrow \bar{n} = d^2$, this approach significantly reduces the required computations compared to the naive implementation.

In the experiments in Sec. 3.3.1 - Sec. 3.3.3, we have used a maximal dimensionality of $d = 10^2$ and no cutoff, $\bar{\lambda} = 0$. For the MNIST data set with $d = 784$, we used a cutoff of $\bar{\lambda} = 10^{-2}$. In the absence of any cutoff, we find a scaling of our algorithm roughly equal to the simpler of two schemes proposed in [102]: there, it was shown that the number of floating point operations needed to compute general contractions of the type of Eq. (H.1) can be reduced by exploiting the symmetry of the tensors. The authors introduce a simple scheme to reduce the number of floating point operations (using $\omega = 3$) to roughly $\mathcal{O}(d^{k+1})$, and a more complex structure of saving these tensors, which further speeds up the computations at the expense of storing more intermediate entries.

To study data sets of even higher dimension, we hence suggest the combination of a cutoff, more efficient storing of symmetric tensors as suggested in [102], or restricting the number of free components in χ directly.

I Random generation of multi-modal actions

I.1 Coefficient distributions for random actions

In Sec. 3.3.2, we use multi-modal actions S_R constructed from randomly drawn coefficients $R^{(k)}$. Here, we describe the way in which these coefficients are sampled in order to obtain a balanced and non-trivial probability distribution.

A basic condition the coefficients must satisfy is that the resulting action be normalizable: $\int S_R(x) dx < \infty$. For large enough x , the action is dominated by the highest order terms:

$$S(x) \xrightarrow{\|x\| \rightarrow \infty} (R^{(4)}) \cdot x^{\otimes 4}.$$

It is therefore necessary and sufficient for normalizability that $R^{(4)}$ be negative definite, which we ensure by choosing $R^{(4)}$ to be a diagonal tensor with negative coefficients:

$$R_{i_1 i_2 i_3 i_4}^{(4)} = \begin{cases} -\frac{x_r^{-4}}{d} & \text{if } i_1 = i_2 = i_3 = i_4; \\ 0 & \text{otherwise.} \end{cases} \quad (\text{I.1})$$

Here d is the dimensionality of the data, and $x_r \in \mathbb{R}$ is a length scale which we are free to choose. Hence, one can view $R^{(4)}$ as a regulator term, and x_r as the value for which it becomes strongly suppressing. For our experiments we used $x_r = 1.0$.

We then define the probability via

$$p_R(x | \{R^{(k)}\}_{k \leq 4}) \sim \exp(S_R(x)),$$

where we have omitted the normalization. Computing the latter is difficult for high dimensional x . Fortunately however, we need not do so to sample from the distribution, therefore we only define S_R up to the normalization.

We choose the coefficients of S_R such that the data can be described as a perturbation of a Gaussian theory. This means that $R^{(2)}$ must be negative definite. We define $R^{(2)}$ as follows:

$$\begin{aligned} W_{ij} &\sim \mathcal{N}(0, 1/d^2) \\ R_{ij}^{(2)} &= -c\delta_{ij} - \frac{1}{2} \sum_a W_{ia} W_{ja}, \end{aligned} \quad (\text{I.2})$$

for all $i, j = 1, \dots, d$ and $c = 0.1$ in our experiments. Equation (I.2) is equivalent to transforming a Gaussian variable $z \sim \mathcal{N}(0, 1)$ by a linear transform $x = W^{-1}z$ and then computing the action of x (compare to Eq. (3.25)). We then add the diagonal term $c\delta_{ij}$ to ensure that $R^{(2)}$ is negative definite even in the unlikely case that W does not have full rank.

Finally, the coefficients $R^{(1)}$ and $R^{(3)}$ are chosen as follows ($i, j, k = 1, \dots, d$):

$$\begin{aligned} R_i^{(1)} &\sim \mathcal{N}(0, \sigma_i^2) & \sigma_i &= \frac{x_r^{-1}}{d}; \\ R_{ijk}^{(3)} &\sim \mathcal{N}(0, \sigma_{ijk}^2) & \sigma_{ijk} &= \frac{x_r^{-3}}{s_{ijk} \gamma_{ijk}}. \end{aligned}$$

The variable $\gamma_\alpha = |\mathcal{P}(\alpha)|$ in the denominator of σ_α is the multiplicity of the index α . This is the number of times the component $R_{ijk}^{(3)}$ appears in $R^{(3)}$; since coefficients are symmetric, it is equal to the number of distinct permutations of (i, j, k) . We scale the multiplicity by s_α , the number of different components which have the same number of permutations – for example the permutations of the indices $(2, 1, 1)$ and $(5, 3, 3)$ appear $\gamma_{ijj} = 3$ times each, and there are $s_{ijj} = d(d-1)$ distinct entries of such indices. Since there are far more off-diagonal components in $R^{(3)}$, we scale those by s_{ijk} to ensure that both the on-diagonal and off-diagonal components of $R^{(2)}$, $R^{(3)}$ are significant, and neither dominates the other completely. Finally, the scaling with respect to x_r^{-1} and x_r^{-3} ensures that neither linear and cubic terms are negligible within the region where the regulator term $R^{(4)}$ is non-suppressing.

I.2 Multimodality of random actions

After sampling S_R , we ensure that it is multimodal using the following procedure: we initialize an optimization algorithm at random points $x \in \mathbb{R}^d$ and attempt to find the maximum of the action $S_R(x^*)$ from there. Depending on where the initialized points are, the optimization algorithm may find different maxima. After the algorithm has converged, we check whether a maximum has been found by computing the eigenvalues of the Hessian of S_R at this point. If all eigenvalues are below zero, the Hessian is negative definite, meaning that the action S_R is locally convex down at the given point.

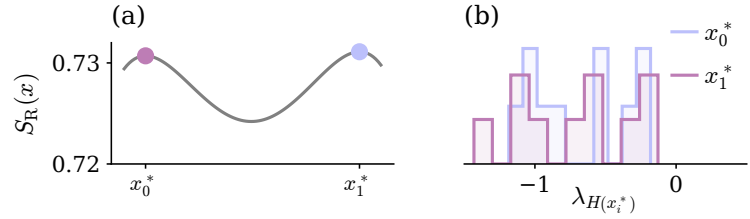


FIG. I.1: Multiple local maxima in S_R . (a) S_R along the straight line connecting two local maxima x_0^* , x_1^* of S_R found by the optimization algorithm. (b) Eigenvalues of the Hessian H of S_R at local maxima x_0^* , x_1^* . All eigenvalues $\lambda_{H(x_i^*)}$ are negative, therefore the action is convex down in all directions.

Here, we initialized at 10^3 different values. In Fig. I.1 we show the maximal values of $S_R(x^*)$ found by the algorithm as well as the eigenvalues of the Hessian for selected, distinct final values x^* . The Hessian indeed is negative definite at both points. The action S_R therefore has at least two local maxima.

J Sampling actions with MCMC

We use Markov chain Monte Carlo (MCMC) sampling to create a data set from an unnormalized action. MCMC is well suited to this task since it computes only update probabilities, which rely on differences between actions at different points, ΔS , where the normalization drops out. We used the No-U-Turn Sampler (NUTS) [103] implementation provided by PyMC3 [56]. The sampler parameters followed the recommended defaults, with 10^3 tuning steps and a mass matrix initialized to unity. The target acceptance rate was increased to 0.95 to increase the sensitivity to small features of the probability distribution.

K Lattice model in low dimensions

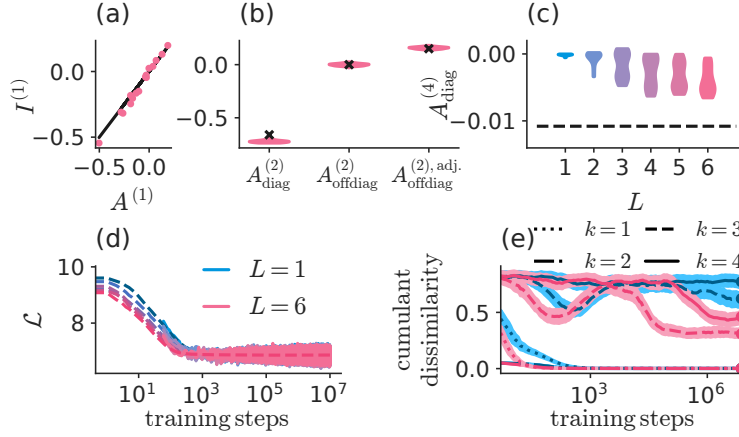


FIG. K.1: Coefficients of lattice model for networks of varying depth trained on a $d = 16$ dimensional data set with $D = 10^5$ samples. **(a)** Learned ($L^{(1)}$) over true ($A^{(1)}$) first order coefficients. **(b)** Distribution of learned coefficient entries $A^{(2)}$ compared to target values (black crosses). Self-interaction terms are labeled $A_{\text{diag}}^{(2)}$, off-diagonal entries $A_{\text{offdiag}}^{(2)}$. Among the off-diagonal entries $A_{\text{offdiag}}^{(2)}$, entries belonging to adjacent lattice sites $A_{\text{offdiag}}^{(2),\text{adj}}$ are shown separately. **(c)** Training loss (full curves) and test loss (dashed curves). Colors distinguish different network depths L . **(d)** Distribution of learned fourth order self-interactions over network depth. The dashed line marks the target value. **(e)** Dissimilarity of true and learned cumulants: $1 - \cos \angle (\langle \langle x^{\otimes k} \rangle \rangle_A, \langle \langle x^{\otimes k} \rangle \rangle_R)$ over training steps. We record the cumulants at logarithmically spaced intervals during training. The curves are smoothed by averaging over ten adjacent recording steps. Shaded areas show the variation due to the estimation of the cumulants from samples. Dots indicate training stage of coefficients shown in (a,b).

To further elucidate the relationship between the dimensionality of the model and the depth of the network, we here examine two lower dimensional versions of the lattice model introduced in Sec. 3.3.3.

First, we study the same system, but on a smaller, 4×4 lattice. Thus the dimensionality is reduced, $d = 16$. Here, the combined number of independent entries in the first four action coefficients is only 4844, which corresponds to roughly 6 network layers. We again train a model on samples from the system. Figure (K.1) shows a comparison of true vs. learned coefficients. As for the higher dimensional model, we find that also here, independent of network depth, $A^{(1)}$ and the off-diagonal entries $A_{ij}^{(2)}$ with $i \neq j$ are recovered correctly (see Fig. K.1 (a,b)). The magnitudes of the diagonal entries in the fourth order coefficient $A_{\text{diag}}^{(4)}$ increase with the depth (see Fig. K.1 (d)). We also observe that the depth L speeds up learning (see Fig. K.1 (c)). From the evolution equation for the coefficients, Eq. (3.28), this must be due to the

derivatives $\partial_\theta A_\theta^{(k)}$ changing: increasing the depth means that there are more terms contributing to the derivative. Furthermore, we find that up to the fourth order, the cumulants of the learned distribution increasingly align with those of the true distribution as we increase network depth. In conclusion, we find that increasing the depth of the network increases the accuracy of the learned distribution, both in terms of its coefficients and of its cumulants.

Second, we repeated the experiment on a 3×3 lattice, thus $d = 9$, where we set the external field to zero $h = 0 \Rightarrow A^{(1)} = 0$. We do this to check whether any of the results rely on the system's symmetry being broken by h . This is not the case: again, we find an alignment of most entries in $A^{(1)}, A^{(2)}$, with $A_{\text{diag}}^{(2)}$ slightly lower than expected and $A_{\text{diag}}^{(4)}$ larger than expected (see Fig. K.2 (a,b)). Since here, the action is symmetric under a global sign flip of x , the first and third cumulants vanish. We therefore only check whether the second and fourth cumulants of true and learned models match. As in the previous cases, this alignment increases with depth, as shown in Fig. K.2 (d). Therefore, neither the learning of the off-diagonal coefficients of $A^{(2)}$, nor the effective nature of the learned theory depends on the system's symmetry being broken.

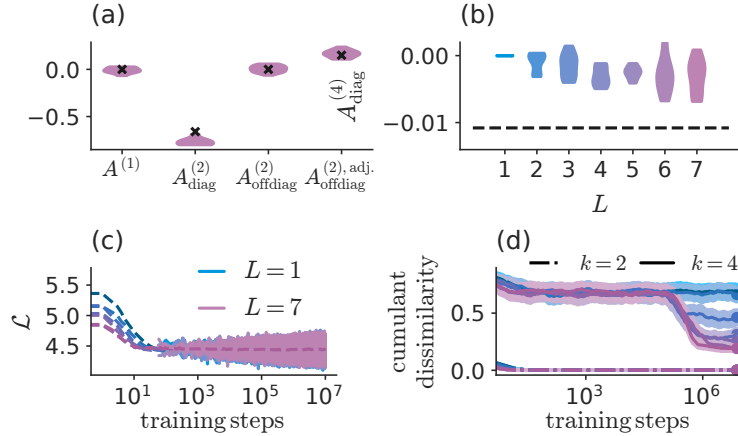


FIG. K.2: Coefficients of lattice model without external field for networks of varying depth trained on a $d = 9$ dimensional data set with $D = 10^5$ samples. **(a)** Distribution of learned coefficient entries $A^{(1)}, A^{(2)}$ compared to target values (black crosses). Self-interaction terms $A_{\text{diag}}^{(2)}$ are shown separately from off-diagonal entries $A_{\text{offdiag}}^{(2)}$. Among the off-diagonal entries $A_{\text{offdiag}}^{(2)}$, those entries belonging to adjacent lattice sites $A_{\text{offdiag}}^{(2), \text{adj.}}$ are shown separately. **(b)** Distribution of learned fourth order self-interactions compared to network depth. The dashed line marks the target value. **(c)** Training loss (full curves) and test loss (dashed curves). Colors distinguish different network depths L . **(d)** Dissimilarity of true and learned cumulants: $1 - \cos \angle (\langle \langle x^{\otimes k} \rangle \rangle_A, \langle \langle x^{\otimes k} \rangle \rangle_R)$ over training steps. We record the cumulants at logarithmically spaced intervals during training. The curves are smoothed by averaging over ten adjacent recording steps. Shaded areas show the variation due to the estimation of the cumulants from samples. Dots indicate training stage of coefficients shown in (a,b).

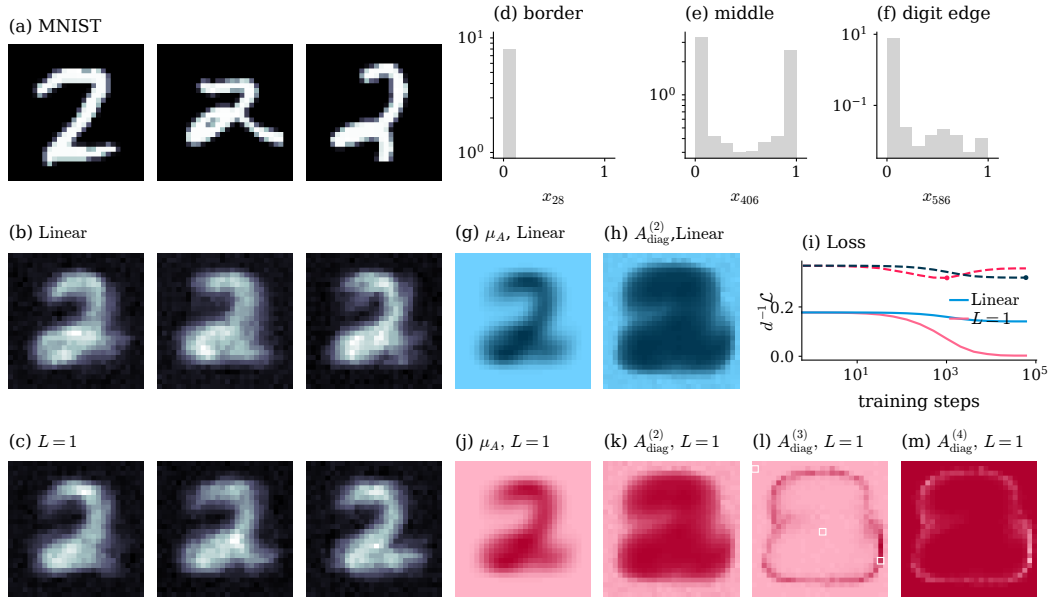


FIG. L.1: Inference of interactions on MNIST for digit two. (a-c) Images from the data set, the linear model, and an $L = 1$ layer nonlinear model, respectively. (d-f) Single pixel activation statistics from three distinct locations in the image. (g) Entries of the mean μ_A of the Gaussian theory (linear model). (h) Entries on the diagonal of the second order coefficient $A_{\text{diag}}^{(2)}$ of the linear model. (i) Training loss (full lines) and test loss (dashed lines) over training steps. Dots mark the training stages from which the coefficients of both models were extracted. (j) Mean μ_A for the nonlinear model if $A^{(3)}, A^{(4)}$ were not present. (k-m) Entries on the diagonals of the remaining coefficients of the $L = 1$ layer nonlinear model. White squares in (l) mark the locations of the single pixel statistics shown in (d-f).

L Training on MNIST digits

Here, we specify details of the training of the networks on the MNIST data set and present further results on the inference of interactions. A principal component analysis of the MNIST data set shows that several eigenvalues of the covariance matrices of the MNIST data set are very small $\lambda_{\text{min}}^{\text{MNIST}} \sim 10^{-28}$. This means that there are eigenvectors the high-dimensional covariance matrix where the data points do not spread at all but are confined almost perfectly to a single point along the eigenvector direction. This confirms the well-known assumption that the MNIST data lie on a lower dimensional manifold.

Training invertible neural networks on lower dimensional data sets can lead to diverging eigenvalues in the Jacobian of the network mapping (see e.g. [104, 105] for a detailed treatment). One way to circumvent the problem would be to train on a lower dimensional space, spanned by the eigenvectors of the covariance matrix with

finite eigenvalues. In this scenario, the degrees of freedom are the projections of the data set onto the corresponding eigenvectors. However, here, we are interested in extracting interactions on the level of pixels, and it is not clear how interactions learned on the data set with reduced dimensionality translate back to the original space: the relevant mapping is not invertible, therefore the change of variables formula doesn't apply.

Instead, we find that the problem of diverging $\ln|\det J_{f_\theta}|$ can be mitigated in two steps: In the first step, we add a small, i.i.d Gaussian noise ξ with mean zero and variance $\sigma_\xi^2 = 10^{-2}$ to each pixel value in the data set. The small noise ensures that the eigenvalues of the covariance matrix of the noised data set are all of order σ_ξ^2 or larger. We then perform a full PC decomposition on this noised data set, retaining all principal components. We then train a network on the data in PC space. As PC decomposition is a linear mapping, we can compute the action coefficients by including this linear transform as a final step in the transformation of coefficients obtained from the trained network. As we have a limited number of samples, we also add a small i.i.d. Gaussian noise ξ_{train} with variance $\sigma_{\xi_{\text{train}}}^2 = 10^{-2} = \sigma_\xi^2$ to each training batch to prevent over-fitting. We find that this procedure both prevents divergences of the training loss during training as well as speeds up the training process.

Naturally, adding Gaussian noise changes the probability distribution of the data set. For example, if the MNIST data were Gaussian distributed, then the noised data set would follow the same Gaussian distribution as well, but with an additional term σ_ξ^2 added to the diagonal covariance matrix. This also means that the noising process cannot induce higher-order interactions in the noised data set, any such higher-order interactions we find must stem from the original data set. We here choose σ_ξ^2 as small as possible in order to ensure that the noised data set stays as close as possible to the original data set.

We now showcase the detection of edges also for the digit two in Fig. L.1. The results are qualitatively similar to the case of digit three, shown in the main text: again, we find an increased magnitude in $A_{\text{diag}}^{(3)}$ and $A_{\text{diag}}^{(4)}$ along the location of the typical digit edges in Fig. L.1 (l,m), while the first and second order coefficients of the linear and non-linear model are very similar (see Fig. L.1 (g,h) and(j,k)). The nonlinear model

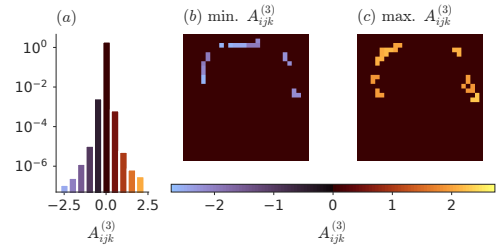


FIG. L.2: Three-point interactions in MNIST for digit two. (a) Histogram of all entries of the third-order coefficient $A_{ijk}^{(3)}$ for $i \neq j, j \neq k$ and $i \neq k$, color coded according to their value. (b) - (c) Triplets corresponding to the ten most negative (b) or most positive (c) values of $A_{ijk}^{(3)}$. For each triplet, we color pixels i, j and k , according to the value of the interaction coefficient in $A_{ijk}^{(3)}$. Thus triplets of pixels corresponding to the same entry in $A^{(3)}$ have the same color.

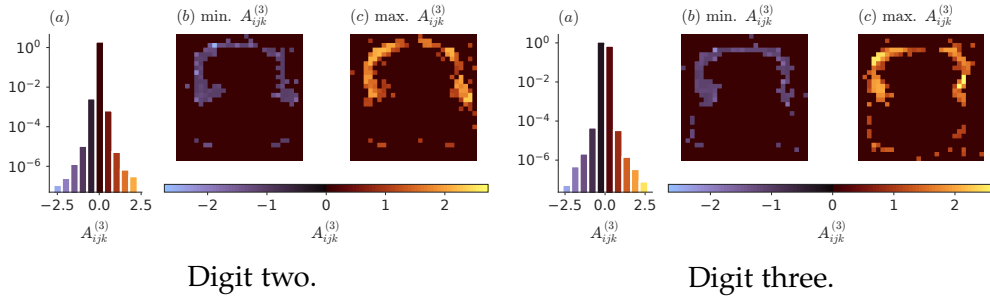


FIG. L.3: Three-point interactions in MNIST for digits two and three. (a) Histogram of all entries of the third-order coefficient $A_{ijk}^{(3)}$ for $i \neq j, j \neq k$ and $i \neq k$, color coded according to their value. (b) - (c) Triplets corresponding to the 10^2 most negative (b) or most positive (c) values of $A_{ijk}^{(3)}$. For each triplet, we color pixels i, j and k , according to the value of the interaction coefficient in $A_{ijk}^{(3)}$. Thus triplets of pixels corresponding to the same entry in $A^{(3)}$ have the same color.

overfits the training set after around 10^4 training steps, there, the test error begins to increase again (see Fig. L.1 (i)). The three-point interactions in $A_{ijk}^{(3)}$ for the digit two are shown in Fig. L.2. The three-point interactions here again couple pixels localized in patches at typical edge locations.

Finally, in Fig. L.3 we show the first one-hundred most positive and most negative three-point interactions of digits two and three. This shows that the higher-order interactions consistently trace the edges of the digits.

Bibliography

1. Broder, A. *et al.* Graph Structure in the Web. *Computer Networks* **33**, 309–320. ISSN: 1389-1286. (2023) (June 2000).
2. Zachary, W. W. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research* **33**, 452–473. ISSN: 0091-7710. (2023) (1977).
3. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and Centrality in Protein Networks. *Nature* **411**, 41–42. ISSN: 1476-4687. (2023) (May 2001).
4. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nature Communications* **10**, 1017. ISSN: 2041-1723. <http://www.nature.com/articles/s41467-019-08746-5> (2021) (Dec. 2019).
5. Holme, P. Rare and everywhere: Perspectives on scale-free networks. *Nature Communications* **10**. Number: 1 Publisher: Nature Publishing Group, 1016. ISSN: 2041-1723. <https://www.nature.com/articles/s41467-019-09038-8> (2021) (Mar. 2019).
6. Bianconi, G. Higher-Order Networks. *Elements in Structure and Dynamics of Complex Networks*. (2021) (Nov. 2021).
7. Battiston, F. *et al.* Networks beyond Pairwise Interactions: Structure and Dynamics. *Physics Reports. Networks beyond Pairwise Interactions: Structure and Dynamics* **874**, 1–92. ISSN: 0370-1573. (2023) (Aug. 2020).
8. Battiston, F. *et al.* The Physics of Higher-Order Interactions in Complex Systems. *Nature Physics* **17**, 1093–1098. ISSN: 1745-2481. (2023) (Oct. 2021).
9. Ghorbanchian, R., Restrepo, J. G., Torres, J. J. & Bianconi, G. Higher-Order Simplified Synchronization of Coupled Topological Signals. *Communications Physics* **4**, 1–13. ISSN: 2399-3650. (2023) (June 2021).
10. Calmon, L., Restrepo, J. G., Torres, J. J. & Bianconi, G. Dirac Synchronization Is Rhythmic and Explosive. *Communications Physics* **5**, 1–17. ISSN: 2399-3650. (2023) (Oct. 2022).
11. Petri, G. *et al.* Homological Scaffolds of Brain Functional Networks. *Journal of The Royal Society Interface* **11**, 20140873. (2023) (Dec. 2014).
12. McInnes, L., Healy, J. & Melville, J. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* Sept. 2020. arXiv: 1802.03426 [cs, stat]. (2023).
13. Dinh, L., Krueger, D. & Bengio, Y. NICE: Non-linear Independent Components Estimation. eprint: arXiv:1410.8516 (2014).
14. Dinh, L., Sohl-Dickstein, J. & Bengio, S. Density Estimation Using Real NVP. *arXiv:1605.08803 [cs, stat]*. arXiv: 1605.08803 [cs, stat]. (2020) (Feb. 2017).

-
15. Kingma, D. P. & Dhariwal, P. *Glow: Generative Flow with Invertible 1x1 Convolutions* in *Advances in Neural Information Processing Systems* (eds Bengio, S. et al.) **31** (Curran Associates, Inc., 2018).
 16. Vaswani, A. et al. Attention is all you need. **30** (2017).
 17. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. *Deep Unsupervised Learning Using Nonequilibrium Thermodynamics* in *Proceedings of the 32nd International Conference on Machine Learning* (PMLR, June 2015), 2256–2265. (2023).
 18. Refinetti, M., Ingrosso, A. & Goldt, S. *Neural Networks Trained with SGD Learn Distributions of Increasing Complexity* Nov. 2022. arXiv: 2211.11567 [cond-mat, stat]. (2023).
 19. Fischer, K. et al. *Decomposing Neural Networks as Mappings of Correlation Functions* Feb. 2022. arXiv: 2202.04925 [cond-mat, stat]. (2022).
 20. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**. Publisher: American Association for the Advancement of Science, 509–512. ISSN: 0036-8075. <https://science.sciencemag.org/content/286/5439/509> (1999).
 21. Ising, E. Beitrag zur Theorie des Ferromagnetismus. **31**, 253–258 (1925).
 22. Kermack, W. O. & McKendrick, A. G. Contributions to the Mathematical Theory of Epidemics—I. *Bulletin of Mathematical Biology* **53**, 33–55. ISSN: 1522-9602. (2021) (Mar. 1991).
 23. Mezard, M, Parisi, G & Virasoro, M. *Spin Glass Theory and Beyond World Scientific Lecture Notes in Physics Volume 9*. ISBN: 978-9971-5-0116-7. <https://www.worldscientific.com/doi/abs/10.1142/0271> (2021) (WORLD SCIENTIFIC, Nov. 1986).
 24. Bianconi, G. Mean field solution of the Ising model on a Barabási–Albert network. *Physics Letters A* **303**, 166–168. ISSN: 0375-9601. <https://www.sciencedirect.com/science/article/pii/S037596010201232X> (2021) (Oct. 2002).
 25. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. Ising model on networks with an arbitrary distribution of connections. *Phys. Rev. E* **66**. Publisher: American Physical Society, 016104. <https://link.aps.org/doi/10.1103/PhysRevE.66.016104> (July 2002).
 26. Leone, M., Vázquez, A., Vespignani, A. & Zecchina, R. Ferromagnetic ordering in graphs with arbitrary degree distribution. *The European Physical Journal B - Condensed Matter and Complex Systems* **28**, 191–197. ISSN: 1434-6036. <https://doi.org/10.1140/epjb/e2002-00220-0> (2021) (July 2002).
 27. Herrero, C. P. Ising model in scale-free networks: A Monte Carlo simulation. *Physical Review E* **69**. Publisher: American Physical Society (APS). ISSN: 1550-2376. <http://dx.doi.org/10.1103/PhysRevE.69.067109> (June 2004).
 28. Aleksiejuk, A., Hołyst, J. A. & Stauffer, D. Ferromagnetic phase transition in Barabási–Albert networks. *Physica A: Statistical Mechanics and its Applications* **310**, 260–266. ISSN: 0378-4371. <http://www.sciencedirect.com/science/article/pii/S0378437102007409> (2002).

-
29. Fischer, K. H. & Hertz, J. A. *Spin Glasses* ISBN: 978-0-511-62877-1 (Cambridge University Press, 1991).
 30. Vasiliev, A. N. & Radzhabov, R. A. Legendre transforms in the Ising model. **21**, 963–970. ISSN: 1573-9333. <https://doi.org/10.1007/BF01035593> (1974).
 31. Thouless, D. J., Anderson, P. W. & Palmer, R. G. Solution of ‘Solvable model of a spin glass’. *Philosophical Magazine* **35**, 593–601. ISSN: 0031-8086. <http://www.tandfonline.com/doi/abs/10.1080/14786437708235992> (2021) (Mar. 1977).
 32. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic Processes in Complex Networks. *Reviews of Modern Physics* **87**, 925–979. (2023) (Aug. 2015).
 33. Herrero, J. L. *et al.* Acetylcholine contributes through muscarinic receptors to attentional modulation in V1. **454**, 1110–1113 (2008).
 34. Goh, K.-I., Kahng, B. & Kim, D. Spectra and eigenvectors of scale-free networks. *Physical Review E* **64**. Publisher: American Physical Society (APS). ISSN: 1095-3787. <http://dx.doi.org/10.1103/PhysRevE.64.051903> (Oct. 2001).
 35. Vasiliev, A. N. & Radzhabov, R. A. Legendre transforms in the Ising model. *Theoretical and Mathematical Physics* **21**, 963–970. ISSN: 1573-9333. <https://doi.org/10.1007/BF01035593> (Oct. 1974).
 36. Plefka, T. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *Journal of Physics A: Mathematical and General* **15**, 1971–1978. ISSN: 0305-4470, 1361-6447. <https://iopscience.iop.org/article/10.1088/0305-4470/15/6/035> (2021) (June 1982).
 37. Goldenfeld, N. *Lectures on phase transitions and the renormalization group* (Perseus books, Reading, Massachusetts, 1992).
 38. Landau, D. P. & Binder, K. *A guide to Monte Carlo simulations in statistical physics* 2nd ed. ISBN: 978-0-521-84238-9 (Cambridge University Press, Cambridge ; New York, 2005).
 39. Mata, A. S. & Ferreira, S. C. Pair Quenched Mean-Field Theory for the Susceptible-Infected-Susceptible Model on Complex Networks. *Europhysics Letters* **103**, 48003. ISSN: 0295-5075. (2023) (Sept. 2013).
 40. Karrer, B. & Newman, M. E. J. Message Passing Approach for General Epidemic Models. *Physical Review E* **82**, 016101. (2023) (July 2010).
 41. Lokhov, A. Y., Mézard, M., Ohta, H. & Zdeborová, L. Inferring the Origin of an Epidemic with a Dynamic Message-Passing Algorithm. *Physical Review E* **90**, 012801. (2023) (July 2014).
 42. Shrestha, M., Scarpino, S. V. & Moore, C. Message-Passing Approach for Recurrent-State Epidemic Models on Networks. *Physical Review E* **92**, 022821. (2023) (Aug. 2015).

-
43. Roudi, Y. & Hertz, J. Dynamical TAP Equations for Non-Equilibrium Ising Spin Glasses. *Journal of Statistical Mechanics: Theory and Experiment* **2011**, P03031. ISSN: 1742-5468. arXiv: 1103.1044 [cond-mat]. (2023) (Mar. 2011).
 44. Gewaltig, M.-O. & Diesmann, M. NEST (NEural Simulation Tool). **2**, 1430. <https://doi.org/10.4249/scholarpedia.1430> (2007).
 45. Erdős, P. & Rényi, A. On random graphs. **6**, 290–297 (1959).
 46. Castellano, C. & Pastor-Satorras, R. Relevance of Backtracking Paths in Recurrent-State Epidemic Spreading on Networks. *Physical Review E* **98**, 052313. (2023) (Nov. 2018).
 47. Silva, J. C. M., Silva, D. H., Rodrigues, F. A. & Ferreira, S. C. Comparison of Theoretical Approaches for Epidemic Processes with Waning Immunity in Complex Networks. *Physical Review E* **106**, 034317. ISSN: 2470-0045, 2470-0053. (2023) (Sept. 2022).
 48. Pastor-Satorras, R. & Vespignani, A. Epidemic Spreading in Scale-Free Networks. *Phys. Rev. Lett.* **86**, 3200–3203. <https://link.aps.org/doi/10.1103/PhysRevLett.86.3200> (14 2001).
 49. Martin, T., Zhang, X. & Newman, M. E. J. Localization and Centrality in Networks. *Physical Review E* **90**, 052808. ISSN: 1539-3755, 1550-2376. (2023) (Nov. 2014).
 50. Newman, M. E. J. in *Handbook of Graphs and Networks* 35–68 (John Wiley & Sons, Ltd, 2002). ISBN: 978-3-527-60275-9. (2023).
 51. Krasilnikov, A., Beregun, V. & Harmash, O. *Analysis of Estimation Errors of the Fifth and Sixth Order Cumulants in 2019 IEEE 39th International Conference on Electronics and Nanotechnology (ELNANO)* (Apr. 2019), 754–759.
 52. Beregun, V. & Harmash, O. *Application of Cumulant Coefficients for Solving the Problems of Testing and Diagnostics in Control Systems in 2018 IEEE 5th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC)* (Oct. 2018), 210–213.
 53. Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. Hybrid Monte Carlo. *Physics Letters B* **195**, 216–222. ISSN: 0370-2693. (2023) (Sept. 1987).
 54. Neal, R. M. MCMC Using Hamiltonian Dynamics. *arXiv:1206.1901 [physics, stat]*. arXiv: 1206.1901 [physics, stat]. (2017) (June 2012).
 55. Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [stat]*. arXiv: 1701.02434 [stat]. (2018) (Jan. 2017).
 56. Salvatier, J., Wiecki, T. V. & Fonnesbeck, C. Probabilistic Programming in Python Using PyMC3. *PeerJ Computer Science* **2**, e55. ISSN: 2376-5992. (2017) (Apr. 2016).
 57. Amit, D. J. *Field theory, the renormalization group, and critical phenomena* (World Scientific, 1984).
 58. Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine* **29**, 141–142 (2012).

-
59. Yu, J. J., Derpanis, K. G. & Brubaker, M. A. *Wavelet Flow: Fast Training of High Resolution Normalizing Flows* Oct. 2020. arXiv: 2010.13821 [cs]. (2023).
 60. Guth, F., Coste, S., De Bortoli, V. & Mallat, S. *Wavelet Score-Based Generative Modeling* Aug. 2022. arXiv: 2208.05003 [cs, stat]. (2023).
 61. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering Governing Equations from Data by Sparse Identification of Nonlinear Dynamical Systems. *Proceedings of the National Academy of Sciences* **113**, 3932–3937. (2023) (Apr. 2016).
 62. Casadiego, J., Nitzan, M., Hallerberg, S. & Timme, M. Model-Free Inference of Direct Network Interactions from Nonlinear Collective Dynamics. *Nature Communications* **8**, 2192. ISSN: 2041-1723. (2023) (Dec. 2017).
 63. Miller, J., Tang, S., Zhong, M. & Maggioni, M. *Learning Theory for Inferring Interaction Kernels in Second-Order Interacting Agent Systems* Oct. 2020. arXiv: 2010.03729 [cs, math, stat]. (2023).
 64. Lu, F., Maggioni, M. & Tang, S. Learning Interaction Kernels in Heterogeneous Systems of Agents from Multiple Trajectories. *ArXiv*. (2023) (Oct. 2019).
 65. Jordan, J., Schmidt, M., Senn, W. & Petrovici, M. A. Evolving Interpretable Plasticity for Spiking Networks. *eLife* **10** (eds van Rossum, M. C., Frank, M. J. & Sprekeler, H.) e66273. ISSN: 2050-084X. (2023) (Oct. 2021).
 66. Confavreux, B., Zenke, F., Agnes, E., Lillicrap, T. & Vogels, T. *A Meta-Learning Approach to (Re)Discover Plasticity Rules That Carve a Desired Function into a Neural Network* in *Advances in Neural Information Processing Systems* **33** (Curran Associates, Inc., 2020), 16398–16408. (2023).
 67. Opper, M. & Sanguinetti, G. *Variational Inference for Markov Jump Processes* in *Advances in Neural Information Processing Systems* **20** (Curran Associates, Inc., 2007). (2023).
 68. Opper, M. Variational Inference for Stochastic Differential Equations. *Annalen der Physik* **531**, 1800233. ISSN: 1521-3889. (2023) (2019).
 69. Vrettas, M. D., Opper, M. & Cornford, D. Variational Mean-Field Algorithm for Efficient Inference in Large Systems of Stochastic Differential Equations. *Physical Review E* **91**, 012148. (2023) (Jan. 2015).
 70. Ruttor, A. & Opper, M. Efficient Statistical Inference for Stochastic Reaction Processes. *Physical Review Letters* **103**, 230601. (2023) (Dec. 2009).
 71. Zache, T. V., Schweigler, T., Erne, S., Schmiedmayer, J. & Berges, J. Extracting the Field Theory Description of a Quantum Many-Body System from Experimental Data. *Physical Review X* **10**, 011020. ISSN: 2160-3308. (2023) (Jan. 2020).
 72. Mastromatteo, I. Beyond Inverse Ising Model: Structure of the Analytical Solution. *Journal of Statistical Physics* **150** (Sept. 2012).
 73. Decelle, A., Hwang, S., Rocchi, J. & Tantari, D. Inverse Problems for Structured Datasets Using Parallel TAP Equations and Restricted Boltzmann Machines. *Scientific Reports* **11**, 19990. ISSN: 2045-2322. (2023) (Oct. 2021).

-
74. Cocco, S., Monasson, R. & Sessak, V. High-Dimensional Inference with the Generalized Hopfield Model: Principal Component Analysis and Corrections. *Physical Review E* **83**, 051123. ISSN: 1539-3755, 1550-2376. (2023) (May 2011).
 75. Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. A Learning Algorithm for Boltzmann Machines. *Cognitive Science* **9**, 147–169. ISSN: 0364-0213. (2023) (Jan. 1985).
 76. Mezard, M. & Sakellariou, J. Exact Mean Field Inference in Asymmetric Kinetic Ising Systems. *Journal of Statistical Mechanics: Theory and Experiment* **2011**, L07002. ISSN: 1742-5468. arXiv: 1103.3433 [cond-mat]. (2023) (July 2011).
 77. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population. *Nature* **440**, 1007–1012. ISSN: 1476-4687. (2023) (Apr. 2006).
 78. Zeng, H.-L., Aurell, E., Alava, M. & Mahmoudi, H. Network Inference Using Asynchronously Updated Kinetic Ising Model. *Physical Review E* **83**, 041135. ISSN: 1539-3755, 1550-2376. (2023) (Apr. 2011).
 79. Agliari, E. *et al.* A Statistical Inference Approach to Reconstruct Intercellular Interactions in Cell Migration Experiments. *Science Advances* **6**, eaay2103. (2023) (Mar. 2020).
 80. Zeng, H.-L., Aurell, E., Alava, M. & Mahmoudi, H. Network inference using asynchronously updated kinetic Ising model. **83**, 041135 (2011).
 81. Roudi, Y. & Hertz, J. Mean field theory for nonequilibrium network reconstruction. **106**, 048702 (2011).
 82. Baldassi, C., Gerace, F., Saglietti, L. & Zecchina, R. From Inverse Problems to Learning: A Statistical Mechanics Approach. *Journal of Physics: Conference Series* **955**, 012001. ISSN: 1742-6588, 1742-6596. (2023) (Jan. 2018).
 83. Sakellariou, J., Roudi, Y., Mezard, M. & Hertz, J. Effect of Coupling Asymmetry on Mean-Field Solutions of Direct and Inverse Sherrington-Kirkpatrick Model. *Philosophical Magazine* **92**, 272–279. ISSN: 1478-6435, 1478-6443. arXiv: 1106.0452 [cond-mat]. (2023) (Jan. 2012).
 84. Braunstein, A., Ingrosso, A. & Muntoni, A. P. Network Reconstruction from Infection Cascades. *Journal of The Royal Society Interface* **16**, 20180844. ISSN: 1742-5689, 1742-5662. (2023) (Feb. 2019).
 85. Mora, T., Walczak, A. M., Bialek, W. & Callan, C. G. Maximum Entropy Models for Antibody Diversity. *Proceedings of the National Academy of Sciences* **107**, 5405–5410. (2023) (Mar. 2010).
 86. Bachschmid-Romano, L. & Opper, M. A Statistical Physics Approach to Learning Curves for the Inverse Ising Problem. *Journal of Statistical Mechanics: Theory and Experiment* **2017**, 063406. ISSN: 1742-5468. (2023) (June 2017).
 87. Berg, J. Statistical Mechanics of the Inverse Ising Problem and the Optimal Objective Function. *Journal of Statistical Mechanics: Theory and Experiment* **2017**, 083402. ISSN: 1742-5468. (2023) (Aug. 2017).

-
88. Aurell, E. & Ekeberg, M. Inverse Ising Inference Using All the Data. *Physical Review Letters* **108**, 090201. (2023) (Mar. 2012).
 89. Abbara, A., Kabashima, Y., Obuchi, T. & Xu, Y. Learning Performance in Inverse Ising Problems with Sparse Teacher Couplings. *Journal of Statistical Mechanics: Theory and Experiment* **2020**, 073402. ISSN: 1742-5468. (2023) (July 2020).
 90. Lokhov, A. Y., Vuffray, M., Misra, S. & Chertkov, M. Optimal Structure and Parameter Learning of Ising Models. *Science advances* **4**. ISSN: 2375-2548. (2023) (Mar. 2018).
 91. Gonçalves, P. J. *et al.* Training Deep Neural Density Estimators to Identify Mechanistic Models of Neural Dynamics. *eLife* **9** (eds Huguenard, J. R., O’Leary, T. & Goldman, M. S.) e56261. ISSN: 2050-084X. (2023) (Sept. 2020).
 92. Ardizzone, L. *et al.* Analyzing Inverse Problems with Invertible Neural Networks. *arXiv:1808.04730 [cs, stat]*. arXiv: 1808.04730 [cs, stat]. (2021) (Feb. 2019).
 93. Gökmen, D. E., Ringel, Z., Huber, S. D. & Koch-Janusz, M. Statistical Physics through the Lens of Real-Space Mutual Information. *arXiv:2101.11633 [cond-mat]*. arXiv: 2101.11633 [cond-mat]. (2021) (Mar. 2021).
 94. Cranmer, M. *et al.* *Discovering Symbolic Models from Deep Learning with Inductive Biases in Advances in Neural Information Processing Systems* (eds Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H.) **33** (Curran Associates, Inc., 2020), 17429–17442.
 95. Castellano, C. & Pastor-Satorras, R. Cumulative Merging Percolation and the Epidemic Transition of the Susceptible-Infected-Susceptible Model in Networks. *Physical Review X* **10**, 011070. ISSN: 2160-3308. (2023) (Mar. 2020).
 96. Katzgraber, H. G., Trebst, S., Huse, D. A. & Troyer, M. Feedback-optimized parallel tempering Monte Carlo. *Journal of Statistical Mechanics: Theory and Experiment* **2006**. Publisher: IOP Publishing, P03018–P03018. <https://doi.org/10.1088%2F1742-5468%2F2006%2F03%2Fp03018> (Mar. 2006).
 97. Georges, A. & Yedidia, J. S. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General* **24**, 2173. <http://stacks.iop.org/0305-4470/24/i=9/a=024> (1991).
 98. Kühn, T. & Helias, M. Expansion of the effective action around non-Gaussian theories. *Journal of Physics A: Mathematical and Theoretical* **51**, 375004. <http://stacks.iop.org/1751-8121/51/i=37/a=375004> (2018).
 99. Kappen, H. J. & Spanjers, J. J. Mean field theory for asymmetric neural networks. *pre* **61**, 5658–5663 (2000).
 100. Grytskyy, D., Tetzlaff, T., Diesmann, M. & Helias, M. A unified view on weakly correlated recurrent networks. **7**, 131 (2013).
 101. Strassen, V. Gaussian Elimination Is Not Optimal. *Numerische Mathematik* **13**, 354–356. ISSN: 0945-3245. (2023) (Aug. 1969).

102. Schatz, M. D., Low, T. M., van de Geijn, R. A. & Kolda, T. G. Exploiting Symmetry in Tensors for High Performance: Multiplication with Symmetric Tensors. *SIAM Journal on Scientific Computing* **36**, C453–C479. ISSN: 1064-8275. (2023) (Jan. 2014).
103. Hoffman, M. D. & Gelman, A. The No-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623. (2017) (2014).
104. Loaiza-Ganem, G., Ross, B. L., Cresswell, J. C. & Caterini, A. L. *Diagnosing and Fixing Manifold Overfitting in Deep Generative Models* Apr. 2022. arXiv: 2204.07172 [cs, stat]. (2022).
105. Cornish, R., Caterini, A. L., Deligiannidis, G. & Doucet, A. *Relaxing Bijectivity Constraints with Continuously Indexed Normalising Flows* in *International Conference on Machine Learning* (Sept. 2019). (2023).

Acknowledgements

It is hardly surprising to find that writing a PhD thesis is a highly challenging task. It is more surprising to me, that I found it to be one of the most interesting and most rewarding times of my life. For this, I am thankful first and foremost to Moritz Helias and Carsten Honerkamp while providing invaluable advice on all levels. In particular, I wish to thank Moritz Helias for being involved in all aspects of the scientific work, from defeating computational challenges to asking the fundamental questions. I also thank both of you for for affording me a great deal of scientific freedom, and support throughout the thesis and the pandemic.

It is an immense privilege it is to be able to pursue scientific research even in a global crisis. This experience also allows me to appreciate much more all the interactions, pairwise and higher-order, which I have since had with my colleagues both in the Physics Department at RWTH Aachen and INM-6. I wish to express my appreciation of all members of the two institutions, and the following list is surely incomplete.

I am grateful to Lars Schutzeichel, Javed Lindner, Bastian Epping, Geonho Han, Alexeander Herbort and Matthias Klaus, who trained models in the same office, laughed at our jokes, asked for "Erklärungen", and reminded me that it is better to do everything right from the beginning. In addition to Michael Dick, Jonas Oberste-Frielingshaus, for many funny bus rides and to José Villamar for not throwing us out the window. Furthermore to Peter Bouss and Sandra Nestler, for normalizing flows of thought, to David Dahmen, Christian Keup, Tobias Kühn, Lorenzo Tiberi, Moritz Layer, Jakob Stubenrauch, Matthieu Gilson and Alex van Meegen, for sharing their deep knowledge and intuition, to Anno Kurth for mental sparring, to Kirsten Fischer for intriguing conversations and advice, to Max Wollgarten and Jannik Grundler for enthusiasm, creativity, and resilience, to Kai Segadlo for patience, to Heather More and Alex Kleinjohann for inspiring bike rides, to Mara Caltapandides for "Verkantung", to Florian Kischel and Matt Bunney for random walks, to Benedikt Kalthoff for limits of humor, and to Jonas Reimann, Paul Brehmer and Takuya Okugawa for diving into cold water.

Finally, I wish to thank Nicola, Lio, Roland and Christel for everything.

Eidesstattliche Erklärung

Claudia Lioba Merger erklärt hiermit, dass diese Dissertation und die darin dargelegten Inhalte die eigenen sind und selbstständig, als Ergebnis der eigenen originären Forschung, generiert wurden. Hiermit erkläre ich an Eides statt

1. Diese Arbeit wurde vollständig oder größtenteils in der Phase als Doktorand dieser Fakultät und Universität angefertigt;
2. Sofern irgendein Bestandteil dieser Dissertation zuvor für einen akademischen Abschluss oder eine andere Qualifikation an dieser oder einer anderen Institution verwendet wurde, wurde dies klar angezeigt;
3. Wenn immer andere eigene- oder Veröffentlichungen Dritter herangezogen wurden, wurden diese klar benannt;
4. Wenn aus anderen eigenen- oder Veröffentlichungen Dritter zitiert wurde, wurde stets die Quelle hierfür angegeben. Diese Dissertation ist vollständig meine eigene Arbeit, mit der Ausnahme solcher Zitate;
5. Alle wesentlichen Quellen von Unterstützung wurden benannt;
6. Wenn immer ein Teil dieser Dissertation auf der Zusammenarbeit mit anderen basiert, wurde von mir klar gekennzeichnet, was von anderen und was von mir selbst erarbeitet wurde;
7. Ein Teil oder Teile dieser Arbeit wurden zuvor veröffentlicht und zwar in:
 - Merger, C., Reinartz, T., Wessel, S., Honerkamp, C., Schuppert, A., Helias, M., 2021. Global hierarchy vs local structure: Spurious self-feedback in scale-free networks. *Phys. Rev. Res.* 3, 033272. <https://doi.org/10.1103/PhysRevResearch.3.033272>
 - Merger, C., René, A., Fischer, K., Bouss, P., Nestler, S., Dahmen, D., Honerkamp, C., Helias, M., 2023. Learning Interacting Theories from Data. <https://doi.org/10.48550/arXiv.2304.00599>
 - Merger, C., Albers, J., Honerkamp, C., Helias, M., 2023. Spurious self-feedback of mean-field predictions inflates infection curves, *in preparation*

Datum:

Claudia Merger