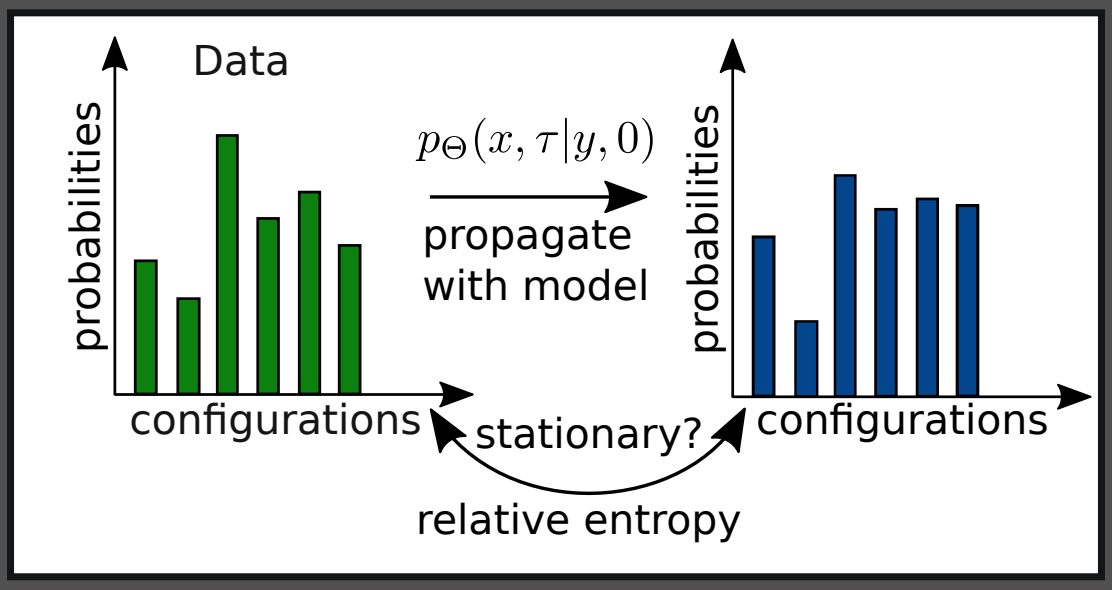
Stochastic inference from snapshots of the non-equilibrium steady state: the asymmetric Ising model and beyond

Dissertation
Simon Lee Dettmer



Köln 2017

Stochastic inference from snapshots of the non-equilibrium steady state: the asymmetric Ising model and beyond

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Simon Lee Dettmer

aus Oberhausen

Köln, 2017

Berichterstatter: Prof. Dr. Johannes Berg
Prof. David Gross, PhD

Tag der mündlichen Prüfung: 05. Oktober 2017

Kurzzusammenfassung

In dieser Arbeit untersuchen wir das Problem der Parameter-Inferenz für ergodische Markov-Prozesse, die gegen einen stationären Zustand konvergieren, der nicht durch die Boltzmann-Verteilung beschrieben wird. Unser Hauptergebnis ist, dass wir die Parameter verschiedener Modelle auf der Grundlage von unabhängigen Stichproben aus dem stationären Zustand lernen können, obwohl wir die stationäre Wahrscheinlichkeitsverteilung nicht kennen. Genauer: für die untersuchten Modelle in stetiger Zeit konnten wir die Parameter bis auf einen Skalierungsfaktor inferieren, welcher die Zeitskala bestimmt, die natürlich nicht aus statischen Messungen ermittelt werden kann; bei Modellen in diskreter Zeit ist die Zeitskala bereits implizit durch die Diskretisierung gewählt und wir konnten alle Parameter der untersuchten Modelle inferieren. Als Paradigma für Nicht-Gleichgewichts-Prozesse untersuchen wir das asymmetrische Ising-Modell mit Glauber-Dynamik. Es beschreibt binäre Spinvariablen mit asymmetrischen paarweisen Wechselwirkungen unter dem Einfluss äußerer Magnetfelder. Diese Magnetfelder und Wechselwirkungsstärken wollen wir lernen. Zu diesem Zweck haben wir in dieser Arbeit zwei verschiedene Inferenzmethoden entwickelt: die erste Methode basiert auf der Berechnung von Magnetisierungen, sowie Zwei- und Dreipunkt-Spin-Korrelationen, zum einen in einer selbstkonsistenten Form, die exakt ist, und zum anderen in einer geschlossenen Form innerhalb einer Molekularfeld-Näherung; die zweite Methode beruht auf der Maximierung einer Funktion, die wir "propagator likelihood" nennen. Diese betrachtet fiktive Übergänge zwischen allen gemessenen Konfigurationen und ist verwandt mit der bekannten Log-Likelihood-Funktion für Gleichgewichts-Systeme. Der Vorteil des Molekularfeld-Ansatzes ist sein vergleichbar geringer numerischer Aufwand, während der Vorteil des "propagator likelihood"-Verfahrens darin besteht, dass es die gesamte empirische Verteilung verwendet und leicht auf jeden ergodischen Markov-Prozess angewandt werden kann. Insbesondere wenden wir die "propagator likelihood"-Methode auf weitere bekannte Nicht-Gleichgewichtsmodelle aus der Statistischen Physik und der Theoretischen Biologie an: den einfachen asymmetrischen Exklusionsprozess (ASEP) in stetiger Zeit mit diskreten Konfigurationen, sowie die Replikatordynamik in stetiger Zeit mit kontinuierlichen Konfigurationen. Die Allgemeingültigkeit des "propagator likelihood"-Ansatzes wird dadurch betont, dass er direkt aus dem Prinzip hergeleitet werden kann, dass die gemessene Verteilung stationär unter der Dynamik sein soll, das heißt wir minimieren die relative Entropie zwischen der empirischen Verteilung und einer Verteilung, die durch eine Zeitentwicklung dieser empirischen Verteilung erzeugt wird. Schließlich untersuchen wir noch eine etwas andere Situation und zeigen wie die Inferenz im asymmetrischen Ising-Modell verbessert werden kann, wenn wir mehrere Datensätze aus unabhängigen Stichproben von verschiedenen stationären Zuständen haben, die durch kontrollierte Störungen der zugrunde liegenden Modellparameter erzeugt werden.

Abstract

In this thesis we study the problem of inferring the parameters of ergodic Markov processes that converge to a non-equilibrium steady state. Our main result is that for many models, we can learn the parameters based on independent samples taken from the steady state, even though we do not know the stationary probability distribution. To be more precise: for the investigated models in continuous time, we could infer the parameters up to a factor that defines the time scale, which, naturally, cannot be determined from static measurements; for the investigated models in discrete time, the time scale is already chosen implicitly by the discretisation and we could infer all parameters. As our main paradigm for non-equilibrium inference problems, we study the asymmetric Ising model with Glauber dynamics. It consists of binary spins subject to external fields and asymmetric pairwise spin-couplings, which we seek to infer. For this purpose we have developed two different inference methods: the first method is based on computing magnetisations, two- and three-point spin correlations, either in a self-consistent form that is exact, or in a closed form within a mean field approximation; the second method is based on maximising a "propagator likelihood", which considers fictitious transitions between all sampled configurations and is akin to the well-known log-likelihood function used for equilibrium systems. The advantage of the mean field approach is its computational efficiency, while the advantage of the propagator likelihood method is that it uses information from the full sampled distribution and can easily be applied to any ergodic Markov process. In particular, we apply the propagator likelihood method to other prominent non-equilibrium models from statistical physics and theoretical biology: (i) the asymmetric simple exclusion process (ASEP) in continuous time with discrete configurations and (ii) replicator dynamics in continuous time with continuous configurations. The generality of this approach is emphasised by the fact that we can derive the propagator likelihood directly from the principle that the sampled distribution should be stationary under the model dynamics: we minimise the relative entropy between the sampled distribution and a distribution generated by propagating the sampled distribution in time. Finally, we investigate a slightly different setting: we show how inference can be improved in the asymmetric Ising model by considering multiple sets of independent samples taken from several steady states, which are generated by controlled perturbations of the underlying model parameters.

Acknowledgements

I would like to acknowledge Chau Nguyen, not only for his collaboration and thoughtful comments concerning the project on mean field inference, but also for the enjoyable experience of teaching statistical physics classes together, where in numerous discussions he was so kind to share with me his thorough understanding of the subject. Further, I owe many thanks to my supervisor, Johannes Berg, not only for presenting me with the problem of non-equilibrium inference, but also for his continuous support, time, and faith, all of which he gave generously.

Contents

1 Introduction

To begin at the beginning

Dylan Thomas

1.1 Thesis overview

This thesis addresses the problem of inferring the parameters of ergodic Markov processes based on independent samples taken from the non-equilibrium steady state.

In this first chapter, we recall the standard results on ergodic Markov processes concerning the convergence to steady states and their classification into equilibrium and non-equilibrium steady states. We then formulate our stochastic inference problem and give a brief overview of established inference methods for equilibrium steady states and for time-series data. First, we motivate the maximum likelihood method within the framework of Bayesian reasoning, before briefly mentioning the equilibrium mean field approximation and the pseudo-likelihood method. Second, we discuss inference based on maximum likelihood for time series.

In chapter 2, we give some background on our main paradigm for non-equilibrium inference problems: the asymmetric Ising model. We will introduce Glauber dynamics and show that this dynamics converges to a non-equilibrium steady state for the case of asymmetric couplings between spins; we motivate the consideration of asymmetric couplings by briefly discussing the connection of the asymmetric Ising model with neural networks. In the following, we describe Callen's identities characterising the spin moments, since they will be used for inference in chapters 3 and 5. We discuss maximum likelihood inference based on time-series data and present some minor results we found for inference in sequential Glauber dynamics, before presenting the Gaussian mean field theory of Mézard and Sakellariou (2011), which we will use for inference in chapter 5.

In chapter 3, we develop our first method for stochastic inference from snapshots of the steady state, which is based on fitting sampled observables to self-consistent equations, which we derive as generalisations of Callen's identities. We show how these self-consistent equations can be used to infer model parameters by replacing steady state expectation values with sample averages. In the following, we discuss how to approximately evaluate the self-consistent equations in a closed-form within an expansion around non-equilibrium mean field theory. The presentation of this expansion was inspired by the work of Kappen

and Spanjers (2000), who developed the non-equilibrium mean field theory for the asymmetric Ising model. Here, we provide a straightforward generalisation of their theory and formulate it for a wider class of ergodic Markov processes. Finally, we use these methods to address the stochastic inference problem for the asymmetric Ising model.

In chapter 4, we develop our second inference method, based on maximising a function we call the propagator likelihood. We give a derivation of this function based on minimising relative entropy and illustrate the method for several toy models spanning the different classes of Markov processes, including the Ornstein-Uhlenbeck process and the asymmetric simple exclusion process (ASEP). Then we use the method to infer the parameters of more challenging models: the asymmetric Ising model (again) and replicator dynamics.

In chapter 5, we consider a slightly different setting and investigate how inference in the asymmetric Ising model can be improved by considering multiple sets of independent samples, which are taken from several steady states generated by known perturbations of the underlying parameters. We begin with some general considerations concerning the observables required for a well-defined inference problem and discuss the different roles of perturbations of the external fields and perturbations of the couplings. Next, we develop a simple inference algorithm based on the expressions for magnetisations and two-point correlations obtained in the non-equilibrium mean field theory of chapter 3 and discuss some basic properties of the approach. We follow with a more powerful inference method based on the Gaussian mean field theory of Mézard and Sakellariou (2011), which we use to derive self-consistent equations for the equal-time two-point correlations. In the case of vanishing external fields, these equations become linear in the couplings and allow for a computationally highly efficient inference algorithm that can easily be scaled to large system sizes. We investigate the performance of this method by considering an example where half of the couplings is set to zero in the perturbation and compare the approach to the setting considered chapters 3 and 4.

Finally, in chapter 6 we summarise and interpret our results in addition to giving a perspective on possible future directions for research. Section 3.2, appendix B, and parts of appendix A were previously published in (Dettmer et al., 2016); chapter 4 has appeared in (Dettmer and Berg, 2017).

1.2 Markov processes

A stochastic process $\{X(t)\}$ is a sequence of random variables $\{X(t)\}_{t\in I}$, where the index t denotes time, which could be discrete, $I = \{0, 1, \ldots, T\}$, or continuous, $I = [0, T]$, with a possibly infinite time horizon $T = \infty$. Examples for the ran-

dom variables could be the continuous-time positions of a set of gas molecules, the daily temperature at noon on the roof of Cologne Cathedral, or the weekly draw of lottery numbers. Due to the randomness of the variables we cannot make definite predictions about outcomes, but instead have to content ourselves with statements about the probabilities of different outcomes. This probability may be interpreted as a subjective belief concerning different events occurring (Bayesian interpretation) or as their relative frequencies in the limit of a large ensemble of copies of the process, each taking a different (random) realisation (frequentist interpretation).

In general, there will be relationships connecting the different variables, e.g. given that today the temperature at Cologne Cathedral is $21°C$, it is highly unlikely that tomorrow the temperature will be $-10°C$. These relationships can be captured by conditional probabilities, which tell us how the observed realisation of the stochastic process until time t_1 influences the probability of some event A taking place at a later time $t_2 > t_1$. The inter-dependence of the random variables may be arbitrarily complicated, however, for many applications we can focus on classes of processes with very simple relationships. The simplest case is when the variables are statistically independent, e.g. knowledge of the past draws of lottery numbers does not influence the probabilities of particular numbers appearing in next week's draw. This case will not be discussed in this thesis. For processes with real inter-dependencies between the variables, the simplest case is when the probability of the future event A depends on the past history of the process only via the present state $X(t_1)$, i.e.

Figure 1.1: *Andrei Andreyevich Markov, who researched the stochastic processes nowadays named after him, was less interested in physical applications of these processes but instead preferred to use them for studying poetry.*

$$P(X(t_2) \in A | \{X(s)\}_{s \leq t_1}) = P(X(t_2) \in A | X(t_1)) \ \forall t_2 > t_1 \ , \tag{1.1}$$

which is known as the **Markov property**. Sequences of random variables obeying the Markov property are known as Markov processes, in recognition of Andrei Markov who studied these processes for the purpose of extending the weak law of large numbers to random variables that are not statistically inde-

pendent (Markov, 2006; Seneta, 2006). The reason for the widespread use of Markov processes is not just their simplicity, but for many real-world processes it can be argued that the Markov property should hold true with a high degree of accuracy. For example, in the kinetic theory of gases (see e.g. Redner et al. (2010)) the molecular chaos assumption argues that the trajectories $(x(t), v(t))$ of gas particles are effectively Markov processes, due to the great number of particle collisions occurring on time-scales much shorter than the observation time.

Of course not all real-world random processes are Markovian and whether a process obeys the Markov property depends also on the choice of variables. Consider a point-mass in classical mechanics described by its position x and momentum p. We know that knowledge of the current position $x(t)$ is not sufficient to predict the future trajectory of the particle so the position process $\{x(t)\}$ does not obey the Markov property. However, adding the momentum $p(t)$ yields the necessary information and the joint process $\{(x(t), p(t))\}$ is indeed a Markov process[1]. In fact, many probabilistic models with memory can be made Markov processes by adding auxiliary variables (see e.g. Lei et al. (2016)).

1.2.1 Discrete configurations: Markov chains

Markov processes can be classified by (i) whether time is discrete or continuous, and (ii) whether the configuration space $\Omega \ni X(t)$ is discrete or continuous. Markov processes with discrete configurations are called Markov chains. The theory is simplest for these Markov chains and for this reason we pick them as our starting point for an exposition of the standard results on Markov processes (see e.g. Feller (1968); Gardiner (2009); Grimmett and Stirzaker (2001); Klenke (2013); Levin and Peres (2008)) most pertinent to the framing our stochastic inference problem.

1.2.1.1 Discrete time

We consider a set of n possible configurations $\Omega = \{\omega_1, \ldots, \omega_n\}$, assumed by the random variables $X_0, X_1, X_2, \ldots$, where X_t is a short-hand for $X(t)$. As examples we can think of the energy levels assumed by a quantum harmonic oscillator, the number of particles present in a subsystem connected to a reservoir of chemical potential μ. The discretisation of time could correspond to measurements taking place at fixed time intervals. In this case, we can use the Markov

[1] A free point-mass would of course obey deterministic dynamics, which can be considered a limiting case of random processes. By adding interactions with a heat bath we can introduce randomness into the process and the same statement applies.

property to iteratively rewrite the joint probability distribution of a set of random variables $X_0, X_1, \ldots, X_k$ in terms of the single-step conditional probabilities $P(X_t = x_t | X_{t-1} = x_{t-1})$ as

$$\begin{aligned} P(X_k = x_k, \ldots, X_0 = x_0) =& P(X_k = x_k | X_{k-1} = x_{k-1}, \ldots, X_0 = x_0) \\ & \times P(X_{k-1} = x_{k-1}, \ldots, X_0 = x_0) \\ =& P(X_k = x_k | X_{k-1} = x_{k-1}) P(X_{k-1} = x_{k-1}, \ldots, X_0 = x_0) \\ =& \ldots = P(X_0 = x_0) \prod_{t=1}^{k} P(X_t = x_t | X_{t-1} = x_{t-1}) \ . \end{aligned} \tag{1.2}$$

The conditional probabilities $P(X_t = x_t | X_{t-1} = x_{t-1})$ are also called transition probabilities. The most commonly studied Markov chains are **time-homogeneous chains**, where the transition probabilities do not depend on the time t of the transition. Hence, the process is fully described by the initial condition $P(X_0 = x_0)$ and the **matrix of transition probabilities**

$$T_{ij} := P(X_1 = \omega_j | X_0 = \omega_i) \ . \tag{1.3}$$

In particular, by summing over intermediate time-steps, the distribution of the random variable at time t, $p_i(t) := P(X_t = \omega_i)$, can be written as the matrix product of the initial distribution $p(0)$ and the transition matrix T taken to the power t

$$p_i(t) = [p(0)T^t]_i = \sum_{j=1}^{n} p_j(0)(T^t)_{ji} \ . \tag{1.4}$$

For the special case of deterministic initial conditions where the process starts in a configuration $x_0 \in \Omega$, i.e. $p_j(0) = \delta_{\omega_j, x_0}$, we reserve the notation

$$p(x,t|x_0,0) := P(X(t) = x | X(0) = x_0) \ , \tag{1.5}$$

which is called the **propagator**, since it takes the probability distribution at time 0 (concentrated in configuration x_0) and propagates this distribution forward in time to create the probability distribution at time t. Due to the linearity of the equations, we can write the solution for an arbitrary initial condition $p(0)$ as a sum over propagators

$$p_i(t) = \sum_{j=1}^{n} p(\omega_i, t | \omega_j, 0) p_j(0) \ . \tag{1.6}$$

THE STEADY STATE AND CONVERGENCE OF THE MARKOV CHAIN

Under certain conditions on the transition matrix T, the single-time distribution $p(t)$ converges to a unique distribution π, called the steady state (or stationary

distribution), independent of the initial probability distribution $p(0)$. We call these chains **ergodic**. It is clear that the steady state has the property of remaining unchanged when propagating the distribution with the transition matrix

$$\pi = \pi T \ , \tag{1.7}$$

which we use as the definition of a steady state.

In some cases one or more steady states may exist, but the Markov chain might not converge for arbitrary initial conditions. For chains with finite configuration space, $|\Omega| < \infty$, the existence and uniqueness of the steady state is guaranteed by the condition of **irreducibility**, stating that the chain can reach any configuration from any starting point in a finite number of steps with positive transition probability. These are the most studied chains and all Markov chains considered in this thesis will be irreducible. Markov chains that are not irreducible can be divided into irreducible sub-chains and then only the transitions between the subclasses have to be accounted for additionally. For infinite configuration spaces, $|\Omega| = \infty$, irreducibility is not sufficient to ensure the existence of a normalisable steady state. In addition, we require that the average time of return to any initial configuration is finite. These chains are called **positive recurrent**. The Markov chain actually converges to the unique steady state, independent of the initial condition, if the chain is **aperiodic**. In an aperiodic chain, the possible paths that start from an initial configuration and return to the starting point must not have a common divisor to their number of steps. An example for a periodic chain is the simple random walk on $\mathbb{Z}$, where the chain hops one place to the left or one place to to right in every time step, so the chain can return to a configuration only after an even number of steps.

The results above on the convergence of Markov chains have been known for a long time. More recently, people have studied how long the Markov chain actually takes to converge to the steady state and how this time depends on the size of the configuration space; this field is known as Markov mixing times (Levin and Peres, 2008).

EXAMPLE: BIASED RANDOM WALK ON $\mathbb{N}_0$

The biased random walk on $\mathbb{N}_0$ is most simply defined by a picture of the chain's configurations and transition probabilities (Fig. 1.2).

In each time step, the chain moves one place to the right with probability r, or one place to the left with probability $1-r$, except when the chain is in 0 where instead of moving to -1, the chain remains in 0. The chain is obviously irreducible and aperiodic, since the chain can remain in 0 for an arbitrary number of time steps. The chain therefore converges to a unique steady state if and only if the chain is positive recurrent. We can check this condition by actually

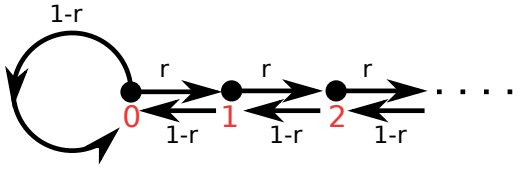


Figure 1.2: *Schematic view of the transition rules for the biased random walk on* $\mathbb{N}_0$.

computing the steady state. The steady state defining equation (1.7) becomes

$$r\pi_{i-1} + (1-r)\pi_{i+1} = \pi_i \ , \ i = 1,2,\ldots \tag{1.8}$$

$$(1-r)\pi_0 + (1-r)\pi_1 = \pi_0 \ . \tag{1.9}$$

These equations can be solved iteratively and we obtain

$$\pi_i = \left(\frac{r}{1-r}\right)^i \pi_0 \ . \tag{1.10}$$

The steady state is normalisable if and only if $r/(1-r) < 1 \Leftrightarrow r < 1/2$:

$$1 = \sum_{i=0}^{\infty} \pi_i = \sum_{i=0}^{\infty} \left(\frac{r}{1-r}\right)^i \pi_0 \overset{\frac{r}{1-r}<1}{=} \frac{1-r}{1-2r}\pi_0 \ . \tag{1.11}$$

Thus, for $r < 1/2$ the chain is positive recurrent and we have a normalisable steady state. It can be shown (Klenke, 2013) that (i) for $r > 1/2$ the chain is transient and wanders off to infinity so any configuration is visited only finitely many times; (ii) for $r = 1/2$ we have a null-recurrent chain, i.e. each configuration is visited infinitely often, but the mean return time is infinite. ■

1.2.1.2 Continuous time

Markov chains in continuous time have no memory of how long they have remained in a certain configuration. The transitions are therefore described by instantaneous **transition rates** $K_{ij}(t)$, giving the probability that the chain jumps from configuration ω_i to configuration ω_j within the infinitesimal time interval $[t, t+\mathrm{d}t)$. We can define them as the limit

$$K_{ij}(t) := \lim_{\delta t \to 0} P(X(t+\delta t) = \omega_j | X(t) = \omega_i)/\delta t \ . \tag{1.12}$$

As for the discrete-time Markov chains, we focus on time-homogeneous chains where the transition rates do not depend on time, $K_{ij}(t) \equiv K_{ij}$. The random time τ that the chain remains in a given configuration ω_i is then exponentially

distributed with parameter $\lambda_i = \sum_{j\neq i} K_{ij} > 0$ and we can define a corresponding transition matrix in discrete time as

$$T_{ij} = \begin{cases} K_{ij}/\lambda_i & i \neq j \\ 0 & i = j \end{cases} , \tag{1.13}$$

where we count time in units of the (random) jump times $T_1, T_2, \ldots$.

Let us consider how the single-time probability $P(X(t) = \omega_i) =: p_i(t)$ changes within an infinitesimal time interval $\mathrm{d}t$. First, the probability $\sum_{j\neq i} p_i(t)K_{ij}\mathrm{d}t$ flows out via the jumps from ω_i to other configurations. Second, the probability $\sum_{j\neq i} p_j(t)K_{ji}\mathrm{d}t$ flows in due to jumps from other configurations to ω_i. Adding the two and dividing by $\mathrm{d}t$ we find the **Master equation**

$$\frac{\mathrm{d}}{\mathrm{d}t} p_i(t) = -\sum_{j\neq i} p_i(t)K_{ij} + \sum_{j\neq i} p_j(t)K_{ji} , \tag{1.14}$$

which is a set of ordinary differential equations describing the time-evolution of the vector of single-time probabilities $p_i(t)$. Defining the new matrix

$$\tilde{K}_{ij} := \begin{cases} K_{ij} & i \neq j \\ -\lambda_i & i = j \end{cases} , \tag{1.15}$$

the Master equation can be written as $\frac{\mathrm{d}}{\mathrm{d}t} p = p\tilde{K}$ and the solution takes the form

$$p_i(t) = \left[p(0) e^{\tilde{K}t} \right]_i , \tag{1.16}$$

analogous to the case of discrete time. We only take the matrix exponential rather than the matrix power. Again, we define the propagator $p(x,t|x_0,0) = P(X(t) = x|X(0) = x_0)$ as the solution for the deterministic initial condition with the chain starting in $x_0 \in \Omega$. In the steady state there should be no net flow of probability in or out of any configuration. The steady state is therefore characterised by the equation

$$\pi\tilde{K} = 0 . \tag{1.17}$$

Since the jumping times are continuous, the Markov chain is automatically aperiodic. The chain is irreducible if and only if $P(X(t) = j|X(0) = i) > 0$ for all pairs $i \neq j$ and any time $t > 0$, which is equivalent to the statement

$$\left(e^{\tilde{K}} \right)_{ij} > 0 \; \forall \; i \neq j . \tag{1.18}$$

The existence of a normalisable steady state and convergence of the chain are equivalent to the chain being positive recurrent, as in discrete time.

EXAMPLE: RANDOM TELEGRAPH PROCESS

In the random telegraph process, the Markov chain can take only two possible configurations $X(t) \in \Omega = \{0,1\}$. The chain jumps from 0 to 1 with rate $\alpha := K_{01}$ and from 1 to 0 with rate $\beta := K_{10}$. The Master equation reads

$$\frac{\mathrm{d}}{\mathrm{d}t}p_0(t) = -\alpha p_0(t) + \beta p_1(t) \tag{1.19}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}p_1(t) = -\beta p_1(t) + \alpha p_0(t)\ . \tag{1.20}$$

For $\alpha > 0$ and $\beta > 0$ the process is irreducible and we know the chain must converge to a steady state. We can compute the full time-dependent solution for this simple process: due to normalisation we have $p_1(t) = 1 - p_0(t)$ and it suffices to solve the single differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}p_0(t) = -\alpha p_0(t) + \beta[1 - p_0(t)]\ , \tag{1.21}$$

which gives

$$p_0(t) = \left(p_0(0) - \frac{\beta}{\alpha+\beta}\right) e^{-(\alpha+\beta)t} + \frac{\beta}{\alpha+\beta}\ . \tag{1.22}$$

For $t \to \infty$ the probabilities $(p_0(t), p_1(t) = 1 - p_0(t))$ given by (1.22) converge to the steady state $\pi_0 = \frac{\beta}{\alpha+\beta}, \pi_1 = \frac{\alpha}{\alpha+\beta}$. ■

1.2.2 Continuous configurations

A second category of Markov processes describes the time-evolution of continuous variables $X(t) \in \Omega \subset \mathbb{R}^d$. An example would be the positions and momenta of N gas molecules in a box. In discrete time, we might consider the random walk $X_n = \sum_{i=1}^{n} Y_i$ with statistically independent, identically distributed increments Y_i, which take continuous values. In the case where the increments have an infinite variance, these random walks are called Lévy flights. In this thesis, we will focus on processes in continuous time with continuous sample paths and finite variance[1], since they have a simple characterisation, which we describe below.

[1]This restriction could be relaxed by including jump rates $K(x'|x,t) = \lim_{\delta t \to 0} P(X(t+\delta t) = x'|X(t) = x)/\delta t$ analogous to the Markov chains in continuous time.

1.2.2.1 Brownian motion

The study of Markov processes with continuous configurations has been pioneered by the study of a particular process known as Brownian motion. Mathematically, it was first studied by Louis Bachelier in the context of stock markets (Bachelier, 1900) and later by Albert Einstein (Einstein, 1905), Marian Smoluchowski (von Smoluchowski, 1906) and Paul Langevin (Lemons and Gythiel, 1997) in the context of diffusing molecules, which we refer to as physical Brownian motion. Today, Brownian motion forms a major pillar on which the theory of more general continuous Markov processes is founded. Its mathematical basis has been made rigorous by Norbert Wiener (Wiener, 1923).

In short, we can characterise the mathematical (one-dimensional) Brownian motion as a Markov process $\{W(t)\}$ that

- starts at the origin $W(0) = 0$,
- has continuous sample paths, and
- increments $W(t+s) - W(t)$ that are statistically independent from the process $(W(\tau))_{\tau<t}$ and normally distributed with zero mean and variance s.

A d-dimensional Brownian motion is simply defined as a vector with d components that are independent one-dimensional Brownian motions.

There are two equivalent approaches to continuous Markov processes. The first approach, known as Langevin equations, generalises the Newtonian equations of motion to include a random-force emanating from the dynamics of a large number of unobserved microscopic degrees of freedom. The second approach directly describes the time-evolution of the probability density of configurations in terms of a partial differential equation: the Fokker-Planck equation. We will explore both approaches and their connection in the following.

1.2.2.2 Langevin equations

Historically, Langevin's development of his stochastic differential equations succeeded the treatment of Brownian motion by Einstein and Smoluchowski. However, because of its intuitive simplicity, we first take a look at Langevin's equations. This simplicity was bought at the price of lacking mathematical rigour, which was later provided by the stochastic calculus of Kiyosi Itô (Itô, 1944, 1946).

At its heart, Langevin's treatment is based on the separation of variables into slowly varying ones, which we track, and rapidly varying ones, which we do not track explicitly. In the context of physical Brownian motion, the slow variables are the position x and velocity v of a colloidal particle, which has a

size on the order of microns; the fast variables are the positions and velocities of an immense number of water molecules, which have a size on the order of angstroms. The collective effect of the collisions of water molecules with the colloid is a random force, which can be separated into its deterministic mean F and random fluctuations $\tilde{\eta}$ around the mean. The (one-dimensional) dynamics of the Brownian particle are then described by the set of differential equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = v(t) \tag{1.23}$$

$$m\frac{\mathrm{d}v}{\mathrm{d}t} = F(x(t), v(t)) + \tilde{\eta}(t) \ . \tag{1.24}$$

The first equation is simply the definition of the particle position as the time-integral over its velocity, the second equation is the generalisation of Newton's third law to a stochastic differential equation known as Langevin equation.

Figure 1.3: *Paul Langevin, inventor of stochastic differential equations, is also known for his work on paramagnetism, ultrasonic detection of submarines, and creating the twin paradox of special relativity. Besides his courageous step into mathematically murky waters when devising his stochastic differential equations, he also boldly challenged the editor Téry to a duel in response to the latter publicising Langevin's affair with his former PhD supervisor's widow, Marie Curie. Luckily, no one was hurt, since Téry retreated in the last minute.*

THE OVERDAMPED LIMIT

In Langevin's and Einstein's treatment of Brownian motion, the particle is assumed to experience a viscous drag described by Stokes' law $F(x,v) = F(v) = -\zeta v$, where for spherical particles the drag coefficient is given by $\zeta = 6\pi\mu a$ with μ the viscosity of the solvent and a the radius of the Brownian particle. In the limit $m/\zeta \ll 1$ inertia becomes negligible compared to friction and the particle velocity directly follows the random force $\tilde{\eta}(t)$. Formally, by setting $m\frac{\mathrm{d}v}{\mathrm{d}t} = 0$ in (1.24) we obtain $v = \tilde{\eta}(t)/\zeta =: \eta(t)$ and therefore the particle position is described by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \eta(t) \ . \tag{1.25}$$

It is straightforward to generalise this argument to the case where the force has a second, position-dependent component $F(x,v) = \tilde{f}(x) - \zeta v$.

The dynamics of the particle position is then described by the equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{\tilde{f}(x)}{\zeta} + \eta(t) =: f(x) + \eta(t) \ . \tag{1.26}$$

When the force derives from a potential, $\tilde{f}(x) = -\partial_x \tilde{U}(x)$, we define the corresponding effective potential $U(x) = \tilde{U}(x)/\zeta$ that produces the effective force $f(x) = -\partial_x U(x)$.

Even though Langevin's treatment includes the case of finite mass, when diffusion or a Brownian particle are discussed, it is common to implicitly assume the overdamped limit.

THE STOKES-EINSTEIN RELATION AND GAUSSIAN WHITE NOISE

The irregularity of the random fluctuating force $\eta(t)$ makes this object somewhat pathological mathematically. While it is clear that the force must have zero mean, $\langle \eta(t) \rangle$, due to its definition as fluctuation around the mean force, it turns out that demanding that the force be uncorrelated from the Brownian particle position $x(t)$ and at the same time produce a finite variance of the particle position, its time-correlation should be a Dirac-Delta function

$$\langle \eta(t)\eta(t') \rangle = \sigma^2 \delta(t-t') \ . \tag{1.27}$$

A random fluctuating force $\xi(t) = \eta(t)/\sigma$ that has unit magnitude, i.e. $\langle \xi(t)\xi(t-t') \rangle = \delta(t-t')$, is known as Gaussian white noise. For a Brownian particle in equilibrium, the magnitude σ of the fluctuations is determined by the equipartition theorem, $\langle mv^2 \rangle = k_B T$. To this end, one can show that for a particle initially at rest, $v(0) = 0$, the velocity solving (1.24) has zero mean and a variance given by

$$\langle v^2(t) \rangle = \frac{\sigma^2}{2} \frac{\zeta}{m} \left(1 - e^{-2\frac{\zeta}{m}t} \right) \ . \tag{1.28}$$

Taking the limit $t \to \infty$ and inserting the result into the equipartition theorem, we obtain the Stokes-Einstein relation $\frac{1}{2}\sigma^2 = k_B T/\zeta$ with absolute temperature T and Boltzmann's constant k_B. In terms of the white noise and the result on the magnitude of the random force, we can rewrite the overdamped Langevin equation in its standard form involving Gaussian white noise

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x) + \sqrt{2D}\xi(t) \ , \tag{1.29}$$

where we introduced the diffusion constant $D = \frac{1}{2}\sigma^2 = k_B T/\zeta$, since the solution of (1.29) for a free Brownian particle, $f(x) = 0$, results in the mean squared displacement increasing linearly in time with the proportionality constant defined as twice the diffusion constant, $\langle (x(t) - x(0))^2 \rangle = 2Dt$.

ITÔ CALCULUS AND STOCHASTIC INTEGRALS

Mathematicians have made Langevin's equations rigorous by "multiplying with dt" , resulting in something called a (Itô) stochastic differential equation for the process $X(t)$:

$$\mathrm{d}X(t) = f(X(t),t)\mathrm{d}t + \sigma(X(t),t)\mathrm{d}W(t) \ , \tag{1.30}$$

where $\mathrm{d}W(t) = \xi(t)\mathrm{d}t$ is an infinitesimal increment of the mathematical Brownian motion, f is called the drift and σ the volatility. This equation is understood in the sense that the process $X(t)$ satisfies the integral equation

$$X(t) = X(0) + \int_0^t f(X(s),s)\mathrm{d}s + \int_0^t \sigma(X(s),s)\mathrm{d}W(s) \ . \tag{1.31}$$

To interpret the random variable on the right-hand side, the stochastic integral $Y(t) := \int_0^t \sigma(X(s),s)\mathrm{d}W(s)$ has to be defined. The common definition is due to Kiyoshi Itô and for this reason we speak of the Itô stochastic integral. Under certain regularity conditions on the integrand $\sigma(X(s),s)$, the stochastic integral can be defined as the limit of a Riemann sum

$$\int_0^t \sigma(X(s),s)\mathrm{d}W(s) := \lim_{N\to\infty} \sum_{i=1}^{N} \sigma\left(X((i-1)t/N),(i-1)t/N\right) \times \left[W(it/N) - W((i-1)t/N)\right] \ . \tag{1.32}$$

This representation as a Riemann sum directly suggests a way to simulate the process on a computer: draw a sequence of statistically independent standard normal random variables and multiply then with the square root of a discrete time step Δt; this creates the increments of the Brownian motion, which in turn can be multiplied with the integrand $\sigma(X(s),s)$ and finally added to the deterministic motion. This algorithm is known as the Euler scheme.

For the mathematical properties of the stochastic integral, it is important that the integrand is evaluated at the beginning of the sub-interval $[(i-1)t/N \ , \ it/N)$ so that the integrand is independent of the increment of the Brownian motion. A different interpretation of the stochastic integral is the **Stratonovich convention**, where the integrand is evaluated at the mid-point $(i-1/2)t/N$ of the sub-intervals. This convention gives a different value of the integral and therefore the stochastic differential equation (1.30) has to be augmented with the information of how the stochastic integral should be evaluated.

There is no rule connecting the Itô and Stratonovich stochastic integrals for general stochastic processes $\{X(t)\}$. For processes that are the solutions of a stochastic differential equation, however, the two integrals can be easily transformed into each other.

Consider the Itô stochastic differential equation for a d-dimensional process $X(t)$ driven by an M-dimensional Brownian motion

$$\mathrm{d}X_i(t) = f_i(X(t),t)\mathrm{d}t + \sum_{j=1}^{M} \sigma_{ij}(X(t),t)\mathrm{d}W_j(t) \ , \tag{1.33}$$

with $i = 1,\ldots,d$ and where f_i is called the drift-vector and σ_{ij} the volatility matrix. By using (1.33) to expand the integrand of the Stratonovich stochastic integral, it can be shown that the equivalent Stratonovich stochastic differential equation is given by

$$\mathrm{d}X_i(t) \overset{(S)}{=} \left[f_i - \frac{1}{2} \sum_{k=1}^{d} \sum_{j=1}^{M} \sigma_{kj} \partial_{x_k} \sigma_{ij} \right] \mathrm{d}t + \sum_{j=1}^{M} \sigma_{ij} \, \mathrm{d}W_j(t) \ , \tag{1.34}$$

with $i = 1,\ldots,d$ and where we omitted the arguments of $f_i(X(t),t)$ and $\sigma_{ij}(X(t),t)$.

Thus, both interpretations have the same volatility matrix but there is a correction to the drift vector. The conversion in the reverse direction from Stratonovich convention to Itô convention is then given by adding $\frac{1}{2}\sum_{k=1}^{d}\sum_{j=1}^{M} \sigma_{kj}\partial_{x_k}\sigma_{ij}$ to the drift vector.

The Stratonovich convention ensures that the rules of ordinary calculus apply to variable transformations, i.e. $\mathrm{d}g(X(t)) = \partial_x g(X(t))\mathrm{d}X(t)$, while for the Itô stochastic integral one has to apply **Itô's lemma** for variable transformations

$$\begin{aligned} \mathrm{d}g(X(t),t) =& \partial_t g(X(t),t)\mathrm{d}t + \sum_{i=1}^{d} \partial_{x_i}(X(t),t) \left[f_i \, \mathrm{d}t + \sum_{j=1}^{M} \sigma_{ij} \, \mathrm{d}W_j(t) \right] \\ &+ \frac{1}{2} \sum_{i,j=1}^{d} \partial_{x_i}\partial_{x_j} g(X(t),t) \sum_{k=1}^{M} \sigma_{ik} \, \sigma_{jk} \, \mathrm{d}t \ . \end{aligned} \tag{1.35}$$

CONVERGENCE TO THE STEADY STATE

We will consider only time-homogeneous processes, where the drift and volatility do not depend on time, $f_i(X(t),t) \equiv f_i(X(t)), \sigma_{ij}(X(t),t) \equiv \sigma_{ij}(X(t))$. Whether the Markov chain converges to a steady state with probability density $\pi(x)$ is not simple to ascertain for general processes. A special case are Martingales, characterised by a vanishing drift term, i.e. $\mathrm{d}X(t) = \sigma(X(t))\mathrm{d}W(t)$. If the Martingale is a non-negative process, the process converges to an integrable random variable X_∞ with probability density $\pi(x)$.

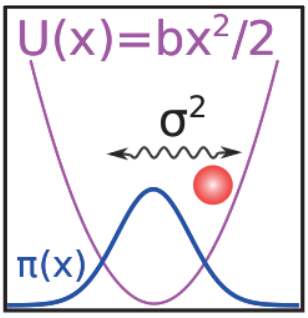


Figure 1.4: *Schematic view of a system described by the Ornstein-Uhlenbeck process. A particle with diffusion constant $\sigma^2/2$ performs Brownian motion in an effective harmonic potential $U(x) = bx^2/2$. Superimposed in blue is the stationary distribution $\pi(x) \sim \exp[-x^2/(\sigma^2/b)]$.*

Figure 1.5: *George E. Uhlenbeck is perhaps most famous for developing the idea of the electron spin together with Samuel Goudsmit. He also made many contributions to statistical mechanics. Uhlenbeck had a penchant for clarity and mathematical rigour. As an undergraduate student, during his laboratory courses, he derived all the employed electromagnetic formulae directly from Maxwell's equations. Later, he rejected Einstein's argument showing the existence of the Bose-Einstein condensation on the grounds that Einstein had replaced finite sums with integrals. At the time, phase transitions had not been properly understood from the point of statistical mechanics. Years later, it was Hendrik Kramers who pointed out that a phase transition could only occur in the thermodynamic limit of infinitely large systems.*

EXAMPLE: ORNSTEIN-UHLENBECK PROCESS

The Ornstein-Uhlenbeck process is the solution of the simplest Langevin equation admitting a steady state. The process describes a single particle with volatility σ, diffusing in an effective one-dimensional harmonic potential $U(x) = \frac{b}{2}x^2$ with $b > 0$ (see Fig. 1.4). A physical realisation is a colloid in solution being held in place by optical tweezers and confined to a one-dimensional channel. The effective deterministic force acting on the particle is the gradient of the potential $f(x) = -\partial_x U(x) = -bx$.

The dynamics of the particle position $X(t) \in \mathbb{R}$ in the overdamped limit is then described by the Langevin equation

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = -bX(t) + \sigma\xi(t) \;, \tag{1.36}$$

where the random force $\xi(t)$ constitutes δ-correlated white noise interpreted in the Itô convention, i.e. we have the equivalent Itô stochastic differential equation

$$\mathrm{d}X(t) = -bX(t)\mathrm{d}t + \sigma\mathrm{d}W(t) \;. \tag{1.37}$$

This stochastic differential equation can be solved by applying Itô's lemma to the transformed variable $Y(t) = X(t)e^{bt}$, which obeys the simpler Itô stochastic differential equation

$$\mathrm{d}Y(t) = e^{bt}\sigma\mathrm{d}W(t) \tag{1.38}$$

with the solution

$$Y(t) = Y(0) + \sigma\int_0^t e^{bs}\mathrm{d}W(s) \tag{1.39}$$

$$\Rightarrow X(t) = X(0)e^{-bt} + e^{-bt}\sigma\int_0^t e^{bs}\mathrm{d}W(s) \;. \tag{1.40}$$

One can show that the resulting process is Gaussian, i.e. for any time points $0 \le t_1 < t_2 < \ldots < t_k$, the joint probability distribution of $X(t_1), X(t_2), \ldots, X(t_k)$ is a k-dimensional Gaussian with means

$$\langle X(t_i)\rangle = \langle X(0)\rangle e^{-bt_i} \tag{1.41}$$

and covariances

$$\langle X(t_i)X(t_j)\rangle - \langle X(t_i)\rangle\langle X(t_j)\rangle = \frac{\sigma^2}{2b}\left(e^{-b|t_i-t_j|} - e^{-b(t_i+t_j)}\right) \;. \tag{1.42}$$

In particular, it follows that the process $X(t)$ converges to a steady state described by a Gaussian probability distribution $\pi(x)$ with mean 0 and variance $\sigma^2/(2b)$,

$$\pi(x) = \frac{1}{\sqrt{\pi\sigma^2/b}}e^{-x^2/(\sigma^2/b)} \;. \tag{1.43}$$

■

1.2.2.3 Fokker-Planck equations

Figure 1.6: *In his PhD thesis on Brownian motion, Adriaan Fokker derived the Fokker-Planck equation for the orientational distribution of rotating dipoles in an electromagnetic field; he later published the results in (Fokker, 1914). Today, his equation is commonly known as the Fokker-Planck equation because Max Planck was asked by colleagues to explain Fokker's work, which he eventually did, but not without adding his own version describing the velocity distribution of Brownian particles (Planck, 1917). Before transferring to physics under the supervision of Hendrik Lorentz, Fokker briefly studied engineering because "my mother always wanted me to become an engineer, and I never objected.". Besides his contributions to physics, Adriaan Fokker also built the 31-tone equal-tempered Fokker organ. He did not, however, build the famous aeroplanes - that was his cousin Anton Fokker.*

The second approach to the description of continuous Markov processes considers the probability density $p(x,t)$ of the variable $X(t)$ and characterises it as the solution of a partial differential equation known as the Fokker-Planck equation. We can derive the Fokker-Planck equation from the principle of local probability conservation

$$\partial_t p(x,t) = -\nabla \cdot j(x,t) \;\; \forall \; (x,t) \in \Omega \times [0,\infty) \; , \tag{1.44}$$

where the **probability current** $j(x,t)$ has one part arising from drift and a second part associated with diffusion

$$j_i(x,t) = a_i(x,t)p(x,t) - \sum_{j=1}^{d} \frac{\partial}{\partial x_j}[D_{ij}(x,t)p(x,t)] \tag{1.45}$$

with drift vector a_i and positive semi-definite diffusion matrix D_{ij} . We will consider only time-homogeneous processes, where the drift and diffusion coefficients do not depend on time, $a_i(x,t) \equiv a_i(x), D_{ij}(x,t) \equiv D_{ij}(x)$. This hyperbolic partial differential equation has to be augmented with the initial condition $p(x,0)$ and appropriate boundary conditions. Two standard boundary conditions are (i) absorbing boundary conditions: $p(x,t) = 0 \; \forall \; x \in \partial\Omega$, corresponding to particles exiting the domain Ω without ever returning (e.g. consider molecules crossing a membrane channel), and (ii) reflecting boundary conditions $j(x,t) \cdot n = 0 \; \forall \; x \in \partial\Omega$, corresponding to particles being reflected at the domain boundary and where n is the vector normal to the domain surface. Again, we define the propagator $p(x,t|x_0,0)$ as the solution for deterministic initial condition $p(x,0) = \delta(x - x_0)$ and the general solution for an arbitrary initial condition can be found by integrating over the propagators

$$p(x,t) = \int_\Omega \mathrm{d}y \; p(x,t|y,0)p(y,0) \; . \tag{1.46}$$

EQUIVALENCE TO THE LANGEVIN EQUATION

If we have the Itô stochastic differential equation (1.33) and consider the average of an arbitrary function $g(X(t),t)$: $\langle g(X(t),t)\rangle = \int \mathrm{d}x p(x,t)g(x,t)$, we can derive the corresponding Fokker-Planck equation by applying Itô's lemma (1.35) and integrating by parts (Gardiner, 2009). We find that the Itô stochastic differential equation (1.33) and the Fokker-Planck equation (1.44) are connected by the relations

$$a_i(x,t) = f_i(x,t) \tag{1.47}$$

$$D_{ij}(x,t) = \frac{1}{2}\sum_{k=1}^{M} \sigma_{ik}(x,t)\sigma_{jk}(x,t) \; . \tag{1.48}$$

STEADY STATE AND CONVERGENCE

A steady state of the Fokker-Planck equation (1.44) is described by a probability density $\pi(x)$ satisfying

$$0 = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}[a_i(x)\pi(x)] + \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}[D_{ij}(x)\pi(x)] \; . \tag{1.49}$$

Whether the solution $p(x,t)$ of the Fokker-Planck equation converges, for any initial condition, to a unique steady state, must be answered by the theory of partial differential equations. The absorbing boundary condition does in general not support the existence of a steady state, since the probability of the domain

$P_\Omega(t) = \int_\Omega \mathrm{d}x\, p(x,t)$ is not conserved. In the following section, we will discuss conditions guaranteeing the existence of a steady state for the simple class of equilibrium processes on infinite domains $\Omega = \mathbb{R}^d$.

EXAMPLE: DIFFUSION IN A GRAVITATIONAL FIELD

Similar to the Ornstein-Uhlenbeck process, we consider a particle with diffusion constant $D = \sigma^2/2 = k_B T/\zeta$ performing overdamped Brownian motion in the gravitational potential $\tilde{U}(x) = mgx$ with $m, g > 0$. The corresponding effective potential is $U(x) = mgx/\zeta$ and the effective force is $f(x) = -\partial_x U(x) = -mg/\zeta$. The Fokker-Planck equation for the particle position $x \in [0,\infty)$ reads

$$\frac{\partial}{\partial t} p(x,t) = +\frac{\partial}{\partial x}\frac{mg}{\zeta} p(x,t) + \frac{k_B T}{\zeta}\frac{\partial^2}{\partial x^2} p(x,t)\ . \tag{1.50}$$

We seek the steady state distribution $\pi(x)$ by solving

$$0 = -\frac{\mathrm{d}}{\mathrm{d}x} j(x) = \frac{mg}{\zeta}\frac{\mathrm{d}}{\mathrm{d}x}\pi(x) + \frac{k_B T}{\zeta}\frac{\mathrm{d}^2}{\mathrm{d}x^2}\pi(x) \tag{1.51}$$

subject to the reflecting boundary condition at $x = 0$

$$0 = j(0) = -\frac{mg}{\zeta}\pi(0) - \frac{k_B T}{\zeta}\frac{\mathrm{d}\pi}{\mathrm{d}x}(0)\ . \tag{1.52}$$

The solution is given by

$$\pi(x) = \frac{1}{Z} e^{-mgx/k_B T} = \frac{1}{Z} e^{-\tilde{U}(x)/k_B T} \tag{1.53}$$

with the normalisation constant $Z = \int_0^\infty \mathrm{d}x e^{-mgx/k_B T} = \frac{k_B T}{mg}$. The reflecting boundary condition is automatically fulfilled, since $j(x) \equiv 0$. ■

1.2.3 Equilibrium versus non-equilibrium steady states

Steady states come in two varieties: equilibrium steady states and non-equilibrium steady states. Equilibrium steady states form a subset of steady states that is characterised by some additional constraints, which make the stationary distribution relatively easy to compute. It is no coincidence that all the simple examples for Markov processes with steady states, given above, have equilibrium steady states. Non-equilibrium steady states are, as the name suggests, all steady states that are not equilibrium steady states. Non-equilibrium steady states are much harder to compute and analytical solutions are available mainly for one-dimensional systems.

1.2.3.1 Equilibrium and detailed balance

Loosely speaking, an equilibrium steady state is characterised by the property that when watching a film of the system, you cannot tell whether the film is played forwards or backwards. More formally, equilibrium is characterised by the condition of detailed balance, demanding that the net probability flow between each pair of configurations vanishes in the steady state. In other words: in the steady state any transition has the same probability as its time-reversed transition [1]. For Markov chains with discrete configurations and time, the condition of detailed balance corresponds to the steady state satisfying

$$\pi_i T_{ij} = \pi_j T_{ji} \; \forall \; i, j = 1, \ldots, |\Omega| \; . \tag{1.54}$$

For Markov chains in continuous time, the corresponding equation has the same form, with the transition probabilities T_{ij} replaced by the transition rates K_{ij}

$$\pi_i K_{ij} = \pi_j K_{ji} \; \forall \; i, j = 1, \ldots, |\Omega| \; . \tag{1.55}$$

Thus, the probability flows between configurations balance each other pairwise in the steady state. It is straightforward to verify that a vector π satisfying the detailed balance condition (1.54) automatically fulfils the definition of a steady state (1.7). Likewise, for continuous time (1.55) implies (1.17). The detailed balance condition for Markov chains can be checked without actually computing the steady state. **Kolmogorov's criterion** asserts that (for ergodic chains) detailed balance is equivalent to the transition probability of any closed loop being independent of the direction it is traversed in

$$T_{i_0,i_1} T_{i_1,i_2} \ldots T_{i_n,i_0} = T_{i_0,i_n} T_{i_n,i_{n-1}} \ldots T_{i_1,i_0} \; , \forall \; i_0, i_1, \ldots, i_n \in \{1, 2, \ldots, |\Omega|\} \tag{1.56}$$

and likewise for the matrix of transition rates K_{ij}. For one-dimensional configuration spaces, in a closed loop each transition has to occur also in the reverse direction (unless periodic boundary condition are imposed); hence, Kolmogorov's criterion is trivially fulfilled. An example is the biased random walk on $\mathbb{N}_0$ discussed above.

For Markov processes with continuous configurations $x \in \Omega \subset \mathbb{R}^d$, we consider the Fokker-Planck formulation, in which the steady state is characterised by a vanishing divergence of the probability flow, $\nabla \cdot j(x) \equiv 0$. An equilibrium steady state obeying detailed balance satisfies the stronger condition that the

[1] We consider only variables that are even under time-reversal. For variables changing sign when reversing time, like physical velocity, the detailed balance conditions have to be modified slightly but take the same form.

probability current itself vanishes

$$0 \equiv j_i(x) = a_i(x)\pi(x) - \sum_{j=1}^{d} \frac{\partial}{\partial x_j}[D_{ij}(x)\pi(x)] \ , \forall i \in \{1,\ldots,d\} \ . \tag{1.57}$$

In the following section, we will give sufficient conditions for the existence of an equilibrium steady state with vanishing probability current. Again, these conditions can be verified without actually computing the steady state. However, when these conditions are met, we find a straightforward procedure for computing the equilibrium steady state.

1.2.3.2 Energy functions and the Boltzmann distribution

The property of detailed balance is strongly linked with the existence of an energy function

$$\begin{aligned} E : \Omega &\to \mathbb{R} \\ x &\mapsto E(x) \ , \end{aligned} \tag{1.58}$$

which we will characterise in more detail for different processes below.

DISCRETE CONFIGURATIONS

Since the steady state of an ergodic Markov chain has strictly positive weights, $\pi(\omega_i) > 0$, we can write the detailed balance equation (1.55) in a slightly more suggestive way [1]

$$-\ln \pi(\omega_j) = -\ln \pi(\omega_i) + \ln\left(\frac{T_{ji}}{T_{ij}}\right) \ . \tag{1.59}$$

This form motivates the following construction of an energy function $E(\omega_i) = -\ln \pi(\omega_i) + \text{const.}$: we start with some configuration ω_k and assign to it an arbitrary energy E_0. Next, we define the energy of a neighbouring configuration[2] ω_j as $E(\omega_j) = E_0 + \ln\left(T_{ji}/T_{ij}\right)$. We iterate this definition procedure until the energies of all configuration have been defined (since the chain is irreducible, we can reach all configurations). Kolmogorov's criterion ensures that the procedure described above gives an unambiguous definition of the configuration energy (up

[1] In the following, we consider discrete time. The same procedure can be carried out for a continuous -time chain by replacing the transition probabilities T_{ij} with the transition rates K_{ij}.

[2] We call configurations ω_j neighbours of ω_i, if they have a non-zero transition probability $T_{ij} > 0$

to an additive constant given by the arbitrary choice of E_0). It is simple to check that the steady state is described by the (discrete) Boltzmann distribution

$$\pi(\omega_i) = \frac{1}{Z} e^{-E(\omega_i)} \ , \ Z = \sum_{\omega_i \in \Omega} e^{-E(\omega_i)} \ , \tag{1.60}$$

where the normalising factor Z, is called the **partition function**. The Boltzmann distribution is well known from statistical mechanics as the distribution describing the canonical ensemble. For physical systems, like a mono-atomic gas in a container connected to a heat bath of temperature T, the energy function in (1.60) corresponds to the physical energy measured in units of the thermal energy $k_B T$, i.e. here we set $k_B T \equiv 1$. In analogy to statistical mechanics, we define the thermodynamic potential called **free energy**

$$F = -\ln Z \tag{1.61}$$

and the **configuration entropy**

$$\phi(\omega_i) = E(\omega_i) - F = -\ln \pi(\omega_i) \ . \tag{1.62}$$

Another thermodynamic potential, **Gibbs' entropy**[1], is then defined as the mean configuration entropy

$$S = -\sum_{\omega_i \in \Omega} \pi(\omega_i) \ln \pi(\omega_i) = \sum_{\omega_i \in \Omega} \pi(\omega_i) \phi(\omega_i) = \sum_{\omega_i \in \Omega} \pi_i (E(\omega_i) - F) = \langle E \rangle - F \ . \tag{1.63}$$

The usefulness of introducing these thermodynamic potentials in the context of equilibrium inference will become clear in the next section, when we consider how the energy function depends on a set of system parameters.

Example: biased random walk on $\mathbb{N}_0$

We consider again the biased random walk on $\mathbb{N}_0$ discussed above (see Fig.1.2). Since the configuration space is one-dimensional, Kolmogorov's criterion is fulfilled and detailed balance holds. We define the energy function starting from configuration 0 with $E(0) = 0$ and sequentially work our way up:

$$E(i+1) = E(i) + \ln\left(\frac{T_{i+1,i}}{T_{i,i+1}}\right) = E(i) + \ln\left(\frac{1-r}{r}\right) = i \ln\left(\frac{1-r}{r}\right) \ . \tag{1.64}$$

[1] This definition of entropy was derived by Willard Gibbs as an extension of Boltzmann's definition of entropy $S = -k_B \ln N$ valid for systems with uniform distribution (the microcanonical ensemble). He considered a limit of configuration counting in an ensemble of identical systems exchanging energy between them.

Therefore, we have the equilibrium steady state

$$\pi(i) = \frac{1}{Z} e^{-E(i)} = \frac{1}{Z} e^{-i \ln\left(\frac{1-r}{r}\right)} = \frac{1}{Z} \left(\frac{r}{1-r}\right)^i \tag{1.65}$$

with partition function $Z = \sum_{i=0}^{\infty} \left(\frac{r}{1-r}\right)^i = \left(\frac{r}{1-2r}\right)$. The linear energy function $E(i) \sim i$ shows that the biased random walk on $\mathbb{N}_0$ is in fact a discretised version of diffusion in a gravitational field, discussed above in section 1.2.2.3. ■

Vice versa, given the energy function $E(\omega_i)$, we can define a steady state in terms of the Boltzmann distribution (1.60) and find many different transition probabilities (or rates respectively) that are compatible with this Boltzmann distribution as an equilibrium steady state. The conditions imposed on the transition probabilities are found by inserting the Boltzmann distribution (1.60) into the detailed balance condition (1.54), yielding

$$\frac{T_{ij}}{T_{ji}} = e^{E(\omega_i) - E(\omega_j)} \ . \tag{1.66}$$

CONTINUOUS CONFIGURATIONS

In the case of continuous configurations $x \in \Omega \subset \mathbb{R}^d$, we find a similar relationship between the energy and the steady state. We begin by considering a configuration-independent, isotropic diffusion matrix $\mathbf{D} = D\mathbb{1} = (k_B T/\zeta)\mathbb{1}$ and a conservative effective force $a(x) = -\nabla E(x)/\zeta$ that derives from a potential energy. Hence, we can rewrite the detailed balance condition (1.57) as

$$\frac{k_B T}{\zeta} \nabla \pi(x) = a(x)\pi(x) = -\frac{\nabla E(x)}{\zeta} \pi(x) \ . \tag{1.67}$$

The solution of this equation is given by the continuous version of the Boltzmann distribution

$$\pi(x) = \frac{1}{Z} e^{-E(x)/k_B T} \tag{1.68}$$

with the normalising partition function $Z = \int_{\Omega \subset \mathbb{R}^d} \mathrm{d}x \ e^{-E(x)/k_B T}$. Analogous to the discrete case, we can define the free energy $F = -\ln Z$ and Gibbs' entropy[1]

$$S = -\int_{\Omega \subset \mathbb{R}^d} \mathrm{d}x \ \pi(x) \ln \pi(x) \ . \tag{1.69}$$

[1] However, the continuous-configuration version of entropy loses the positivity property of the discrete entropy.

Example: Ornstein-Uhlenbeck process

Interpreting the Ornstein-Uhlenbeck process as describing a physical Brownian particle diffusing in the potential $\tilde{U}(x) = \kappa x^2/2$, the volatility is determined by the diffusion constant $\sigma^2 = 2D = 2k_BT/\zeta$ and we can write the steady state (1.43) in the well-known Boltzmann form (1.68)

$$\pi(x) = \frac{1}{\sqrt{2\pi k_BT/\kappa}} e^{-\kappa x^2/2k_BT} = \frac{1}{Z} e^{-\tilde{U}(x)/k_BT} \ . \tag{1.70}$$

■

If the diffusion coefficient is anisotropic or configuration-dependent, things get a little more complicated, but we can still find a **generalised energy** $U(x)$ such that the steady state is described by the Boltzmann distribution $\pi(x) = \exp(-U(x))/Z$. In any case, we can take this exponential form for the steady state as an ansatz, since the steady state distribution of an ergodic Markov process is strictly positive $\pi(x) > 0$. Inserting the Boltzmann form into the detailed balance condition (1.57), we find the equation

$$\sum_{j=1}^{d} D_{ij}(x) \frac{\partial U}{\partial x_j} = -a_i(x) + \sum_{j=1}^{d} \frac{\partial D_{ij}}{\partial x_j} =: \lambda_i(x) \ . \tag{1.71}$$

Next, we assume an invertible diffusion matrix $\mathbf{D}$ to rewrite this equation as

$$\frac{\partial U}{\partial x_i} = \sum_{j=1}^{d} D^{-1}_{ij}(x) \lambda_j(x) =: \gamma_i(x) \ . \tag{1.72}$$

In a simply connected domain $\Omega \subset \mathbb{R}^d$ this equation has a solution $U(x)$, if and only if the curl of the auxiliary vector $\gamma(x)$ vanishes,

$$\frac{\partial \gamma_i}{\partial x_j}(x) \equiv \frac{\partial \gamma_j}{\partial x_i}(x) \ , \ \forall i,j = 1,\ldots,d \ . \tag{1.73}$$

This condition can be verified directly on the basis of the coefficients of the Fokker-Planck equation, without the need of computing the steady state. Given the generalised energy exists, we can write it as a line integral starting from an arbitrary point $x' \in \Omega$:

$$U(x) = \int_{x'}^{x} \mathrm{d}z \cdot \gamma(z) \ . \tag{1.74}$$

Again, the partition function must be chosen such that the steady state is normalised, i.e.

$$Z = \int_{\Omega \subset \mathbb{R}^d} \mathrm{d}x \ e^{-U(x)} \ . \tag{1.75}$$

In summary, we have seen that equilibrium Markov processes, characterised by dynamics obeying detailed balance, allow for an easy computation of the steady state, at least in principle. The main difficulty in computing the steady state and any derived observables lies in the computational complexity of computing the partition function Z, which usually involves high-dimensional sums or integrals. For example, consider a system consisting of N binary spin variables $\omega = (s_1, \ldots, s_N) \in \{\pm 1\}^N$. The configuration space has a size $|\Omega| = 2^N$ that grows exponentially in the number spin variables. The partition function $Z = \sum_{\omega\in\Omega} e^{-E(\omega)} = \sum_{s_1=\pm 1} \cdots \sum_{s_N=\pm 1} e^{-E(s_1,\ldots,s_N)}$ is hard to compute if the energy function contains interaction terms between different spin variables that prevent a factorisation of the high-dimensional sum. In order to deal with this computational problem, many different approximations to the equilibrium steady state or its associated observables have been developed and a few of these will be mentioned in the next section.

Non-equilibrium Markov processes, on the other hand, are much harder to characterise even in principle. While we could still define a generalised potential $\Phi(x)$ such that $\pi(x) \sim e^{-\Phi(x)}$, we do not have a general procedure to compute this potential.

1.3 Stochastic inference

So far, we have suppressed in our notation that we are dealing with entire families of ergodic Markov processes with transition probabilities characterised by a set of k parameters $\Theta \in \Lambda \subset \mathbb{R}^k$. Correspondingly, we will denote the steady state belonging to a specific parameter set by $\pi(x;\Theta)$ and the propagators by $p_\Theta(x,t|y,0)$. In the examples above, we encountered the random walk on $\mathbb{N}_0$ with the single parameter $\Theta = r$, the random telegraph process with parameters $\Theta = (\alpha,\beta)$, the Ornstein-Uhlenbeck process with parameters $\Theta = (b,\sigma)$, and diffusion in a gravitational field with parameters $\Theta = (mg/\zeta, k_BT/\zeta)$.

Markov processes are frequently used to model real-world processes such as the movement of particles in a molecular gas. The general form of the transition probabilities is based on qualitative modelling, which pins down a family of models. The actual quantitative predictions derived from this model, however, will depend on the numerical values of the free parameters Θ characterising the model family. For example, we might want to make statements about some observable like the particle position in Brownian motion. Since the process modelled is stochastic, these statements will be concerned with statistical quantities.

Simple examples are single-time means of the process[1]

$$m(t;\Theta) := \langle X(t)\rangle = \sum_x xP(X(t)=x;\Theta)\ , \tag{1.76}$$

or time-correlations

$$C(t+\tau,t;\Theta) := \langle X(t+\tau)X(t)\rangle = \sum_{x,y} xyP(X(t+\tau)=x,X(t)=y;\Theta)\ . \tag{1.77}$$

Since the distribution of $X(t)$ depends on the values of the parameter Θ, it is clear that our statistical predictions are also functions of the parameters Θ. A special class of processes are **stationary processes**, where the process $\{X_{t+s}\}$ shifted by a time $s>0$ has the same distribution, implying that the means become time-independent, $m(t;\Theta)\equiv m(\Theta)$ and time-correlations depend only on the time-difference $C(t+\tau,t;\Theta)\equiv C(\tau,0;\Theta)$. A stationary process can be created by initialising with the steady state: $P(X(0))=\pi(x;\Theta)$. Similarly, a general ergodic Markov process becomes asymptotically stationary in the long time limit, since

$$\lim_{t\to\infty} P(X(t)=x;\Theta) = \pi(x;\Theta) \tag{1.78}$$

$$\lim_{t\to\infty} P(X(t+\tau)=x,X(t)=y;\Theta) = \pi(y;\Theta)\left(T^\tau(\Theta)\right)_{yx}\ , \tag{1.79}$$

independent of the initial condition X_0. The task of computing observable statistics as functions of the parameters is called the **statistical forward problem**. In some processes, the quantitative values of the parameters Θ are directly accessible experimentally. For the particle diffusing in a gravitational field for example, we can directly measure the absolute temperature T, Boltzmann's constant k_B, drag coefficient ζ, particle mass m and gravitational constant g by means of separate experiments. In many other processes, however, the parameters might not be easy to measure, or in the case where the model is only an effective description of the underlying process, impossible to measure independently of the stochastic process, even in principle. Stochastic inference is concerned with the **inverse statistical problem** of inferring the parameters Θ from some data set D of observations sampled from the process. Solving this problem requires the ability to solve the forward problem, so the statistical predictions made for some fixed parameter values Θ may be compared to the observed statistics of the data set D. In addition, it is necessary to find suitable observables that allow to calibrate their statistics to the data set D such that we obtain a unique model. In the simplest cases, these observables might be means $\langle X(t)\rangle$ or (time-)correlations $\langle X(t+\tau)X(t)\rangle$. A different approach is based on Bayes' theorem, which uses the full sampled distribution.

[1] We consider processes with discrete configurations. For continuous random variables, we replace the sums with integrals and the probabilities with probability densities.

1.3.1 The Bayesian framework and maximum likelihood

A widely used framework for stochastic inference is based on Bayes' theorem and is hence called Bayesian inference (see e.g. Barber (2012); MacKay (2003)). We will motivate the use of Bayesian inference by analysing a fictitious scene described by Nassim Nicholas Taleb (Taleb, 2010), supposedly exhibiting the superiority of street-smarts, represented by Brooklyn-born Fat Tony, over the in-the-box thinking of science and engineering graduates, represented by Dr. John:

```
NNT [...]: Assume that a coin is fair, i.e., has an equal
probability of coming up heads or tails when flipped. I flip it
ninety-nine times and get heads each time. What are the odds of my
getting tails on my next throw?
Dr. John: Trivial question. One half, of course, since you
are assuming 50 percent odds for each and independence between draws.
NNT: What do you say, Tony?
Fat Tony: I'd say no more than 1 percent, of course.
NNT: Why so? I gave you the initial assumption of a fair coin,
meaning that it was 50 percent either way.
Fat Tony: You are either full of crap or a pure sucker to buy
that "50 pehcent" business. The coin gotta be loaded. It can't
be a fair game. (Translation: It is far more likely that your
assumptions about the fairness are wrong than the coin delivering
ninety-nine heads in ninety-nine throws.)
NNT: But Dr. John said 50 percent.
Fat Tony [...]: I know these guys with the nerd examples from the
bank days. They think way too slow. And they are too commoditized.
You can take them for a ride.
```

What Taleb is asking of Dr. John and Fat Tony is to make a prediction about the statistics of a random variable $X_{100} \in \{\pm 1\}$, based on an observation of the process $D = \{X_1 = +1, X_2 = +1, \ldots, X_{99} = +1\}$. The process is Markovian, since we can reasonably assume statistically independent coin tosses. Their prediction will depend on the value they assign to the parameter p, the probability of the coin to show heads. The probability of the observed data set under Taleb's assumption of $p = 1/2$ is $P(D|p = 1/2) = (1/2)^{99} \approx 1.6 \times 10^{-30}$. Thus, we have ample reason to doubt the validity of Taleb's assumption and should switch to a probabilistic description of p. We choose a simple model of our uncertainty and model our **prior information** about p as an exponential probability density $P(p)$ symmetric around its peak at $p = 1/2$

$$P(p) = \begin{cases} \frac{\beta}{2(e^{\beta/2}-1)} e^{\beta p} & 0 \leq p \leq 1/2 \\ \frac{\beta}{2(e^{\beta/2}-1)} e^{\beta(1-p)} & 1/2 < p \leq 1 \end{cases}, \tag{1.80}$$

where $\beta > 0$ quantifies our belief in Taleb's hypothesis. This prior is shown for different values of β in the left panel of Fig. 1.7.

Now in contrast to Dr. John, we should actually take into account the data D and update our belief about p by conditioning on the observation of the data. We do this with the help of Bayes' theorem, which gives us the **posterior probability** of p

$$P(p|D) = \frac{P(D|p)P(p)}{P(D)} \tag{1.81}$$

in terms of the prior information $P(p)$, the **likelihood** $P(D|p)$ and the normalising factor $P(D) = \int_0^1 \mathrm{d}p P(D|p)P(p)$ called the **evidence** or **marginal likelihood**. The first thing we have to do is solve the forward problem of computing $P(D|p)$, which is easy, since the different sample variables are statistically independent. We find

$$P(D|p) = p^{99} \,. \tag{1.82}$$

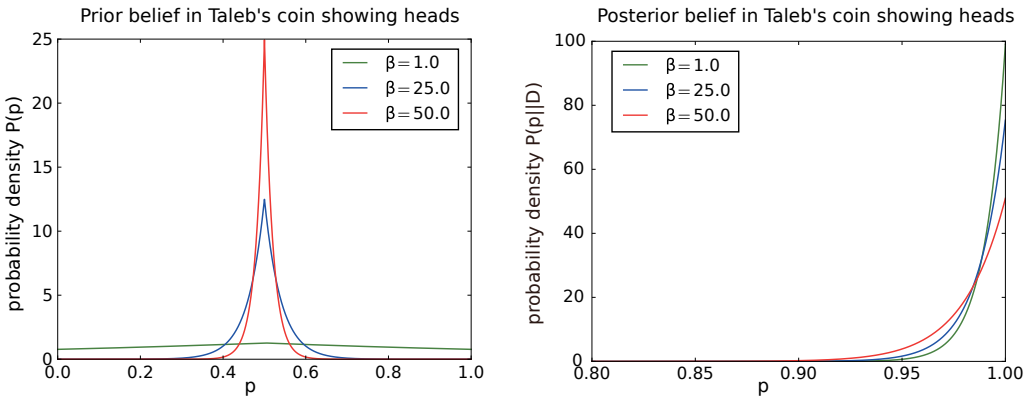


Figure 1.7: *Prior and posterior probability densities for the probability p of Taleb's coin showing heads.*

Inserting the likelihood (1.82) and the prior (1.80) into Bayes' theorem (1.81), we can compute the evidence (by numerical integration) and posterior probability density $P(p|D)$, which is shown in the right panel of Fig. 1.7 for different values of our trust β in Taleb's hypothesis of $p = 1/2$. We find that even for very high degrees of belief in Taleb's hypotheses, the data convincingly suggests that p is very close to 1. The probability of the coin showing tails at the next toss, is obtained via the law of total probability in terms of the average over the posterior

density

$$P(X_{100} = -1|D) = \int_0^1 \mathrm{d}p P(X_{100} = -1|D,p)P(p|D) = \int_0^1 \mathrm{d}p(1-p)P(p|D)$$
$$= 1 - \int_0^1 \mathrm{d}p\ pP(p|D)\ , \tag{1.83}$$

which is roughly 10^{-13} for $\beta = 50$ (high trust in Taleb). So in fact our street-smart Fat Tony is having an incredible faith in the information Taleb gave him about the coin. Well, to be fair, he only gave an upper bound claiming $P(X_{100} = -1|D) \leq 1\%$, however, had Dr. John been trained in Bayesian inference, he could have made a much better estimate! ■

The procedure we have followed in this coin tossing example is in fact the general reasoning in Bayesian inference. Given some data D and some prior information $P(\Theta)$ about the model parameters Θ, we seek to compute the statistics of some observable $\mathcal{O}(\{X(t)\})$ given some data, by averaging over the posterior probability

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)}\ . \tag{1.84}$$

The mean, for example, would be given by

$$\langle \mathcal{O}(\{X(t)\})|D\rangle = \int \mathrm{d}\Theta \langle \mathcal{O}(\{X(t)\})|\Theta\rangle P(\Theta|D)\ , \tag{1.85}$$

where $\langle \mathcal{O}(\{X(t)\})|\Theta\rangle$ denotes the average of the observable, given a fixed set of parameters Θ.

In practice, the models are far more complicated than independent coin tosses, having many different parameters. Computing the high-dimensional integrals over the posterior density or computing the evidence becomes infeasible. In the particular case studied in this thesis, the data consists of **snapshots of the steady state**, i.e. $D = \{x_\mu\}_{\mu=1}^M$ with the samples x_μ drawn independently from the steady state distribution $\pi(x;\Theta^{\text{true}})$ (the other case of interest is time-series data, which is discussed below). The likelihood

$$P(D|\Theta) = \exp\left(M\frac{1}{M}\sum_{\mu=1}^{M}\pi(x_\mu;\Theta)\right) \overset{M\to\infty}{\approx} \exp\left(M\langle\pi(X;\Theta)\rangle\right) \tag{1.86}$$

scales exponentially in the number of samples M and becomes sharply peaked in the limit $M \to \infty$. If the prior is sufficiently smooth and is non-zero around this maximum, the likelihood dominates the posterior probability (1.84), which also becomes sharply peaked around roughly the same value. Therefore, the integral over the posterior can be approximated by taking only the parameter value at the

peak of the likelihood (more formally, we make a saddle-point-approximation of the integral), yielding the **maximum likelihood estimate** of the parameters

$$\Theta^{ML} = \underset{\Theta}{\operatorname{argmax}}\, P(D|\Theta) \,. \tag{1.87}$$

Due to the exponential scaling of the likelihood for independent samples, the **log-likelihood function**

$$\mathcal{L}(\Theta;D) := \frac{1}{M} \ln P(D|\Theta) = \frac{1}{M} \sum_{\mu=1}^{M} \ln \pi(x_\mu;\Theta) \tag{1.88}$$

is considered instead. The log-likelihood has the same maximiser Θ^{ML}, since the logarithm is a monotonic function.

Interestingly, for Markov chains with discrete configuration space, the likelihood maximisation principle has a connection to information theory: maximising the likelihood is equivalent to **minimising the relative entropy**, or Kullback-Leibler divergence (Kullback and Leibler, 1951), between the **sampled distribution**, characterised by the probability mass function

$$\hat{p}(x) = \frac{1}{M} \sum_{\mu=1}^{M} \delta_{x,x_\mu} \,, \tag{1.89}$$

and the steady state distribution $\pi(x;\Theta)$,

$$\begin{aligned} D_{\mathrm{KL}}(\hat{p}(x)||\pi(x;\Theta)) &= \sum_{x\in\Omega} \hat{p}(x) \ln\left(\frac{\hat{p}(x)}{\pi(x;\Theta)}\right) \\ &= \sum_{x\in\Omega} \hat{p}(x) \ln \hat{p}(x) - \sum_{x\in\Omega} \hat{p}(x) \ln \pi(x;\Theta) \\ &= -S(\hat{p}(x)) - \frac{1}{M} \sum_{\mu=1}^{M} \ln \pi(x_\mu;\Theta) \\ &= -S(\hat{p}(x)) - \mathcal{L}(\Theta;D) \,, \end{aligned} \tag{1.90}$$

where the Gibbs entropy (or **Shannon entropy** in the terminology of information theory) of the sampled distribution $S(\hat{p}(x)) = -\sum_x \hat{p}(x) \ln \hat{p}(x)$ does not depend on the model parameters. Thus, we can think of the maximum likelihood approach as trying to minimise a distance measure[1] between the model distribution

[1] Note that relative entropy $D_{\mathrm{KL}}(p||q)$ is not a true metric. While it is non-negative and zero only if the two distributions are identical, it is neither symmetric, nor does it obey the triangle inequality.

and the sampled one. The link between the maximum likelihood approach and minimising relative entropy is also of practical utility: due to the non-negativity of relative entropy, the log-likelihood function has the negative Shannon entropy as an upper bound, which is saturated if and only if the sampled distribution and the test steady state $\pi(x;\Theta)$ are identical (which we can only expect to hold in the limit $M \to \infty$, due to sampling errors for finite M).

For equilibrium processes, the steady state is characterised by the energy function $E(x;\Theta)$ [cp. Eq.(1.58)] and we can maximise the likelihood (or minimise the relative entropy), at least in principle. However, the computation of the partition function $Z(\Theta) = \sum_x e^{-E(x;\Theta)}$, required in the evaluation of the likelihood, is too costly for most models, which has led to the development of various computationally efficient approximations.

For most non-equilibrium processes, on the other hand, we lack an explicit characterisation of the steady state $\pi(x;\Theta)$ and the maximum likelihood approach is infeasible. However, the variational principles and approximations used for equilibrium inference can be modified in a way that allows us to solve also the non-equilibrium inference problem. For this reason, we will continue with a short overview of the methods used in equilibrium inference. For a comprehensive review of equilibrium inference in the Ising model (and non-equilibrium inference from time-series) see Nguyen et al. (2017).

1.3.2 Equilibrium inference from snapshots of the steady state

For equilibrium systems, the steady state is described by the Boltzmann distribution $\pi(x;\Theta) = \frac{1}{Z(\Theta)} e^{-E(x;\Theta)}$ (recall that we measure the energy in units of the thermal energy, i.e. we set $k_B T \equiv 1$, since in stochastic inference the overall energy scale and temperature of the system cannot be inferred separately). Inserting the Boltzmann distribution into the log-likelihood function (1.88), we find

$$\begin{aligned} \mathcal{L}(\Theta;D) &= -\ln Z(\Theta) - \frac{1}{M}\sum_{\mu=1}^{M} E(x_\mu;\Theta) \\ &= F(\Theta) - \langle E(\Theta)\rangle_{\hat{p}} \,, \end{aligned} \tag{1.91}$$

which becomes difficult to evaluate for a high-dimensional configuration space $\Omega \subset \mathbb{R}^N$, $N \gg 1$, due to the computation of the free energy $F(\Theta) = -\ln Z(\Theta)$ involving high-dimensional sums. For this reason, it becomes necessary to find suitable approximations for the free energy. To this end, our framing of the inference problem in terms of minimising relative entropy becomes useful in providing a framework for measuring the quality of such approximations. For

a fixed value of parameters Θ, we approximate the free energy $F(\Theta)$ by the **variational free energy** (cp. (1.63))

$$F[q] = \langle E \rangle_q - S(q) = \sum_x q(x) E(x;\Theta) + \sum_x q(x) \ln q(x) \tag{1.92}$$

evaluated for distributions $q \in \mathcal{Q}$ in particular families $\mathcal{Q}$, for which we can actually calculate the free energy. Particular choices are factorising distributions, which lead to mean field theory, or distributions factorising into one- and two-point marginals, which lead to the Bethe-Peierls approximation (Bethe, 1935; Peierls, 1936). From the set $\mathcal{Q}$ of tractable distributions, we choose the distribution q^* with the smallest relative entropy to the Boltzmann distribution $\pi(x;\Theta)$

$$F(\Theta) \approx F[q^*(\Theta)] \tag{1.93}$$

$$\begin{aligned} q^*(\Theta) &= \operatorname*{argmin}_{q\in\mathcal{Q}} D(q||\pi(x;\Theta)) \\ &= \operatorname*{argmin}_{q\in\mathcal{Q}} \left\{ \sum_x \ln q(x) + \sum_x q(x) E(x;\Theta) + \ln Z(\Theta) \right\} \\ &= \operatorname*{argmin}_{q\in\mathcal{Q}} \{F[q] - F(\Theta)\} \ . \end{aligned} \tag{1.94}$$

The true free energy $F(\Theta)$ gives a lower bound on the variational free energy $F[q]$ and this bound is saturated if and only if the test distribution $q(x)$ equals the Boltzmann distribution $\pi(x;\Theta)$. Furthermore, the minimisation of the variational free energy does not require the (often intractable) computation of the true free energy $F(\Theta)$.

1.3.2.1 Equilibrium mean field theory

There is a large class of approximation methods usually grouped in the category of mean field theory (see e.g. Opper and Saad (2001)). Their common theme is using distributions that factorise into simpler distributions that are tractable. In the simplest form of mean field theory, the different components of the configuration $x \in \Omega \subset \mathbb{R}^N$ are assumed statistically independent so the distribution $q(x)$ factorises into one-dimensional distributions

$$q(x) = \prod_{i=1}^{N} q^{(i)}(x_i) \ , \tag{1.95}$$

where

$$q^{(i)}(x_i) = \sum_{\{x_j, j\neq i\}} q(x_1, \ldots, x_N) \ , \tag{1.96}$$

is the marginal distribution of the component x_i of configuration x. This factorisation allows a simple computation of the variational free energy of the mean field distribution

$$\begin{aligned} F[q] &= \sum_x q(x)E(x;\Theta) + \sum_x q(x)\ln q(x) \\ &= \left(\prod_{i=1}^N \sum_{x_i} q^{(i)}(x_i)\right) E(x;\Theta) + \sum_{i=1}^N \sum_{x_i} q^{(i)}(x_i)\ln q(x_i) \\ &= \left(\prod_{i=1}^N \sum_{x_i} q^{(i)}(x_i)\right) E(x;\Theta) + \sum_{i=1}^N S[q^{(i)}] \ . \end{aligned} \tag{1.97}$$

EXAMPLE: BINARY VARIABLES

For a system described by binary variables $x = (x_1,\ldots,x_N) \in \{\pm 1\}^N$, the one-dimensional marginals can be parametrised by their means

$$q^{(i)}(x_i) = \frac{1+m_i x_i}{2} \ , \tag{1.98}$$

with

$$m_i = \sum_x x_i \ q(x) = \sum_{x_i} x_i \ q^{(i)}(x_i) = q^{(i)}(+1) - q^{(i)}(-1) \ . \tag{1.99}$$

This parametrisation allows a straightforward computation of the mean field values of Gibbs' entropy

$$S[q(\{m_i\})] = \sum_{i=1}^d \left[\left(\frac{1+m_i}{2}\right) \ln \left(\frac{1+m_i}{2}\right) + \left(\frac{1-m_i}{2}\right) \ln \left(\frac{1-m_i}{2}\right) \right] \tag{1.100}$$

and the average energy

$$E[q(\{m_i\})] = \left(\prod_{i=1}^d \sum_{x_i=\pm 1} q^{(i)}(x_i)\right) E(x;\Theta) = E(\{m_i\},\Theta) \ , \tag{1.101}$$

which together give the mean field variational free energy. Next, we have to minimise the free energy over the parameters m_i by solving

$$0 = \frac{\partial}{\partial m_i} F[q(\{m_i\})] = \frac{\partial E[q(\{m_i\})]}{\partial m_i} - \frac{\partial S[q(\{m_i\})]}{\partial m_i} \ , \tag{1.102}$$

which gives the mean-field equations

$$\operatorname{atanh}(m_i) = -\frac{\partial E}{\partial m_i}(\{m_i^*\},\Theta) \ . \tag{1.103}$$

For the equilibrium Ising problem, the parameters are the magnetic fields $\{h_i\}$ and pairwise interactions $\{J_{ij}\}_{j>i}$ and the energy would be $E(x;\Theta) = -\sum_i h_i x_i - \sum_{i<j} J_{ij} x_i x_j$, giving the mean-field energy

$$E[q(\{m_i\})] = -\sum_i h_i m_i - \sum_{i<j} J_{ij} m_i m_j \ , \tag{1.104}$$

yielding the self-consistent mean field equations

$$\begin{aligned} \text{atanh} &= h_i + \sum_{j\neq i} J_{ij} m_j \\ &\Leftrightarrow \\ m_i &= \tanh\left(h_i + \sum_{j\neq i} J_{ij} m_j \right) \ . \end{aligned} \tag{1.105}$$

However, since the factorising mean field distribution is fully characterised by the vector of means m_i, it generally does not have enough degrees of freedom to fully describe the steady state. To reconstruct the original model parameters, more elaborate techniques must be used to find non-trivial predictions for the correlations $\langle x_i x_j \rangle$, as can be done by exploiting linear response relations (Kappen and Rodríguez, 1998) or by expanding the free energy or its Legendre transforms (called variational thermodynamic potentials) around the factorising distribution (see e.g. Georges and Yedidia (1991); Plefka (1982)). ■

1.3.2.2 Pseudolikelihood

A different approach to inference in the Ising model (and other models) that is not based on Bayes' theorem is called the pseudolikelihood method (Besag, 1974). Originally, Besag conceived of a a random vector X, where the different components X_i are associated with positions on a lattice with direct interactions between the variables restricted to neighbouring lattice sites. For the Ising model, the components would be individual spin variables $X_i \in \{\pm 1\}$, which could, for example, sit on a 2D square lattice and interact only with their four nearest neighbours. However, the restriction to a lattice structure with nearest-neighbour interactions is not crucial and the method can be used even in infinite-dimensional models where all variables interact. The idea is to approximate the distribution of the random vector X as a product of distributions of the variables X_i conditioned on the values x_j of the neighbouring variables. Hence, instead of maximising the log-likelihood function (1.88) one aims to maximise a quantity

called the (log-)pseudolikelihood

$$\begin{aligned}\mathcal{L}_{\text{Pseudo}}(D;\Theta) &= \frac{1}{M}\sum_{\mu=1}^{M}\ln\prod_{i=1}^{N}\pi_\Theta(X_i = x_i^\mu|\{X_j = x_j^\mu\}_{j\neq i}) \\ &= \sum_{i=1}^{N}\frac{1}{M}\sum_{\mu=1}^{M}\ln\pi_\Theta(X_i = x_i^\mu|\{X_j = x_j^\mu\}_{j\neq i})\ . \end{aligned} \tag{1.106}$$

The advantage of using the pseudolikelihood is that the conditional distribution $\pi_\Theta(X_i|X_j(j\neq i))$ is relatively easy to compute, since its partition function is only a one-dimensional sum. Hence, computing the pseudolikelihood requires only N one-dimensional sums, i.e. a number of computational steps polynomial in N, rather than computing a single N-dimensional sum involving a number of steps growing exponentially in N.

1.3.3 Non-equilibrium inference from time-series data

If we have time-series data of the process, the inference becomes simpler than for the case of independent samples taken from the steady state (equilibrium or non-equilibrium). For a Markov process in discrete time, the data might consist of M independent trajectories of L time-steps, $D = \{x^\mu(0), x^\mu(1), \ldots, x^\mu(L)\}_{\mu=1}^{M}$. For sufficiently long trajectories, we can neglect the probability of the starting point and consider only the transitions. Then, the trajectories need not even be sampled from a stationary process and we do not need to compute the steady state distribution, since we can directly write down and maximise the (log-)likelihood of the trajectories

$$\begin{aligned}\mathcal{L}(D;\Theta) &= \frac{1}{M}\frac{1}{L}\ln P(D|\Theta) \\ &= \frac{1}{M}\sum_{\mu=1}^{M}\frac{1}{L}\ln P_\Theta(X(L) = x^\mu(L), \ldots, X(1) = x^\mu(1)|X(0) = x^\mu(0)) \\ &= \frac{1}{M}\sum_{\mu=1}^{M}\frac{1}{L}\sum_{t=0}^{L-1}\ln p_\Theta(x^\mu(t+1), t+1|x^\mu(t), t)\ , \end{aligned} \tag{1.107}$$

where the single-step propagators $p_\Theta(x^\mu(t+1), t+1|x^\mu(t), t)$ define the Markov process and are easily available.

Similarly, for processes in continuous time, we might have measurements of the process for discrete times $0, \Delta t, 2\Delta t, \ldots, L\Delta t$ and compute the (log-)likelihood

of the time-series

$$\mathcal{L}(D;\Theta) = \frac{1}{M}\sum_{\mu=1}^{M}\frac{1}{L}\sum_{n=0}^{L-1}\ln p_\Theta(x^\mu((n+1)\Delta t),(n+1)\Delta t|x^\mu(n\Delta t),n\Delta t)\ .$$

In continuous-time processes, the propagators $p_\Theta(x^\mu(n+1),(n+1)\Delta t|x^\mu(n\Delta t),n\Delta t)$ are not directly available, but for sufficiently short measurement time-intervals $\Delta t \ll 1$ we can approximate them as

$$p_\Theta(x^\mu(n+1),(n+1)\Delta t|x^\mu(n\Delta t),n\Delta t) \approx \Delta t\frac{\partial p_\Theta}{\partial t}(x^\mu(n\Delta t),n\Delta t)\ , \qquad (1.108)$$

where the infinitesimal generators $\frac{\partial}{\partial t}p_\Theta(x,t)$ define the Markov process and are easily available.

EXAMPLE: BIASED RANDOM WALK ON $\mathbb{N}_0$

We return to the example of the biased random walk on $\mathbb{N}_0$. The propagator is particularly simple:

$$p_r(x,t+1|y,t) = r\delta_{x,y+1} + (1-r)\delta_{x,y-1} + (1-r)\delta_{x,y}\delta_{y,0}\ . \qquad (1.109)$$

Defining the fraction of observed jumps to the right $\hat{r} = \frac{1}{M}\frac{1}{L}\sum_{\mu,t}\delta_{x_\mu(t+1),x_\mu(t)+1}$, the log-likelihood of the time-series becomes

$$\mathcal{L}(D|r) = \hat{r}\ln(r) + (1-\hat{r})\ln(1-r)\ ,$$

which gives the maximum likelihood estimate

$$r^{\mathrm{ML}} = \operatorname*{argmax}_r \mathcal{L}(D;r) = \hat{r}\ . \qquad (1.110)$$

Note that the maximum likelihood inference from time-series data is possible even in the case $r \geq 1/2$, where no steady state exists. ■

Since the evaluation of the likelihood of the time-series is easy and does not require knowledge of the steady state $\pi(x;\Theta)$, it is the first choice for inferring the parameters of a Markov process. However, for many systems, classical as well as quantum, time series data is not available. An extreme case is whole-genome single-cell gene expression profiling, where cells are destroyed by the measurement process. In such cases, we have only independent samples as data from which to infer the model parameters. This brings us to the problem of inferring the parameters of an ergodic Markov process based on independent samples taken from the non-equilibrium steady state $\pi(x;\Theta)$, which is unknown. We will address this problem by looking at specific models, with the asymmetric Ising model being the main paradigm, which we introduce in the following chapter before turning to the actual inference methods in chapters 3, 4, and 5.

2 The asymmetric Ising model

> The most successful elaboration of technique in statistical mechanics exists in connection with the Ising model.
>
> Gregory Hugh Wannier

In this chapter, we give some background on our main paradigm for Markov processes with a non-equilibrium steady state, the asymmetric Ising model. While the inference problem for the Ising model has been discussed extensively in the equilibrium case and also in the non-equilibrium case with time-series data, the problem of inferring the parameters from snapshots of the non-equilibrium steady state has not been addressed so far. We begin by giving a short overview of the model's history, before introducing Glauber dynamics and showing that this dynamics converges to a non-equilibrium steady state for the case of asymmetric couplings between spins. After briefly discussing the connection of the asymmetric Ising model with neural networks, which motivates the consideration of asymmetric couplings, we describe Callen's identities characterising the spin moments, since they will be used for inference in chapters 3 and 5. In the following, we present some minor results we found for the maximum likelihood inference based on time-series data for sequential Glauber dynamics, before discussing the Gaussian mean field theory of Mézard and Sakellariou (2011), which we will use for inference in chapter 5.

2.1 The model and its history

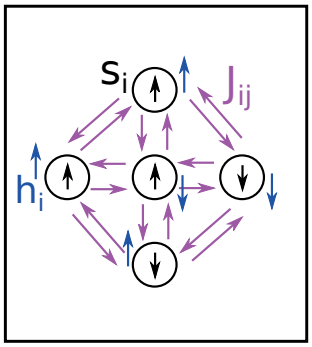


Figure 2.1: *Schematic view of the Ising model. Binary spins s_i interact via pairwise couplings J_{ij} and are subject to external magnetic fields h_i.*

The Ising model consists of a set of N binary spin variables $\mathbf{S} = (S_1, \ldots, S_N) \in \Omega = \{\pm 1\}^N$, which interact with each other via couplings J_{ij} and are subject to external fields h_i (see Fig. 2.1). The model was developed by Wilhelm Lenz (Lenz, 1920) for the purpose of explaining ferromagnetism and first solved in one dimension by Ernst Ising (Ising, 1925), showing that no phase transition exists. That we nowadays primarily associate Ising's name with the model is attributed to Rudolf Peierls (Peierls, 1936). Later, Heisenberg (Heisenberg, 1928) gave the quantum mechanical interpretation of the binary variables as atomic spins and the couplings J_{ij} as emanating from exchange interactions. For a two-dimensional square lattice, the Ising ferromagnet was solved by Lars Onsager (Onsager, 1944), while in three and higher dimensions, the ferromagnetic model remains unsolved. Later, non-uniform interactions were considered in order to model disordered spin glasses (Sherrington and Kirkpatrick, 1975). These Ising models used to model magnetic materials are equilibrium systems characterised by the Boltzmann distribution

$$\pi(\boldsymbol{s}; \mathbf{h}, J) = \frac{1}{Z(\mathbf{h}, J)} e^{-E(s; \mathbf{h}, J)/k_B T} \tag{2.1}$$

with energy function

$$E(\boldsymbol{s}; \mathbf{h}, J) = -\sum_{i=1}^{N} h_i s_i - \frac{1}{2} \sum_{i,j=1}^{N} J_{ij} s_i s_j \,, \tag{2.2}$$

with a symmetric coupling matrix J_{ij} without self-interactions, i.e. $J_{ii} \equiv 0$, and a factor $1/2$ ensuring that the bonds (i, j) are not counted twice. The stochastic element of the model arises from energy exchange with a heat bath of temperature T, i.e. the Ising model is considered in the canonical ensemble. From now on, we will measure the energy in units of the thermal energy $k_B T$ and subsume the factor $1/k_B T$ into the external fields and couplings.

2.2 Glauber dynamics

We have seen that we can define many different dynamics that have the Boltzmann distribution (2.1) as an equilibrium steady state. One common choice is **sequential Glauber dynamics** (Glauber, 1963): a spin i is chosen in each time step and its value $s_i(t+1)$ updated according to the probability distribution

$$\begin{aligned} p(s_i(t+1)|\boldsymbol{s}(t)) &:= p(S_i(t+1) = s_i(t+1)|\mathbf{S}(t) = \boldsymbol{s}(t)) \\ &= \frac{\exp\{s_i(t+1)\psi_i(t)\}}{2\cosh(\psi_i(t))} \,, \end{aligned} \tag{2.3}$$

Figure 2.2: *Ernst Ising, who was born 1900 in Cologne, studied his now famous model (originally developed by his PhD supervisor Wilhelm Lenz) in his PhD thesis, of which an excerpt was published in "Zeitschrift für Physik" (Ising, 1925). Ironically, this remained his only scientific publication, since he quit academia and, after a short stint at AEG, took several different positions as school teacher. Later, himself a Jew, he became the headmaster of a Jewish boarding school close to Potsdam, which was destroyed in the "Kristallnacht". However, Ising managed to escort his students home safely and a short while later he emigrated to Luxembourg, not without first being "interviewed" be the Gestapo. On his fortieth birthday, Ernst Ising got a surprise party of the very unpleasant kind: Luxembourg was invaded by the Nazis and in turn Ernst Ising became a forced labourer dismantling rail-roads on the Maginot line. After the war, he emigrated to America, where in 1947 he finally heard of the interest his model had attracted, when, looking for a job at a physics convention in Boston, he was asked whether he was the Ising of the "Ising model". Later, he became a physics professor at Bradley University in Peoria. In this position he did not carry out any research, but instead excelled at teaching with his charming motto that a lecture had failed if the students did not laugh at least once.*

where the effective local field at time t is

$$\psi_i(t) = h_i + \sum_{j=1}^{N} J_{ij} S_j(t) \; . \tag{2.4}$$

For a symmetric coupling matrix without self-couplings, the sequential Glauber dynamics (2.3) converges to the equilibrium state characterised by the Boltzmann distribution (2.1) as can be easily verified by inserting the transition rule (2.3) and Boltzmann distribution (2.1) with Hamiltonian (2.2) into the detailed balance condition (1.54).

In **parallel Glauber dynamics**, all spins are updated simultaneously according to (2.3), so we find

$$p(\boldsymbol{s}(t+1)|\boldsymbol{s}(t)) = \prod_{i=1}^{N} \frac{\exp\{s_i(t+1)\psi_i(t)\}}{2\cosh(\psi_i(t))} \; . \tag{2.5}$$

The parallel Glauber dynamics (2.5) converges to a different equilibrium steady state (Peretto, 1984), characterised by the Boltzmann distribution with the energy function

$$\tilde{E}(\boldsymbol{s};\mathbf{h},J) = -\sum_{i=1}^{N} h_i s_i - \sum_{i=1}^{N} \ln\left\{ 2\cosh\left(h_i + \sum_{j=1, j\neq i}^{N} J_{ij} s_j \right) \right\} \; . \tag{2.6}$$

2.2.1 Interaction symmetry and detailed balance

Besides its classical application in the study of the magnetic properties of solids, the Ising model has also been used to model gene regulatory and neural networks (Bailly-Bechet et al., 2010; Coolen, 2000a,b; Derrida et al., 1987; Hertz et al., 1991). In this biological context, there is no reason to assume symmetric couplings in the effective local fields (2.4). For example, a gene A might produce a protein that regulates the gene expression of another gene B, but this does not imply that gene B necessarily also regulates gene A. For asymmetric couplings, both sequential and parallel Glauber dynamics (2.3) converge to non-equilibrium steady states, which lack detailed balance and are hard to characterise (Coolen, 2000a). Since this statement is not directly obvious, we recall below the argument for sequential Glauber dynamics given by Coolen (2000a).

For checking the detailed balance condition (1.54), we consider an arbitrary configuration $\boldsymbol{s}$ and a spin-flip leading to the new configuration $\boldsymbol{s}' = F_k\boldsymbol{s} := (s_1, \ldots, s_{k-1}, -s_k, s_{k+1}, \ldots, s_N)$, where we defined the spin-flip operator F_k, which flips spin k but leaves the remaining spins unchanged[1]. The detailed balance

[1] All possible transitions can be described by a single spin-flip operator, since the transition probabilities in sequential Glauber dynamics are non-zero only for single spin-flips.

condition for the spin-flip transition $\boldsymbol{s} \to \boldsymbol{s}'$ and its reverse transition $\boldsymbol{s}' \to \boldsymbol{s}$ reads

$$p(\boldsymbol{s}' = F_k\boldsymbol{s}|\boldsymbol{s})\pi(\boldsymbol{s};\mathbf{h},J) = p(\boldsymbol{s}|\boldsymbol{s}' = F_k\boldsymbol{s})\pi(\boldsymbol{s}' = F_k\boldsymbol{s};\mathbf{h},J) \;, \tag{2.7}$$

which must be satisfied for any spin configuration $\boldsymbol{s} \in \{\pm 1\}^N$ and flipped spin $k \in \{1,\ldots,N\}$. Since the sequential Glauber dynamics (2.3) is obviously ergodic, all states have a non-zero probability in the steady state and without loss of generality we may write the steady state distribution in an exponential form

$$\pi(\boldsymbol{s};\mathbf{h};J) = \exp\left\{\sum_i h_i s_i + \frac{1}{2}\sum_{i\neq j} J_{ij}s_i s_j + K(\boldsymbol{s})\right\} \tag{2.8}$$

with some unknown function $K(\boldsymbol{s})$. Inserting this ansatz into the detailed balance condition (2.7) we find that the transition probabilities must satisfy

$$\begin{aligned} 0 =& \ln\left(\frac{p(\boldsymbol{s}'|\boldsymbol{s})\pi(\boldsymbol{s};\mathbf{h},J)}{p(\boldsymbol{s}|\boldsymbol{s}')\pi(\boldsymbol{s}';\mathbf{h},J)}\right) \\ =& K(\boldsymbol{s}) - K(\boldsymbol{s}') + \frac{1}{2}\sum_{i\neq j} J_{ij}(s_i s_j - s'_i s'_j) + \sum_i h_i(s_i - s'_i) \\ &+ (s'_k\psi_k(\boldsymbol{s}) - s_k\psi_k(\boldsymbol{s}')) + \ln\frac{2\cosh(\psi_k(\boldsymbol{s}'))}{2\cosh(\psi_k(\boldsymbol{s}))} \;. \end{aligned} \tag{2.9}$$

Since the transition involves only a single spin-flip, $s'_k = -s_k$ and $s_j = s'_j \;\; \forall j \neq k$, and since we excluded self-couplings, the effective local fields of the forward and reverse transition are identical, $\psi_k(\boldsymbol{s}) = \psi_k(\boldsymbol{s}')$, and the expression (2.9) simplifies to

$$\begin{aligned} 0 &= K(\boldsymbol{s}) - K(F_k\boldsymbol{s}) + \sum_j (J_{jk} + J_{kj})s_k s_j + 2h_k s_k - 2s_k\psi_k(\boldsymbol{s}) \\ &= K(\boldsymbol{s}) - K(F_k\boldsymbol{s}) + \sum_j (J_{jk} + J_{kj})s_k s_j - 2\sum_j J_{kj}s_k s_j \\ &= K(\boldsymbol{s}) - K(F_k\boldsymbol{s}) + \sum_j (J_{jk} - J_{kj})s_k s_j \;. \end{aligned} \tag{2.10}$$

It is easy to see from (2.10) that for symmetric couplings, $J_{jk} = J_{kj}$, detailed balance is fulfilled with $K(\boldsymbol{s}) = \text{const.}$, which becomes the equilibrium free energy. In order to see that symmetry is not only sufficient but also a necessary condition for detailed balance, we sort the spin-flip operator F_k and coupling terms to the two sides of the equation

$$(F_k - 1)K(\boldsymbol{s}) = \sum_j (J_{jk} - J_{kj})s_j s_k \;. \tag{2.11}$$

Applying a second spin-flip to a spin $i \neq k$ we find

$$(F_i - 1)(F_k - 1)K(\boldsymbol{s}) = (J_{ik} - J_{ki})s_i s_k \ . \tag{2.12}$$

Since the left-hand side is symmetric under the permutation of i and k, the right-hand side must also be, implying symmetric interactions if detailed balance holds.

2.2.2 Callen's identities

Even though the non-equilibrium steady state of the asymmetric Ising model is hard to characterise, we can derive exact self-consistent relationships, called Callen's identities (Callen, 1963) by averaging over the transition probabilities (2.3) for sequential Glauber dynamics or (2.5) for parallel Glauber dynamics. These will turn out to be useful in mean field theories of the Ising model. We begin by considering parallel Glauber dynamics, since the equations take a slightly simpler form than for sequential dynamics.

PARALLEL GLAUBER DYNAMICS

First, we define the time-dependent magnetisation as

$$m_i(t) := \langle S_i(t) \rangle = \sum_{\boldsymbol{s} \in \{\pm 1\}^N} p(\mathbf{S}(t) = \boldsymbol{s}) s_i \ . \tag{2.13}$$

This expectation value can be rewritten by conditioning on the value of the spin variables at time $t-1$, which expresses the magnetisation in terms of an average over the effective local field at the previous time-step $\psi_i(\boldsymbol{s}(t-1)$

$$\begin{aligned} m_i(t) &= \sum_{\boldsymbol{s} \in \{\pm 1\}^N} \sum_{\boldsymbol{s}' \in \{\pm 1\}^N} p(\mathbf{S}(t) = \boldsymbol{s} | \mathbf{S}(t-1) = \boldsymbol{s}') p(\mathbf{S}(t-1) = \boldsymbol{s}') s_i & (2.14) \\ &= \sum_{\boldsymbol{s}' \in \{\pm 1\}^N} \left\{ \sum_{s_i = \pm 1} s_i \frac{\exp\{s_i \psi_i(\boldsymbol{s}(t-1))\}}{2\cosh(\psi_i(\boldsymbol{s}(t-1))} \right\} p(\mathbf{S}(t-1) = \boldsymbol{s}') & (2.15) \\ &= \sum_{\boldsymbol{s}' \in \{\pm 1\}^N} \tanh(\psi_i(\boldsymbol{s}')) p(\mathbf{S}(t-1) = \boldsymbol{s}') \\ &= \langle \tanh(\psi_i(\mathbf{S}(t-1)) \rangle \ . & (2.16) \end{aligned}$$

Next, we define the fluctuations of a spin variable around its mean $\delta S_i(t) := (S_i(t) - m_i(t))$ and consider correlations of these fluctuations. The equal-time two-point connected correlations $C_{ij}(t) = \langle \delta S_i(t) \delta S_j(t) \rangle$ $(i < j)$ can be com-

puted by the same conditioning procedure as for the magnetisations and we obtain

$$\begin{aligned} C_{ij}(t) &:= \langle \delta S_i(t) \delta S_j(t) \rangle \\ &= \left\langle [\tanh(\psi_i(\mathbf{S}(t-1))) - m_i(t)] \, [\tanh(\psi_j(\mathbf{S}(t-1))) - m_j(t)] \right\rangle \ . \end{aligned} \tag{2.17}$$

Similarly, we the time-shifted connected correlations are given by

$$D_{ij}(t) := \langle \delta S_i(t+1) \delta S_j(t) \rangle = \left\langle [\tanh(\psi_i(\mathbf{S}(t)) - m_i(t)] \, \delta S_j(t) \right\rangle \ . \tag{2.18}$$

More generally, we can define n-point connected correlations

$$\begin{aligned} C_{i_1,i_2,\ldots,i_n}(t) &:= \langle \delta S_{i_1}(t) \delta S_{i_2}(t) \cdots \delta S_{i_n}(t) \rangle \\ &= \left\langle \prod_{k=1}^{n} [\tanh(\psi_{i_k}(\mathbf{S}(t-1)) - m_{i_k}(t)] \right\rangle \ , \end{aligned} \tag{2.19}$$

where $\{i_1 < \ldots < i_n\} \subset \{1, \ldots N\}$ is a subset of n spins.

In the steady state these expectation values become time-independent and we find the self-consistent expressions of the spin statistics

$$m_i = \langle S_i \rangle = \langle \tanh(\psi_i) \rangle = \left\langle \tanh\left(h_i + \sum_j J_{ij} S_j \right) \right\rangle \tag{2.20}$$

$$C_{ij} = \left\langle \left[\tanh\left(h_i + \sum_k J_{ik} S_k \right) - m_i \right] \left[\tanh\left(h_j + \sum_l J_{jl} S_l \right) - m_j \right] \right\rangle \ , \tag{2.21}$$

$$D_{ij} = \left\langle \left[\tanh\left(h_i + \sum_k J_{ik} S_k \right) - m_i \right] \delta S_j \right\rangle \ , \tag{2.22}$$

where the expectation values $\langle . \rangle$ are taken with respect to the steady state distribution $\pi(\boldsymbol{s}; \mathbf{h}, J)$.

Sequential Glauber dynamics

For sequential updates, we have to take into account that any given spin has only a chance of $1/N$ to be flipped in a specific time-step, hence

$$m_i(t) = \frac{1}{N} \langle \tanh(\psi_i(\mathbf{S}(t-1)) \rangle \ + \frac{N-1}{N} m_i(t-1) \ , \tag{2.23}$$

$$\begin{aligned} C_{ij}(t) =& \frac{1}{N}\left\langle [\tanh(\psi_i(\mathbf{S}(t-1)) - m_i(t)]\, \delta S_j(t-1)\right\rangle \\ &+ \frac{1}{N}\left\langle \delta S_i(t-1)\left[\tanh(\psi_j(\mathbf{S}(t-1)) - m_j(t)\right]\right\rangle \\ &+ \frac{N-2}{N} C_{ij}(t-1)\ , \end{aligned} \tag{2.24}$$

$$D_{ij}(t) = \frac{1}{N}\left\langle [\tanh(\psi_i(\mathbf{S}(t)) - m_i(t)]\, \delta S_j(t)]\right\rangle + \frac{N-1}{N} C_{ij}(t)\ , \tag{2.25}$$

$$C_{i_1,i_2,\ldots,i_n} = \left\langle \prod_{k=1}^{n} \left[\tanh\left(h_{i_k} + \sum_l J_{i_k l} S_l \right) - m_{i_k} \right] \right\rangle\ . \tag{2.26}$$

In the steady state these expectation values become time-independent and simplify to

$$m_i = \langle S_i \rangle = \langle \tanh(\psi_i) \rangle = \left\langle \tanh\left(h_i + \sum_j J_{ij} S_j \right) \right\rangle \tag{2.27}$$

$$\begin{aligned} C_{ij} =& \frac{1}{2}\left\langle \left[\tanh\left(h_i + \sum_k J_{ik} S_k \right) - m_i \right] \delta S_j \right\rangle \\ &+ \frac{1}{2}\left\langle \delta S_i \left[\tanh\left(h_j + \sum_l J_{jl} S_l \right) - m_j \right] \right\rangle\ , \end{aligned} \tag{2.28}$$

$$D_{ij} = \frac{1}{N}\left\langle \left[\tanh\left(h_i + \sum_k J_{ik} S_k \right) - m_i \right] \delta S_j \right\rangle + \frac{N-1}{N} C_{ij}\ , \tag{2.29}$$

$$C_{i_1,i_2,\ldots,i_n} = \frac{1}{n}\sum_{k=1}^{n} \left\langle \left(\prod_{j=1, j\neq k}^{n} \delta S_{i_j} \right) \left[\tanh\left(h_{i_k} + \sum_l J_{i_k l} S_l \right) - m_{i_k} \right] \right\rangle\ . \tag{2.30}$$

2.3 Connection with neural networks

Having mentioned that the Ising model has been used to model neural networks, we now want to shortly elucidate this connection for the purpose of motivating the consideration of asymmetric couplings. In a simplistic, physicist-style view, the brain may be considered as consisting of a large number of identical building blocks - the neurons. These are connected by synapses that transfer electrical signals from one neuron to another. A neuron's configuration may be classified

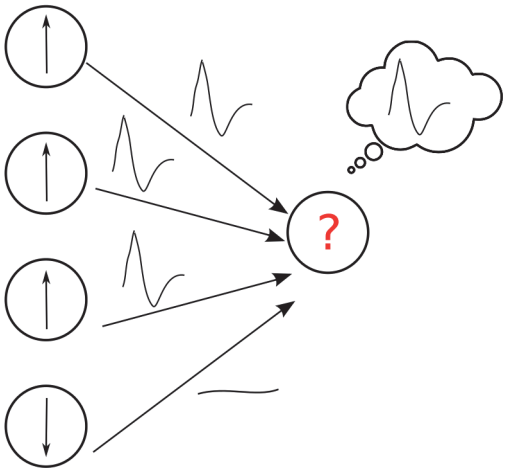


Figure 2.3: *A simplistic view of neural computation as an Ising model.*

by whether it is active (producing electrical signals - the action potential) or inactive. In this simplistic model, we may represent the configuration of the brain by a vector $\boldsymbol{s}(t)$ of Ising spins. At the lowest level, information processing in the brain is performed by the individual neurons, which execute a nonlinear transformation of incoming electrical signals transferred via the synapses from other neurons. The result of this transformation is a decision whether the neuron itself becomes active active and sends electrical signals. Non-linearity is an essential feature, since it produces much richer and complexer behaviour than linear dynamics. A simple way to model this non-linearity is to use a step function, i.e. the neuron fires when the incoming signals exceed a certain threshold (corresponding to the local fields h_i). Modelling time in discrete steps, which should be sufficiently short so we can assume that only one neuron updates its configuration at a time, we may describe neural computation by the deterministic dynamics

$$S_i(t+1) = \operatorname{sgn}\left(\sum_{j=1,j\neq i}^{N} J_{ij}S_j(t) + h_i\right) = \begin{cases} 1 & \text{if } \sum_j J_{ij}S_j(t) > -h_i \\ -1 & \text{if } \sum_j J_{ij}S_j(t) \leq -h_i \end{cases} . \quad (2.31)$$

The neural couplings J_{ij} may be of varying strength with a positive sign corresponding to a stimulative synapse and a negative sign corresponding to an inhibitory one. In the context of modelling associative memory, this Markov process is known as the **Hopfield model** when the neural couplings are chosen symmetrically according to the Hebbian learning rule (Hertz et al., 1991)

$$J_{ij} = \frac{1}{p}\sum_{\nu=1}^{p} \xi_i^{(\nu)}\xi_j^{(\nu)} \ , \ (i \neq j) \ , \quad (2.32)$$

where the configurations $\xi^\nu \in \{\pm 1\}^N$ are neural patterns that should be learned. In the Hopfield model, a learned configuration corresponds to an attractor of the dynamics (2.31). If not too many patterns are to be stored, there exist separate basins of attractions, i.e. the Ising spin system initialised to a configuration similar to a stored configuration ξ^ν, will converge to this attractor, i.e. the pattern is recalled (Hertz et al., 1991).

Since biological systems are intrinsically noisy, we should add a random force η to the incoming signal. Choosing a random force with probability density $p(z) = (1 - \tanh^2(z))$ recovers sequential Glauber dynamics (2.3) (for a proof see Coolen (2000a)). For symmetric couplings the resulting Markov process is known as a **Boltzmann machine** (Ackley et al., 1985).

2.4 Stochastic inference from time-series data

We now seek to infer the parameters $\Theta = (\{h_i\}, \{J_{ij}\})$ from some data set D. The inference problem for independent samples taken from the steady state has been successfully addressed for equilibrium systems, characterised by symmetric couplings and detailed balance. A comprehensive review can be found in Nguyen et al. (2017). Next, we sketch the state-of-the art inference methods for non-equilibrium inference in the asymmetric Ising model, which rely on time-series data (see also Nguyen et al. (2017) for a review). We will add some simple results we found for the maximum likelihood inference in the case of discrete time sequential Glauber dynamics and discuss the Gaussian mean field theory of Mézard and Sakellariou (2011) and the role of higher-order correlations. These higher-order correlations will become important for our work addressing the inference problem when only snapshots of the steady state are available, which will be discussed in chapters 3 and 4. Additionally, in chapter 5 we will adapt the Gaussian mean field theory in order to infer the parameters based on independent samples drawn from several steady states generated by perturbing the couplings.

2.4.1 Maximum likelihood of time-series

As discussed in section 1.3.3, inferring the parameters of a Markov process is a comparatively easy task when we can observe time series of consecutive configurations of the system $\boldsymbol{s}(t), \boldsymbol{s}(t+1), \boldsymbol{s}(t+2), \ldots$. Using the dynamical rule (2.3), the probability of a particular trajectory $\prod_t p(\boldsymbol{s}(t+1)|\boldsymbol{s}(t))$ can be written down explicitly and be maximised with respect to the couplings and fields in polynomial time in N and the length of the trajectory. This is easiest for parallel Glauber dynamics and was done by Roudi and Hertz (2011).

2.4.1.1 Paralell Glauber dynamics

We consider time-series data consisting of M trajectories of length L, $D = \{\boldsymbol{s}^\mu(0), \boldsymbol{s}^\mu(t = 1), \ldots \boldsymbol{s}^\mu(t = L)\}_{\mu=1}^M$. With the dynamical rule (2.5) we can write the log-likelihood of the time-series (1.107) as

$$\begin{aligned}\mathcal{L}(D;\mathbf{h},J) &= \frac{1}{M}\sum_{\mu=1}^{M}\frac{1}{L}\sum_{t=0}^{L-1}\ln p_{\mathbf{h},J}(\boldsymbol{s}^\mu(t+1)|\boldsymbol{s}^\mu(t))) \\ &= \frac{1}{M}\sum_{\mu=1}^{M}\frac{1}{L}\sum_{t=0}^{L-1}\sum_{i=1}^{N}\ln\left(\frac{\exp\{s_i^\mu(t+1)\psi_i(\boldsymbol{s}^\mu(t))\}}{2\cosh(\psi_i(\boldsymbol{s}^\mu(t))}\right) \\ &= \frac{1}{M}\sum_{\mu=1}^{M}\frac{1}{L}\sum_{t=0}^{L-1}\sum_{i=1}^{N}\left[s_i^\mu(t+1)\psi_i(\boldsymbol{s}^\mu(t)) - \ln 2\cosh(\psi_i(\boldsymbol{s}^\mu(t))\right] .\end{aligned} \tag{2.33}$$

The likelihood (2.33) is concave with respect to the parameters h, J, so it has a unique maximum, which can be found by simply climbing up the gradient (Roudi and Hertz, 2011). For equilibrium Ising models, this gradient ascent procedure is known as Boltzmann machine learning. The trial fields h_i and couplings J_{ij} are updated by increments

$$\delta h_i \sim \frac{\partial\mathcal{L}(D;\mathbf{h},J)}{\partial h_i} = \langle S_i(t+1)\rangle_D - \left\langle \tanh\left(h_i + \sum_j J_{ij}S_j(t)\right)\right\rangle_D \tag{2.34}$$

and

$$\begin{aligned}\delta J_{ij} &\sim \frac{\partial\mathcal{L}(D;\mathbf{h},J)}{\partial J_{ij}} \\ &= \left\langle S_i(t+1)S_j(t)\right\rangle_D - \left\langle S_j(t)\tanh\left(h_i + \sum_j J_{ij}S_j(t)\right)\right\rangle_D ,\end{aligned} \tag{2.35}$$

where the average $\langle f(\mathbf{S})\rangle_D = \frac{1}{M}\frac{1}{L}\sum_{\mu=1}^M\sum_{t=0}^{L-1} f(\boldsymbol{s}^\mu(t))$ is taken over the sampled time-series data D and the proportionality constant is known as the learning rate. These learning steps have a nice intuitive interpretation: the fields h_i are determined such that Callen's identities (2.16) for the magnetisations $m_i(t+1)$ are fulfilled in the sample average and the couplings J_{ij} are determined such that Callen's identities (2.16) for the time-shifted correlations $D_{ij}(t)$ are fulfilled in the sample average.

2.4.1.2 Sequential Glauber dynamics

For sequential Glauber dynamics the procedure is slightly more complicated, since only one (random) spin is updated in a given time-step. If the updated

spin is not flipped, then one has no way to know which spin was chosen for updating. The problem for sequential Glauber dynamics in continuous time was thoroughly treated by Zeng et al. (2013). In this thesis we consider discrete-time Glauber dynamics. We find that the likelihood (2.33) can be split into a part $\mathcal{L}_F$, involving time-steps where a spin was flipped, and a part $\mathcal{L}_U$, involving time-steps where no spin was flipped; both parts are required for a correct maximum likelihood estimate of the parameters h and J. Interestingly, we found that the antisymmetric part of the coupling matrix $J^{\text{as}}_{ij} = (J_{ij} - J_{ji})/2$ can be inferred from $\mathcal{L}_F$ alone, which suggests different roles of the symmetric and antisymmetric part in controlling the dynamics of spin-flips (see Fig. 2.4).

We define F as the set of time-steps (μ, t) in which a spin was flipped and U as the set of time-steps where no spin was flipped. Further, we decompose F into subsets F_i consisting of time-steps where spin i was updated. With these definitions we write the log-likelihood function (1.107)

$$\mathcal{L}(D;\mathbf{h},J) = \underbrace{\sum_{(t,\mu)\in F} \ln P(\boldsymbol{s}^\mu(t+1)|\boldsymbol{s}^\mu(t))}_{=:\mathcal{L}_F} + \underbrace{\sum_{(t,\mu)\in U} \ln P(\boldsymbol{s}^\mu(t+1)|\boldsymbol{s}^\mu(t))}_{=:\mathcal{L}_U}, \quad (2.36)$$

In the case where no spin was flipped, we simply take a uniform average over all the N spins that could have been updated, yielding the learning steps

$$\begin{aligned}\delta h_i &\sim \frac{\partial}{\partial h_i}\mathcal{L}(D;\mathbf{h},J) \\ &= \sum_{(\mu,t)\in D} \phi_i^\mu(t)\left[s_i^\mu(t+1) - \tanh\left(h_i + \sum_j J_{ij}s_j^\mu(t)\right)\right] \end{aligned} \quad (2.37)$$

$$\begin{aligned}\delta J_{ij} &\sim \frac{\partial}{\partial J_{ij}}\mathcal{L}(D;\mathbf{h},J) \\ &= \sum_{(\mu,t)\in D} \phi_i^\mu(t)\left[s_i^\mu(t+1)s_j^\mu(t) - \tanh\left(h_i + \sum_k J_{ik}s_k^\mu(t)\right)s_j^\mu(t)\right] \end{aligned} \quad (2.38)$$

where we defined the auxiliary quantity

$$\phi_i^\mu(t) = \begin{cases} 1 & (t,\mu)\in F_i \\ q_i^\mu(t) & (t,\mu)\in U \\ 0 & (t,\mu)\in F\setminus F_i \end{cases} \quad (2.39)$$

and where

$$q_i^\mu(t) = \frac{\exp(s_i^\mu(t+1)\psi_i^\mu(t))}{2\cosh(\psi_i^\mu(t))}\left(\sum_{k=1}^N \frac{\exp(s_k^\mu(t+1)\psi_k^\mu(t))}{2\cosh(\psi_k^\mu(t))}\right)^{-1} \quad (2.40)$$

denotes the probability that spin i was updated but not flipped in step $t \to t+1$ of trajectory μ, given that no spin was flipped.

In Fig. 2.4 we show numerical results for the coupling inference in an asymmetric version of the Sherrington-Kirkpatrick-Model, i.e. J_{ij} and J_{ji} are independent quenched random variables drawn form a normal distribution with mean 0 and variance $1/N$.

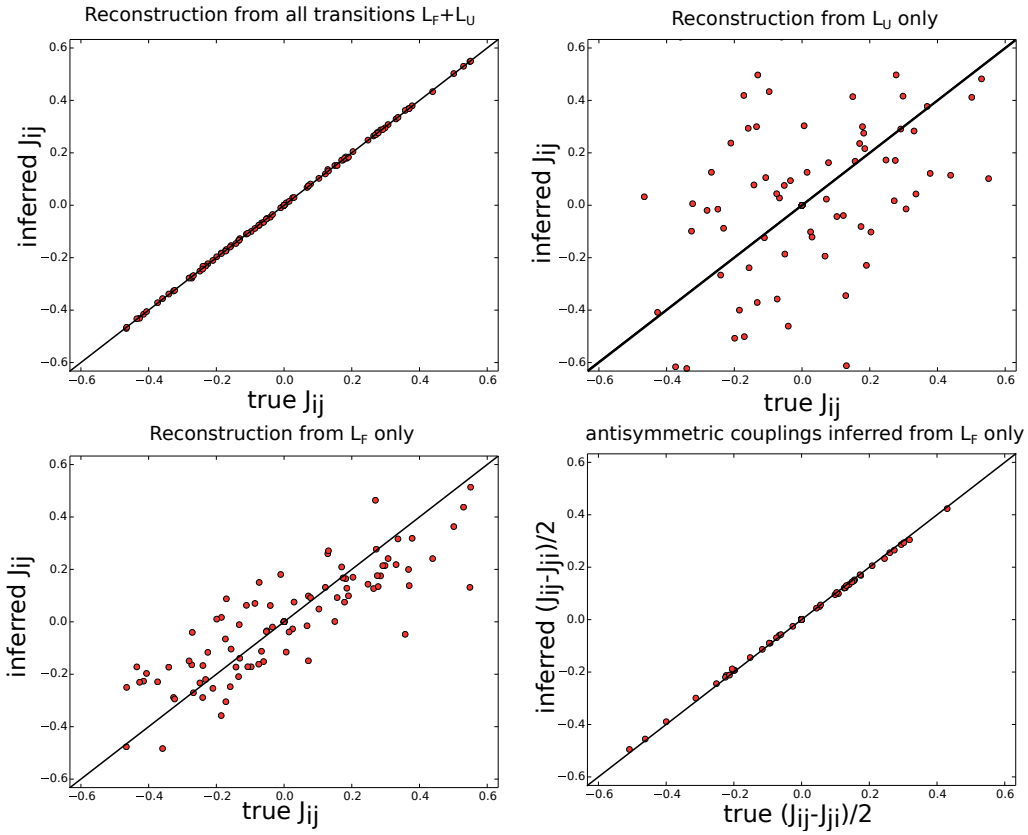


Figure 2.4: *Coupling reconstruction from maximising the likelihood of time-series in sequential Glauber dynamics. $\mathcal{L}_F$ denotes the likelihood of all time-steps where a spin was flipped, $\mathcal{L}_U$ denotes the likelihood of all time-steps where no spin was flipped. The underlying couplings were quenched random variables drawn form a normal distribution with mean 0 and variance $1/N$ with full asymmetry, i.e. J_{ij} and J_{ji} were drawn independently without self-couplings ($J_{ii} \equiv 0$). The system consisted of 10 Ising spins and the data set consisted of 100 trajectories of 10^5 time-steps each.*

An estimate of the couplings and fields that can be computed even faster has been derived using different variations of mean field theory involving the computation of time-shifted correlations $D_{ij} = \langle \delta S_i(t+1) \delta S_j(t) \rangle$ (Mézard and Sakellariou, 2011; Roudi and Hertz, 2011). In the following, we will take a

closer look at the Gaussian mean field theory that becomes exact in the thermodynamic limit for the completely asymmetric Sherrington-Kirkpatrick-model. The reasons for this are twofold: first, we will adapt the theory in chapter 5 in order to infer the parameters from independent samples taken from several, perturbed steady states; second, we will show that three-point correlations are small compared to magnetisations and two-point correlations, which explains the main difficulty in inferring the full coupling matrix from independent samples taken from the steady state (see chapter 3).

2.4.2 The Gaussian mean field theory and time-shifted correlations

The Gaussian mean-field theory was developed by Mézard and Sakellariou (2011) for the asymmetric Ising model with parallel Glauber dynamics. The theory allows to compute the averages in the right-hand sides of Callen's identities for the magnetisations (2.20) and time-shifted correlations (2.22) in a way that is exact in the thermodynamic limit for the fully asymmetric Sherrington-Kirkpatrick-Model. This model is characterised by couplings J_{ij} that are quenched random variables with J_{ij} and J_{ji} drawn independently from a Gaussian with zero mean and variance β/N, where β describes the interaction strength. The intuition is that the effective local fields $\psi_i = h_i + \sum_{j=1}^{N} J_{ij} S_j$ become Gaussian random variables in the limit $N \to \infty$. While there is only a heuristic argument but no proof of the applicability of a central limit theorem, empirically, the effective fields become normally distributed in the thermodynamic limit. When the asymmetry is broken by correlating the entries J_{ij} and J_{ji}, the theory can still be used as a reasonable approximation (Sakellariou et al., 2012). In fact, the term mean field theory is a bit of a misnomer, since the full distribution of the effective fields ψ_i is considered and not just their means. However, in the literature (including the original paper) this approach is commonly referred to simply as "mean field theory". Hence, we make a compromise and add the prefix "Gaussian", in order to distinguish it from the standard mean field theory.

The theory proceeds by decomposing the effective fields ψ_i into their deterministic mean g_i and a Gaussian fluctuation x_i around this mean

$$\psi_i = g_i + x_i \ , \ x_i \sim \mathcal{N}(0, \Delta_i) \tag{2.41}$$

where the mean is given by

$$g_i = \langle \psi_i \rangle = h_i + \sum_j J_{ij} m_j \tag{2.42}$$

and the variance by

$$\begin{aligned}\Delta_i &= \langle(\psi_i - g_i)^2\rangle = \sum_{j,k} J_{ij}J_{ik}\langle\delta S_j \delta S_k\rangle \\ &\approx \sum_j J_{ij}^2(1-m_j^2)\ ,\end{aligned} \tag{2.43}$$

where the expectation values are taken in the steady state and we have used that the last double-sum is dominated by the diagonal entries, since equal-time pairwise spin-correlations scale as $1/\sqrt{N}$ and hence become small in the thermodynamic limit (Mézard and Sakellariou, 2011).

Since the magnetisations are determined solely by the statistics of their corresponding effective field, they can be obtained by integrating over the Gaussian fields

$$m_i = \langle\tanh(\psi_i)\rangle = \int_{-\infty}^{\infty}\frac{\mathrm{d}y}{\sqrt{2\pi}}e^{-y^2/2}\tanh\left(g_i + y\sqrt{\Delta_i}\right)\ . \tag{2.44}$$

For the time-shifted correlations (2.22)

$$D_{ij} = \langle[\tanh(\psi_i) - m_i]\delta S_j\rangle\ , \tag{2.45}$$

we need the distribution of δS_j, which can be expressed in terms of the field statistics by multiplying the equation with the coupling matrix

$$\sum_j J_{kj}D_{ij} = \left\langle[\tanh(\psi_i) - m_i]\sum_j J_{kj}\delta S_j\right\rangle = \langle[\tanh(g_i + x_i) - m_i]x_k\rangle\ . \tag{2.46}$$

In order to evaluate this average, we approximate the joint distribution of x_i and x_k as a two-dimensional Gaussian with correlation

$$\rho_{ik} = \frac{\langle x_i x_k\rangle}{\sqrt{\Delta_i}\sqrt{\Delta_j}} = \frac{\left\langle\sum_j J_{ij}\delta S_j \sum_l J_{kl}\delta S_l\right\rangle}{\sqrt{\Delta_i}\sqrt{\Delta_j}} = \frac{(JCJ^T)_{ik}}{\sqrt{\Delta_i}\sqrt{\Delta_j}}\ . \tag{2.47}$$

Since the correlations are of order $1/\sqrt{N}$, the density of the joint distribution can be expanded to first order in the correlations ρ_{ik}

$$\begin{aligned}P(x_i,x_k) &= \frac{1}{2\pi\sqrt{\Delta_i\Delta_k}}\exp\left\{-\frac{x_i^2}{2\Delta_i} - \frac{x_k^2}{2\Delta_k} + \rho_{ik}\frac{x_i x_k}{\sqrt{\Delta_i}\sqrt{\Delta_k}}\right\} \\ &\approx \frac{1}{2\pi\sqrt{\Delta_i\Delta_k}}\exp\left\{-\frac{x_i^2}{2\Delta_i} - \frac{x_k^2}{2\Delta_k}\right\}\left(1 + \rho_{ik}\frac{x_i x_k}{\sqrt{\Delta_i}\sqrt{\Delta_k}}\right)\ .\end{aligned} \tag{2.48}$$

Equipped with the joint distribution (2.48), we compute the average on the right-hand side of (2.46)

$$\begin{aligned}\langle[\tanh(g_i+x_i)-m_i]x_k\rangle =& \int_{-\infty}^{\infty}\frac{\mathrm{d}y}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{\mathrm{d}z}{\sqrt{2\pi}}[\tanh(g_i+y\sqrt{\Delta_i})-m_i]z\sqrt{\Delta_k}\\ &\times e^{-y^2/2}e^{-z^2/2}\,(1+\rho_{ik}yz)\\ =&\rho_{ik}\sqrt{\Delta_k}\int_{-\infty}^{\infty}\frac{\mathrm{d}y}{\sqrt{2\pi}}e^{-y^2/2}y[\tanh(g_i+y\sqrt{\Delta_i})-m_i]\\ =&\rho_{ik}\sqrt{\Delta_k}\sqrt{\Delta_i}\int_{-\infty}^{\infty}\frac{\mathrm{d}y}{\sqrt{2\pi}}e^{-y^2/2}[1-\tanh^2(g_i+y\sqrt{\Delta_i})]\ .\end{aligned} \tag{2.49}$$

Inserting this result and the expression (2.47) for ρ_{ik} into (2.46) we find

$$(DJ^T)_{ik}=\lambda_i(JCJ^T)_{ik}\ , \tag{2.50}$$

or equivalently by multiplying with $(J^T)^{-1}$ from the right (J is invertible with probability one)

$$D_{ij}=\lambda_i(JC)_{ij}\ , \tag{2.51}$$

where we defined the auxiliary quantity

$$\lambda_i:=\int_{-\infty}^{\infty}\frac{\mathrm{d}y}{\sqrt{2\pi}}e^{-y^2/2}[1-\tanh^2(g_i+y\sqrt{\Delta_i})]\ . \tag{2.52}$$

SOLVING THE INVERSE PROBLEM WITH TIME-SERIES DATA

Given time-series data from a stationary Ising model with parallel Glauber dynamics, we can compute the empirical estimates of the magnetisations m_i, equal-time connected correlations C_{ij} and time-shifted correlations D_{ij}. In the first order mean field approximation of Roudi and Hertz (2011), we would obtain an equation identical to (2.51) with $\lambda_i=1-m_i^2$, so we could directly invert (2.51) to obtain the coupling matrix J. In the Gaussian inference, a direct inversion is possible only for the case of vanishing fields and magnetisations. In the more general case, λ_i is a non-linear function of the couplings and external fields and we need to jointly solve (2.51) and (2.44) by an iteration procedure (Mézard and Sakellariou, 2011).

HIGHER ORDER CORRELATIONS

As the inference procedure described above requires the time-shifted correlations D_{ij}, we require time-series data for this method. If we have only independent samples taken from the steady state, magnetisations m_i and correlations C_{ij} are not sufficient for inferring an asymmetric coupling matrix J_{ij} (see chapter 3

for a detailed discussion). For this reason, we would be interested in computing higher-order correlations. Unfortunately, however, the Gaussian mean-field theory as described above is not able to compute non-trivial higher-order correlations. For example, consider the three-point connected correlations (3.53)

$$C_{ijk} = \langle \delta S_i \delta S_j \delta S_k \rangle = \langle [\tanh(\psi_i) - m_i][\tanh(\psi_j) - m_j][\tanh(\psi_k) - m_k] \rangle \ . \quad (2.53)$$

Above, we assumed a joint Gaussian distribution for a pair of effective fields ψ_i, ψ_j to compute the time-shifted correlation D_{ij}. If we also assume a joint Gaussian distribution for the triplet of effective fields ψ_i, ψ_j, ψ_k, it is easy to see that the average (2.53) becomes zero. Similarly, four-point correlations will be formed by linear combinations of the two-point correlations so no additional information is gained. Since the Gaussian mean field theory is exact up to order $\mathcal{O}(1/N)$, this shows that the three-point correlations are of order $o(1/N)$. Hence, in order to infer the asymmetric coupling matrix from snapshots of the steady state, we need to consider corrections of order $o(1/N)$ to the spin-correlations. We will do this in chapter 3 within the mean field expansion of Kappen and Spanjers (2000) and show that the three-point correlations are in fact of order $\mathcal{O}(1/N^2)$.

3

Self-consistent equations and non-equilibrium mean field theory

> Elegance should be left to shoemakers and tailors.
>
> Ludwig Boltzmann

In this chapter, we develop our first method for stochastic inference from snapshots of the steady state, which is based on self-consistent equations that link the model parameters to observable statistics. We begin by showing how to derive the self-consistent equations, which can be considered generalisations of Callen's identities (cp. section 2.2.2), for the different classes of Markov processes. Next, we describe how the obtained expressions can be used to infer model parameters from independent samples taken from the steady state by replacing the steady state expectation values with sample averages and fitting the self-consistent equations to the data. In the following, we show how the self-consistent expressions can be evaluated approximately in a closed-form within a non-equilibrium mean field theory, thus yielding a computationally more efficient inference algorithm. The presentation of the non-equilibrium mean field theory was inspired by the work of Kappen and Spanjers (2000), who suggested how to extend the equilibrium free energy expansion of Plefka (1982) to the non-equilibrium case of the asymmetric Ising model and computed the magnetisations and two-point correlations to second order in the couplings, corresponding to the equilibrium TAP equations. Here, we provide a straightforward generalisation of their approach and formulate the theory for a wider class of ergodic Markov processes.

As an application, we consider inferenence in the asymmetric Ising model with sequential Glauber dynamics. First, we argue that for a successful inference, three-point correlations are needed in addition to the magnetisations and two-point correlations. We recall the relevant Callen identities for the magnetisations, two- and three-point correlations, which we will to fit to the data. Next, we turn to their mean field approximation and retrace the calculations of Kappen and Spanjers (2000) to compute the magnetisations and two-point correlations in a mean field expansion to second order in the couplings, before continuing with an expansion of the three-point correlations to second order in the couplings. Analysing the symmetries exhibited by the mean field equations, we argue that the expansion needs to be continued to third order in the couplings in order to successfully infer the model parameters. Hence, we also compute the third order corrections for the magnetisations, two-point, and three-point correlations. Finally, we use the two approaches of exact sample averaging and mean field approximation to infer the external fields and couplings of a fully asymmetric Sherrington-Kirkpatrick model consisting of $N = 10$ spins and show that we can distinguish the steady states generated by parallel and sequential Glauber dynamics based on the three-point correlations.

3.1 The general theory

We consider a family of ergodic Markov processes characterised by the parameters $\Theta \in \Lambda \subset \mathbb{R}^k$ and (non-equilibrium) steady state distributions $\pi(x;\Theta)$ on a configuration space Ω. The problem is that we do not know the steady state distribution but want to infer the parameters Θ^{true} underlying a specific process based on samples $D = \{x^\mu\}_{\mu=1}^M$, which were drawn independently from the steady state.

3.1.1 Deriving self-consistent equations

In contrast to equilibrium inference, the steady state is not described by the Boltzmann distribution, hence we do not have an explicit expression for the likelihood to maximise. Instead, we compute steady state statistics of different observables like means $m_i = \langle X_i \rangle_\pi$ and link them to the model parameters. Since we do not know the steady state distribution $\pi(x;\Theta)$, we cannot determine this link directly. The idea is to use self-consistent equations, like Callen's identities for the Ising model (cp. section 2.2.2). In the following, we show how these self-consistent equations can be obtained for the different types of Markov processes and give toy examples, before explaining how these equations can be used for inferring the model parameters.

3.1.1.1 Discrete-time Markov chains

For the asymmetric Ising model, we saw in section 2.2.2 that self-consistent equations can be derived by averaging over the transition probabilities, giving us Callen's identities. These can be in the time-dependent form or in their simpler steady state version. Since we are interested in inference from snapshots of the steady state, we focus on single-time averages in the steady state. Generalising the approach of averaging over the single-step transition probabilities, we can derive a self-consistent equation for the steady state mean of an observable $O(X)$

$$\begin{aligned}\langle O(X)\rangle_\pi &= \sum_x O(x)\pi(x;\Theta) = \sum_x \sum_y O(x) p_\Theta(x|y)\pi(y;\Theta) \\ &= \sum_y G(y,\Theta)\pi(y;\Theta) = \langle G(y,\Theta)\rangle_\pi =: g(\Theta)\end{aligned} \tag{3.1}$$

for some function $G(y;\Theta) = \sum_x O(x) p_\Theta(x|y)$ that involves the model parameters. For example, in the asymmetric Ising model with sequential Glauber dynamics, we considered the magnetisation of spin i, $O(X) = X_i$ and found the corresponding function was $G(y;h,J) = \tanh(h_i + \sum_j J_{ij} y_j)$.

3.1.1.2 Continuous-time Markov chains

In continuous-time Markov chains, the derivation of self-consistent equations is not quite as straightforward, but follows the same procedure. In the steady state, the mean of an observable $O(X(t))$ should become time-independent, hence

$$\begin{aligned} 0 &= \lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}t}\langle O(X(t))\rangle = \lim_{t\to\infty}\sum_x O(x)\frac{\mathrm{d}}{\mathrm{d}t}P(X(t)=x) \\ &= \lim_{t\to\infty}\sum_x O(x)\sum_y \tilde{K}_{yx}(\Theta)P(X(t)=y) = \sum_y\sum_x O(x)\tilde{K}_{yx}(\Theta)\pi(y;\Theta) \\ &= \langle G(y;\Theta)\rangle_\pi := g(\Theta) \end{aligned} \tag{3.2}$$

with some function $G(y;\Theta) = \sum_x O(x)\tilde{K}_{yx}(\Theta)$ and the transition rate matrix $\tilde{K}$ as defined in (1.15). Note that, in contrast to the discrete-time chains, we have not self-consistently re-expressed the mean of the observable $O(x)$, but instead found the relationship $0 = \langle G(y;\Theta)\rangle_\pi$ for some different observable G. Nonetheless, these relationships are similarly useful.

EXAMPLE: RANDOM TELEGRAPH PROCESS

In the random telegraph process from section 1.2.1.2, we can consider the mean

$$m(t) = \langle X(t)\rangle = p_1(t) \tag{3.3}$$

and its stationarity

$$\begin{aligned} 0 &= \lim_{t\to\infty}\frac{\mathrm{d}}{\mathrm{d}t}m(t) = \lim_{t\to\infty}\frac{\mathrm{d}}{\mathrm{d}t}p_1(t) = \lim_{t\to\infty}\frac{\mathrm{d}}{\mathrm{d}t}(-\beta p_1(t) + \alpha(1-p_1(t)) \\ &= -\beta\pi_1 + \alpha(1-\pi_1) = \alpha - (\alpha+\beta)m\,. \end{aligned} \tag{3.4}$$

3.1.1.3 Markov processes with continuous configurations

For the purpose of deriving the self-consistent equations for Markov processes with continuous configurations, it is most natural to consider the Langevin formulation (in the Itô convention)

$$\frac{\mathrm{d}}{\mathrm{d}t}X_i(t) = f_i(X(t),t;\Theta) + \sum_j \sigma_{ij}(X(t),t;\Theta)\xi_j(t)\,. \tag{3.5}$$

Again, the observable means are independent of time in the steady state. For example, considering that the mean $m_i(t) := \langle X_i(t)\rangle$ should become stationary, we may write

$$0 = \lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}t} m_i(t) = \lim_{t\to\infty} \left\langle \frac{\mathrm{d}}{\mathrm{d}t} X_i(t) \right\rangle$$
$$= \langle f_i(X;\Theta)\rangle_\pi =: g(\Theta) \,. \tag{3.6}$$

More generally, we can do the same for any observable $O(X(t))$, by considering $\lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}t}\langle O(X(t))\rangle = 0$ and applying Itô's lemma (1.35) to compute $\frac{\mathrm{d}}{\mathrm{d}t} O(X(t))$.

EXAMPLE: DIFFUSION IN A NON-CONSERVATIVE FORCE-FIELD I

For illustration, we consider a particle diffusing with unit diffusivity in a non-conservative force-field. It is described by the simple two-dimensional system of Langevin equations

$$\frac{\mathrm{d}}{\mathrm{d}t} X_1(t) = -\left(X_1(t) - e^{-J[X_2(t)]^2}\right) + \xi_1(t) =: f_1(X_1(t), X_2(t)) + \xi_1(t) \tag{3.7}$$
$$\frac{\mathrm{d}}{\mathrm{d}t} X_2(t) = -\left(X_2(t) - e^{-J[X_1(t)]^2}\right) + \xi_2(t) =: f_2(X_1(t), X_2(t)) + \xi_2(t) \,, \tag{3.8}$$

with $J \geq 0$ and independent white noise random forces $\xi_1(t)$ and $\xi_2(t)$.

We are considering a non-equilibrium system, since the detailed balance condition (1.73) reduces to

$$\frac{\partial f_1}{\partial x_2} \stackrel{?}{=} \frac{\partial f_2}{\partial x_1} \,, \tag{3.9}$$

which is obviously not fulfilled for $J > 0$. Furthermore, we cannot determine the steady state distribution analytically for $J > 0$. As described above, for finding self-consistent relationships, we consider that the means become stationary

$$0 = \lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}t} m_1(t) = \lim_{t\to\infty} \left\langle \frac{\mathrm{d}}{\mathrm{d}t} X_1(t) \right\rangle = \langle X_1\rangle_\pi - \langle e^{-J(X_2)^2}\rangle_\pi \tag{3.10}$$
$$0 = \lim_{t\to\infty} \frac{\mathrm{d}}{\mathrm{d}t} m_2(t) = \lim_{t\to\infty} \left\langle \frac{\mathrm{d}}{\mathrm{d}t} X_2(t) \right\rangle = \langle X_2\rangle_\pi - \langle e^{-J(X_1)^2}\rangle_\pi \,. \tag{3.11}$$

Hence, we have found the self-consistent characterisations of the means

$$\langle X_1\rangle_\pi = \langle e^{-J(X_2)^2}\rangle_\pi =: g_1(J) \tag{3.12}$$
$$\langle X_2\rangle_\pi = \langle e^{-J(X_1)^2}\rangle_\pi =: g_2(J) \tag{3.13}$$

■

3.1.2 Exact inference based on direct sample averages of the self-consistent equations

At first, these self-consistent equations do not appear to be of much use, since we do not know the steady state distribution $\pi(x;\Theta)$ required to compute the averages. However, in the context of stochastic inference, we have M independent samples $D = \{x^\mu\}_{\mu=1}^M$ drawn from the steady state and may replace the average over the steady state distribution with the average over the samples[1]. In the limit $M \to \infty$ this replacement becomes exact. Our problem now reduces to finding a sufficient number of these self-consistent equations relating the sample statistics to the model parameters. Then we can numerically compute the parameter-dependent functions $g_k(\Theta)$ $(k = 1,\ldots,K)$ by averaging over the samples and fit the functions $g_k(\Theta)$ to the directly observable means they characterise (it is clear that for a well-posed problem we should have at least as many self-consistent equations as there are parameters to infer).

EXAMPLE: DIFFUSION IN A NON-CONSERVATIVE FORCE-FIELD I: EXACT INFERENCE

The two-dimensional diffusion example discussed above is particularly simple, since it is described by the single parameter J. We can determine J from numerically solving the equation for $g_1(J)$

$$\begin{aligned} \langle X_1 \rangle_D &= \langle e^{-JX_2^2} \rangle_D \\ \Leftrightarrow \frac{1}{M}\sum_{\mu=1}^M x_1^\mu &= \frac{1}{M}\sum_{\mu=1}^M e^{-J(x_2^\mu)^2} \,. \end{aligned} \tag{3.14}$$

If we have sufficiently many samples, the empirical mean of X_1 should be close to the empirical mean of $e^{-JX_2^2}$, hence we can assume that the left-hand side of (3.14) lies in $(0,1)$. Since the right-hand side of (3.14) is monotonic in J and approaches zero for $J \to \infty$, while approaching one for $J \to 0$, there exists a unique value of J such that (3.14) holds. ■

For more complicated models, we will have many different self-consistent equations that depend on several model parameters. The self-consistent equations will in general be non-linear in the parameters. Thus, we have to solve a

[1] Using the sample average instead of the steady state average is reminiscent of the pseudolikelihood method for equilibrium inference. However, we do not know the pseudo-likelihood function to maximise and hence it is unclear which observables to match. Instead, we have to make an ad hoc choice of the observables we want to calibrate.

non-linear optimisation problem. Many efficient optimisation algorithms exist, however, the problem becomes harder when the number of parameters increases and the functions must be evaluated many times before convergence is achieved. The sample averaging approach is associated with a considerable computational cost, since the evaluation of the right-hand side functions $g(\Theta)$ requires $\mathcal{O}(M)$ computational steps. While this is still much better than the exact computation of the free energy in equilibrium problems, which scales exponentially in the dimension of the configuration space, we might want to reduce the computational effort in evaluating the functions $g(\Theta)$. For this we turn to non-equilibrium mean field theory.

3.1.3 Expanding the self-consistent equations with non-equilibrium mean field theory

We consider the case of random variables X with several components X_i ($i = 1,\ldots,N$), which are correlated. These correlations generally prevent analytical computations of observable expectation values like the means $m_i(\Theta) = \langle X_i \rangle_{\pi(x;\Theta)}$ or correlations $C_{ij}(\Theta) = \langle (X_i - m_i)(X_j - m_j) \rangle_{\pi(x;\Theta)}$. Hence, we need to find approximations for these expectations in order to obtain a closed form for the self-consistent functions $g_k(\Theta)$ described above. For this purpose, we compute the expectations in a series expansion around a factorising steady state distribution, i.e. the case where the components X_i are statistically independent. This approach is the non-equilibrium extension of Plefka's expansion (Plefka, 1982) of the equilibrium free energy. This extension was proposed for the asymmetric Ising model by Kappen and Spanjers (2000). Inspired by their approach, we will describe a straightforward generalisation of their non-equilibrium mean field theory to other Markov processes. We begin by characterising the family of factorising distributions used to approximate the steady state, before invoking the principle of minimising relative entropy in order to pin down the optimal factorising distribution, which will be used as the starting point of the expansion. Finally, we show how to expand the self-consistent expressions $g_k(\Theta)$ around this optimal factorising distribution.

3.1.3.1 Approximating the steady state with a factorising distribution

We assume that the parameters can be divided into two sets $\Theta = (\Theta_h, \Theta_J) \in \Lambda$, where the parameters Θ_J model the interactions between the different components X_i, i.e. setting $\Theta_J = 0$ makes the different components of the random vector X statistically independent. Further, we assume that the ergodicity of the Markov process is preserved when setting $\Theta_J = 0$. As factorising approximations $q(x)$ to the steady state $\pi(x)$, we consider the family Ω of steady states

produced by processes without interactions

$$\mathcal{Q} = \{q(x;\Theta_h) = \pi(x;\Theta_h,\Theta_J = 0) \mid (\Theta_h,0) \in \Lambda\} \ . \tag{3.15}$$

In general, the family $\mathcal{Q}$ consists of many factorising distributions, characterised by the parameters Θ_h, and we need to ask ourselves which distribution out of this family should be used to approximate the steady state distribution $\pi(x;\Theta_h,\Theta_J)$. As proposed by Kappen and Spanjers (2000), we take an information theoretic approach and seek the distribution $q^* \in \mathcal{Q}$ that minimises the relative entropy between the factorising distribution and the steady state

$$q^* := \underset{q\in\mathcal{Q}}{\operatorname{argmin}}\, D_{\mathrm{KL}}(\pi(x;\Theta_h,\Theta_J)||q(x)) = q(\Theta_h^q[\Theta_h,\Theta_J]) \ . \tag{3.16}$$

The minimising condition usually translates into specific observables that should be reproduced exactly with the factorising distribution, thus fixing the parameter values Θ_h^q, characterising q^*, as functions of the original parameters Θ_h and Θ_J.

EXAMPLE: DIFFUSION IN A NON-CONSERVATIVE FORCE-FIELD II: THE OPTIMAL FACTORISING DISTRIBUTION

We modify our previous example by including two additional parameters $h_1 > 0, h_2 > 0$ and write the Langevin equations as

$$\frac{\mathrm{d}}{\mathrm{d}t}X_1(t) = -\frac{1}{h_1+J}\left(X_1(t) - e^{-J[X_2(t)]^2}\right) + \xi_1(t) \tag{3.17}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}X_2(t) = -\frac{1}{h_2+J}\left(X_2(t) - e^{-J[X_1(t)]^2}\right) + \xi_2(t) \ . \tag{3.18}$$

Again, the choice $J = 0$ results in two independent Ornstein-Uhlenbeck processes. Hence we have $\Theta_h = (h_1,h_2)$ and $\Theta_J = J$ and the family $\mathcal{Q}$ of factorising distributions is characterised by the parameters $\Theta_h^q = (h_1^q,h_2^q)$

$$\begin{aligned} q(x_1,x_2;h_1^q,h_2^q) &:= \pi(x_1,x_2;h_1 = h_1^q,h_2 = h_2^q,J = 0) \\ &= \left(\frac{\exp\left\{-(x_1-1)^2/h_1^q\right\}}{\sqrt{\pi h_1^q}}\right)\left(\frac{\exp\left\{-(x_2-1)^2/h_2^q\right\}}{\sqrt{\pi h_2^q}}\right) \ . \end{aligned} \tag{3.19}$$

Next, we seek the optimal distribution q^* and determine h_1^q and h_2^q by minimising the relative entropy between the factorising distribution $q(x_1,x_2;h_1^q,h_2^q)$ and the

steady state $\pi(x_1,x_2;h_1,h_2,J)$. For the relative entropy we find

$$\begin{aligned} D_{\mathrm{KL}}(\pi(x)||q(x))[h_1^q,h_2^q]) &= \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \pi(x_1,x_2;h_1,h_2,J) \ln\left(\frac{\pi(x_1,x_2;h_1,h_2,J)}{q(x_1,x_2;h_1^q,h_2^q)}\right) \\ &= \langle \ln \pi(x_1,x_2;h_1,h_2,J)\rangle_\pi - \left\langle \ln q(x_1,x_2;h_1^q,h_2^q)\right\rangle_\pi \\ &= -S[\pi(x)] - \left\langle \frac{1}{2}\ln(\pi h_1^q) + (X_1-1)^2/h_1^q \right\rangle_\pi \\ &\quad + \left\langle \frac{1}{2}\ln(\pi h_2^q) + (X_2-1)^2/h_2^q \right\rangle_\pi , \end{aligned} \tag{3.20}$$

with the Shannon entropy of the steady state $S[\pi(x)] = -\langle \ln \pi(x_1,x_2)\rangle_\pi$.

The parameter values h_1^q, h_2^q are then fixed by demanding that the gradient of the relative entropy vanishes

$$0 = \frac{\partial D_{\mathrm{KL}}}{\partial h_1^q} = \frac{1}{2h_1^q} - \frac{1}{(h_1^q)^2}\langle (X_1-1)^2\rangle_\pi \tag{3.21}$$

$$0 = \frac{\partial D_{\mathrm{KL}}}{\partial h_2^q} = \frac{1}{2h_2^q} - \frac{1}{(h_2^q)^2}\langle (X_2-1)^2\rangle_\pi , \tag{3.22}$$

which fixes h_1^q and h_2^q in terms of steady state expectation values

$$h_1^q = 2\langle (X_1-1)^2\rangle_\pi \tag{3.23}$$

$$h_2^q = 2\langle (X_2-1)^2\rangle_\pi . \tag{3.24}$$

Since we do not know the steady state distribution $\pi(x_1,x_2;h_1,h_2,J)$, we cannot directly determine the right-hand sides of (3.23) and (3.24) as functions of h_1, h_2, and J. Instead, we compute them in a series expansion around the factorising distribution q^*, which leads to self-consistent equations that fix the parameters $h_1^q = h_1^q(h_1,h_2,J)$ and $h_2^q = h_2^q(h_1,h_2,J)$. ■

3.1.3.2 Expanding around the factorising distribution

As seen in the example above, after having identified the observables that should be matched exactly, we are still at a loss how to compute the parameters Θ_h^q as functions of the actual parameters Θ_h, Θ_J, since we cannot evaluate the observable means in the actual steady state distribution $\pi(x;\Theta_h,\Theta_J)$. The solution is to compute these means self-consistently in an expansion around the optimal factorising distribution q^*. Formally, we use self-consistent relations as derived in section 3.1.1 to express the expectation value of an observable $O(X)$ in terms of the expectation value of a function $G(X,\Theta_h,\Theta_J)$ involving the parameters and then expand its expectation value $g(\Theta_h,\Theta_J) = \langle O(X)\rangle_{\pi(x;\Theta_h,\Theta_J)} =$

$\langle G(X,\Theta)\rangle_{\pi(x;\Theta_h,\Theta_J)}$ around the parameters $(\Theta_h = \Theta_h^q, \Theta_J = 0)$. To this end, we introduce an expansion parameter λ and consider the model parameters $\Theta_h(\lambda) = \Theta_h^q + \lambda\,\delta\Theta_h, \Theta_J(\lambda) = \lambda\Theta_J$ with $\delta\Theta_h = \Theta_h - \Theta_h^q(\Theta_h,\Theta_J)$. Hence, we can smoothly interpolate from the factorising distribution q^*, characterised by $\lambda = 0$, to the steady state distribution π, characterised by $\lambda = 1$, and write the function $g(\Theta_h,\Theta_J)$ as a Taylor series in λ

$$g(\Theta_h,\Theta_J) \stackrel{!}{=} \sum_{k=0}^{\infty} \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}\lambda^k} g(\Theta_h^q + \lambda\,\delta\Theta_h, \lambda\Theta_J)\Big|_{\lambda=0} \tag{3.25}$$

$$= g(\Theta_h^q,0) + \frac{\partial g}{\partial \Theta_h}(\Theta_h^q,0)\delta\Theta_h + \frac{\partial g}{\partial \Theta_J}(\Theta_h^q,0)\Theta_J + \dots$$

$$=: g\Big|_{q^*} + \frac{\partial g}{\partial \Theta_h}\Big|_{q^*}\delta\Theta_h + \frac{\partial g}{\partial \Theta_J}\Big|_{q^*}\Theta_J + \dots\ . \tag{3.26}$$

For the means that should be matched exactly, the resulting equations can be solved for $\Theta_h^q = \Theta_h^q(\Theta_h,\Theta_J)$. This is done successively for each expansion order, yielding closed expressions for the observable means as functions of the original parameters $\Theta = (\Theta_h,\Theta_J)$.

EXAMPLE: DIFFUSION IN A NON-CONSERVATIVE FORCE-FIELD II: MEAN-FIELD EXPANSION AND PARAMETER INFERENCE

Recalling our last example, we have the family of factorising distributions described by the probability density (3.19) and the self-consistent expressions for the first moments

$$m_1(h_1,h_2,J) = \langle e^{-J(X_2)^2}\rangle_{\pi(h_1,h_2,J)} \tag{3.27}$$

$$m_2(h_1,h_2,J) = \langle e^{-J(X_1)^2}\rangle_{\pi(h_1,h_2,J)}\ . \tag{3.28}$$

In addition, we need self-consistent expressions for the second moments $\alpha_1(h_1,h_2,J) = \langle (X_1)^2\rangle_\pi$ and $\alpha_2(h_1,h_2,J) = \langle (X_2)^2\rangle_\pi$, in order to determine the free parameters $h_1^q = h_1^q(h_1,h_2,J), h_2^q = h_2^q(h_1,h_2,J)$ of the factorising distribution from conditions (3.23) and (3.24). By considering $\lim_{t\to\infty}\frac{\mathrm{d}}{\mathrm{d}t}\langle X_i^2(t)\rangle = 0$ we find the self-consistent expressions

$$\alpha_1(h_1,h_2,J) = \frac{h_1+J}{2} + \left\langle X_1 e^{-J(X_2)^2}\right\rangle_{\pi(h_1,h_2,J)} \tag{3.29}$$

$$\alpha_2(h_1,h_2,J) = \frac{h_2+J}{2} + \left\langle X_2 e^{-J(X_1)^2}\right\rangle_{\pi(h_1,h_2,J)}\ . \tag{3.30}$$

We perform the expansion to first order in λ and begin by considering the first moment

$$m_1(h_1,h_2,J) \approx m_1(h_1^q,h_2^q,0) + \frac{\partial m_1}{\partial h_1}\Big|_{q^*}\delta h_1 + \frac{\partial m_1}{\partial h_2}\Big|_{q^*}\delta h_2 + \frac{\partial m_1}{\partial J}\Big|_{q^*} J\ . \quad (3.31)$$

The first term is simply the mean in the factorising approximation

$$m(h_1^q,h_2^q,0) = \int \mathrm{d}x_1 \int \mathrm{d}x_2\ q(x_1,x_2;h_1,h_2)x_1 = 1\ , \quad (3.32)$$

where we used expression (3.19) for the factorising distribution. The derivatives can be computed from the self-consistent expressions. For m_1 (3.27) we find

$$\begin{aligned}
\frac{\partial m_1}{\partial J}\Big|_{q^*} &= \frac{\partial}{\partial J}\Big|_{q^*} \int \mathrm{d}x_1 \int \mathrm{d}x_2\ \pi(x_1,x_2;h_1,h_2;J) e^{-J(x_2)^2} \\
&= \int \mathrm{d}x_1 \int \mathrm{d}x_2\ \left\{ \frac{\partial \pi}{\partial J}\Big|_{q^*} e^{-0(x_2)^2} + q(x_1,x_2;h_1^q,h_2^q)\frac{\partial}{\partial J}\Big|_{q^*} e^{-J(x_2)^2} \right\} \\
&= \int \mathrm{d}x_1 \int \mathrm{d}x_2\ \left\{ \frac{\partial \pi}{\partial J}\Big|_{q^*} + q(x_1,x_2;h_1^q,h_2^q)(-(x_2)^2 e^{0(x_2)^2}) \right\} \\
&= \frac{\partial}{\partial J}\Big|_{q^*} \Big\{ \underbrace{\int \mathrm{d}x_1 \int \mathrm{d}x_2\ \pi(x_1,x_2;h_1,h_2,J)}_{=1} \Big\} - \langle (X_2)^2 \rangle_{q^*} \\
&= -\langle (X_2)^2 \rangle_{q^*} = -\langle (X_2 - 1 + 1)^2 \rangle_q^* = -(1 + h_2^q/2)\ . \quad (3.33)
\end{aligned}$$

With the same procedure we find $\frac{\partial m_1}{\partial h_1}\Big|_{q^*} = \frac{\partial m_1}{\partial h_2}\Big|_{q^*} = 0$. Hence, in total we arrive at

$$m_1(h_1,h_2,J) \approx 1 - \left(1 + \frac{h_2^q}{2}\right) J. \quad (3.34)$$

By symmetry, we also find

$$m_2(h_1,h_2,J) \approx 1 - \left(1 + \frac{h_1^q}{2}\right) J. \quad (3.35)$$

Next, we expand $\alpha_1(h_1,h_2,J)$

$$\alpha_1(h_1,h_2,J) \approx \alpha_1(h_1^q,h_2^q,0) + \frac{\partial \alpha_1}{\partial h_1}\Big|_{q^*}\delta h_1 + \frac{\partial \alpha_1}{\partial h_2}\Big|_{q^*}\delta h_2 + \frac{\partial \alpha_1}{\partial J}\Big|_{q^*} J\ . \quad (3.36)$$

The first term is again easily computed in the factorising distribution q^*

$$\begin{aligned}
\alpha_1(h_1^q,h_2^q,0) &= \langle (X_1)^2 \rangle_q^* = \langle (X_1 - 1 + 1)^2 \rangle_q^* \\
&= 1 + h_1^q/2 \quad (3.37)
\end{aligned}$$

For the derivative with respect to J, we need the derivative of m_1 with respect to J, which we have already computed

$$\begin{aligned}
\frac{\partial \alpha_1}{\partial J}\Big|_{q^*} &= \frac{1}{2} + \frac{\partial}{\partial J}\Big|_{q^*} \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \pi(x_1,x_2;h_1,h_2;J) x_1 e^{-J(x_2)^2} \\
&= \frac{1}{2} + \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \frac{\partial \pi}{\partial J}\Big|_{q^*} x_1 e^{-0(x_2)^2} + \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; q(x_1,x_2;h_1^q,h_2^q) x_1 \frac{\partial}{\partial J}\Big|_{q^*} e^{-J(x_2)^2} \\
&= \frac{1}{2} + \frac{\partial}{\partial J}\Big|_{q^*} \Big\{ \underbrace{\int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \pi(x_1,x_2) x_1}_{=m_1(h_1,h_2,J)} \Big\} - \underbrace{\langle X_1 (X_2)^2 \rangle_{q^*}}_{=1+h_2^q/2} \\
&= \frac{1}{2} + \frac{\partial m_1}{\partial J}\Big|_{q^*} - (1+h_2^q/2) = \frac{1}{2} - (1+h_2^q/2) - (1+h_2^q/2) = -\frac{3}{2} - h_2^q \; .
\end{aligned} \tag{3.38}$$

For the other derivatives we proceed similarly and find in total

$$\alpha_1(h_1,h_2,J) \approx 1 + \frac{h_1}{2} - \left(\frac{3}{2} + h_2^q\right) J \; . \tag{3.39}$$

By symmetry we also find

$$\alpha_2(h_1,h_2,J) \approx 1 + \frac{h_2}{2} - \left(\frac{3}{2} + h_1^q\right) J \; . \tag{3.40}$$

Now we can determine h_1^q and h_2^q (to first order in λ) from inserting the expansion into (3.23) and (3.24)

$$\begin{aligned}
h_1^q &= 2\langle (X_1 - 1)^2 \rangle_\pi = 2\alpha_1 - 4m_1 + 2 \approx h_1 + J \\
h_2^q &= 2\langle (X_2 - 1)^2 \rangle_\pi = 2\alpha_2 - 4m_2 + 2 \approx h_2 + J \; .
\end{aligned} \tag{3.41}$$

Inserting $h_1^q(h_1,h_2,J)$ and $h_2^q(h_1,h_2,J)$ back into (3.34) and (3.35), we obtain the closed form expressions (to first order in λ) for the first moments as functions of the original model parameters

$$m_1(h_1,h_2,J) \approx 1 - \left(1 + \frac{h_2^q}{2}\right) J = 1 - \left(1 + \frac{h_2 + J}{2}\right) J \tag{3.42}$$

$$m_2(h_1,h_2,J) \approx 1 - \left(1 + \frac{h_1^q}{2}\right) J = 1 - \left(1 + \frac{h_1 + J}{2}\right) J \; . \tag{3.43}$$

Similarly, by inserting h_1^q and h_2^q into (3.39) and (3.39), we obtain the second moments

$$\alpha_1(h_1,h_2,J) \approx 1 + \frac{h_1 - 3J}{2} - h_2^q J = 1 + \frac{h_1 - 3J}{2} - (h_2 + J) J \tag{3.44}$$

$$\alpha_2(h_1,h_2,J) \approx 1 + \frac{h_2 - 3J}{2} - h_1^q J = 1 + \frac{h_2 - 3J}{2} - (h_1 + J) J \; . \tag{3.45}$$

Now we can return to the inference problem: given sample estimates $\hat{\alpha}_1 = \frac{1}{M}\sum_{\mu=1}^{M}(x_1^\mu)^2$, $\hat{\alpha}_2 = \frac{1}{M}\sum_{\mu=1}^{M}(x_2^\mu)^2, \hat{m}_1 = \frac{1}{M}\sum_{\mu=1}^{M}x_1^\mu, \hat{m}_2 = \frac{1}{M}\sum_{\mu=1}^{M}x_2^\mu$ we can invert the closed expressions (3.42), (3.43), (3.44), and (3.45) to solve for the parameters h_1, h_2, J (for this particular system the equations are so simple that we can invert analytically), giving the (first order) estimates

$$J^{\text{inf}} = \left(\frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\hat{m}_1 - \hat{m}_2} - 2\right)^{-1} \tag{3.46}$$

$$h_1^{\text{inf}} = 2(1-\hat{m}_2)/J^{\text{inf}} - J^{\text{inf}} - 2 \tag{3.47}$$

$$h_2^{\text{inf}} = 2(1-\hat{m}_1)/J^{\text{inf}} - J^{\text{inf}} - 2 \tag{3.48}$$

In principle, this expansion can be continued to higher orders, which would result in an increased accuracy in the reconstruction. ■

Having illustrated the general theory for toy examples, we now apply the same methods to the more challenging problem of inferring the parameters of our main paradigm - the asymmetric Ising model.

3.2 Inference in the asymmetric Ising model

We consider the asymmetric Ising model consisting of N binary spins $\boldsymbol{s} = (s_1, \ldots, s_N)$ introduced in chapter 2. Already elementary arguments show that, unlike in the equilibrium inverse Ising problem, pairwise spin-correlations are insufficient to infer the model parameters: the matrix of pairwise correlations $\langle s_i s_j \rangle$ is symmetric and has only $N(N-1)/2$ independent entries, whereas there are $N(N-1)$ entries of the asymmetric coupling matrix J_{ij} to be determined (self-interactions $J_{ii} \neq 0$ are excluded). Thus we expect that at least three-point correlations $\langle s_i s_j s_k \rangle$ are required. On the other hand, the information one can extract from single-time measurements in the non-equilibrium steady state $\pi(\boldsymbol{s}; \mathbf{h}, J)$ is limited to the frequencies of the 2^N different spin configurations. Taking into account the normalisation constraint, there are thus at most $2^N - 1$ independent observables available to determine the $N(N-1)+N$ parameters of couplings and external fields. This implies that the parameters can only be inferred for $N \geq 5$.

3.2.1 Callen's identities and their mean field expansion

In the following reconstruction, we will use Callen's identities (cp. section 2.2.2) for the magnetisations $m_i = \langle s_i \rangle$, two-point connected spin-correlations $C_{ij} = \langle \delta s_i \delta s_j \rangle$ $(i < j)$, and three-point connected spin-correlations $C_{ijk} = \langle \delta s_i \delta s_j \delta s_k \rangle$

$(i < j < k)$ in sequential Glauber dynamics

$$m_i = \langle \tanh(\psi_i) \rangle \equiv \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h}, J) \tanh\left(h_i + \sum_j J_{ij} s_j \right) , \tag{3.49}$$

$$C_{ij} = \frac{1}{2} \langle \delta S_i [\tanh(\psi_j) - m_j] \rangle + \frac{1}{2} \langle \delta S_j [\tanh(\psi_i) - m_i] \rangle , \tag{3.50}$$

$$\begin{aligned} C_{ijk} =& \frac{1}{3} \langle \delta S_i \delta S_j [\tanh(\psi_k) - m_k] \rangle + \frac{1}{3} \langle \delta S_i \delta S_k [\tanh(\psi_j) - m_j] \rangle \\ &+ \frac{1}{3} \langle \delta S_j \delta S_k [\tanh(\psi_i) - m_i] \rangle , \end{aligned} \tag{3.51}$$

and in parallel Glauber dynamics

$$C_{ij}^{\text{par}} = \langle [\tanh(\psi_i) - m_i][\tanh(\psi_j) - m_j] \rangle , \tag{3.52}$$

$$C_{ijk}^{\text{par}} = \langle [\tanh(\psi_i) - m_i][\tanh(\psi_j) - m_j][\tanh(\psi_k) - m_k] \rangle . \tag{3.53}$$

Next, we consider the mean field expansion of Callen's identities for sequential Glauber dynamics, which was originally developed by Kappen and Spanjers (2000), who performed the expansion up to second order in the couplings for the magnetisations and two-point spin-correlations. In the following, we will retrace the steps of Kappen and Spanjers (2000), going a little more into detail, for illustrating how the calculations work, before adding the expansion of the three-point correlations and third order corrections, which are required for inference. We also derived the corresponding expansions for parallel Glauber dynamics, which are given in appendix A.

3.2.1.1 The optimal factorising distribution

The distribution factorises in the different spin variables if we set $J = 0$. Since the spins are binary variables, the factorising distribution q^* can be characterised in terms of the means $m_i^q = \langle s_i \rangle_{q^*}$ as

$$q^*(\boldsymbol{s}) = q(\boldsymbol{s}; \mathbf{h}^q) = \prod_{i=1}^{N} \left(\frac{1 + m_i^q(h_i^q) s_i}{2} \right) . \tag{3.54}$$

The choice $\mathbf{h} = \mathbf{h}^q, J = 0$ makes the effective local fields $\psi_i = h_i^q$ deterministic, hence the magnetisations under q^* are connected to the external fields h_i^q via the relation

$$m_i^q = \tanh(h_i^q) . \tag{3.55}$$

In equilibrium statistical physics, this distribution is the well-known mean field ansatz. As described above in section 3.1.3.1, the mean field $\mathbf{h}^q$ is fixed by demanding that it minimises the relative entropy between the factorising distribution $q^*(\boldsymbol{s}) = q(\boldsymbol{s};\mathbf{h}^q)$ and the actual steady state $\pi(\boldsymbol{s}) = \pi(\boldsymbol{s};\mathbf{h},J)$

$$\begin{aligned}\mathbf{h}_q &= \underset{\tilde{\mathbf{h}}_q}{\operatorname{argmin}}\, D_{\mathrm{KL}}(\pi(\boldsymbol{s})||q(\boldsymbol{s};\tilde{\mathbf{h}}_q)) \\ &= \underset{\tilde{\mathbf{h}}_q}{\operatorname{argmin}} \left\{ -S[\pi(\boldsymbol{s})] - \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}) \sum_{i=1}^{N} \ln\left(\frac{1+m_i^q(\tilde{h}_q^i)s_i}{2}\right)\right\} . \end{aligned} \tag{3.56}$$

By differentiating with respect to $\tilde{h}_i^q$ we find

$$\begin{aligned} 0 &= \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}) \left(\frac{s_i}{1+m_i^q(h_i^q)s_i}\right) \frac{\mathrm{d}m_i^q}{\mathrm{d}\tilde{h}_i^q}(h_i^q) \\ &= \left\{ \pi(s_i=+1)\frac{1}{1+m_i^q} - (1-\pi(s_i=+1))\frac{1}{1-m_i^q}\right\}(1-(m_i^q)^2) \\ &= m_i^q(h_i^q) - \langle s_i\rangle_\pi = m_i^q(h_i^q) - m_i(\mathbf{h},J) \; . \end{aligned} \tag{3.57}$$

Hence, as in equilibrium mean field theory, the external fields $\mathbf{h}^q$ characterising the mean-field distribution are chosen such that the mean-field distribution yields the same magnetisations as the original model with couplings J and fields $\mathbf{h}$.

3.2.1.2 Expanding Callen's identities around the optimal factorising distribution

As described above in section 3.1.3.2, we introduce the expansion parameter λ and consider external fields $\mathbf{h}^q + \lambda\,\delta\mathbf{h}$ and couplings λJ with $\delta\mathbf{h} = \mathbf{h} - \mathbf{h}^q$, such that the choice $\lambda = 0$ corresponds to the factorising distribution and $\lambda = 1$ to the steady state, which we want to approximate. Next, we expand Callen's identities for the magnetisations and correlations in powers of λ

$$\mathbf{m}(\mathbf{h},J) = \sum_{k=0}^{\infty} \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}\lambda^k}\mathbf{m}(\mathbf{h}_q + \lambda\,\delta\mathbf{h}, \lambda J)\Big|_{\lambda=0} \tag{3.58}$$

$$C_{ij}(\mathbf{h},J) = \sum_{k=0}^{\infty} \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}\lambda^k}C_{ij}(\mathbf{h}_q + \lambda\,\delta\mathbf{h}, \lambda J)\Big|_{\lambda=0} \tag{3.59}$$

$$C_{ijk}(\mathbf{h},J) = \sum_{k=0}^{\infty} \frac{1}{k!}\frac{\mathrm{d}^k}{\mathrm{d}\lambda^k}C_{ijk}(\mathbf{h}_q + \lambda\,\delta\mathbf{h}, \lambda J)\Big|_{\lambda=0} . \tag{3.60}$$

The external fields $\mathbf{h}_q = \mathbf{h}_q(\mathbf{h},J)$ characterising the factorising distribution q^* are fixed by setting $m_i(\mathbf{h},J) = m_i^q(\mathbf{h}_q) = \tanh(h_i^q)$, i.e. $\mathbf{h}_q$ is determined from

the expansion of $\mathbf{m}(\mathbf{h},J)$. Inserting $\mathbf{h}_q(\mathbf{h},J)$ into the expansions (3.58)-(3.60), the expansion in λ in fact becomes an expansion in the couplings J. In the following, the diagonal couplings are always understood to be zero, $J_{ii}=0$.

EXPANSION OF THE MAGNETISATIONS TO SECOND ORDER IN λ

For keeping track of the different terms in the expansion (3.58) of the magnetisations, we introduce some notation and write the second order expansion of component m_i as

$$\begin{aligned} m_i(\mathbf{h},J) =& m_i^q + \sum_{j=1}^N \frac{\partial m_i}{\partial h_j}\Big|_{q^*} \delta h_j + \sum_{j,k=1}^N \frac{\partial m_i}{\partial J_{jk}}\Big|_{q^*} J_{jk} + \frac{1}{2}\sum_{j,k=1}^N \frac{\partial^2 m_i}{\partial h_j \partial h_k}\delta h_j \delta h_k + \\ &+ \sum_{j,k,l=1}^N \frac{\partial^2 m_i}{\partial h_j \partial J_{kl}}\delta h_j J_{kl} + \frac{1}{2}\sum_{j,k,l,n=1}^N \frac{\partial^2 m_i}{\partial J_{jk}\partial J_{ln}} J_{jk} J_{ln} + \dots \\ =:& m_i^q + m_i^{\mathbf{h}} + m_i^J + \frac{1}{2} m_i^{\mathbf{hh}} + m_i^{\mathbf{h}J} + \frac{1}{2} m_i^{JJ} + \dots \end{aligned} \tag{3.61}$$

For actually computing the derivatives, we consider the self-consistent equation (3.49) and start by considering the derivative with respect to the external fields

$$\begin{aligned} \frac{\partial m_i}{\partial h_k}\Big|_{q^*} =& \frac{\partial}{\partial h_k}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J) \tanh\left(h_i + \sum_j J_{ij}s_j\right) \\ =& \sum_{\boldsymbol{s}} \frac{\partial \pi(\boldsymbol{s}|\mathbf{h},J)}{\partial h_k}\Big|_{q^*} \tanh\left(h_i + \sum_j J_{ij}s_j\right)\Big|_{q^*} \\ &+ \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J)\Big|_{q^*} \frac{\partial \tanh\left(h_i + \sum_j J_{ij}s_j\right)}{\partial h_k}\Big|_{q^*} \\ =& \frac{\partial}{\partial h_k}\Big|_{q^*} \underbrace{\sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J)}_{=1} \tanh\left(h_i^q\right) + \underbrace{\sum_{\boldsymbol{s}} q^*(\boldsymbol{s})}_{=1} \underbrace{\left(1-\tanh^2(h_i^q)\right)}_{=(1-m_i^2)} \delta_{i,k} \\ =& (1-m_i^2)\delta_{i,k} \,. \end{aligned} \tag{3.62}$$

Hence, we find

$$m_i^{\mathbf{h}} = \sum_{k=1}^N \frac{\partial m_i}{\partial h_k}\Big|_{q^*} \delta h_k = (1-m_i)^2 \delta h_i \,. \tag{3.63}$$

Next, we consider the derivative with respect to the couplings

$$\begin{aligned}
\frac{\partial m_i}{\partial J_{kl}}\Big|_{q^*} =& \frac{\partial}{\partial J_{kl}}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J) \tanh\left(h_i + \sum_j J_{ij}s_j\right) \\
=& \sum_{\boldsymbol{s}} \frac{\partial \pi(\boldsymbol{s}|\mathbf{h},J)}{\partial J_{kl}}\Big|_{q^*} \tanh\left(h_i + \sum_j J_{ij}s_j\right)\Big|_{q^*} \\
&+ \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J)\Big|_{q^*} \frac{\partial \tanh\left(h_i + \sum_j J_{ij}s_j\right)}{\partial J_{kl}}\Big|_{q^*} \\
=& \frac{\partial}{\partial J_{kl}}\Big|_{q^*} \underbrace{\sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J)}_{=1} \tanh\left(h_i^q\right) + \sum_{\boldsymbol{s}} q^*(\boldsymbol{s})(1-\tanh^2(h_i^q))\delta_{i,k}s_l \\
=& (1-m_i^2)\delta_{i,k}m_l \ , \qquad (3.64)
\end{aligned}$$

which gives

$$m_i^J = \sum_{k,l=1}^{N} \frac{\partial m_i}{\partial J_{kl}}\Big|_{q^*} J_{kl} = (1-m_i)^2 \sum_{j=1}^{N} J_{ij}m_j \ . \qquad (3.65)$$

If we truncated the expansion at the first order in λ, we would have $m_i = m_i^q + m_i^{\mathbf{h}} + m_i^J$ and since the optimal factorising distribution reproduces the magnetisations exactly, $m_i^q = m_i$, we can find h_q (to first expansion order) by solving

$$0 = m_i - m_i^q \approx m_i^{\mathbf{h}} + m_i^J = (1-m_i^2)\left\{\delta h_i + \sum_{j=1}^{N} J_{ij}m_j\right\} \qquad (3.66)$$

for $h_i^q = h_i - \delta h_i$, which yields

$$h_i^q \approx h_i - \delta h_i^{\mathrm{nMF}} = h_i + \sum_{j=1}^{N} J_{ij}m_j \ , \qquad (3.67)$$

where we have defined the shorthand

$$\delta h_i^{\mathrm{nMF}} = -\sum_{j=1}^{N} J_{ij}m_j \ , \qquad (3.68)$$

since it will turn up quite often in the rest of the expansion.

The resulting magnetisations expanded to first order are then given by

$$m_i = \tanh(h_i^q) \approx \tanh\left(h_i + \sum_{j=1}^{N} J_{ij}m_j\right) \ , \qquad (3.69)$$

which is the same result as the well-known mean field equations of the equilibrium Ising model.

For the second order corrections, we compute the double derivative with respect to the external fields

$$\begin{aligned}
\frac{\partial^2 m_i}{\partial h_k h_l}\Big|_{q^*} =& \frac{\partial^2}{\partial h_k h_l}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J) \tanh\left(h_i + \sum_j J_{ij} s_j\right) \\
=& \sum_{\boldsymbol{s}} \frac{\partial^2 \pi(\boldsymbol{s}|\mathbf{h},J)}{\partial h_k \partial h_l}\Big|_{q^*} \tanh\left(h_i^q\right) + \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J)\Big|_{q^*} \frac{\partial^2 \tanh\left(h_i + \sum_j J_{ij} s_j\right)}{\partial h_k \partial h_l}\Big|_{q^*} \\
&+ \sum_{\boldsymbol{s}} \frac{\partial \pi(\boldsymbol{s}|\mathbf{h},J)}{\partial h_k}\Big|_{q^*} \frac{\partial \tanh\left(h_i + \sum_j J_{ij} s_j\right)}{\partial h_l}\Big|_{q^*} + \sum_{\boldsymbol{s}} \frac{\partial \pi(\boldsymbol{s}|\mathbf{h},J)}{\partial h_l}\Big|_{q^*} \frac{\partial \tanh\left(h_i + \sum_j J_{ij} s_j\right)}{\partial h_k}\Big|_{q^*} \\
=& 0 + \sum_{\boldsymbol{s}} q^*(\boldsymbol{s})(-2)\tanh(h_i^q)(1-\tanh^2(h_i^q))\delta_{i,k}\delta_{i,l} \\
&+ \frac{\partial}{\partial h_k}\Big|_{q^*} \underbrace{\sum_{\boldsymbol{s}} \pi(\boldsymbol{s};\mathbf{h},J)}_{=1} (1-\tanh^2(h_i^q))\delta_{i,l} + \frac{\partial}{\partial h_l}\Big|_{q^*} \underbrace{\sum_{\boldsymbol{s}} \pi(\boldsymbol{s};\mathbf{h},J)}_{=1} (1-\tanh^2(h_i^q))\delta_{i,k} \\
=& -2m_i(1-m_i^2)\delta_{i,k}\delta_{i,l} \,, \qquad (3.70)
\end{aligned}$$

hence

$$m_i^{\mathbf{hh}} = \sum_{k,l=1}^{N} \frac{\partial^2 m_i}{\partial h_k \partial h_l}\Big|_{q^*} \delta h_k \delta h_l = -2m_i(1-m_i)^2(\delta h_i)^2 \,. \qquad (3.71)$$

The mixed derivative with respect to the external field h_k and coupling J_{ln} is

given by

$$\begin{aligned}
\frac{\partial^2 m_i}{\partial h_k J_{ln}}\Big|_{q^*} =& \frac{\partial^2}{\partial h_k J_{ln}}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s}|\mathbf{h},J)\tanh\left(h_i+\sum_j J_{ij}s_j\right)\\
=& \sum_{\boldsymbol{s}} \frac{\partial^2\pi(\boldsymbol{s}|\mathbf{h},J)}{\partial h_k \partial J_{ln}}\Big|_{q^*}\tanh\left(h_i^q\right)+\sum_{\boldsymbol{s}}\pi(\boldsymbol{s}|\mathbf{h},J)\Big|_{q^*}\frac{\partial^2\tanh\left(h_i+\sum_j J_{ij}s_j\right)}{\partial h_k\partial J_{ln}}\Big|_{q^*}\\
&+\sum_{\boldsymbol{s}}\frac{\partial\pi(\boldsymbol{s}|\mathbf{h},J)}{\partial h_k}\Big|_{q^*}\frac{\partial\tanh\left(h_i+\sum_j J_{ij}s_j\right)}{\partial J_{ln}}\Big|_{q^*}\\
&+\sum_{\boldsymbol{s}}\frac{\partial\pi(\boldsymbol{s}|\mathbf{h},J)}{\partial J_{ln}}\Big|_{q^*}\frac{\partial\tanh\left(h_i+\sum_j J_{ij}s_j\right)}{\partial h_k}\Big|_{q^*}\\
=& \frac{\partial^2}{\partial h_k J_{ln}}\Big|_{q^*}\tanh(h_i^q)+\sum_{\boldsymbol{s}} q^*(\boldsymbol{s})(-2)\tanh(h_i^q)(1-\tanh^2(h_i^q))\delta_{i,k}\delta_{i,l}s_n\\
&+\frac{\partial}{\partial h_k}\Big|_{q^*}\sum_{\boldsymbol{s}}\pi(\boldsymbol{s};\mathbf{h},J)(1-\tanh^2(h_i^q))\delta_{i,l}s_n\\
&+\frac{\partial}{\partial J_{ln}}\Big|_{q^*}\sum_{\boldsymbol{s}}\pi(\boldsymbol{s};\mathbf{h},J)(1-\tanh^2(h_i^q))\delta_{i,k}\\
=& 0-2m_i(1-m_i^2)m_n\delta_{i,k}\delta_{i,l}+(1-m_i^2)\delta_{i,l}\frac{\partial m_n}{\partial h_k}\Big|_{q^*}+0\\
=& (1-m_i^2)\left\{-2m_i m_n\delta_{i,k}\delta_{i,l}+(1-m_n^2)\delta_{i,l}\delta_{n,k}\right\}\ .
\end{aligned} \tag{3.72}$$

We see that this result involves the first order derivative $\frac{\partial m_n}{\partial h_k}\Big|_{q^*}$, which we have already calculated[1]. Summing, we find

$$m_i^{\mathbf{h}J}=\sum_{k,l,n}\frac{\partial^2 m_i}{\partial h_k J_{ln}}\Big|_{q^*}\delta h_k J_{ln}=(1-m_i^2)\left(2m_i\delta h_i^{\mathrm{nMF}}\delta h_i+\sum_j J_{ij}(1-m_j^2)\delta h_j\right)\ . \tag{3.73}$$

Proceeding similarly, we also find

$$m_i^{JJ}=-2(1-m_i^2)\left(\sum_{j=1}^{N}[J_{ij}(1-m_j^2)\delta h_j^{\mathrm{nMF}}]+m_i\left(\delta h_i^{\mathrm{nMF}}\right)^2+m_i\sum_{j=1}^{N}J_{ij}^2(1-m_j^2)\right)\ . \tag{3.74}$$

[1]This happens quite frequently in the expansion. The expressions of the higher order derivatives involve lower order derivatives and similarly, the derivatives of correlations involve the derivatives of the magnetisations. Hence, the expansion has a hierarchical structure

Now, we can determine the fields h_i^q to second expansion order by solving

$$0 = m_i - m_i^q \approx m_i^J + m_i^{\mathbf{h}} + \frac{1}{2} m_i^{\mathbf{hh}} + m_i^{\mathbf{h}J} + \frac{1}{2} m_i^{JJ} \ , \tag{3.75}$$

which yields $h_i^q = h_i - \delta h_i$ with

$$\delta h_i \approx \delta h_i^{\mathrm{nMF}} + m_i \sum_{j=1}^{N} J_{ij}^2 (1 - m_j^2) =: \delta h_i^{\mathrm{TAP}} \tag{3.76}$$

and hence the second order magnetisations as derived by Kappen and Spanjers (2000) are given by

$$\begin{aligned} m_i = \tanh(h_i^q) &\approx \tanh\left(h_i - \delta h_i^{\mathrm{TAP}}\right) \\ &= \tanh\left(h_i + \sum_{j=1}^{N} J_{ij} m_j - m_i \sum_{j=1}^{N} J_{ij}^2 (1 - m_j^2)\right) \ . \end{aligned} \tag{3.77}$$

Surprisingly, the magnetisations (3.77) computed to second order in λ agree with the TAP equations for the equilibrium magnetisations of a spin glass (Thouless et al., 1977). This agreement appears to be a coincidence; our result for the magnetisations to third order in λ (see Eq. (A.1) in appendix A) does not agree with the corresponding third order result found via Plefka's expansion of the equilibrium free energy (Georges and Yedidia, 1991; Plefka, 1982).

Expansion of the two-point correlations to second order in λ

For the two-point correlations in sequential Glauber dynamics, we consider Callen's identity (3.50). Due to symmetry, it is sufficient to expand the auxiliary quantities

$$\chi_{ij} := \left\langle \delta S_i \left[\tanh\left(h_j + \sum_{k=1}^{N} J_{jk} S_k \right) - m_j \right] \right\rangle \ . \tag{3.78}$$

Analogous to (3.61), we define the short-hands $\chi_{ij}^{\mathbf{h}} = \sum_k \frac{\partial \chi_{ij}}{\partial h_k}\Big|_{q^*} \delta h_k$, $\chi_{ij}^{J} = \sum_{k,l} \frac{\partial \chi_{ij}}{\partial J_{kl}}\Big|_{q^*} J_{kl}$, etc. and write the second order expansion of the (asymmetric) auxiliary quantities χ_{ij} as

$$\chi_{ij} = \chi_{ij}^{\mathbf{h}} + \chi_{ij}^{J} + \frac{1}{2}\chi_{ij}^{\mathbf{hh}} + \chi_{ij}^{\mathbf{h}J} + \frac{1}{2}\chi_{ij}^{JJ} + \dots \ , \tag{3.79}$$

where $\chi_{ij}\Big|_{q^*} = 0$, since q^* is a factorising distribution. From this we obtain the expansion of the symmetric two-point correlations as

$$C_{ij} = \frac{1}{2}\chi_{ij} + \frac{1}{2}\chi_{ji} =: C_{ij}^{\mathbf{h}} + C_{ij}^{J} + \frac{1}{2} C_{ij}^{\mathbf{hh}} + C_{ij}^{\mathbf{h}J} + \frac{1}{2} C_{ij}^{JJ} + \dots \ . \tag{3.80}$$

For the derivative with respect to the external field h_k we find

$$\begin{aligned}
\frac{\partial \chi_{ij}}{\partial h_k}\Big|_{q^*} =& \frac{\partial}{\partial h_k}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s};\mathbf{h},J)(s_i-m_i)\left[\tanh\left(h_j+\sum_{l=1}^{N} J_{jl}s_l\right)-m_j\right] \\
=& \sum_{\boldsymbol{s}} \frac{\partial\left\{\pi(\boldsymbol{s};\mathbf{h},J)(s_i-m_i)\right\}}{\partial h_k}\Big|_{q^*} \underbrace{\left[\tanh\left(h_j^q\right)-m_j\right]}_{=0} \\
&+ \sum_{\boldsymbol{s}} q^*(\boldsymbol{s})(s_i-m_i)\frac{\partial}{\partial h_k}\Big|_{q^*}\left\{\left[\tanh\left(h_j+\sum_{l=1}^{N} J_{jl}s_l\right)-m_j\right]\right\} \\
&= \left\langle \delta S_i \frac{\partial}{\partial h_k}\Big|_{q^*}\left[\tanh\left(h_j+\sum_{l} J_{jl}S_l\right)-m_j\right]\right\rangle_{q^*} \\
=& \underbrace{\langle \delta S_i\rangle_{q^*}}_{=0}\left(\left[1-\tanh^2\left(h_j^q\right)\right]\delta_{j,k}-\frac{\partial m_j}{\partial h_k}\Big|_{q^*}\right) \\
&= 0\,. \qquad (3.81)
\end{aligned}$$

Similarly, we find that all higher order derivatives involving only external fields vanish, since we can always factor out terms like $\langle \delta S_i\rangle$ or $[\tanh(h_j^q)-m_j]$ that are zero.

For the derivative with respect to the coupling J_{kl} we find

$$\begin{aligned}
\frac{\partial \chi_{ij}}{\partial J_{kl}}\Big|_{q^*} =& \frac{\partial}{\partial J_{kl}}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s};\mathbf{h},J)(s_i-m_i)\left[\tanh\left(h_j+\sum_{n=1}^{N} J_{jn}s_n\right)-m_j\right] \\
=& \sum_{\boldsymbol{s}} \frac{\partial\left\{\pi(\boldsymbol{s};\mathbf{h},J)(s_i-m_i)\right\}}{\partial J_{kl}}\Big|_{q^*} \underbrace{\left[\tanh\left(h_j^q\right)-m_j\right]}_{=0} \\
&+ \sum_{\boldsymbol{s}} q^*(\boldsymbol{s})\delta s_i \frac{\partial}{\partial J_{kl}}\Big|_{q^*}\left[\tanh\left(h_j+\sum_{n=1}^{N} J_{jn}s_n\right)-m_j\right] \\
&= \left\langle \delta S_i\left[(1-m_j^2)\delta_{j,k}S_l-\frac{\partial m_j}{\partial J_{kl}}\Big|_{q^*}\right]\right\rangle_{q^*} \\
&= \delta_{j,k}\left\langle \delta S_i(1-m_j^2)\left[S_l-m_l\right]\right\rangle_{q^*} = \delta_{j,k}(1-m_j^2)\left\langle \delta S_i \delta S_l\right\rangle_{q^*} \\
&= \delta_{j,k}\delta_{i,l}(1-m_i^2)(1-m_j^2)\,, \qquad (3.82)
\end{aligned}$$

where $\langle \delta S_i \delta S_l\rangle_{q^*}$ is non-zero only for $i=l$, since the distribution q^* factorises. Summing over the couplings, we find

$$\chi_{ij}^J = (1-m_i^2)(1-m_j^2)J_{ji}\,, \qquad (3.83)$$

and hence by symmetry we have

$$C^J_{ij} = (\chi^J_{ij} + \chi^J_{ji})/2 = (1-m_i)^2(1-m_j^2)(J_{ij}+J_{ji})/2\ . \tag{3.84}$$

Proceeding similarly, we find

$$C^{\mathbf{h}J}_{ij} = -2m_i(1-m_i^2)(1-m_j^2)\frac{J_{ij}+J_{ji}}{2}(m_i\delta h_i + m_j\delta h_j) \tag{3.85}$$

and

$$\begin{aligned} C^{JJ}_{ij} =& m_i(1-m_i^2)(1-m_j^2)\Big\{2(J_{ij}+J_{ji})(m_i\delta h_i^{\mathrm{nMF}} + m_j\delta h_j^{\mathrm{nMF}}) + 2m_im_j(J_{ij}^2+J_{ji}^2) \\ &+ \sum_{k\neq i} J_{jk}\frac{J_{ik}+J_{ki}}{2}(1-m_k^2) + \sum_{k\neq j} J_{ik}\frac{J_{jk}+J_{kj}}{2}(1-m_k^2)\Big\}\ . \end{aligned} \tag{3.86}$$

Inserting the individual terms (3.84)-(3.86) into (3.80), we find the second order expansion of the two-point correlations given by Kappen and Spanjers (2000)

$$\begin{aligned} C_{ij} \quad = \quad & (1-m_i^2)(1-m_j^2)\Bigg(J_{ij}^{\mathrm{sym}} + m_im_j\left(J_{ij}^2+J_{ji}^2\right) \\ & + \sum_{\substack{k=1\\k\neq i}}^{N} \frac{J_{jk}J_{ik}^{sym} + J_{ik}J_{jk}^{sym}}{2}(1-m_k^2)\Bigg)\ , \end{aligned} \tag{3.87}$$

where $J^{\mathrm{sym}} = \frac{1}{2}(J+J^T)$ and $J^{\mathrm{asym}} = \frac{1}{2}(J-J^T)$ are the symmetric and antisymmetric parts of the coupling matrix respectively.

Expansion of the three-point correlations to second order in λ

Since they are needed for parameter inference, we now derive the second order expansion of the connected three-point correlations (3.51) in sequential Glauber dynamics

$$C_{ijk} = C^{\mathbf{h}}_{ijk} + C^J_{ijk} + \frac{1}{2}C^{\mathbf{hh}}_{ijk} + C^{\mathbf{h}J}_{ijk} + \frac{1}{2}C^{JJ}_{ijk} + \dots\ . \tag{3.88}$$

By symmetry, is is sufficient to consider the expansion of the auxiliary quantities

$$L_{ijk} := \left\langle \delta S_i \delta S_j \left[\tanh\left(h_k + \sum_l J_{kl}S_l\right) - m_k\right]\right\rangle \tag{3.89}$$

and then compute the symmetric connected correlations as

$$C_{ijk} = (L_{ijk} + L_{jki} + L_{kij})/3\ . \tag{3.90}$$

For the same reasons as explained for the two-point correlations, we have $L_{ijk}\Big|_{q^*} = 0$ and all derivatives involving only the external fields vanish

$$0 = \frac{\partial L_{ijk}}{\partial h_l}\Big|_{q^*} = \frac{\partial^2 L_{ijk}}{\partial h_l \partial h_n}\Big|_{q^*} = \dots \,. \tag{3.91}$$

For the first order derivative with respect to the couplings we find

$$\begin{aligned}
\frac{\partial L_{ijk}}{\partial J_{ln}}\Big|_{q^*} =& \frac{\partial}{\partial J_{ln}}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s};\mathbf{h},J)\delta s_i \delta s_j \left[\tanh\left(h_k + \sum_p J_{kp}s_p\right) - m_k\right] \\
=& \sum_{\boldsymbol{s}} \frac{\partial \pi(\boldsymbol{s};\mathbf{h},J)\delta s_i \delta s_j}{\partial J_{ln}}\Big|_{q^*} \underbrace{\left[\tanh(h_k^q) - m_k\right]}_{=0} \\
&+ \sum_{\boldsymbol{s}} q^*(\boldsymbol{s})\delta s_i \delta s_j \frac{\partial}{\partial J_{ln}}\Big|_{q^*} \left[\tanh\left(h_k + \sum_p J_{kp}s_p\right) - m_k\right] \\
=& \left\langle \delta S_i \delta S_j \left([1-\tanh^2(h_k^q)]\delta_{k,l} S_n - \frac{\partial m_k}{\partial J_{ln}}\Big|_{q^*}\right)\right\rangle_{q^*} \\
=& (1-m_k^2)\delta_{k,l}\left\langle \delta S_i \delta S_j \delta S_n \right\rangle_{q^*} \qquad (3.92) \\
=& 0 \,, \qquad (3.93)
\end{aligned}$$

since q^* factorises and there will be at least one unpaired spin-fluctuation δS. Similarly, we find

$$\frac{\partial^2 L_{ijk}}{\partial J_{ln}\partial h_p}\Big|_{q^*} = 0 \tag{3.94}$$

for the mixed derivative, since we again end up with an expectation value over q^* involving at least one unpaired spin-fluctuation δS.

The first non-zero term in the expansion is found by considering the double derivative with respect to the couplings

$$\begin{aligned}
\frac{\partial^2 L_{ijk}}{\partial J_{ln}\partial J_{vr}}\Big|_{q^*} =& \frac{\partial^2}{\partial J_{ln}\partial J_{vr}}\Big|_{q^*} \sum_{\boldsymbol{s}} \pi(\boldsymbol{s})\delta s_i \delta s_j \left[\tanh\left(h_k + \sum_p J_{kp}s_p\right) - m_k\right] \\
=& \sum_{\boldsymbol{s}} \frac{\partial\left\{\pi(\boldsymbol{s})\delta s_i \delta s_j\right\}}{\partial J_{ln}}\Big|_{q^*} \frac{\partial}{\partial J_{vr}}\Big|_{q^*} \left[\tanh\left(h_k + \sum_p J_{kp}s_p\right) - m_k\right] + (ln)\leftrightarrow(vr) \\
&+ \sum_{\boldsymbol{s}} q^*(\boldsymbol{s})\delta s_i \delta s_j \frac{\partial^2}{\partial J_{ln}\partial J_{vr}}\Big|_{q^*} \left[\tanh\left(h_k + \sum_p J_{kp}s_p\right) - m_k\right] \\
=& \frac{\partial}{\partial J_{ln}}\Big|_{q^*} (1-m_k^2)\delta_{k,v}\left\langle \delta S_i \delta S_j \delta S_r \right\rangle_\pi + (ln)\leftrightarrow(vr) \\
&- 2m_k(1-m_k^2)\delta_{k,l}\delta_{k,v}\left\langle \delta S_i \delta S_j S_n S_r \right\rangle_{q^*} \,. \qquad (3.95)
\end{aligned}$$

Since $\left\langle \delta S_i \delta S_j \delta S_r \right\rangle_{q^*} = 0$ and $\frac{\partial C_{ijr}}{\partial J_{ln}}\Big|_{q^*} = 0$, the first term gives a non-zero contribution only if $r = i$ or $r = j$. The second term gives a non-zero contribution only if there is no unpaired spin fluctuation, i.e. $i = n$ and $j = r$, or $i = r$ and $j = n$. Hence, we find

$$\begin{aligned}
\frac{\partial^2 L_{ijk}}{\partial J_{ln}\partial J_{vr}}\Big|_{q^*} =& (1-m_k^2)\delta_{k,v}\frac{\partial}{\partial J_{ln}}\Big|_{q^*}\left[\delta_{i,r}\left\langle \delta S_i^2 \delta S_j\right\rangle_\pi + \delta_{j,r}\left\langle \delta S_i (\delta S_j)^2\right\rangle_\pi\right] + (ln)\leftrightarrow(vr) \\
&- 2m_k(1-m_k^2)\delta_{k,l}\delta_{k,v}[\delta_{i,n}\delta_{j,r}+\delta_{i,r}\delta_{j,n}]\left\langle S_i\delta S_i\right\rangle_{q^*}\left\langle S_j \delta S_j\right\rangle_{q^*} \\
=& (1-m_k^2)\delta_{k,v}\frac{\partial}{\partial J_{ln}}\Big|_{q^*}[\delta_{i,r}(-2m_i C_{ij}) + \delta_{j,r}(-2m_j)C_{ij}] + (ln)\leftrightarrow(vr) \\
&- 2m_k(1-m_k^2)\delta_{k,l}\delta_{k,v}[\delta_{i,n}\delta_{j,r}+\delta_{i,r}\delta_{j,n}](1-m_i^2)(1-m_j^2) \\
=& (1-m_k^2)\delta_{k,v}\left[\delta_{i,r}\left(-2m_i\frac{\partial C_{ij}}{\partial J_{ln}}\Big|_{q^*}\right) + \delta_{j,r}\left(-2m_j\frac{\partial C_{ji}}{\partial J_{ln}}\Big|_{q^*}\right)\right] + (ln)\leftrightarrow(vr) \\
&- 2m_k(1-m_i^2)(1-m_j^2)(1-m_k^2)\delta_{k,l}\delta_{k,v}[\delta_{i,n}\delta_{j,r}+\delta_{i,r}\delta_{j,n}]\,.
\end{aligned} \tag{3.96}$$

Since we already computed $\frac{\partial C_{ij}}{\partial J_{ln}}$ above, we can now sum up to find

$$\begin{aligned}
L^{JJ}_{ijk} =& \sum_{l,n,v,r}\frac{\partial^2 L_{ijk}}{\partial J_{ln}\partial J_{vr}}\Big|_{q^*} J_{ln}J_{vr} \\
=& 2(1-m_k^2)(-2m_iJ_{ki}-2m_jJ_{kj})C^J_{ij} - 4m_k(1-m_k^2)(1-m_i^2)(1-m_j^2)\left(J_{ki}J_{kj}\right) \\
=& 2(1-m_k^2)(-2m_iJ_{ki}-2m_jJ_{kj})(1-m_i^2)(1-m_j^2)(J_{ij}+J_{ji})/2 \\
&- 4m_k(1-m_k^2)(1-m_i^2)(1-m_j^2)\left(J_{ki}J_{kj}\right)\,.
\end{aligned} \tag{3.97}$$

Hence, by considering

$$C_{ijk} = (L_{ijk}+L_{jki}+L_{kij})/3 = \frac{1}{2}(L^{JJ}_{ijk}+L^{JJ}_{jki}+L^{JJ}_{kij})/3 + \dots\,, \tag{3.98}$$

we find the expansion of the connected three-point correlations to second order in the couplings

$$\begin{aligned}
C_{ijk} = \quad & \tfrac{1}{3}(1-m_i^2)(1-m_j^2)(1-m_k^2)\times \\
& [-6A_{ijk}(J^{\mathrm{sym}},\mathbf{m}) - 2A_{ijk}(J^{\mathrm{asym}},\mathbf{m})]\,,
\end{aligned} \tag{3.99}$$

where

$$A_{ijk}(J,\mathbf{m}) = J_{ij}J_{kj}m_j + J_{ji}J_{ki}m_i + J_{jk}J_{ik}m_k\,, \tag{3.100}$$

again with the definitions of $J^{\mathrm{sym}} = \frac{1}{2}(J+J^T)$ and $J^{\mathrm{asym}} = \frac{1}{2}(J-J^T)$ as the symmetric and antisymmetric parts of the coupling matrix.

SYMMETRIES OF THE MEAN FIELD EXPRESSIONS FOR THE CORRELATIONS

In this expansion, the spin correlations exhibit particular symmetries, which affect the reconstruction of model parameters. Already the two-point correlations (3.87) depend, to first order in J, only on the symmetric part of the coupling matrix. However, also the three-point correlations show a symmetry; (3.99) is unchanged when the coupling matrix is replaced with its transpose so J^{asym} transforms to $-J^{\text{asym}}$, since $A_{ijk}(J,\mathbf{m})$ is quadratic in the couplings. Thus jointly solving (3.87) and (3.99) for the coupling matrix J either yields the reconstruction of the original coupling matrix, or its transpose. This binary symmetry is lifted only at third order in the couplings, see Eq. (A.9) in appendix A. Therefore, the third order terms of the expansion (3.58)-(3.60) give not only a quantitative improvement on the second order expressions, but are in fact necessary for successful inference. Since calculating the third order terms of the expansion (3.58)-(3.60) is straightforward but tedious, we give only the results in appendix A.

3.2.2 Parameter inference for sequential Glauber dynamics

Given empirical samples from the non-equilibrium steady state π we can now solve the inverse problem in two ways: (i) *exact inference.* We jointly solve the self-consistent equations (3.49)-(3.51) for the couplings J and external fields $\mathbf{h}$ (the approach from section 3.1.2). (ii) *mean-field inference.* We jointly solve the explicit correlation expressions (3.87) and (3.99) (taken to third order in λ) for the coupling matrix J. Subsequently solving the magnetisation equations (3.77) (also taken to third order in λ) for the external fields $\mathbf{h}$ completes the parameter reconstruction.

To test these inference schemes, we numerically simulated a system of $N = 10$ spins with random asymmetric couplings. Off-diagonal entries of the matrix of couplings were chosen independently from a Gaussian distribution with zero mean and standard deviation $\beta/\sqrt{N}$ (self-interactions were excluded: $J_{ii} \equiv 0$), and external fields independently from a Gaussian distribution with zero mean and standard deviation β. Samples of the spin configurations $\boldsymbol{s}$ under sequential Glauber dynamics (2.3) were recorded at each update after an initial settling-in period of $10^5 N$ updates to reach the steady state. Based on these measurements, we reconstructed the parameters by fitting the self-consistent equations to the data. Specifically, we minimised the sum of the relative squared prediction errors of the magnetisations, two- and three-point correlations by using the Levenberg-Marquardt algorithm (for details on the inference algorithm see appendix B).

Figure 3.1 shows the reconstruction of the couplings for different numbers of samples and coupling strengths. Three-point correlations are small and as a result the inference is affected by sampling noise. For the exact inference, the re-

construction improves significantly with the number of samples (left hand plots). For the mean-field inference, the correlations (3.87)-(3.99) computed to finite order in the couplings become inaccurate in the limit of strong couplings, which can also limit the reconstruction quality. As a result, the mean-field reconstruction performs best for intermediate coupling strengths (right hand plots). Also, the reconstruction error for the symmetric part of the couplings J^{sym} is smaller than for the antisymmetric part J^{asym}, since the former is primarily determined by the connected two-point correlations (3.87), which are considerably larger than the three-point correlations. For this reason, fewer samples are required for the accurate inference of the symmetric part of the couplings. The reconstruction of the external fields exhibits a similar behaviour, although at much lower errors, and is shown in Fig. 3.2.

3.2.3 Model selection

Beyond estimating the parameters of a particular dynamical model, an important question is what *type* of dynamics produced a particular steady state. In inference, this question is known as the model selection problem. Here, we compare three different dynamics: (i) Glauber dynamics with sequential updates, (ii) Glauber dynamics with parallel updates, and (iii) equilibrium dynamics (sequential updates with $J^{\text{asym}} = 0$). We start by taking independent samples from the steady state produced by a model with sequential Glauber dynamics as described above and calculate magnetisations and correlations. Next, we solve the exact self-consistent equations for the magnetisations, two- and three-point correlations for the different dynamics by minimising the relative prediction error as above. This gives the model parameters for a particular dynamics that best reproduce the sampled correlations. In Fig. 3.3 we compare the three-point correlations predicted by these best fits of the three different dynamical models with the sampled correlations. Indeed, the sequential model shows the best match with the sampled data, leading to the conclusion that out of the three alternatives, the data was indeed most likely produced by a model with sequential Glauber dynamics. We find analogous results for samples generated by parallel updates (2.5), see Fig. 3.4. This shows that one can distinguish the different types of dynamics based on independent samples from their steady state.

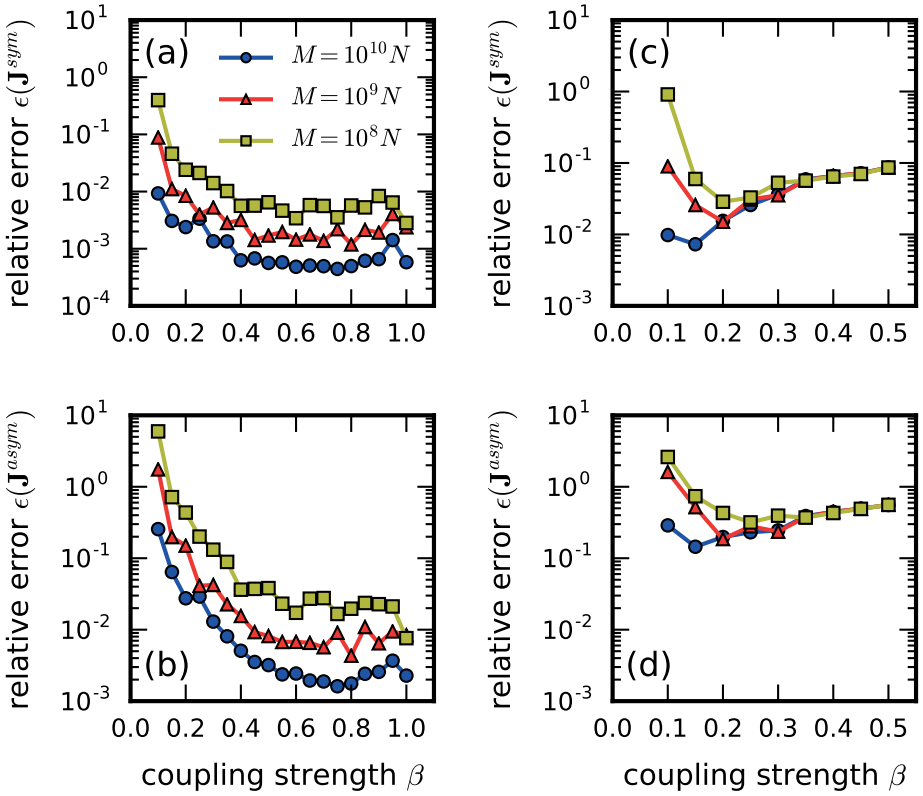


Figure 3.1: Couplings inferred from the non-equilibrium steady state. *We consider a system of $N = 10$ spins with random asymmetric couplings under Glauber dynamics* (2.3)*, see text. We plot the relative root-mean-squared reconstruction error ε between the inferred and the true couplings against the coupling strength β for different numbers of samples M. (a) and (b) show the reconstruction errors for the exact inference. (c) and (d) show the reconstruction errors for the mean-field inference. The symmetric part of the couplings $J^{sym} = (J + J^T)/2$ [(a) and (c)] generally has a lower reconstruction error than the antisymmetric part $J^{asym} = (J - J^T)/2$ [(b) and (d)].*

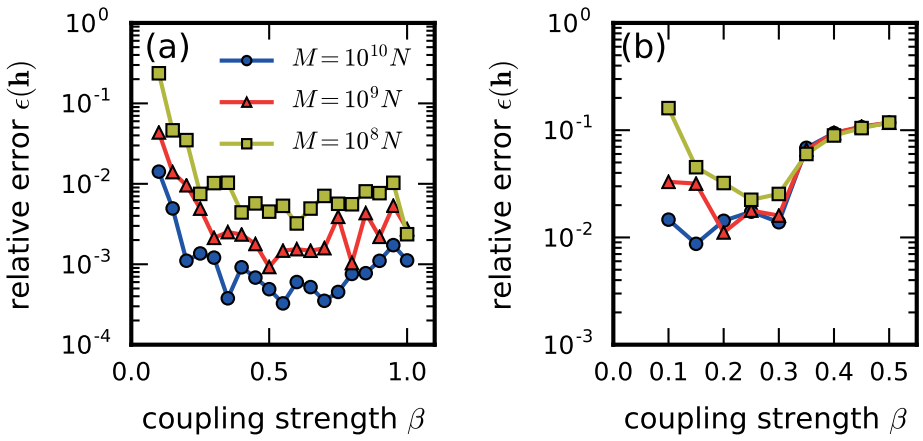


Figure 3.2: External fields inferred from the non-equilibrium steady state. *We consider a system of* $N = 10$ *spins with random asymmetric couplings under Glauber dynamics* (2.3)*, see text. We plot the relative root-mean-square reconstruction error* ε *between the inferred and true external fields against the coupling and external field strength* β *for the exact inference (a) and the mean-field inference (b).*

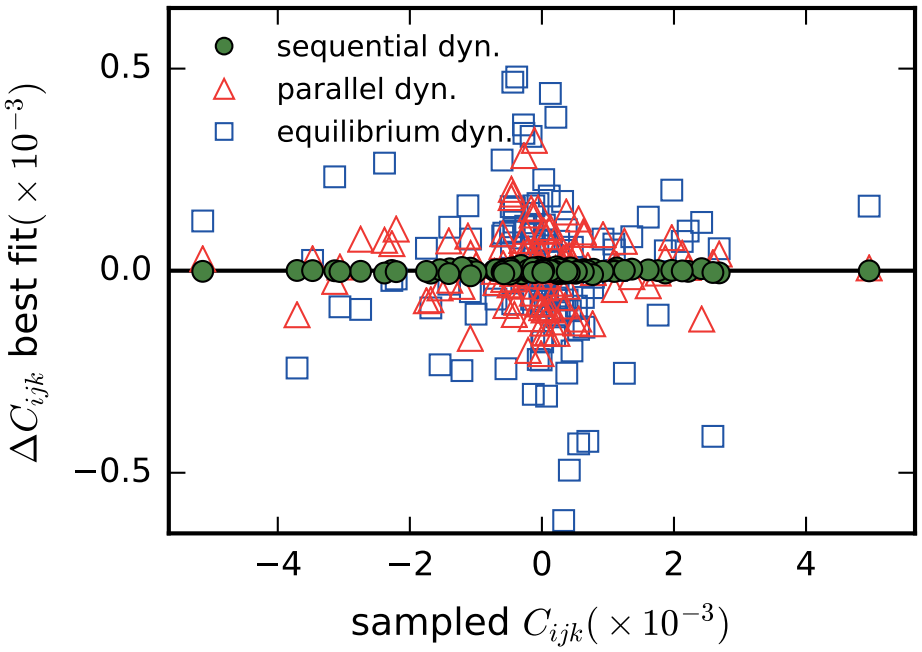


Figure 3.3: Detecting sequential dynamics on the basis of three-point correlations. *We sample configurations from the non-equilibrium steady state of Glauber dynamics with sequential updates, see text. Next, we reconstruct the model parameters from the magnetisations, two- and three-point correlations, assuming the data was generated by Glauber dynamics with sequential updates (circles), parallel updates (triangles), or equilibrium dynamics (squares). The deviations* $\Delta C_{ijk} = C_{ijk}(\mathbf{h}, J) - C_{ijk}^{sampled}$ *of the three-point correlations predicted by these three models from the corresponding correlations seen in the original samples are plotted against the sampled correlations. The relative root-mean-squared prediction errors* $\|\Delta C_{ijk}\|_2 / \|C_{ijk}^{sampled}\|_2$ *are* 0.003, 0.06, *and* 0.13 *for the sequential, parallel, and equilibrium dynamics respectively, clearly favouring the dynamics with sequential updates. The horizontal line is a guide to the eye representing a perfect fit. We used* $N = 10$ *spins, a coupling and external field strength of* $\beta = 0.2$ *and* $M = 10^{10} N$ *samples.*

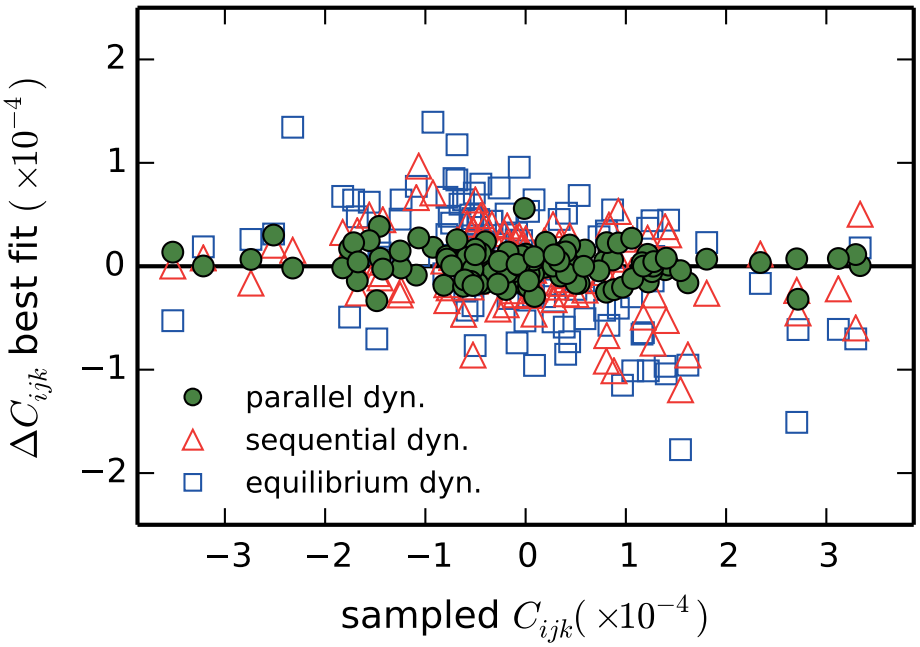


Figure 3.4: Detecting parallel dynamics on the basis of three-point correlations. *We sample magnetisations and correlations from the non-equilibrium steady state of Glauber dynamics with parallel updates* (2.5) *in the same way as was done for sequential updates, see text. Again model parameters are obtained under the three alternative models with parallel (circles), sequential (triangles), and equilibrium dynamics (squares). Their predictions for the three-point correlations C_{ijk} are compared to those observed in the original data. Shown are the deviations $\Delta C_{ijk} = C_{ijk}(\mathbf{h},J) - C_{ijk}^{sampled}$ of the predicted correlations from the sampled correlations against the sampled correlations. The relative prediction errors $\|\Delta C_{ijk}\|_2/\|C_{ijk}^{sampled}\|_2$ are* 0.10, 0.25, *and* 0.41 *for parallel, sequential, and equilibrium dynamics respectively, showing that indeed parallel dynamics best describes the data. The horizontal line is a guide to the eye representing a perfect fit. We used a coupling and external field strength of $\beta = 0.2$ and $M = 10^{10}N$ samples.*

4 The propagator likelihood

> The drive to propagate our race has also propagated a lot of other things.
>
> Georg Christoph Lichtenberg

In this chapter, we introduce an inference method that is closely linked to the maximum likelihood approach used in equilibrium inference. We propose to maximise a function, which we call propagator likelihood, that considers fictitious transitions between all sampled configurations. First, we give an intuitive justification of our function, before we derive it from an information theoretic argument by minimising relative entropy. Second, we illustrate the features of the propagator likelihood approach by inferring the model parameters for simple toy models that span the different categories of Markov processes. We begin with Markov chains in discrete time and consider a simple two-state model before proceeding to the asymmetric simple exclusion process (ASEP), representing Markov chains in continuous time. Next, we consider the Ornstein-Uhlenbeck process as an example of Markov processes with continuous configurations that are slightly more complicated to treat. Finally, we apply the propagator likelihood method to solve the more challenging inference problems for the asymmetric Ising model and replicator dynamics.

4.1 The concept

As in the last chapter, we consider a family of ergodic Markov processes with configurations $x \in \Omega$ and propagators $p_\Theta(x,\tau|y,0)$ characterised by a set of parameters Θ such that the process converges to a steady-state $\pi(x;\Theta)$, which in general will be a non-equilibrium steady state. We are given samples $D = \{x^\mu\}_{\mu=1}^M$ drawn independently from a steady state distribution $\pi(x;\Theta^{\text{true}})$ and want to infer the underlying parameter Θ^{true}. Suppose we knew the functional dependence of the steady-state distribution $\pi(x;\Theta)$ on the model parameters Θ. Then a standard approach would be to maximise the (log-) likelihood of the samples (cp. section 1.3.1),

$$\mathcal{L}(\Theta;D) = \frac{1}{M}\sum_{\mu=1}^{M} \log \pi(x^\mu;\Theta) = \sum_x \hat{p}(x) \log \pi(x;\Theta) \;, \tag{4.1}$$

where the set of sampled configurations characterises the empirical distribution $\hat{p}(x)$ with probability mass function

$$\hat{p}(x) = \frac{1}{M}\sum_{\mu=1}^{M} \delta_{x^\mu,x} \,. \tag{4.2}$$

$\delta_{x^\mu,x}$ denotes a Kronecker-δ.

However, in non-equilibrium systems we frequently do not know the steady-state distribution. Non-equilibrium systems lack detailed balance, so the steady state is not described by the Boltzmann distribution, and lacks a simple characterisation. Our solution to this inference problem in such cases is based on exploiting one elementary fact: since the distribution $\pi(x;\Theta)$ is stationary, it remains unchanged if we propagate forward in time by an arbitrary time τ. Thus, we can replace the steady-state distribution $\pi(x;\Theta)$ in the log-likelihood function (4.1) with a version propagated in time, $\sum_y p_\Theta(x,\tau|y,0)\pi(y;\Theta)$. The propagator $p_\Theta(x,\tau|y,0)$ is the conditional probability of observing the system in configuration x at time $t=\tau$, given it was in configuration y at time $t=0$. By further replacing the unknown steady-state distribution $\pi(y;\Theta)$ with the empirical distribution $\hat{p}(y)$, we arrive at the propagator likelihood

$$\begin{aligned}\mathcal{PL}(\Theta,\tau;D) &= \sum_x \hat{p}(x)\log\sum_y p_\Theta(x,\tau|y,0)\hat{p}(y) \\ &= \frac{1}{M}\sum_{\mu=1}^{M}\log\left(\frac{1}{M}\sum_{\nu=1}^{M} p_\Theta(x^\mu,\tau|x^\nu,0)\right) \,. \end{aligned}\tag{4.3}$$

In this way, we have moved the parameter-dependence from the (unknown) steady-state distribution $\pi(x;\Theta)$ to the (known) propagator $p_\Theta(x,\tau|y,0)$. Although using only configurations sampled independently with no sense of temporal order, the propagator likelihood effectively considers fictitious transitions $x^\nu \xrightarrow{\tau} x^\mu$ between all pairs of sampled configurations. For models with continuous configurations, $p_\Theta(x^\mu,\tau|x^\nu,0)$ is the transition probability density.

4.1.1 Minimising relative entropy

We rephrase the inference problem as finding a set of parameters Θ such that the propagator $p_\Theta(x,\tau|y,0)$ is compatible with the empirical distribution $\hat{p}$ being stationary (see Fig. 4.1). Demanding stationarity corresponds to requiring that $\hat{p}$ is in some sense close to a distribution $q_{\Theta,\tau}$ generated by propagating the empirical distribution for an arbitrary time τ,

$$q_{\Theta,\tau}(x) = \sum_y p_\Theta(x,\tau|y,0)\hat{p}(y) \,. \tag{4.4}$$

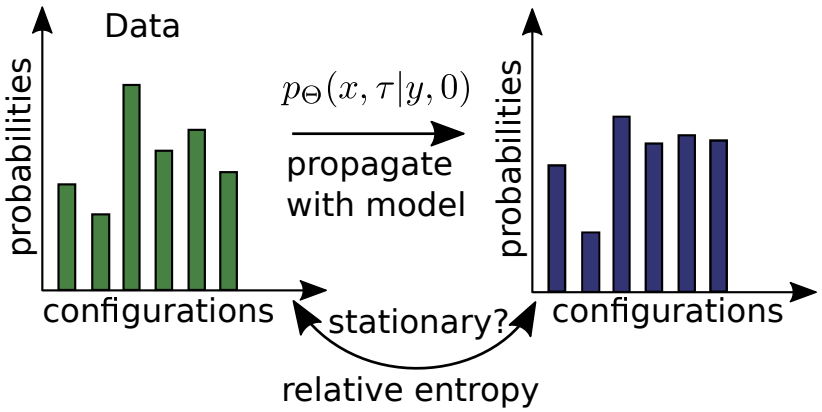


Figure 4.1: *The concept of the propagator likelihood. The independent samples* $\{x^\mu\}_{\mu=1}^M$ *define the empirical distribution* $\hat{p}$ *define by equation* (4.2)*, shown on the left. We fix an arbitrary time* τ *and use the known transition probabilities* $p_\Theta(x,\tau|y,0)$ *to propagate* $\hat{p}$ *in time and generate a new distribution* $q_{\Theta,\tau}$ *(see Eq.*(4.4)*), shown on the right. Demanding stationarity of the model, we can estimate the underlying parameters* Θ^{true} *by finding the parameters* Θ^{inf} *that minimise the distance between* $\hat{p}$ *and* $q_{\Theta,\tau}$ *as measured with relative entropy. This is equivalent to maximising the propagator likelihood (see main text).*

To quantify this notion of closeness for discrete configurations, we use the relative entropy or Kullback-Leibler divergence Kullback and Leibler (1951)

$$D_{\mathrm{KL}}(\hat{p}\|q_{\Theta,\tau}) = \sum_x \hat{p}(x)\log\frac{\hat{p}(x)}{q_{\Theta,\tau}(x)}\ . \tag{4.5}$$

Inserting the probability mass function $q_{\Theta,\tau}(x)$ defined by (4.4) into the relative entropy, we find that the relative entropy can be written as the negative sum of the Shannon entropy of the empirical distribution, $S(\hat{p}) = -\sum_x \hat{p}(x)\log\hat{p}(x)$ and the propagator likelihood (4.3):

$$D_{\mathrm{KL}}(\hat{p}\|q_{\Theta,\tau}) = -S(\hat{p}) - \mathcal{PL}(\Theta;\tau)\ . \tag{4.6}$$

The first term depends only on the sampled configurations and is independent of the model parameters; thus minimising the relative entropy over Θ is equivalent to maximising the propagator likelihood. Furthermore, due to the positivity of relative entropy, the propagator likelihood is bounded from above by the negative Shannon entropy, and this bound will be saturated only for a perfectly stationary model.

For models with continuous configurations, the relative entropy (4.5) cannot be defined, since the propagated distribution $q_{\Theta,\tau}(x)$ is continuous, while the

empirical distribution $\hat{p}(x)$ is discrete. However, the propagator likelihood (4.3) is well-defined for both discrete and continuous state spaces.

For long propagation times τ, the propagator likelihood converges to the standard log-likelihood (4.1), since $\lim_{\tau\to\infty} q_{\Theta,\tau}(x) \equiv p_\Theta(x)$ for ergodic systems. However, in general the complexity of calculating the propagator increases with τ. For models with discrete time, the single-step propagator takes the form of a (usually high-dimensional) matrix, from which propagators for longer times can in principle be computed by taking powers of this matrix.

4.2 Stochastic inference

4.2.1 Models with discrete configurations (Markov chains)

4.2.1.1 Discrete time: a simple two-configuration model

To illustrate the propagator likelihood with a toy example, we consider a system with only two configurations, denoted by 0 and 1 (see inset of Fig. 4.2). At each time step, if the system is in configuration 1, it moves to configuration 0. If it is in configuration 0, it moves to configuration 1 with probability $r \in (0,1)$ or remains in configuration 0 with probability $1-r$. In this simple model, the steady-state distribution is easily computed, giving $p_r(0) = 1/(1+r)$ and $p_r(1) = 1 - p_r(0) = r/(1+r)$.

We are now given samples $\{x^\mu\}_{\mu=1}^M \in \{0,1\}^M$ taken independently from the steady state and want to infer the model parameter r. The empirical distribution is given by the frequencies of the two configurations, $\hat{p}(0) = \frac{1}{M}\sum_{\mu=1}^M \delta_{0,x^\mu}$ and $\hat{p}(1) = 1 - \hat{p}(0)$. In this case, since we know the steady state, we can infer r from the relationship $\langle \hat{p}(0) \rangle = 1/(1+r)$, yielding $r^{\text{inf}} = (1-\hat{p}(0))/\hat{p}(0)$. For comparison, we also use the propagator likelihood (4.3) with the single-step propagator $p_r(x, \tau = 1|y, 0) = \delta_{y,1}\delta_{x,0} + \delta_{y,0}(r\delta_{x,1} + (1-r)\delta_{x,0})$, giving

$$\begin{aligned} \mathcal{PL}(r;1) &= \hat{p}_0 \log(\underbrace{(1-r)\hat{p}_0}_{0\to 0} + \underbrace{\hat{p}_1}_{1\to 0}) + \hat{p}_1 \log(\underbrace{r\hat{p}_0}_{0\to 1}) \\ &= \hat{p}_0 \log(1 - r\hat{p}_0) + (1-\hat{p}_0)\log(r\hat{p}_0)\ . \end{aligned} \tag{4.7}$$

Maximising this propagator likelihood analytically with respect to r, by setting $\frac{\mathrm{d}\mathcal{PL}}{\mathrm{d}r}(r^{\text{inf}}) = 0$, we recover the same result as obtained above from analysing the known steady-state distribution. Indeed, for uneven propagation times, the propagator likelihood shows a unique maximum at the correct value $r^{\text{inf}} = \frac{1-\hat{p}_0}{\hat{p}_0}$ and approaches the log-likelihood for increasing τ, as expected (see Fig. 4.2). For even propagation times, however, a second (global) maximum occurs at the boundary $r = 1$: since the choice $r = 1$ makes the two configurations simply exchange their probabilities in each step, any distribution becomes stationary over

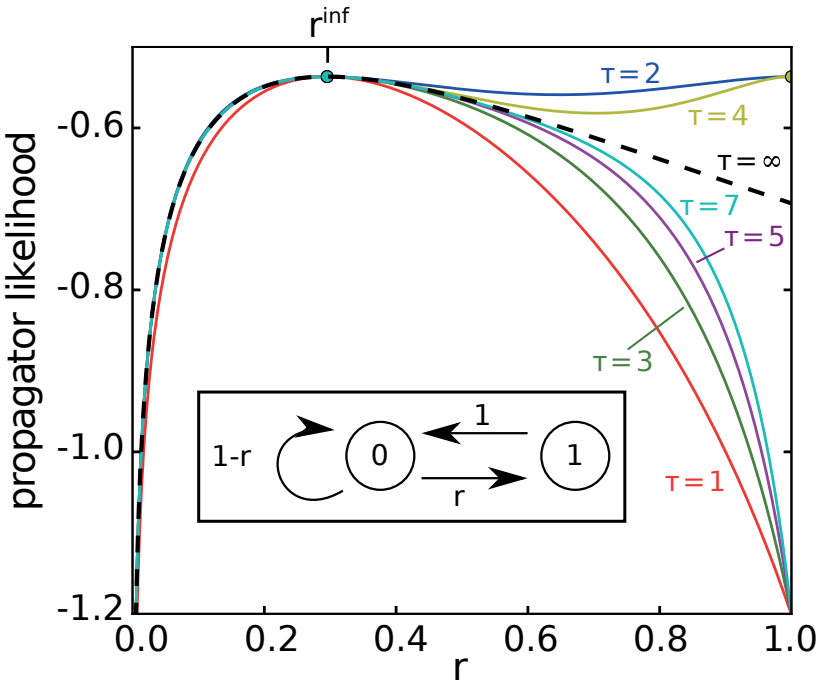


Figure 4.2: *The propagator likelihood for a simple two-configuration system. The inset shows the single-step dynamics of the system with configurations 0 and 1, controlled by the hopping probability $r \in (0,1)$. In the main figure, the solid lines show the propagator likelihood for varying propagation times τ. The dashed line shows the log-likelihood* (4.1)*, which corresponds to an infinite propagation time. The maximum likelihood estimate of the hopping probability, $r^{inf} = \frac{1-\hat{p}(0)}{\hat{p}(0)}$, is marked on the top axis and coincides with the maximum for all propagator likelihoods with an uneven number of time steps τ (see the main text for an explanation of even numbers of time steps).*

an even number of time steps and the system loses its ergodicity. Canonically, we use the single-step propagator for inferring system parameters in models with discrete time; therefore such periodicity issues cannot arise.

4.2.1.2 Continuous time: the asymmetric simple exclusion process (ASEP)

As an example of a model with continuous time, we consider the asymmetric simple exclusion process (ASEP) on a ring with asynchronous updates (see inset of Fig. 4.3). The ASEP is a simple model of a driven lattice gas and has been applied to traffic flow, surface growth, and directed paths in random media (Derrida, 1998; Evans, 1997; Krug and Ferrari, 1996).

The steady-state distribution in 1D can be calculated analytically in terms of matrix products (Derrida, 1998; Evans, 1997). In higher dimensions, however,

there is no such systematic approach and, to the best of our knowledge, the steady-state distribution is unknown.

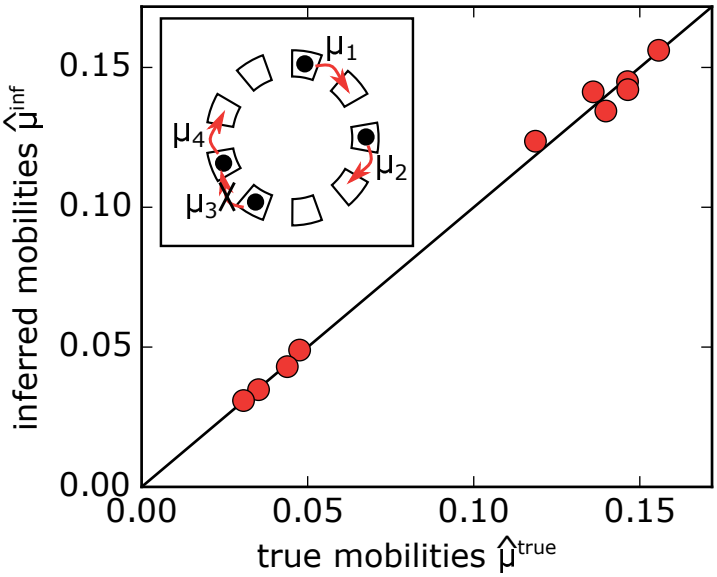


Figure 4.3: *Reconstruction of hopping rates in the asymmetric simple exclusion process (ASEP). The inset schematically shows the dynamics: K particles move on a periodic one-dimensional lattice with $N > K$ lattice sites, see text. In the main figure, we plot the relative mobilities $\hat{\mu}_i^{inf}$ inferred using the propagator likelihood versus the underlying relative mobilities $\hat{\mu}_i^{true} = \mu_i^{true} / \sum_j \mu_j^{true}$ that were used to generate the data. We simulated $K = 10$ particles hopping on a lattice with $N = 15$ sites and took $M = 10^{10}$ samples independently from the steady state. The underlying mobilities μ_i were drawn independently from a uniform distribution on the unit interval $(0,1)$.*

The model consists of K particles moving on a periodic one-dimensional lattice with $N > K$ lattice sites. Each lattice site can be occupied by at most one particle. Particles labelled $i = 1,\ldots,K$ independently attempt to jump one step in the clockwise direction at a rate μ_i called their intrinsic mobility or hopping rate. Transitions between different configurations occur at discrete random times $T_1, T_2, \ldots$, and the waiting time between the jumps is exponentially distributed with parameter $\mu_1 + \mu_2 + \ldots + \mu_K$.

The configuration of the system can be characterised by the number of free lattice sites in front of each particle, $\mathbf{n} = (n_1,\ldots,n_K) \subset (\mathbb{N}_0)^K$, with the restriction that all lattice sites are occupied: $n_1 + n_2 + \ldots + n_K = N - K$. The easiest way of stating the propagator is in terms of the numbers of such free lattice sites.

Since the steady-state distribution $p_\mu(\mathbf{n})$ itself is not associated with a time scale, we need to eliminate one parameter by rescaling time. We choose to measure time in units such that $\mu_1 + \mu_2 + \ldots + \mu_K = 1$. The steady-state distribution

is then determined only by the relative hopping rates $\hat{\mu}_i := \mu_i / \sum_j \mu_j$. In the propagator likelihood we consider the distribution of the next configuration $\mathbf{n}'$ that is visited by the process after starting in configuration $\mathbf{n}$. The transition probabilities are then determined by the relative hopping rates $\hat{\mu}_i$. Alternatively, in the general framework of propagators, we choose as the propagation time the (random) instant of the first jump, $\tau = T_1$, when a (random) particle I attempts to move one step in the clockwise direction. This random variable T_1 varies jointly with the configurations $\mathbf{n}(t)$ over different realisations of the process and its marginal distribution is irrelevant. Choosing $\tau = T_1$ simply corresponds to conditioning on the event that exactly one (attempted) jump has taken place. The single-step propagator is then defined as the probability that this first jump leads to a transition from configuration $\mathbf{n} = (n_1, \ldots, n_K)$ to a new configuration $\mathbf{n}' = (n'_1, \ldots, n'_K)$.

We use the law of total probability to decompose the propagator into the contributions from different particles that may attempt a jump at T_1, giving

$$p_{\hat{\mu}}(\mathbf{n}', T_1 | \mathbf{n}, 0) = \sum_{i=1}^{K} p_{\hat{\mu}}(\mathbf{n}', T_1 | \mathbf{n}, 0, I = i) P(I = i) \; . \tag{4.8}$$

The probability that a specific particle i attempts to jump is equal to its relative hopping rate

$$P(I = i) = \left(\frac{\mu_i}{\sum_{j=1}^{K} \mu_j} \right) \; . \tag{4.9}$$

Given that particle i attempts to jump, two things can happen: if the place in front of particle i is already occupied ($n_i = 0$), the jump is unsuccessful and the system remains in the same configuration, $\mathbf{n}' = \mathbf{n}$. Otherwise, the jump is successful and particle i hops one place forward to a free lattice site, decreasing the gap in front of it by one, $n_i \to n_i - 1$, and increasing the gap behind it by one, $n_{i-1} \to n_{i-1} + 1$ (we define $n_0 \equiv n_K$ due to the periodic boundary condition). All other gaps remain unaffected. The resulting propagator is

$$\begin{aligned} &p_{\hat{\mu}}(\mathbf{n}', T_1 | \mathbf{n}, 0, I = i) = \\ &\delta_{n_i, 0} \prod_j \delta_{n'_j, n_j} + \delta_{n'_i, n_i - 1} \delta_{n'_{i-1}, n_{i-1} + 1} \prod_{j \neq i, i-1} \delta_{n'_j, n_j} \; . \end{aligned} \tag{4.10}$$

We use this result to evaluate the propagator likelihood (4.3) and infer the relative mobilities $\hat{\mu}_i$ of $K = 10$ particles hopping on $N = 15$ lattice sites. The particle mobilities μ_i are drawn uniformly from the interval $(0, 1)$. We generate $M = 10^{10}$ Monte Carlo samples, recorded every 10 jumps after an initial settling time of 10^5 jumps to reach the steady state. We then maximise the propagator likelihood numerically with the sequential least squares programming algorithm,

as implemented in the SciPy library (Jones et al., 2001–). In Fig. 4.3 we plot the inferred relative mobilities versus the relative mobilities used to generate the samples.

4.2.2 Models with continuous configurations

Markov processes with continuous configurations pose an additional challenge: Finite-time propagators are generally not known explicitly. Instead, they are characterised indirectly as the solution of a Fokker-Planck equation. Rather than solving a Fokker-Planck equation, which for systems with a large number of degrees of freedom is generally infeasible, we proceed by approximating the propagator for short times τ via a linearisation of the corresponding Langevin equation that describes the stochastic dynamics of the model.

Again, we illustrate this procedure for a toy model. We consider one of the simplest processes with continuous configurations, the Ornstein-Uhlenbeck process, which we already encountered in section 1.2.2.2. Note that, again, for this particular case the steady-state distribution is known exactly, so we could infer the model parameters using the standard maximum likelihood approach. Nonetheless, we infer the parameters of the Ornstein-Uhlenbeck process using the propagator likelihood before turning to more complex models where the likelihood-based approach is not feasible.

4.2.2.1 The Ornstein-Uhlenbeck process

We consider a single particle diffusing in a one-dimensional harmonic potential $U(x) = \frac{b}{2}x^2$ with diffusion constant σ^2. A physical realisation of this model is a colloidal particle in solution being held in place by optical tweezers and confined to a one-dimensional channel. The dynamics of the particle is modelled by the Langevin equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -bx + \sigma\xi(t) \,, \tag{4.11}$$

where the random force $\xi(t)$ is described by δ-correlated white noise interpreted in the Itô convention.

As for the exclusion process, we must eliminate one parameter by rescaling time, since the steady-state distribution is time-independent. We consider the dimensionless time $t' = t\sigma^2$ such that the particle has unit diffusivity. To calculate the propagator likelihood for short times $\tau \ll 1$, we linearise the Langevin equation (4.11) in time

$$x(\tau) \approx x(0) - \frac{b}{\sigma^2}x(0)\tau + \int_0^\tau \mathrm{d}t'\,\xi(t') \,. \tag{4.12}$$

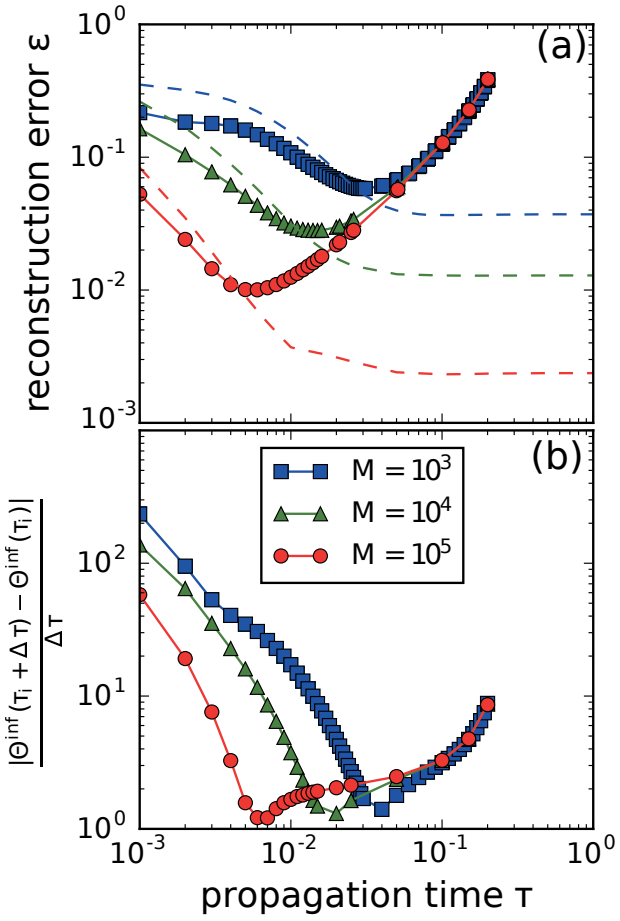


Figure 4.4: *Parameter inference in the Ornstein-Uhlenbeck process. (a) We show the relative reconstruction error* $\varepsilon = |\Theta^{inf} - \Theta^{true}|/\Theta^{true}$ *of the parameter* $\Theta = b/\sigma^2$ *(characterising the steady state) versus the dimensionless propagation time* τ *used in the propagator for sample sizes* $M = 10^3$ *(■),* $M = 10^4$ *(▲), and* $M = 10^5$ *(●). The solid lines with markers show the reconstruction errors for the approximate short-time propagator, the dashed lines indicate the reconstruction errors for the exact finite-time propagator. (b) shows the estimated rate of change of the inferred parameter with respect to the propagation time used as computed with the forward difference quotients:* $|\partial\Theta^{inf}/\partial\tau(\tau_i)| \approx |\Theta^{inf}(\tau_i+\Delta\tau) - \Theta^{inf}(\tau_i)|/\Delta\tau$*, shown on the vertical axis for the differentiation step size* $\Delta\tau = 10^{-3}$*. The minimal rate of change corresponds to the optimal choice of the propagation time (see main text).*
The data was generated by independent sampling from the stationary distribution, i.e. a centred Gaussian with variance $\sigma^2/(2b) = 1/4$*. In order to remove fluctuations between different sample sets* $\{x_\mu\}_{\mu=1}^{M}$ *and demonstrate the dependence of the average error on the sample size M and propagation time* τ*, the results were averaged over 50 independent sample sets. The equivalence of the minima of the reconstruction error and the rate of change hold true on the level of individual sample sets, while the position of the minima may vary between different sample sets.*

Since the integrated white noise $\int_0^\tau \mathrm{d}t' \xi(t')$ is normally distributed with mean 0 and variance τ, we obtain the approximate Gaussian propagator

$$p_{b/\sigma^2}(x,\tau|y,0) \approx \frac{\exp\left(-[x-\bar{x}]^2/2\tau\right)}{\sqrt{2\pi\tau}} , \tag{4.13}$$

where $\bar{x} = y - (b/\sigma^2)y\tau$ is the most likely future position of the particle.

Such a Gaussian form of the propagator emerges for any linearised Langevin equation with white noise and is not specific to the Ornstein-Uhlenbeck process. For coloured, multiplicative noise, $\xi(t) \to f(x(t),t)\eta(t)$, where f is some function and the random force $\eta(t)$ has a finite correlation time, we can proceed similarly. In this case, the normal distribution of the integrated white noise is replaced with the appropriate distribution of the integrated coloured noise $\int_0^\tau \mathrm{d}t' f(x(t'),t')\eta(t') \approx f(x(0),0)\int_0^\tau \mathrm{d}t'\eta(t')$.

After inserting the approximated propagator (4.13) into the propagator likelihood (4.3), we perform a one-dimensional maximisation of the propagator likelihood to infer the parameter $\Theta = b/\sigma^2$. In Fig. 4.4(a) we plot the relative reconstruction errors versus the dimensionless propagation time τ for various sample sizes, both for the approximate short-time propagator and for the exact finite-time propagator, which is known for the Ornstein-Uhlenbeck process. The non-monotonic behaviour of the error for the short-time propagator shows that the optimal choice for τ involves a trade-off between the error made in approximating the propagator (increasing with τ) and the error due to the finite distances between the sampled configurations that are typically crossed in the propagation time, accompanied by numerical instabilities in the exponentially damped tails of the Gaussian propagators (decreasing with τ). Indeed, as the sample size is increased, resulting in lower typical distances between individual samples, both the optimal value of τ and the total reconstruction error decrease. The exact finite-time propagator suffers only from the numerical instabilities in the tails and therefore the error decreases monotonically with τ, converging to the maximum likelihood estimate as expected. Note that the results for the approximate and exact propagators do not converge for $\tau \to 0$, since the relative difference of the propagators converges to 0 only for the peak $x = y$, even though the absolute difference converges to 0 for all values of x.

CHOOSING THE OPTIMAL PROPAGATION TIME

The non-monotonic behaviour of the reconstruction error $\varepsilon = |\Theta^{\mathrm{inf}} - \Theta^{\mathrm{true}}|/\Theta^{\mathrm{true}}$ raises the question how to choose the optimal propagation time without prior knowledge of the underlying parameter Θ^{true}. We find an answer by assuming that the error is a smooth function of the propagation time: we seek the

minimal error by demanding $0 = \partial\varepsilon/\partial\tau = \frac{\mathrm{sgn}(\Theta^{\mathrm{inf}}-\Theta^{\mathrm{true}})}{|\Theta^{\mathrm{true}}|}\frac{\partial\Theta^{\mathrm{inf}}}{\partial\tau} \sim \partial\Theta^{\mathrm{inf}}/\partial\tau$. Realistically, we will not achieve a perfect fit, i.e. $\Theta^{\mathrm{inf}} \neq \Theta^{\mathrm{true}}$ and therefore $\mathrm{sgn}(\Theta^{\mathrm{inf}} - \Theta^{\mathrm{true}}) = \pm 1$. Thus, the error derivative will become small only for $\partial\Theta^{\mathrm{inf}}/\partial\tau \to 0$. The latter quantity can be estimated directly from the data by repeating the inference for a set of propagation times $\{(\tau_i, \tau_i + \Delta\tau)\}$ and computing the forward difference quotients $\partial\Theta^{\mathrm{inf}}/\partial\tau(\tau_i) \approx [\Theta^{\mathrm{inf}}(\tau_i + \Delta\tau) - \Theta^{\mathrm{inf}}(\tau_i)]/\Delta\tau$. Since estimating the derivative from the data will incur numerical errors, we relax the condition $0 = \partial\Theta^{\mathrm{inf}}/\partial\tau$ and demand only that $|\partial\Theta^{\mathrm{inf}}/\partial\tau|$ is minimal. In Fig. 4.4(b) we show that these minima indeed coincide with the optimal choice of τ as judged from the reconstruction error shown in Fig. 4.4(a).

4.2.3 Non-equilibrium models in statistical physics and theoretical biology

We now turn to non-equilibrium applications where the standard maximum likelihood approach is not feasible, as the steady-state distribution is unknown.

4.2.3.1 The asymmetric Ising model

We consider again the asymmetric Ising model from section 2 that consists of a set of N binary spins $s_i = \pm 1$, which interact with each other via couplings J_{ij} and are subject to external fields h_i. In section 3.2 we have shown how the spin couplings J_{ij} and external fields h_i can be inferred from independent samples taken from the steady state by fitting couplings and fields to match the magnetisations, two-, and three-point correlations sampled in the data. In this section we demonstrate that the couplings can be inferred even more accurately with the propagator likelihood (4.3), which uses information from the full empirical distribution. We insert the single-step propagator (2.3) into the propagator likelihood (4.3) and maximise the propagator likelihood with respect to the external fields h_i and off-diagonal couplings J_{ij} (we consider a model without self-interactions: $J_{ii} = 0$). For the last step, we use the Broyden-Fletcher-Goldfarb-Shanno algorithm as implemented in the SciPy library (Jones et al., 2001–), and initialise the algorithm with the naive mean-field parameter estimates as described in appendix B. In Fig. 4.5, we compare the relative errors of coupling reconstruction $\varepsilon = \|\mathbf{J}^{\mathrm{inf}} - \mathbf{J}^{\mathrm{true}}\|_2 / \|\mathbf{J}^{\mathrm{true}}\|_2$ using the single-step propagator likelihood with those of fitting finite spin moments up to three-point correlations.

It turns out that parameter inference in the asymmetric Ising model requires more samples than in the equilibrium inverse Ising problem. To achieve a relative reconstruction error of 10^{-2} for an equilibrium system of $N = 10$ spins, the pseudolikelihood method requires of the order of 10^6 samples (Aurell and

Ekeberg, 2012). In the non-equilibrium model, we require at least 10^8 samples for a similar reconstruction accuracy (see Fig. 4.5). The reason for this is that, in the asymmetric Ising model, couplings are not uniquely determined by pairwise correlations. Instead, many different models can reproduce the same pairwise correlations. For this reason, we need information from higher order spin correlations, which require more samples to determine them accurately.

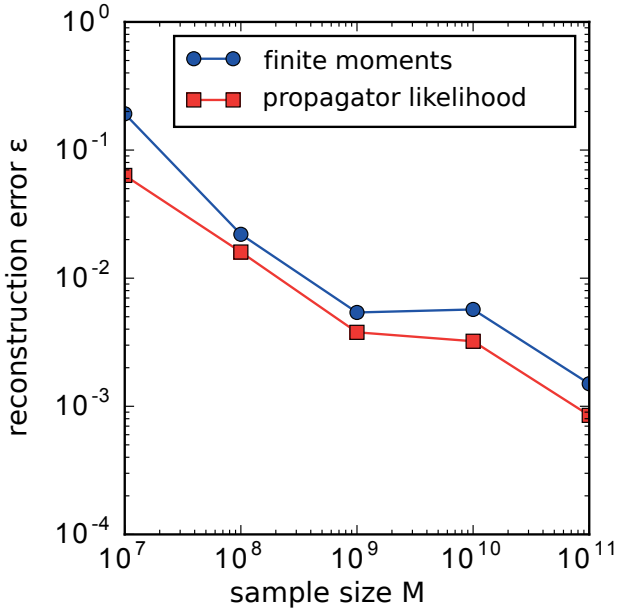


Figure 4.5: *The inference of couplings in the asymmetric Ising model. In the main figure we plot the relative errors of couplings* $\varepsilon = \|\mathbf{J}^{inf} - \mathbf{J}^{true}\|_2 / \|\mathbf{J}^{true}\|_2$ *versus the number of independent samples used for inference, using (i) finite spin moments up to three-point correlations (•) and (ii) the single-step propagator likelihood (■). Both methods are exact, so the relative error decreases with the sample size as* $\varepsilon \sim M^{-1/2}$*. The propagator likelihood (which uses the full set of configurations sampled) performs only a little better than the fit to the first three moments, showing that most information required for reconstruction is already contained in the first three moments. The underlying off-diagonal couplings were drawn independently from a Gaussian distribution with mean* 0 *and standard deviation* $1/\sqrt{N}$ *(we excluded self-interactions,* $J_{ii} = 0$*), the external fields were drawn independently from a Gaussian distribution with mean* 0 *and standard deviation* 1*. The system size was* $N = 10$ *spins.*

SPARSE NETWORKS

We consider a particular situation, where the parameter inference requires fewer samples. We consider sparse coupling matrices and assume that the topology of

the couplings is known, and only the values of the couplings are needed. Specifically, we consider the asymmetric Ising model with sparse couplings (so most interactions are zero) and assume as prior knowledge the pairs (i, j) that have a non-zero coupling between them, i.e. $J_{ij}^{\text{true}} \neq 0$ or $J_{ji}^{\text{true}} \neq 0$, regardless of the direction of the coupling. The problem has been addressed for undirected equilibrium systems like Ising models with ferromagnetic or binary couplings (Aurell and Ekeberg, 2012; Bento and Montanari, 2009). We apply the propagator likelihood to a network of $N = 10$ spins, where each possible directed link J_{ij} from spin i to spin j is non-zero with probability $p = 0.2$. The non-zero couplings are again drawn independently from a Gaussian distribution with mean 0 and variance $1/N$. Self-interactions are excluded and the external fields h_i drawn independently from a Gaussian distribution with mean 0 and variance 1. Figure 4.6 shows that the directed couplings can be inferred with slightly fewer samples when the topology of the couplings is known.

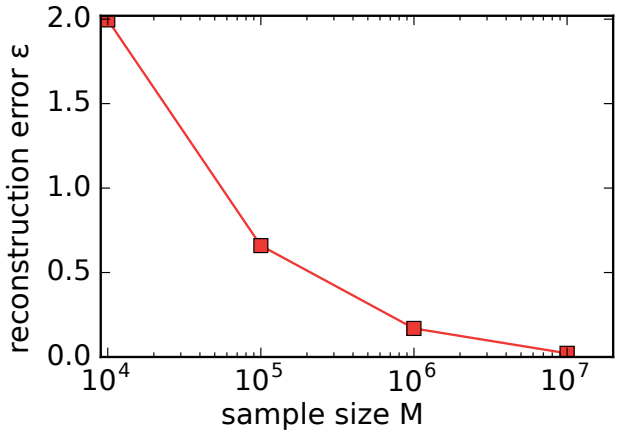


Figure 4.6: *Coupling inference in the sparse asymmetric Ising model. In the main figure, we plot the relative errors of couplings* $\varepsilon = \|\mathbf{J}^{inf} - \mathbf{J}^{true}\|_2 / \|\mathbf{J}^{true}\|_2$ *versus the number of independent samples. The underlying off-diagonal couplings were chosen sparsely: they were set to zero with probability* $1 - p = 0.8$*, and with probability* $p = 0.2$ *were drawn independently from a Gaussian distribution with mean 0 and variance* $1/N$ *(we excluded self-interactions,* $J_{ii} = 0$ *). The external fields were drawn independently from a Gaussian distribution with mean 0 and variance 1. The system size was* $N = 10$ *spins. The couplings were inferred by maximising the single-step propagator likelihood over the set of couplings between directly interacting spin pairs* (i, j)*, i.e. there is at least one true non-zero coupling between the spin pair,* $J_{ij}^{true} \neq 0$ *or* $J_{ji}^{true} \neq 0$*, regardless of the direction.*

INCREASING THE PROPAGATION TIME

So far we have restricted ourselves to the single-step propagator ($\tau = 1$). Next, we address the question whether the inference can be improved by increasing the propagation time. Intuitively, we expect that the single-step propagator is optimal when all configurations have been sampled, since this implies that all transitions over longer propagation times consist of single-step transitions that have already been probed by the single-step propagator likelihood: $x^\nu \xrightarrow{\tau} x^\mu = \sum_{x_1,x_2,\ldots,x_{\tau-1}} x^\nu \xrightarrow{\tau=1} x_1 \xrightarrow{\tau=1} x_2 \ldots \xrightarrow{\tau=1} x_{\tau-1} \xrightarrow{\tau=1} x^\mu$. Indeed, the examples with discrete time considered so far in this article fall into this category and our numerical evidence confirms that increasing the propagation time does not improve the inference. If, however, the configuration space is undersampled, some of the trajectories considered by the longer-time propagator will involve intermediate configurations that are not present in the sample and are therefore not considered by the single-step propagator. In this case, we find that increasing the propagation time does improve the inference for a fixed sample size. In Fig. 4.7 we consider a kinetic Ising model where only a small fraction of system configurations is present in the sample. Increasing the propagation time from $\tau = 1$ to $\tau = 3$ improves the inference markedly, however, the reconstruction error is considerably smaller for the symmetric part of the coupling matrix [shown in Fig. 4.7(a)] than for the antisymmetric part [shown in Fig. 4.7(b)]. This is because the symmetric part of the couplings is governed by the pairwise spin-correlations, while the antisymmetric part is dominated by higher-order spin-correlations, which require more samples for an accurate computation. The benefit of increasing the propagation time is also larger for the symmetric part, suggesting that the reconstruction of the antisymmetric part of the couplings is mainly limited by the sample size and that increasing the propagation time even further will not lead to an accurate reconstruction.

4.2.3.2 Replicator dynamics

The replicator model describes the dynamics of self-replicating entities, for instance genotypes, different animal species, RNA-molecules, or an abstract strategy in the game-theoretic problem. The replicator model has been used in population genetics, ecology, prebiotic chemistry, and sociobiology (Schuster and Sigmund, 1983).

We consider a population consisting of N different species and denote by x_i the fraction of species i in the total population (scaled for convenience by a factor on N so $\sum_i x_i = N$). The growth rate of species i, called its fitness, is denoted by f_i. The population fraction change in time depends on the growth rate f_i and the

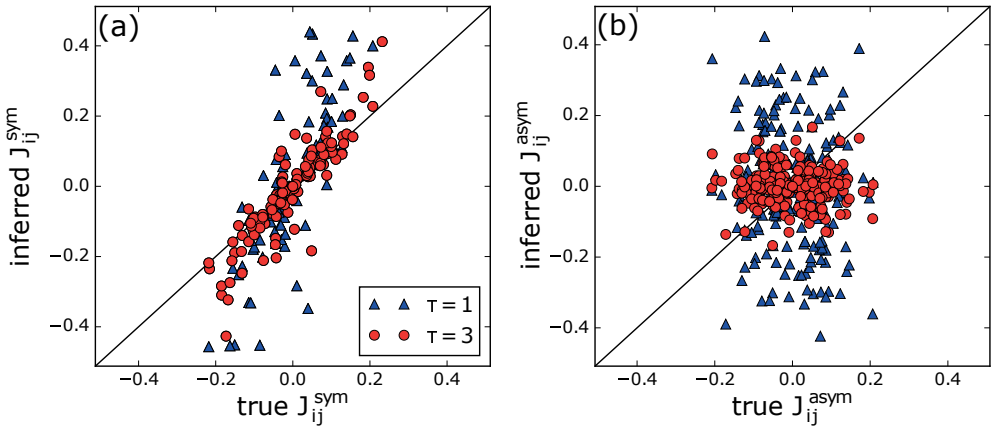


Figure 4.7: *Increasing the propagation time in the undersampled asymmetric Ising model. (a) shows the reconstructed symmetric part of the coupling matrix* $J_{ij}^{sym} = (J_{ij} + J_{ji})/2$ *for inference with the single-step propagator likelihood (▲) and for inference with the longer propagation time* $\tau = 3$ *(●). (b) shows the reconstructed antisymmetric part of the coupling matrix* $J_{ij}^{asym} = (J_{ij} - J_{ji})/2$ *for inference with the single-step propagator likelihood (▲) and for inference with the longer propagation time* $\tau = 3$ *(●). The underlying off-diagonal couplings were drawn independently from a Gaussian distribution with mean* 0 *and standard deviation* $0.5/\sqrt{N}$ *(we excluded self-interactions,* $J_{ii} = 0$*), the external fields were drawn independently from a Gaussian distribution with mean* 0 *and standard deviation* 0.5*. The system size was* $N = 16$ *spins and* $M = 10^4 N$ *samples were used. As a result, less than a third of the* 2^{16} *system configurations were present in the sample.*

average growth rate of the population $\overline{f}$

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i(t)(f_i(\mathbf{x},t) - \overline{f}(\mathbf{x},t)) \ , \tag{4.14}$$

with $\overline{f}(\mathbf{x},t) = \frac{1}{N}\sum_{j=1}^{N} x_j(t) f_j(\mathbf{x},t)$. The set of equations (4.14) defines the replicator model. The average fitness $\overline{f}$ enters to ensure that the fractions remain normalised such that $\sum_i x_i(t) = N$ for all times.

We consider a fitness which for each species i depends linearly on the population fractions of the other species

$$f_i(\mathbf{x}(t)) = \sum_{j \neq i}^{N} J_{ij} x_j(t) \ . \tag{4.15}$$

The inter-species interactions J_{ij} are quenched random variables with mean u (corresponding to a cooperation pressure) and standard deviation $1/\sqrt{N}$. There are no self-interactions, i.e. $J_{ii} = 0$.

For symmetric interactions, $J_{ij} = J_{ji}$, the fitness vector can be written as the gradient of a Lyapunov function. This implies that the system converges to an equilibrium steady state, which can be characterised by methods from statistical physics (Diederich and Opper, 1989). In the socio-biological context, however, there is no reason for the interactions to be symmetric, or in fact to assume deterministic dynamics. Assuming an asymmetric matrix J_{ij} and allowing random fluctuations $\sigma\xi_i(t)$ in the reproduction of species i leads to a set of Langevin equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = x_i(t)\left(f_i(\mathbf{x}(t)) + \sigma\xi_i(t) - \lambda(\mathbf{x},t)\right) \ , \tag{4.16}$$

where the $\xi_i(t)$ are N independent sources of white noise interpreted in the Stratonovich convention, the parameter $\sigma > 0$ controls the overall noise strength, and the factor $\lambda(\mathbf{x}(t),t) = \frac{1}{N}\sum_j x_j(t)(f_j(\mathbf{x}(t)) + \sigma\xi_j(t))$ ensures normalisation, i.e. $\sum_i x_i(t) = N$ for all times. This dynamics converges to a non-equilibrium steady state. Its characteristics for typical realisations of the matrix of couplings have been studied in the limit of a large number of species using dynamical mean field theory (Opper and Diederich, 1992).

We now turn to the problem of inferring the couplings J_{ij} of the replicator model from a set of configurations $\{\mathbf{x}^\mu\}_{\mu=1}^M$ taken independently from the non-equilibrium steady state. For simplicity, we focus on the so-called cooperative regime, in which all species survive in the long-time limit, i.e. $\lim_{t\to\infty} x_i(t) > 0\ \forall i$. This regime is characterised by a sufficiently large value of the cooperation pressure u (Opper and Diederich, 1992). Our results can be generalised to the case where species go extinct by restricting the transitions $\mathbf{x}^\nu \to \mathbf{x}^\mu$ considered in the propagator likelihood to those between configurations with the same set of surviving species.

Again, to make time dimensionless, we rescale time $t' = t\sigma^2$, resulting in a noise-term with unit magnitude. The steady state and the propagator depend only on the rescaled couplings $\hat{J}_{ij} \equiv J_{ij}/\sigma^2$. By linearising the Langevin equation (4.16) for short times and eliminating x_N via the normalisation constraint, $x_N = N - \sum_{i=1}^{N-1} x_i$, we arrive at an approximate Gaussian propagator

$$\begin{aligned} p(\mathbf{x},\tau|\mathbf{y},0) \approx & \frac{1}{\sqrt{2\pi\tau}^{N-1}\sqrt{\mathrm{Det}\Sigma}} \times \\ & \exp\left\{-\frac{1}{2\tau}\sum_{i,j=1}^{N-1}(x_i - y_i - \mu_i\tau)\,\Sigma^{-1}_{ij}\left(x_j - y_j - \mu_j\tau\right)\right\} \end{aligned} \tag{4.17}$$

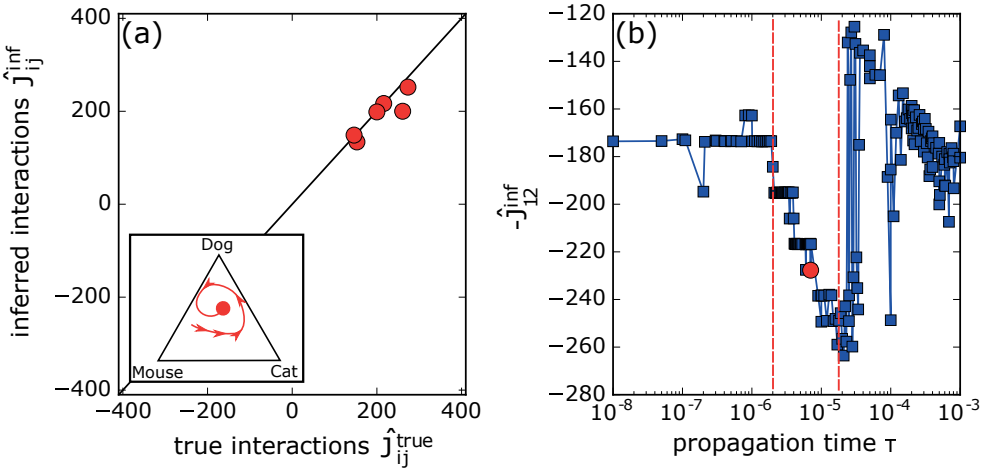


Figure 4.8: *Reconstruction of the inter-species interactions in replicator dynamics.*
(a) The inset schematically shows the dynamical model of different species competing for fractions of the total population size. The population moves on a N-dimensional simplex defined by the normalisation $\sum_i x_i = N$, $x_i \geq 0$. *In the main figure, we plot the inferred rescaled inter-species interactions* $\hat{J}_{ij}^{inf} \equiv J_{ij}^{inf}/\sigma^2$ *versus the rescaled underlying interactions* $\hat{J}_{ij}^{true} = J_{ij}^{true}/\sigma^2$ *for the propagation time* $\tau = 5.0 \times 10^{-6}$. *(b) shows how the propagation time was chosen. An arbitrary parameter (in this case* $\hat{J}_{12}$*) is chosen and its inferred value plotted for several different propagation times* τ_i. *For large* τ, *the error in the linearisation of the Langevin equations becomes large and the inference becomes unstable, as signalled by the erratic changes in the value of the inferred parameter. For small values of* τ, *most transition probabilities are damped exponentially and the numerical inaccuracies in the evaluation of these exponentials results in a saturation of the parameter value. Reasonable propagation times must lie between those two regimes and we choose a propagation time (marked by the red circle) that lies in the centre of this transition region (marked by the dashed lines). The other parameters (not shown) show a similar behaviour with the same transition region.*
The system consisted of $N = 3$ *species, the noise strength was set to* $\sigma = 0.1$, *and the underlying interactions* J_{ij}^{true} *were quenched random variables chosen independently from a Gaussian with mean* $u = 2.0$ *and standard deviation* $1/\sqrt{N}$ *(no self-interactions:* $J_{ii} = 0$*). We used an Euler discretisation of the Langevin equation* (4.16) *with time steps of length* $\Delta t = 10^{-6}/\sigma^2$ *and a total of* $M = 10^3$ *samples were taken every* 10^4 *steps after an initial settling time of* 10^9 *steps.*

with drift [1]

$$\mu_i = y_i(\hat{f}_i(\mathbf{y}) - \overline{\hat{f}}(\mathbf{y})) - \frac{y_i}{N}\left(y_i - \frac{1}{N}\sum_{j=1}^{N} y_j^2\right) \tag{4.18}$$

and covariance matrix $\Sigma = AA^T \in \mathbb{R}^{N-1\times N-1}$ with

$$A_{ij} = y_i(y_j/N - \delta_{i,j}) \ . \tag{4.19}$$

We denote by $\hat{f}_i(\mathbf{y})$ the fitness (4.15) calculated with the rescaled variables $\hat{J}_{ij} = J_{ij}/\sigma^2$, instead of the original interactions J_{ij}, and by $\overline{\hat{f}}(\mathbf{y}) = \frac{1}{N}\sum_j y_j \hat{f}_j(\mathbf{y})$ its species-weighted average.

To reconstruct the rescaled interactions $\hat{J}_{ij}$, we insert the approximate propagator (4.17) into the propagator likelihood (4.3) and maximise it using the Broyden-Fletcher-Goldfarb-Shanno algorithm (see Fig. 4.8). As for the Ornstein-Uhlenbeck process, the reconstruction error depends non-monotonically on the choice of the dimensionless propagation time τ, due to the tradeoff between the error from linearising the Langevin equation and the error from the numerical evaluation of the exponentially damped propagators. Unfortunately, the simple procedure we used for the Ornstein-Uhlenbeck process, i.e. minimising the parameter derivative $|\partial\Theta^{\text{inf}}/\partial\tau|$, cannot easily be generalised to higher dimensions. The reason is that the derivative of the reconstruction error $\partial\varepsilon/\partial\tau$ is a linear combination of the individual parameter entries $(\partial\Theta_i^{\text{inf}}/\partial\tau)_{i=1}^K$, which can cancel each other without vanishing individually (here $K = N(N-1)$ denotes the number of model parameters). To see that not all individual derivatives can vanish simultaneously, we remind ourselves that the inferred parameters must satisfy $0 \equiv \frac{\partial\mathcal{P}\mathcal{L}}{\partial\Theta_i}(\Theta^{\text{inf}}(\tau),\tau)$, $i = 1,\ldots,K$. Additionally demanding $\partial\Theta_i^{\text{inf}}/\partial\tau = 0, i = 1,\ldots,K$, corresponds to solving the system of equations $\{\frac{\partial\mathcal{P}\mathcal{L}}{\partial\Theta_i} = 0, \frac{\partial^2\mathcal{P}\mathcal{L}}{\partial\Theta_i\partial\tau} = 0\}_{i=1}^K$ for the $K+1$ variables (Θ_i,τ). This system of $2K$ nonlinear equations for $K+1$ variables will in general have no solution for $K > 1$. Instead, we can find a good propagation time by plotting a single inferred parameter versus the propagation time τ used for inference [see Fig. 4.8(b)]. The regime where the inference is dominated by the error from the linearisation for large values of τ is characterised by an erratic change of the value of the inferred parameter. The regime dominated by the error from the exponential damping for small values of τ is characterised by a saturation of the inferred parameter. These regimes are connected by a transition region, from which the propagation time should

[1] The second term in the drift arises from the difference between the Itô and Stratonovich convention in the Langevin equation.

be chosen. We choose a propagation time that lies in the centre of this transition region and find this produces a good (although not necessarily optimal) reconstruction quality.

5

Learning from perturbations in the asymmetric Ising model

> If your experiment needs statistics, you ought to have done a better experiment.
>
> Ernest Rutherford

In this chapter, we consider a setting slightly different from the one in chapers 3 and 4; we show how the parameters of the asymmetric Ising model can be inferred from samples drawn from several steady states that are generated by perturbing the couplings (or external fields). First, we define the problem of perturbation inference for the asymmetric Ising model by stating exactly what data is considered and how it is linked to the model parameters. We continue with some simple considerations about when we can infer the parameters from magnetisations and two-point spin-correlations alone, thus avoiding the difficult sampling of the three-point correlations (or higher moments). Next, we consider how the mean field theory from chapter 3 can be used to infer the couplings and external fields from these additional samples generated by couplings with added perturbations. For this purpose, we give a very simple algorithm for inference using the first order mean field equations for the magnetisations and two-point correlations. We follow up by considering the more powerful Gaussian mean field theory presented in section 2.4.2 and use it to derive self-consistent equations for the two-point spin-correlations that are exact in the thermodynamic limit. We employ these equations two infer the couplings of a fully asymmetric Sherrington-Kirkpatrick model without external fields and compare this method to the propagator likelihood inference from chapter 4.

5.1 General setting and considerations

We consider the asymmetric Ising model with sequential Glauber dynamics (2.3) without self-interactions. We are given a data set D consisting of K subsets $D = D_1 \cup D_2 \cup \ldots \cup D_K$, with each subset $D_k = \{\boldsymbol{s}^{\mu,k}\}_{\mu=1}^{M_k}$ consisting of M_k samples drawn independently from a steady state $\pi_k(\boldsymbol{s}) = \pi(\boldsymbol{s}; \mathbf{h}^{(k)}, J^{(k)})$. We assume the first data set D_1 is drawn from the steady state described by model parameters $(\mathbf{h}^{(1)}, J^{(1)}) = (\mathbf{h}, J)$, which we want to infer. The other parameter sets $(\mathbf{h}^{(k)} = \mathbf{h} + \delta\mathbf{h}^{(k)}, J^{(k)} = J + \delta J^{(k)})$ are known transformations of the original parameters; we call them perturbed parameters. The scenario we considered in sections 3.2 and 4.2.3.1 corresponds to a single data set, i.e. $K = 1$.

The motivation for considering such a setting comes from the field of **perturbation biology**. A long standard qualitative approach to identify the functions of genes is to observe what happens when the gene is artificially deactivated; this experimental technique is known as gene-knockout. On the more quantitative side, cellular signalling networks of cancer cells have been inferred by perturbing the cells with the application of targeted drugs (singly and in combination) and measuring the resulting change in protein levels etc. (Molinelli et al., 2013).

For $K = 1$, a parameter counting argument showed that N magnetisations m_i and $N(N-1)/2$ two-point spin-correlations C_{ij} are not sufficient to infer the N external fields h_i and $N(N-1)$ off-diagonal couplings J_{ij} of the asymmetric Ising model, hence we need to consider at least three-point correlations. However, since the three-point correlations are small and therefore difficult to sample, we would like to be able to infer the parameters without them. For the case $K \geq 2$, the KN magnetisations and $KN(N-1)/2$ two-point correlations could be sufficient, at least in principle. However, we have to carefully consider how to perturb the parameters. It is intuitively clear and confirmed by the explicit mean field expressions from section 3.2 that the connected spin-correlations do not depend directly on the external fields h_i[1]. Hence, a data set from a steady state where only the external fields were perturbed, will add at most N new equations, namely those of the magnetisations. Thus, for $K < N$ the inference problem would remain ill-defined when we restrict ourselves to magnetisations and two-point spin-correlations. For this reason, we will focus on steady states with perturbed couplings and for simplicity restrict ourselves to the case where one or more couplings were set to zero. This particular choice is motivated by perturbation biology: the gene regulatory interaction between two genes for example can effectively be turned off by adding molecules that bind a certain transcription factor, thus preventing this transcription factor from binding to the DNA and influencing gene expression.

5.2 Mean field inference

We begin by considering the parameter inference using the mean field theory of chapter 3. For simplicity, we consider the mean field equations expanded to first order in the couplings

[1] They do so only indirectly via the magnesations $m_i(\mathbf{h}, J)$.

$$m_i^{(k)} = \tanh\left(h_i^{(k)} + \sum_j J_{ij}^{(k)} m_j^{(k)}\right) \ , \ (i = 1,\ldots,N) \tag{5.1}$$

$$C_{ij}^{(k)} = \left(1 - (m_i^2)^{(k)}\right)\left(1 - (m_j^2)^{(k)}\right)\left(J_{ij}^{(k)} + J_{ji}^{(k)}\right)/2 \ , \ (N \geq i > j \geq 1) \ , \tag{5.2}$$

where $m_i^{(k)} = \langle S_i \rangle_{\pi_k}$ and $C_{ij}^{(k)} = \langle \delta S_i \delta S_j \rangle_{\pi_k}$ are the magnetisations and two-point spin-correlations computed from the steady state $\pi_k(\boldsymbol{s}) = \pi(\boldsymbol{s}; \mathbf{h}^{(k)}, J^{(k)})$. For the purpose of inference, we will compute the magnetisations and correlations as averages over the samples D_k drawn from the steady state $\pi_k(\boldsymbol{s})$. The structure of these equations is particularly simple and allows a straightforward algorithm for inferring the model parameters:

1. Generate $K = 1 + N(N-1)/2$ data sets: the first with the original parameters and additionally $N(N-1)/2$ sample sets, with a single coupling J_{ij} from the upper triangular part of the coupling matrix ($i > j$) set to zero (so that all couplings from the upper triangular part were set to zero exactly once).

2. To determine the couplings J_{ij}^{inf} from the lower triangular part of the coupling matrix ($i < j$), select the data set $k = k[i,j]$ in which the opposite coupling was set to zero, i.e. $J_{ji}^{(k)} = 0, J_{ij}^{(k)} = J_{ij}$. From (5.2) it is clear that the coupling can be estimated as

$$J_{ij}^{\text{inf}} = \frac{2C_{ji}^{(k[i,j])}}{\left(1 - (m_i^2)^{(k[i,j])}\right)\left(1 - (m_j^2)^{(k[i,j])}\right)} \ . \tag{5.3}$$

3. Determine the couplings J_{ij}^{inf} from the upper triangular part ($i > j$) by inserting the inferred couplings of the upper triangular part into the two-point correlations (5.2) of the original data set $k = 1$, which yields

$$\begin{aligned} J_{ij}^{\text{inf}} &= \frac{2C_{ij}^{(1)}}{\left(1 - (m_i^2)^{(1)}\right)\left(1 - (m_j^2)^{(1)}\right)} - J_{ji}^{\text{inf}} \\ &= \frac{2C_{ij}^{(1)}}{\left(1 - (m_i^2)^{(1)}\right)\left(1 - (m_j^2)^{(1)}\right)} - \frac{2C_{ij}^{(k[j,i])}}{\left(1 - (m_i^2)^{(k[i,j])}\right)\left(1 - (m_j^2)^{(k[i,j])}\right)} \ . \end{aligned} \tag{5.4}$$

4. Determine the external fields from the magnetisation equations (5.1) of the original data set

$$h_i^{\text{inf}} = \operatorname{artanh}(m_i^{(1)}) - \sum_j J_{ij}^{\text{inf}} m_j^{(1)} \ . \tag{5.5}$$

Alternatively, we may choose to generate only one perturbed data set ($K = 2$), in which all the upper triangular couplings are set to zero $J_{ij}^{(2)} \equiv 0$ ($j > i$) (more generally, one of the two couplings between each pair (i, j) must be set to zero). Then all couplings can be inferred from the same data set $k[i, j] \equiv 2$.

The strength of this inference algorithm its computational simplicity, however, it requires specific perturbations and will probably not make optimal use of the data. We can improve on the statistical efficiency of this algorithm in several ways. First, we could include the magnetisation equations (5.1) for the other data sets $k = 2, \ldots, K$, which we have not used in the algorithm described above. Second, we could include higher order terms of the mean field expansion. As can be seen from the second equations (3.87), including higher expansion orders would result in multiple two-point correlations being linked to the perturbation of a single coupling, in contrast to the case above where only a single correlation was altered. Third, we could directly evaluate Callen's identities for the magnetisations (3.49) and two-point correlations (3.50) by averaging over the respective sample sets D_k. This would also link multiple correlations to the perturbed coupling and have the additional advantage that we avoid truncating the mean field expansion [Eqs. (3.58) and (3.59)] at a finite order. However, in all these cases, we face a system of coupled non-linear equations that are much harder to solve than the linear equations encountered in the simple algorithm described above. For the fully asymmetric Sherrington-Kirkpatrick model, however, we can use the Gaussian mean field theory discussed in section 2.4.2. The Gaussian mean field theory combines the advantages of being exact (in the thermodynamic limit) and at the same time producing a system of linear equations from which to infer the couplings (in the case of vanishing external fields).

5.3 Inference with the Gaussian mean field theory

We consider the particular asymmetric Ising model, for which the Gaussian mean-field theory discussed in section 2.4.2 becomes exact: the **fully asymmetric Sherrington-Kirkpatrick model**, characterised by quenched random couplings J_{ij} with J_{ij} and J_{ji} drawn independently from a Gaussian with zero mean and variance $1/N$. However, in contrast to the time-series of parallel Glauber dynamics (2.5) considered in section 2.4.2, we consider samples taken

independently from the steady state produced by sequential or parallel Glauber dynamics (2.3). Hence, we cannot sample the time-shifted correlations $D_{ij}^{\text{par}} = \langle \delta S_i(t+1)\delta S_j(t)\rangle$. However, in the following we will show that we can use the distribution of the effective local fields $\psi_i = h_i + \sum_j J_{ij}S_j$ derived by Mézard and Sakellariou (2011) to find a self-consistent characterisation of the equal-time correlations $C_{ij} = \langle \delta S_i(t)\delta S_j(t)\rangle \equiv \langle \delta S_i \delta S_j\rangle$ in the steady state.

5.3.1 Self-consistent equations for the two-point correlations.

For sequential Glauber dynamics (2.3), we consider Callen's identities for the parallel time-shifted correlations D_{ij}^{par} (2.22) and for the sequential equal-time correlations C_{ij}^{seq} (2.28) and notice that they are linked by the relation

$$C_{ij}^{\text{seq}} \overset{i\neq j}{=} (D_{ij}^{\text{par}} + D_{ji}^{\text{par}})/2 \,. \tag{5.6}$$

Inserting expression (2.51) into (5.6), we find

$$C_{ij}^{\text{seq}} \overset{i\neq j}{=} \frac{\lambda_i}{2}(JC^{\text{seq}})_{ij} + \frac{\lambda_j}{2}(C^{\text{seq}}J^T)_{ij} \,, \tag{5.7}$$

with $\lambda_i = \langle 1 - \tanh^2(\psi_i)\rangle$ as defined in (2.52).

For parallel Glauber dynamics (2.5), we evaluate Callen's identity (2.21) $C_{ij}^{\text{par}} = \langle \tanh(\psi_i)\tanh(\psi_j)\rangle - m_i m_j$ by integrating over the joint distribution (2.48) of the local fields $(\psi_i, \psi_j) = (g_i + x_i, g_j + x_j)$, which yields

$$\begin{aligned}
C_{ij}^{\text{par}} \overset{i\neq j}{=} &\int \frac{\mathrm{d}x_i}{\sqrt{2\pi\Delta_i}} \int \frac{\mathrm{d}x_j}{\sqrt{2\pi\Delta_j}} \exp\left\{-\frac{x_i^2}{2\Delta_i} - \frac{x_j^2}{2\Delta_j}\right\}\left(1 + \rho_{ij}\frac{x_i x_j}{\sqrt{\Delta_i}\sqrt{\Delta_j}}\right) \\
&\times \tanh(g_i + x_i)\tanh(g_j + x_j) - m_i m_j \\
=&\rho_{ij}\left(\int \frac{\mathrm{d}y_i}{\sqrt{2\pi}} e^{-y_i^2/2}\, y_i \tanh(g_i + y_i\sqrt{\Delta_i})\right)\left(\int \frac{\mathrm{d}y_j}{\sqrt{2\pi}} e^{-y_j^2}\, y_j \tanh(g_j + y_j\sqrt{\Delta_j})\right) \\
=&\rho_{ij}\sqrt{\Delta_i}\sqrt{\Delta_j}\left(\int \frac{\mathrm{d}y_i}{\sqrt{2\pi}} e^{-y_i^2/2}\, [1 - \tanh^2(g_i + y_i\sqrt{\Delta_i})]\right) \\
&\times \left(\int \frac{\mathrm{d}y_j}{\sqrt{2\pi}} e^{-y_j^2/2}\, [1 - \tanh^2(g_j + y_j\sqrt{\Delta_j})]\right) \\
=&(JC^{\text{par}}J^T)_{ij}\lambda_i\lambda_j \,.
\end{aligned} \tag{5.8}$$

In the statistical forward problem, we may solve (5.7) (sequential Glauber dynamics) or (5.8) (parallel Glauber dynamics) for the correlation matrix $C = C(\mathbf{h}, J)$, given the coupling matrix J and external fields $\mathbf{h}$ (in the general case, the equations are non-linear, since the $\lambda_i = \lambda_i(\mathbf{h}, J; \mathbf{m}(\mathbf{h}, J), C(\mathbf{h}, J))$ depend on

C). In the inverse statistical problem, we cannot solve for the couplings J_{ij} given the correlations C_{ij}, since the equation system is under-determined. Hence, we need to consider perturbations.

5.3.2 Inference from perturbations in sequential Glauber dynamics

In the following, for simplicity we will only consider sequential Glauber dynamics and vanishing external fields $h_i \equiv 0$. This leads to vanishing magnetisations $m_i \equiv 0$ and makes the variance (2.43) of the effective local fields uniform (in the thermodynamic limit) $\Delta_i \equiv 1$ (Bachschmid-Romano and Opper, 2015). Thus, also the λ_i (2.52) are independent of the couplings

$$\lambda_i \equiv \lambda = \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{\sqrt{2\pi}}\, e^{-x^2/2}(1-\tanh^2(x)) \approx 0.605706\ . \tag{5.9}$$

This simplifies our self-consistent expression (5.7) to

$$C_{ij} \overset{i\neq j}{=} \frac{\lambda}{2}\left((JC)_{ij}+(CJ^T)_{ij}\right) = \frac{\lambda}{2}\left(\sum_k J_{ik}C_{jk}+J_{jk}C_{ik}\right)\ . \tag{5.10}$$

We can use these equation to infer the couplings from the sample sets D_k in the following way: we compute the sample averages $C_{ij}^{(k)} = \langle \delta S_i \delta S_j \rangle_{D_k}$ for the data sets generated by the steady states $\pi_k(\boldsymbol{s}) = \pi(\boldsymbol{s}; \mathbf{h}^{(k)} = 0, J + \delta J^{(k)})$ and insert them into (5.10), which yields a system of $KN(N-1)/2$ linear equations, which we want to solve for the $N(N-1)$ off-diagonal couplings J_{ij}. For $K = 1$ this system is under-determined, for $K \geq 2$ the system could be under-determined, well-determined, or over-determined, depending K and the nature of the perturbations used. Since we are dealing with a set of linear equations, we can find the solution with the Moore-Penrose pseudoinverse, or even faster with computationally very efficient linear least squares algorithms. For the case of non-zero external fields, the auxiliary variables λ_i will depend on the external fields and couplings, thus destroying the linearity of the Gaussian self-consistent correlation equations. However, this happens in a benign way that allows simple iterative algorithms to solve the full correlation equations (5.7) also for large system sizes (Mézard and Sakellariou, 2011).

5.3.2.1 Gaussian inference by deleting upper triangular couplings

For concreteness, we consider one additional data set ($K = 2$) generated by setting all couplings of the upper triangular part of the original coupling matrix to

zero, i.e.

$$J_{ij}^{(2)} = \begin{cases} J_{ij} & i > j \\ 0 & i \leq j \end{cases} . \tag{5.11}$$

We distribute a total number of $M = M_1 + M_2$ samples evenly among the two data sets D_1 and D_2 and compute the connected correlations $C_{ij}^{(1)} = \langle \delta S_i \delta S_j \rangle_{\pi(J)}$ and $C_{ij}^{(2)} = \langle \delta S_i \delta S_j \rangle_{\pi(J^{(2)})}$ with $J^{(2)}$ as in (5.11).

To generate the samples ,we run Monte Carlo simulations of the sequential Glauber dynamics (2.3) for a system consisting of N spins, once for the original couplings J_{ij} and once for the perturbed couplings $J_{ij}^{(2)}$. We let the dynamics run for $10^7 N$ initial spin updates to reach the steady state and then collect samples every $100\,N$ spin updates.

To infer the coupling matrix from the data we solve

$$C_{ij}^{(1)} = \frac{\lambda}{2} \left(\sum_{k=1}^{N} J_{ik} C_{jk}^{(1)} + J_{jk} C_{ik}^{(1)} \right) \quad (1 \leq i < j \leq N) \tag{5.12}$$

$$C_{ij}^{(2)} = \frac{\lambda}{2} \left(\sum_{k=1}^{i-1} J_{ik} C_{jk}^{(2)} + \sum_{k=1}^{j-1} J_{jk} C_{ik}^{(2)} \right) \quad (1 \leq i < j \leq N) \tag{5.13}$$

for the coupling matrix J_{ij}.

The inference results for varying system sizes N and number of samples M are shown in Fig. 5.1. For small sample sizes, the error decreases with the sample size as expected. For large sample sizes, the error saturates at a finite value due to the error of the Gaussian theory for finite system size N. Consequently, the saturation level decreases with N. Numerically, we find that the equation system has full rank, hence, there seems to be no significant overlap between the two data sets D_1 and D_2, which would hinder an accurate coupling reconstruction.

Importantly, with this perturbation inference we are able to accurately reconstruct the couplings from sample sizes that are orders of magnitude smaller than in the case of inference from the unperturbed steady state alone, where we required at least $10^6 N$ samples (see sections 3.2 and 4.2.3.1). In the following, we will take a closer look at the differences between the two inference approaches.

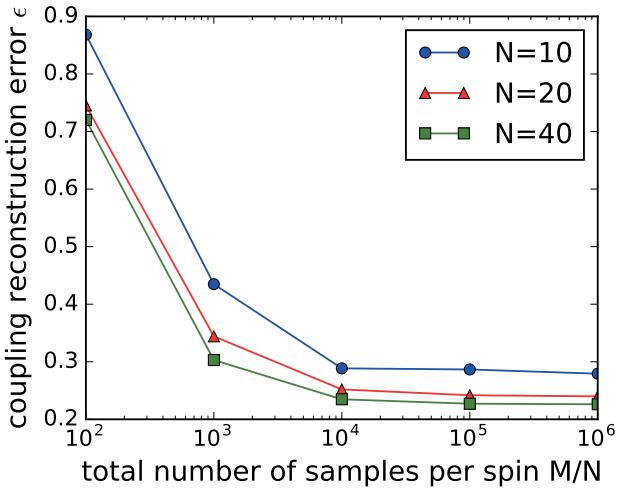


Figure 5.1: *Inference from perturbations with the Gaussian mean field theory. Shown are the relative errors* $\varepsilon = \|J^{inf} - J^{true}\|_2 / \|J^{true}\|_2$ *of the reconstructed couplings* J^{inf} *relative to the underlying couplings* J^{inf} *versus the total number of samples per spin* M/N *for different system sizes* N*. The couplings were inferred by a linear least-squares solution of the Gaussian self-consistent equations* (5.12) *and* (5.13) *for two data sets* D_1, D_2 *consisting of* $M/2$ *samples each. The data set* D_2 *was generated by a coupling matrix where the upper triangular part of the original couplings* J_{ij}^{true} *was set to zero. The samples were generated by a Monte Carlo simulation of the sequential Glauber dynamics* (2.3) *for the fully asymmetric Sherrington-Kirkpatrick model with couplings* J_{ij} *chosen independently from a Gaussian with mean 0 and variance* $1/N$*, in the absence of external fields,* $h_i \equiv 0$*.*

COMPARISON WITH INFERENCE BASED ON THE STEADY STATE ONLY

We ask ourselves how the Gaussian inference approach described above, based on the two-point spin-correlations from the two data sets D_1 and D_2, compares to the propagator likelihood method applied to data sampled from the unperturbed steady state only. To this end, we consider a system consisting of $N = 10$ spins and two different sample sizes $M = 10^3 N$ and $M = 10^6 N$. For the perturbation inference we distribute the samples evenly among the two data sets D_1 and D_2 described above, so each data set consists of $M/2$ samples. Moreover, we consider a second case where we distribute the same perturbations across not one but three additional data sets ($K = 4$), i.e. for generating $\tilde{D}_2$ we set only the first third of the upper triangular couplings to zero, in $\tilde{D}_3$ the second third, and in $\tilde{D}_4$ the last third and each sample set consists of $M/4$ samples. For inference with the propagator likelihood, we draw all M samples from the steady state generated by the unperturbed parameters. Then we infer the couplings J_{ij} with

(i) the Gaussian perturbation method by solving (5.12) and (5.13), (ii) the Gaussian perturbation method with three additional data sets by equations analogous to (5.12) and (5.13), and (iii) the propagator likelihood method as described in section 4.2.3.1.

The results of the coupling inference are shown in Fig. 5.2. The performance of the Gaussian perturbation inference relative to the propagator likelihood depends on the sample size and is determined by different contributions: the Gaussian perturbation inference uses the additional information provided by the perturbations, i.e. for the data sets generated with perturbations, it has prior knowledge concerning the couplings. However, it considers only two-point correlations and thus neglects the additional information contained in higher moments of the sampled distribution. The propagator likelihood inference without perturbations, on the other hand, effectively uses this information contained in the higher moments by considering the full sampled distribution, but it does not benefit from the information contained in the perturbations [1]. For small sample sizes, most of the information contained in a sample is already captured by the two-point correlations, thus the Gaussian inference does not neglect much information, but benefits form the perturbation information and hence outperforms the propagator likelihood [Fig.5.2(a)]. For larger sample sizes, there is more additional information contained in the higher moments and the propagator likelihood outperforms the Gaussian perturbation inference, which becomes limited by errors of the Gaussian theory for finite system size N [Fig.5.2(b)]. By distributing the perturbations over three instead of one additional data set, the information contained in the perturbations is increased and the Gaussian inference performance is increased, but remains limited by errors due to the finite system size [Fig.5.2(c) and (d)].

Finally, we notice that it is easier to scale up the Gaussian inference from perturbations to larger system sizes, as compared to the the propagator likelihood approach. This is due to the linearity of equations (5.12) and (5.13) and due to the fact that the sample average of the two-point correlations C_{ij} needs to be computed only once with $MN(N-1)/2$ steps, in contrast to the sample average of the propagator likelihood which needs to be performed in every iteration step of the nonlinear optimisation algorithm and takes on the order of M^2 elementary computations.

[1] Of course we could also apply the propagator likelihood method to the perturbation data sets, thus using not only the information contained in the higher moments but also the information contained in the perturbations. This approach should perform at least as well as the Gaussian perturbation inference, but is not so easily scalable to larger system sizes.

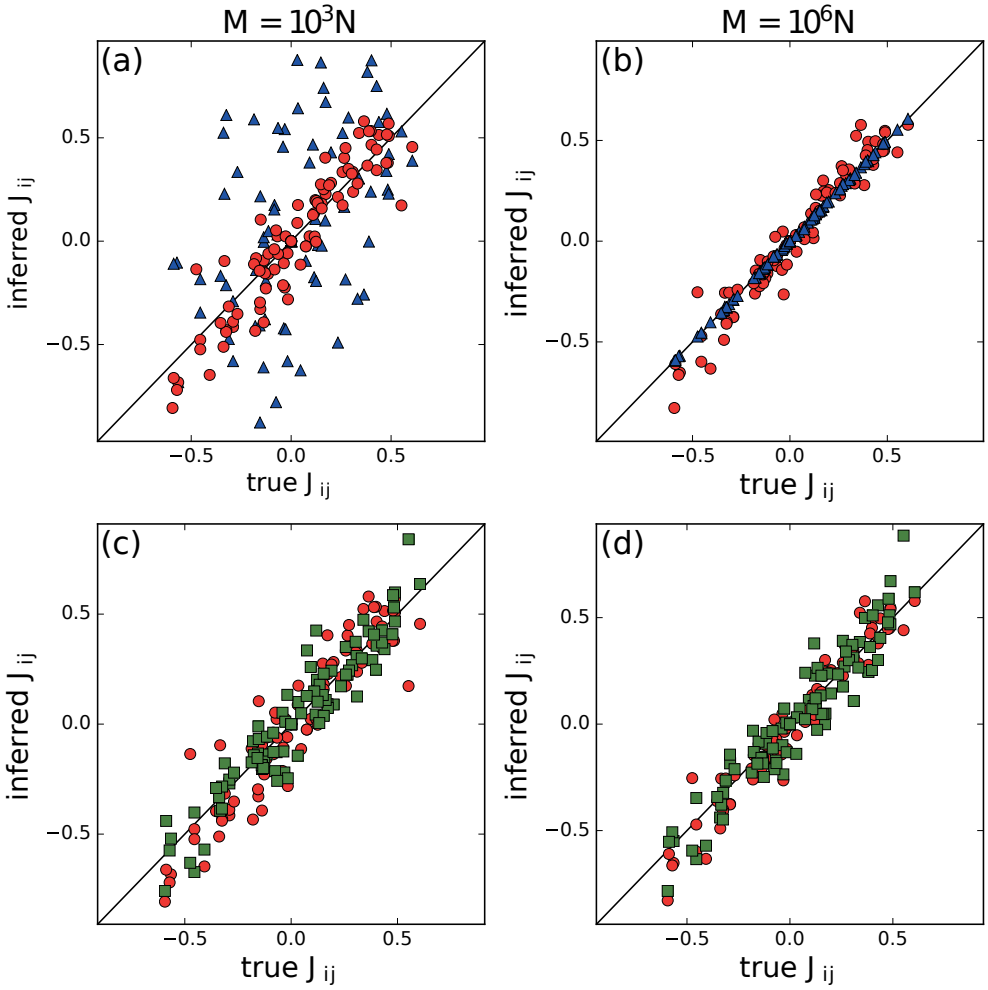


Figure 5.2: *Coupling inference from perturbations with the Gaussian mean field theory versus the propagator likelihood. We consider a fixed number of samples M and consider three scenarios: (i) all samples are taken from the unperturbed steady state and the couplings J_{ij}^{inf} are inferred with the propagator likelihood method from section 4.2.3.1 (▲), (ii) the samples are split evenly among two data sets ($K = 2$) with one set of samples taken from the unperturbed steady state and the second set taken from a steady state generated by deleting all upper triangular couplings ($J_{ij} = 0$ for $j > i$) (●), (iii) the samples are split evenly among four data sets ($K = 4$) where the same perturbations of case (ii) are distributed across three additional data sets instead of one (see text) (■). Shown are the inferred couplings J_{ij}^{inf} versus the underlying, unperturbed couplings J_{ij}^{true} that generated the data, once for $M = 10^3$ samples per spin [(a) and (c)] and once for $M = 10^6$ samples per spin [(b) and (d)]. The samples were generated by a Monte Carlo simulation of the sequential Glauber dynamics* (2.3) *for the fully asymmetric Sherrington-Kirkpatrick model consisting of $N = 10$ spins, with couplings J_{ij} chosen independently from a Gaussian with mean 0 and variance $1/N$, in the absence of external fields, $h_i \equiv 0$.*

6 Conclusions and Outlook

> If all statisticians in the world were laid head to toe, they wouldn't be able to reach a conclusion.
>
> George Bernard Shaw

To summarize, we have developed two different methods to solve the problem of inferring model parameters of ergodic Markov processes based on independent samples taken from their non-equilibrium steady state, which is not accessible to the standard equilibrium inference algorithms, since it is not described by a Boltzmann distribution. In particular, we successfully inferred the parameters of our main paradigm, the asymmetric Ising model, and showed that the developed inference algorithms could be extended to infer the parameters of other ergodic Markov processes such as Ornstein-Uhlenbeck processes, the asymmetric simple exclusion process, or replicator dynamics. Finally, we showed that the couplings in the asymmetric Ising model can be inferred even more efficiently by distributing a fixed number of samples across several steady states, which are linked by perturbations of the parameters controlled by the experimenter.

Self-consistent equations and non-equilibrium mean field theory

Our first method is based on computing exact self-consistent relations that hold true in the steady state and link observable statistics to model parameters, like the well-known Callen identities for the magnetisations $\langle s_i \rangle = \langle \tanh(h_i + J_{ij} s_i) \rangle$ in the Ising model. In the spirit of the pseudolikelihood method for equilibrium inference, these can be evaluated exactly (for the number of samples tending to infinity) by replacing the expectation value over the steady state distribution with the sample average. Fitting sufficiently many of these equations to the sampled observable statistics lead to a successful parameter reconstruction in different models. In contrast to the pseudolikelihood method, however, there is no associated function that is maximised, since the steady state is not described by the Boltzmann distribution and unknown to us. Therefore, the choice of the observables that should be fitted is not determined by a (pseudo-)likelihood function (as are the magnetisations and two-point spin-correlations in the equilibrium Ising model), but instead have to be chosen in an ad hoc way. While this method is computationally more efficient than exact likelihood maximisation for equilibrium systems, which requires an exponential time for computing the normalising partition function, the computation of the self-consistent equations still requires

a time proportional to the number of samples due to the averaging procedure. Since the parameter fitting, in general, constitutes a non-linear optimisation problem, the equations have to be evaluated many times over several iterations of the optimisation algorithm. A way to compute the self-consistent equations even faster and in closed form is to use mean field theory. To this end, we extended the non-equilibrium mean field theory of Kappen and Spanjers (2000), developed originally for the asymmetric Ising model, to more general Markov processes and showed how it can be used to compute the self-consistent equations in a series expansion around a factorising distribution. In particular, for the asymmetric Ising model we argued that magnetisations and two-point spin-correlations cannot be sufficient to infer the full coupling matrix and external fields, but showed that including three-point correlations is already sufficient for parameter reconstruction. For this purpose we computed the three-point spin-correlations to third order in the couplings within the non-equilibrium mean field theory and added third order correction terms to the results for the magnetisations and two-point correlations obtained by Kappen and Spanjers (2000). Interestingly, it turned out that the mean-field expressions obeyed particular symmetries that required the expansion to be carried out to third order for successful parameter inference. Of course, there may be observables that are better suited to the inference of couplings and external fields in the asymmetric Ising model, however, the three-point correlations seem a canonical choice. Unfortunately, a very large number of samples is required for reconstructing the parameters of the asymmetric Ising model. This can be understood quite easily, since compared to the equilibrium, symmetric Ising model, the non-equilibrium, asymmetric Ising model has roughly twice as many free parameters that need to be inferred. While the couplings of the equilibrium Ising model can be uniquely determined from pairwise spin-correlations, the asymmetric Ising model requires at least three-point correlations which are smaller and thus harder to sample.

THE PROPAGATOR LIKELIHOOD METHOD

Our second method avoids the ad hoc choice of observables to be fitted and uses the full sampled distribution for inference. It is based on maximising a function we call propagator likelihood, which considers the likelihood of fictitious transitions between all pairs of sampled configurations, even though the configurations were sampled without sense of temporal order. We showed that the propagator likelihood could be derived from a measure for stationarity of the process: we minimised relative entropy between the sampled distribution and a distribution generated by propagating this sampled distribution in time. Further, we illustrated the wide applicability of this method and used it to infer the model parameters not only of our non-equilibrium paradigm, the asymmetric Ising model, but

also for various models with both discrete and continuous time and configurations, such as the asymmetric simple exclusion process and replicator dynamics. For Markov chains with discrete configurations, considering single-step transitions turned out to be sufficient (and optimal) if the configuration space was fully sampled. In particular, we showed that for the asymmetric Ising model, maximising the single-step propagator likelihood provided a parameter estimate that was slightly better than fitting only magnetisations, two- and three-point correlations. In case that the configuration space is undersampled, we found that the inference could be improved by increasing the propagation time, however, at the price of a more costly computation of the propagators. An interesting open question is whether there exists an optimal propagation time that should be used. The answer could possibly be linked to the time needed by the chain to converge to the steady state, which is investigated by the theory of Markov mixing times. For Markov processes with continuous configurations, there exists an optimal propagation time due to a trade-off between, on the one hand, the error from linearising the Langevin equation that increases with the propagation time, and on the other hand, the error from numerical instabilities in the exponentially damped tails of the propagators that become important when the propagation time becomes small. For one-parameter models like the one-dimensional Ornstein-Uhlenbeck process, we showed how the optimal propagation time could be inferred from the data, for more complex models, however, we could only give a rough guide to finding suitable propagation times. It would be interesting to find out whether there is a general procedure for choosin an optimal propagation time also in high-dimensional models. For large sample sizes, the propagator likelihood method is even more costly than our first approach based on an exact evaluation of the self-consistent equations: evaluating the propagator likelihood in general requires a time that is proportional to the square of the number of samples. For systems with discrete configurations, this can be reduced to a time linear in the samples, provided there are only a small number of neighbouring configurations that can be reached in a single step with non-zero transition probability. In the future, it would be interesting to find approximations that allow for a more efficient computation of the propagator likelihood, thus making parameter inference with this method feasible for larger systems.

Comparison with minimum probability flow learning

Our propagator likelihood method bears some resemblance to an equilibrium inference method called minimum probability flow learning (Sohl-Dickstein et al., 2011). The similarity is that in minimum probability flow learning one also seeks to minimise the relative entropy between the sampled distribution and a distribution that is generated by propagating the sampled distribution in time.

However, minimum probability flow learning is only concerned with equilibrium inference problems described by the Boltzmann distribution. Instead of using a Markovian dynamics intrinsic to the system, an artificial, deterministic continuous-time dynamics is constructed that obeys detailed balance such that it produces the desired Boltzmann distribution as a steady state (as noted in section 1.2.3.1 there are many choices one can make). This artificial dynamics is then run for an infinitesimal time and the relative entropy between the propagated distribution and the sampled distribution minimised. The authors show, by a Taylor expansion in time, that this is equivalent to minimising the probability current out of the sampled configurations (it is assumed that the configuration space is undersampled). This trick allows to circumvent the costly computation of the equilibrium partition function. However, this method is not directly applicable to non-equilibrium steady states, since they are in fact characterised by non-vanishing probability currents and a lack of detailed balance. Our approach, on the other hand, uses the actual Markovian dynamics of the system and runs it for a finite propagation time (although we used a linearised approximation for continuous-time dynamics).

CAN WE ALWAYS INFER ALL PARAMETERS?

In the particular models we studied, we managed to infer all model parameters for the discrete time models and for the continuous time models we could infer all parameters apart from a single parameter that determined the time scale. However, this will not hold true in general cases. A priori, it is not clear whether all model parameters actually enter the steady state distribution or in which combinations they do. As an example, consider a system of N independent Ornstein-Uhlenbeck processes $\mathrm{d}X_i = -b_i X_i \mathrm{d}t + \sigma_i \mathrm{d}W(t)$. The steady state factorises into N independent Gaussian distributions, each with zero mean and variance $\sigma_i^2/2b_i$. Thus, of the originally $2N$ parameters only N (transformed) parameters determine the steady state. In this light, it is a non-trivial result that we could infer all parameters of the asymmetric Ising model from independent samples of the steady state alone.

LEARNING FROM PERTURBATIONS IN THE ASYMMETRIC ISING MODEL

We showed how to use independent samples from several steady states generated by perturbed couplings to infer the underlying couplings in the asymmetric Ising model. To this end, we considered the non-equilibrium mean field theory from chapter 3 in addition to deriving self-consistent equations for the equal-time two-point correlations within the Gaussian mean field theory of Mézard and Sakellariou (2011). By distributing the independent samples evenly among the unperturbed steady state and a perturbed steady state generated by deleting

half of the original couplings, we could use these self-consistent equations to infer the underlying couplings with a sample size that was order of magnitudes smaller than for inference from the unperturbed steady state alone (as was done in chapters 3 and 4). The reason is the following: by considering magnetisations and two-point spin-correlations for several steady states, we could find enough equations to determine the couplings without recourse to higher order correlations. Hence, we avoided the estimation of the small three-point (and higher order) correlations, which requires many samples. In principle, the statistical efficiency of the Gaussian inference approach used in chapter 5 could be improved even further by evaluating the exact Callen identities for magnetisations and n-point spin-correlations via sample averaging, or by computing the propagator likelihood of the different data sets. However, for the practically relevant case of large systems and small sample sizes, we do not expect a significant gain in efficiency that would justify the immense increase in the computational complexity of the inference algorithm. In contrast, the Gaussian inference method based on the two-point correlations from perturbed steady states has the great advantage that it is easily scalable to larger system sizes. Since we investigated only one single kind of perturbation, it would be interesting to further explore the full space of possible perturbations and their combinations in order to identify the optimal way to distribute a fixed number of samples across several perturbed data sets - subject of course to the constraint imposed by experimental limitations in creating the perturbations.

Future research directions

The work presented in this thesis could be extended in several directions. One interesting question would be how ergodicity breaking may effect the inference, i.e. what happens the process explores only a limited region of the configuration space on experimentally accessible time scales? Is it still possible to infer the model parameters? For the equilibrium inverse Ising problem, for example, it is known that even with an exact theory for the statistical forward problem, the inference can be hindered by the emergence of multiple thermodynamic states at low temperatures. In order to successfully infer the parameters, one has to cluster the samples and evaluate the averages of magnetisations and correlations separately for each cluster (Nguyen and Berg, 2012).

A question motivated by biological applications is how inferred coupling topologies in sparse models differ between equilibrium Ising models and non-equilibrium Ising models, i.e. do equilibrium models mostly describe only effective interactions that are not due to physical interactions and can an asymmetric Ising model better explain these connections?

Another interesting line of inquiry would be to ask how the methods de-

veloped in this thesis could be extended to non-Markovian processes that are described by an explicit memory kernel or to Markov processes with hidden variables.

Finally, it would be interesting to see the development of novel inference methods completely different from the ones considered in this thesis. Since we were the first to address the problem of inferring the parameters of the asymmetric Ising model based on independent samples taken from the non-equilibrium steady state, it seems likely that our methods might be considered only a starting point and we can expect that potentially better methods will be developed in the future. For comparison, the equilibrium inverse Ising problem has been addressed since the seventies and still new insights emerge, new methods are developed, and older ones improved (the latest results are as recent as this year).

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cogn. Sci.*, 9(1):147–169, 1985. 46

E. Aurell and M. Ekeberg. Inverse Ising inference using all the data. *Phys. Rev. Lett.*, 108(9):090201, 2012. 95, 97

L. Bachelier. Théorie de la spéculation. *Annales Scientifiques de l'École Normale Supérieure*, 17:21–86, 1900. 10

L. Bachschmid-Romano and M. Opper. Learning of couplings for random asymmetric kinetic Ising models revisited: random correlation matrices and learning curves. *J. Stat. Mech.*, 2015. 110

M. Bailly-Bechet, A. Braunstein, A. Pagnani, M. Weigt, and R. Zecchina. Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach. *BMC Bioinformatics*, 11(1):355, 2010. 40

D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012. 27

J. Bento and A. Montanari. Which graphical models are difficult to learn? *Adv. Neural Inf. Process. Syst.*, 22, 2009. 97

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. B*, 36(2):192–236, 1974. 34

H. Bethe. Statistical theory of superlattices. In *Proc. Roy. Soc. London A*, volume 150, pages 552–575, 1935. 32

H. B. Callen. A note on Green functions and the Ising model. *Phys. Lett.*, 4(3): 161, 1963. 42

A. C. C. Coolen. Statistical Mechanics of Recurrent Neural Networks I. Statics. arXiv:cond-mat/0006010, 2000a. 40, 46

A. C. C. Coolen. Statistical Mechanics of Recurrent Neural Networks II. Dynamics. arXiv:cond-mat/0006011, 2000b. 40

B. Derrida. An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. *Phys. Rep.*, 301(1-3):65–83, 1998. 89

B. Derrida, E. Gardner, and A. Zippelius. An exactly solvable asymmetric neural network model. *Europhys. Lett.*, 4(2):167–173, 1987. 40

S. L. Dettmer and J. Berg. Inferring the parameters of a Markov process from snapshots of the steady state. preprint, `https://arxiv.org/abs/1707.04114v1`, 2017. 2

S. L. Dettmer, H. C. Nguyen, and J. Berg. Network inference in the nonequilibrium steady state. *Phys. Rev. E*, 94(5):052116, 2016. 2

S. Diederich and M. Opper. Replicators with random interactions - a solvable model. *Phys. Rev. A*, 39(8):4333–43336, 1989. 100

A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.*, 322(8):549–560, 1905. 10

M. Evans. Exact steady states of disordered hopping particle models with parallel and ordered sequential dynamics. *J. Phys. A-Math. Gen.*, 30(16):5669–5685, 1997. 89

W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley & Sons, third edition, 1968. 4

A. Fokker. Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Ann. Phys.*, 43:810–820, 1914. 17

C. W. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer, Heidelberg, fourth edition, 2009. 4, 18

A. Georges and J. Yedidia. How to axpand around mean-field theory using high-temperature expansions. *J. Phys. A*, 24(9):2173–2192, 1991. 34, 73

R. J. Glauber. Time-dependent statistics of the Ising model. *J. Math. Phys.*, 4 (2):294–307, 1963. 38

G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, third edition, 2001. 4

W. Heisenberg. Zur Theorie des Ferromagnetismus. *Physik. Z.*, 49:619–636, 1928. 38

J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation*, volume 1. Addison-Wesley Publishing Company, 1991. 40, 45, 46

E. Ising. Beitrag zur Theorie des Ferromagnetismus. *Z. Phys.*, 31(1):253–258, 1925. 38, 39

K. Itô. Stochastic Integral. *Proc. Imp. Acad.*, 20(8):519–524, 1944. 10

K. Itô. On a Stochastic Integral Equation. *Proc. Japan Acad.*, 22(2):32–35, 1946. 10

E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL `http://www.scipy.org/`. [Online; accessed 2016-07-08]. 92, 95, 132

H. J. Kappen and F. B. Rodríguez. Efficient learning in Boltzmann machines using linear response theory. *Neural Comput.*, 10(5):1137–1156, 1998. 34

H. J. Kappen and J. J. Spanjers. Mean-field theory for asymmetric neural networks. *Phys. Rev. E*, 61(5):5658, 2000. 1, 53, 55, 60, 61, 67, 73, 75, 116

A. Klenke. *Wahrscheinlichkeitstheorie*. Springer Berlin Heidelberg, third edition, 2013. 4, 7

J. Krug and P. Ferrari. Phase transitions in driven diffusive systems with random rates. *J. Phys. A-Math. Gen.*, 29(18):L465–L471, 1996. 89

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 03 1951. 30, 87

H. Lei, N. A. Baker, and X. Li. Data-driven parameterization of the generalized Langevin equation. *Proc. Natl. Acad. Sci.*, 113(50):14183–14188, 2016. 4

D. S. Lemons and A. Gythiel. Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530-533 (1908)]. *Am. J. Phys.*, 65:1079–1081, 1997. 10

W. Lenz. Beiträge zum Verständnis der magnetischen Eigenschaften in festen Körpern. *Physik. Z.*, 21:613–615, 1920. 38

K. Levenberg. A method for the solution of certein non-linear problems in least-squares. *Q. Appl. Math.*, 2(2):164–168, 1944. 132

D. A. Levin and Y. Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2008. 4, 6

D. J. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003. 27

A. Markov. Extension of the law of large numbers to quantities, depending on each other (1906). reprint. *Journal Électronique d'Histoire des Probabilités et de la Statistique [electronic only]*, 2(1b):Article 10, 12 p., 2006. 4

D. W. Marquardt. An algorithm for least square estimation of non-linear parameters. *J. Soc. Ind. Appl. Math.*, 11(2):431–441, 1963. 132

M. Mézard and J. Sakellariou. Exact mean field inference in asymmetric kinetic Ising systems. *J. Stat. Mech.*, page L07001, 2011. 1, 2, 37, 46, 49, 50, 51, 52, 109, 110, 118

E. J. Molinelli, A. Korkut, W. Wang, M. L. Miller, N. P. Gauthier, X. Jing, P. Kaushik, Q. He, G. Mills, D. B. Solit, C. A. Pratilas, M. Weigt, A. Braunstein, A. Pagnani, R. Zecchina, and C. Sander. Perturbation biology: inferring signaling networks in cellular systems. *PLOS Comput. Biol.*, 9(12):e1003290, 2013. 106

H. C. Nguyen and J. Berg. Mean-field theory for the inverse Ising problem at low temperatures. *Phys. Rev. Lett.*, 109(5):050602, 2012. 119

H. C. Nguyen, R. Zecchina, and J. Berg. Inverse statistical problems: from the inverse Ising problem to data science. preprint, `http://arxiv.org/abs/1702.01522`, 2017. 31, 46

L. Onsager. Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition. *Phys. Rev.*, 65:117–149, 1944. 38

M. Opper and S. Diederich. Phase transition and 1/f noise in a game dynamical model. *Phys. Rev. Lett.*, 69(10):1616–1619, 1992. 100

M. Opper and D. Saad, editors. *Advanced Mean-field Methods: Theory and Practice*. The MIT Press, 2001. 32

R. Peierls. On Ising's model of ferromagnetism. In *Math. Proc. Cambridge Philos. Soc.*, volume 32, pages 477–481. Cambridge University Press, 1936. 32, 38

P. Peretto. Collective properties of neural networks: a statistical physics approach. *Biol. Cybern.*, 50(1):51–62, 1984. 40

M. Planck. Über einen Satz der statistischen Dynamik und eine Erweiterung der Quantentheorie. *Sitzungsberichte der Preuss. Akademie der Wissenschaften, Berlin*, pages 324–341, 1917. 17

T. Plefka. Convergence condition of the TAP equations for the infinite-ranged Ising spin glass model. *J. Phys. A: Math. Gen.*, 15:1971, 1982. 34, 55, 60, 73

S. Redner, P. L. Krapivsky, and E. Ben-Naim. *A Kinetic View of Statistical Physics*. Cambridge University Press, 2010. 4

Y. Roudi and J. Hertz. Mean-field theory for nonequilibrium network reconstruction. *Phys. Rev. Lett.*, 106(4):048702, 2011. 46, 47, 49, 52

J. Sakellariou, Y. Roudi, M. Mézard, and J. Hertz. Effect of coupling asymmetry on mean-field solutions of the direct and inverse Sherrington–Kirkpatrick model. *Phil. Mag.*, 92(1-3):272–279, 2012. 50

P. Schuster and K. Sigmund. Replicator dynamics. *J. Theor. Biol.*, 100(3):533–538, 1983. 98

E. Seneta. *MAM2006: Markov Anniversary Meeting*, pages 1–20. Boson Books, Raleigh, North Carolina, 2006. 4

D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. *Phys. Rev. Lett.*, 35(26):1792–1796, 1975. 38

J. Sohl-Dickstein, P. B. Battaglino, and M. R. DeWeese. New Method for Parameter Estimation in Probabilistic Models: Minimum Probability Flow. *Phys. Rev. Lett.*, 107(22):220601, 2011. 117

N. N. Taleb. *The Black Swan*, chapter 9, page 124. Penguin Books, third edition, 2010. 27

D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of 'solvable model of a spin glass'. *Phil. Mag.*, 35:593, 1977. 73

M. von Smoluchowski. Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Ann. Phys.*, 326(14):756–780, 1906. 10

N. Wiener. Differential-Space. *J. Math. Phys.*, 2(1-4):131–174, 1923. 10

H.-L. Zeng, M. Alava, E. Aurell, J. Hertz, and Y. Roudi. Maximum likelihood reconstruction for Ising models with asynchronous updates. *Phys. Rev. Lett.*, 110(21):210601, 2013. 48

A

Further mean field equations for the asymmetric Ising model

A.1 Magnetisations to third order

Continuing the mean-field expansion for the magnetisations (3.58) to third order inthe couplings, we find

$$m_i = \tanh\left(h_i + \sum_{j=1}^{N} J_{ij} m_j - m_i a_i + \frac{2}{3}(1-3m_i^2) b_i - m_i c_i \right) , \tag{A.1}$$

where we defined the auxiliary quantities

$$a_i = \sum_{j=1}^{N} J_{ij}^2 (1-m_j^2) \tag{A.2}$$

$$b_i = \sum_{j=1}^{N} J_{ij}^3 m_j (1-m_j^2) \tag{A.3}$$

$$c_i = \sum_{\substack{j,k=1 \\ j\neq k}}^{N} J_{ij}(1-m_j^2) J_{ik} (1-m_k^2) J_{jk}^{\mathrm{sym}} \tag{A.4}$$

and $J_{ij}^{\mathrm{sym}} = (J_{ij} + J_{ji})/2$ denotes the entry of the symmetric part of the coupling matrix.

A.2 Correlations under sequential Glauber dynamics

Here we give the results of expanding the two- and three-point correlations (3.59),(3.60) for sequential Glauber dynamics to third order in the couplings and for the connected four-point correlations to second order in the couplings.

We denote the two- and three-point correlations computed to second order by C_{ij}^{TAP} and C_{ijk}^{TAP}, respectively. Their expressions are given by Eqs.(3.87) and (3.99).

For the third order correction to the two-point correlations we obtain

$$\begin{aligned} C_{ij} - C_{ij}^{\mathrm{TAP}} =& (1-m_i^2)(1-m_j^2) \times \\ & \left\{ \frac{1}{3} J_{ij}^3 (1-3m_i^2)(1-3m_j^2) + 2 m_i m_j J_{ij} A_{ji} \right. \\ & \left. - \frac{1}{2} J_{ij}(1-m_i^2) a_i + \frac{1}{2} m_i F_{ij} + \frac{1}{2} E_{ij} \right\} \\ & + (i \leftrightarrow j), \end{aligned} \tag{A.5}$$

where we defined the auxiliary quantities

$$A_{ij} = \sum_{\substack{k=1 \\ k\neq i}}^{N} J_{ik}^{\mathrm{sym}} J_{jk}(1-m_k^2) \tag{A.6}$$

$$E_{ij} = \sum_{\substack{k=1 \\ k\neq i}}^{N} \frac{A_{ik}+A_{ki}}{2} J_{jk}(1-m_k^2) \tag{A.7}$$

$$F_{ij} = \sum_{\substack{k=1 \\ k\neq i}}^{N} (J_{ik}^2+J_{ki}^2) J_{jk}(1-m_k^2) m_k \tag{A.8}$$

with a_i as defined in (A.2).

For the third order correction to the three-point correlations we find

$$\begin{aligned} C_{ijk} - C_{ijk}^{TAP} =& \frac{(1-m_i^2)(1-m_j^2)(1-m_k^2)}{3} \times \\ & \Bigg\{ -J_{ij}^{\mathrm{sym}} \left[4J_{ki}J_{kj}m_i m_j m_k + 4m_k(J_{ki}^2 m_i^2) \right] \\ & + 2J_{ki}m_i \Big[(1-3m_k^2) J_{ki}J_{kj} \\ & - m_i m_j (J_{ij}^2 + J_{ji}^2) - A_{ij} \Big] \\ & - 2m_k \sum_{\substack{l=1 \\ l\neq i,j}}^{N} J_{kl}(1-m_l^2) J_{ki} J_{jl}^{\mathrm{sym}} \\ & + \sum_{\substack{l=1 \\ l\neq i,j}}^{N} \frac{J_{kl}}{2} \left(\frac{C_{ijl}^{TAP}}{(1-m_i^2)(1-m_j^2)} \right) \Bigg\} \\ & + \text{permutations of } (i,j,k). \end{aligned} \tag{A.9}$$

For the connected four-point correlations C_{ijkl} with $i<j<k<l$ we find to lowest non-vanishing order

$$\begin{aligned} C_{ijkl}^{TAP} =& (1-m_i^2)(1-m_j^2)(1-m_k^2)(1-m_l^2) \\ & \times \left(J_{ij}^{sym} J_{kl}^{sym} + J_{ik}^{sym} J_{jl}^{sym} + J_{il}^{sym} J_{jk}^{sym} \right) \end{aligned} \tag{A.10}$$

A.3 Correlations under parallel Glauber dynamics

To second order in the couplings, we find that the connected two-point correlations C_{ij}^{p} $(i<j)$ are given by

$$C_{ij}^{p,\mathrm{TAP}} = (1-m_i^2)(1-m_j^2)\sum_{k=1}^{N} J_{ik}J_{jk}(1-m_k^2)\ , \tag{A.11}$$

and their third order correction reads

$$C_{ij}^{p} - C_{ij}^{p,\mathrm{TAP}} = (1-m_i^2)(1-m_j^2)\sum_{l} J_{il}J_{jl}2m_l(1-m_l^2)(m_iJ_{il}+m_jJ_{jl})\ . \tag{A.12}$$

For the connected three-point correlations C_{ijk}^{p} with $i<j<k$ we find to lowest non-vanishing order

$$C_{ijk}^{p} = (1-m_i^2)(1-m_j^2)(1-m_k^2)\sum_{l=1}^{N} J_{il}J_{jl}J_{kl}(-2m_l)(1-m_l^2)\ . \tag{A.13}$$

B

Description of the moment-matching inference algorithm

In order to reconstruct the model parameters of the asymmetric Ising model, we consider Callen's identities for the magnetisations (3.49), two- and three-point correlations ((3.50) and (3.51) for sequential dynamics, (3.52) and (3.53) for parallel dynamics, and solve them for the couplings and external fields. Our goal is to minimise the relative squared error between the magnetisations and correlations predicted by Callen's identities for a particular set of parameters and the sampled averages, which leads us to define the cost function

$$\begin{aligned} E(\mathbf{h},J) = \quad & \frac{\|\mathbf{m}^{\text{exact}}(\mathbf{h},J)-\mathbf{m}^{\text{sampled}}\|_2^2}{\|\mathbf{m}^{\text{sampled}}\|_2^2} \\ & + \frac{\|\mathbf{C}_{ij}^{\text{exact}}(\mathbf{h},J)-\mathbf{C}_{ij}^{\text{sampled}}\|_2^2}{\|\mathbf{C}_{ij}^{\text{sampled}}\|_2^2} \\ & + \frac{\|\mathbf{C}_{ijk}^{\text{exact}}(\mathbf{h},J)-\mathbf{C}_{ijk}^{\text{sampled}}\|_2^2}{\|\mathbf{C}_{ijk}^{\text{sampled}}\|_2^2}, \end{aligned} \tag{B.1}$$

and its mean-field approximation

$$\begin{aligned} E_{\text{MF}}(J) = \quad & \frac{\|\mathbf{C}_{ij}^{\text{MF}}(J)-\mathbf{C}_{ij}^{\text{sampled}}\|_2^2}{\|\mathbf{C}_{ij}^{\text{sampled}}\|_2^2} \\ & + \frac{\|\mathbf{C}_{ijk}^{\text{MF}}(J)-\mathbf{C}_{ijk}^{\text{sampled}}\|_2^2}{\|\mathbf{C}_{ijk}^{\text{sampled}}\|_2^2}, \end{aligned} \tag{B.2}$$

where the l_2-norm $\|.\|_2$ for the symmetric correlation tensors is defined as sum over the squared independent entries $\|\mathbf{X}_{ij}\|_2^2 = \sum_{i<j} X_{ij}^2$ and $\|\mathbf{X}_{ijk}\|_2^2 = \sum_{i<j<k} X_{ijk}^2$ and the indices exact and MF denote the exact self-consistent equations averaged over the samples, and the explicit mean-field expressions for the magnetisations (A.1) and correlations (A.5),(A.9) for sequential dynamics and (A.12),(A.9) for parallel dynamics.

These cost functions are non-negative functions of the model parameters, they are zero when the connected two- and three-point correlations (and magnetisations for the exact inference) exactly match the empirically measured correlations (and magnetisations). The mean-field cost function only depends on the coupling matrix J since the reconstruction of the fields is independent from the coupling inference, and is done by solving the magnetisation equation (A.1) with the reconstructed couplings. The cost functions (B.1) and (B.2) can be viewed as formal energies to be minimised by some algorithm.

Our inference problem can now be restated as finding

$$(\mathbf{h}^*, J^*)_{\text{exact}} = \underset{\mathbf{h},J}{\operatorname{argmin}}\, E(\mathbf{h}, J) \tag{B.3}$$

for the exact inference, or

$$J^*_{\text{MF}} = \underset{J}{\operatorname{argmin}}\, E_{\text{MF}}(J)\ ,\ \mathbf{h}^*_{\text{MF}} = \mathbf{m}^{-1}(J^*_{\text{MF}}) \tag{B.4}$$

for the mean-field inference.

Since this requires solving a system of nonlinear equations, this cannot be done analytically. We use the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) as implemented in the Python library SciPy (Jones et al., 2001–) as a numerical solver. We find that the energy landscapes exhibit many local minima. For that reason it is not sufficient to use a single starting point for the solver. Instead we use 100 random starting points centred around the naive mean field estimate of external fields and symmetric couplings (which can be computed analytically). For each of these starting points, the Levenberg-Marquardt algorithm finds a local minimum and of these candidates we choose the one with the lowest energy.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen - noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Johannes Berg betreut worden.

Köln, August 2017

Simon Lee Dettmer

Teilpublikationen:

1. S. L. Dettmer and J. Berg. "Inferring the parameters of a Markov process from snapshots of the steady state". preprint `https://arxiv.org/abs/1707.04114v1` (2017)

2. S. L. Dettmer, H. C. Nguyen, and J. Berg. "Network inference in the nonequilibrium steady state". Phys. Rev. E 94, 052116 (2016)