Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Physics (36^{th}cycle)

Epidemic Inference from a Statistical Physics viewpoint

By

Matteo Mariani

******

Supervisor(s):
Prof. Alfredo Braunstein

Doctoral Examination Committee:
Alejandro Lage-Castellanos, Referee;
Pierfrancesco Urbani, Referee;
Andrea Gamba;
Carlo Lucibello;
Andrea Pagnani.

Politecnico di Torino
2024

A mio zio Massimo.

Acknowledgements

I acknowledge my supervisor, Alfredo Braunstein, for his constant and thorough support. A great thanks goes also to Luca Dall'Asta, who gave me many insights during the entire PhD course.
The discussions with Stefano Crotti, Andrea Pagnani, Andrea Gamba, Elisa Floris, Luca Sesta, Tommaso Barberis, Anna Paola Muntoni, Giovanni Catania, Luise Boudzynski have been extremely beneficial.
A special thanks goes to my family, my friends and to Irene. All of them have constantly motivated me during this period of my life.

Abstract

This manuscript deals with the study of epidemic inference in the framework of Statistical Physics. Epidemics are treated as stochastic processes on graphs. The inference task consists in a probabilistic reconstruction of a specific epidemic cascade (so called *planted*) using partial and noisy knowledge of the contact network and of individuals' infection state. The reconstruction process is reframed in this thesis as the computation of observables over a high-dimensional probability distribution, known as the *posterior*. In the Introduction, connections are drawn between posterior computation and Statistical Physics. Specifically, a parallel is portrayed between inference and spin glass theory. Special attention is given to the Nishimori conditions, which play a central role in both Spin Glass theory and (epidemic) inference. The computation of the epidemic posterior marginals is shown to be an NP-hard problem (as shown in the manuscript). Thus, some approximate methods are required. The Causal Variational Approach is introduced for this purpose. It allows sampling without rejection from a distribution which approximates the posterior. This method surpasses previously existing machine learning-based techniques, as well as some Mean-Field approximations, in terms of accuracy. An attempt to characterize the difficulty of inference tasks involves computing theoretical bounds on algorithmic performance as functions of epidemic parameters. This is the objective of *Epidemble*, introduced in Chapter 3 of this manuscript. Epidemble (Epidemic Ensemble) is a semi-analytical tool based on the Replica Symmetric Cavity Method. This technique allows to compute, in the limit of large-sized graphs, what a perfect (exact) algorithm would find. In particular, Epidemble finds the values of statistical estimators (e.g., Area Under the ROC, Minimum Mean Squared Error, Maximum Mean Overlap) as functions of epidemic parameters such as infection rate, patient zero density, and the quantity and quality of clinical tests. These results are

provided in the form of phase diagrams which can be interpreted as upper bounds to real inference alogrithms' performances.

Contents

Chapter 1

Introduction

Epidemics are hard-to-predict natural events characterized by an infectious disease which transmits from an individual to another. Epidemics can cause a wide range of damages, affecting individuals, communities, and entire societies. The most immediate and tragic consequence is the loss of human lives, but they can also overwhelm healthcare systems, leading to shortages of medical personnel and hospital beds. Economically, there can be dramatic consequences too. Businesses may suffer due to decreased productivity, disrupted supply chains, and lower consumer confidence. Also tourism is strongly damaged as epidemics may result in travel restrictions. The stress, fear, and uncertainty can finally have significant psychological and mental health consequences. In our interconnected society, diseases can easily cross borders. Studying epidemic prevention is therefore crucial for the stability and well being of public health systems, economy and individuals in general. The study of epidemics involves a multidisciplinary approach, combining various scientific methods to understand the causes, the spread, and the impacts of infectious diseases. An example is the genomic analysis, namely the sequencing and interpretation of the genetic material of pathogens responsible for the disease. Understanding the genetic makeup of a pathogen aids in the development of vaccines that are effective against diverse strains of the pathogen and which proved to be of extreme importance in containing the COVID-19 pandemic. Mathematical models, also, play a crucial role. They focus on how infectious diseases spread within a population, trying to understand which are the best solutions to mitigate epidemic outbreaks. One example is compartmental

modeling, for which the population is divided into several compartments based on the individuals' disease status. The classic example is the SIR model, which includes compartments for Susceptible ($\mathcal{S}$), Infectious ($\mathcal{I}$), and Recovered ($\mathcal{R}$) individuals. Agent-based models, instead, simulate the interactions of individual agents (people) within a population, providing a rather detailed representation of the epidemic process. Each agent follows some precise rules: for example, an $\mathcal{I}$ individual can infect with certain infection probability its $\mathcal{S}$ contacts. An advantage of mathematical modeling is that the results are typically not tailored on a specific disease and some results may apply to a large category of epidemics; models and techniques developed during the COVID-19 pandemic might be used in the future for containing other diseases. Mathematical approaches to epidemiology form a vast field. In this thesis, we explore epidemics transmitted from person to person, distinguishing them from diseases (as Dengue or Salmonella), reliant on external elements like water, food, or mosquitoes for transmission. In the realm of epidemics among humans, ranging from COVID-19 to influenza and HIV, the literature is immense. The choice of epidemic modeling and inference approaches depends on the research goal, as described in [1], where more and more structured models of epidemic propagation are progressively introduced. There is the homogeneous hypothesis, in which contacts among individuals have the same probability to happen. A step in refinement is represented by mixed heterogeneous-homogeneous models, for which the population is separated in groups of individuals based on some common features, as the age or the job. Individuals inside the same group have high probability to interact, while contacts among individuals belonging to different groups are more sparse. Finally, epidemic modeling on networks can be introduced, for which each individual has a specific set of contacts. The approach chosen (homogeneous, mixed heterogeneous/homogeneous, fully heterogeneous) to describe the epidemic process depends on the aim of the study and on the available information: for example, if one has low information on the single individual's states and is interested in finding macroscopic laws (as the number of infectious individuals in time), then it may be sufficient to use a homogeneous model, which typically has the advantage of being analytically tractable. The fully heterogeneous models are much harder to deal with, but they allow to develop a more detailed description of the epidemic propagation [2, 3]. Not only models, but also inference techniques depend on the aim of the

specific research path. A guide on some of the most used inference methods is in [4]. In this thesis we want to study epidemics under the hypothesis of having access to single-individuals information, as the contact network and the clinical tests. The research direction we address is to study how observations can be used to extract information on the epidemic cascade which generated them. Understanding this point can have practical applications in the field of automatic contact tracing: being able to reconstruct the epidemic cascade i.e. identifying the patient zero and the infection events) could help some targeted testing on the individuals having the highest risk of being infected by the disease. This would help to optimally use the limited number of available clinical tests. Another possible application of this study is in policy making: interpreting the available information from clinical tests can help orienting decisions of governments. Unfortunately, reconstructing the epidemic history from clinical tests is mathematically challenging (NP-hard), as shown in section 1.3. Physics, however, can help developing good approximations! We are going to review two research paths: developing new algorithms and studying information-theoretic performance bounds. Chapter 2 is devoted to the first aim: some of the most used algorithms are reviewed, their limitations are discussed and possible future developments are presented. Algorithms aim to assign to each individual an estimate of their marginal risk of being infectious, playing a crucial role in possible mitigation of epidemic outbreaks. However, they are not enough; we also need to understand whether the available information (contained e.g. in clinical tests) is sufficient to reconstruct the epidemic process. Chapter 3 is devoted to this study. The remaining part of the introduction deals with the characterization of the epidemic inference problem: the general framework is introduced, the hardness of the problem is quantified and the connections with physics and Bayesian inference are portrayed.

1.1 Framework of epidemic inference

In this paragraph, we will describe the general framework of epidemic inference adopted in this thesis. The hypothesis is that the disease (virus, bacterium, parasite, etc.) spreads throughout the population via a network of contacts among individuals. We model each contact between two individuals, denoted

as i and j, who meet for a certain time interval from t_{start} to t_{end}, with a link represented by the quadruplet $(i, j, t_{\text{start}}, t_{\text{end}})$. The set of all links among individuals forms the dynamic contact network through which the disease spreads. Since networks play a crucial role in the study of epidemic inference, it is worth introducing some notions of graph theory here.

1.1.1 Graphs

A graph $G = (V, \mathcal{E})$ is a mathematical structure that consists of a set V of vertices (also called nodes) and a set $\mathcal{E}$ of edges (also called links or arcs). These edges connect pairs of vertices, representing contacts between them. There are two main types of graphs: directed graphs and undirected graphs. In an undirected graph, the edges have no direction. This means that any edge between two vertices represents a bidirectional connection. In a directed graph, instead each edge has a direction. Namely, if there is an edge from vertex A to vertex B, it does not necessarily mean there is an edge from B to A. Both directed and undirected graphs are used in epidemic study. Graphs can be used to model various real-world connections. In the epidemic context, they are used to model the interaction among individuals. An important property of graphs is related to the presence of cycles. A cycle is a path along the edge set which starts and ends on the same vertex. Some graphs have cycles (sometimes called *loops*), which means that it is possible to start from a vertex A, walk along nodes connected by edges and end up again on A. Some other graphs, instead, have no loops and they are called *acyclic*. Based on their properties, graphs are categorized in several *graph ensembles* or families. Here we give three important examples.

- A tree is an acyclic and connected graph. In simpler terms, it's a collection of vertices and edges where there is exactly one path between any two vertices.
- In a random regular graph (RRG) each vertex has the same *degree*. The degree of a vertex is the number of edges incident to that vertex. So in a RRG each vertex has the same number of edges attached to it.
- The Erdős–Rényi (ER) family of graphs $G(n, p)$ is a set of graphs with n vertices and a random (binomial) number of edges. Each of the possible

$\binom{N}{2}$ edges is in fact assigned to the graph with probability p. On average, therefore, there are $p\frac{N(N-1)}{2}$ edges on a ER graph.

These graphs models are useful concepts to understand the properties of typical networks that emerge from random processes, although they may not capture the structural features found in more complex (realistic) contexts. In Chapter 2 we are going to mention two more specific models for epidemic contact networks. One final yet important definition in graph theory is the one of neighbors (or contacts) of a vertex. A neighbor $j \in V$ of $i \in V$ is a vertex which is connected with an edge $(i, j) \in \mathcal{E}$ to i. The set of all the neighbors of i is called $\partial i = \{j \in V : (i, j) \in \mathcal{E}\}$.

Graphs in epidemic inference We use graphs to model contacts among individuals. Each individual i is in fact associated to a vertex and each contact (i, j) between two individuals i and j is modeled with an edge. In epidemic inference graphs can be directed or undirected. In fact, if individual i wears a surgical mask and individual j wears no face mask, than the contact is protected for j but not for i. It is good in that case to model this contact with an directional edge. Sometimes the hypothesis of symmetrical and time-independent links is made. The epidemic disease spreads throughout the population by jumping among linked individuals. The epidemic inference problem aims at reconstructing the infection chain using the information contained in the clinical tests collected by the hospitals, pharmacies, etc... Even if we exactly knew the contact network, the epidemic inference problem would still be extremely hard to solve. In section 1.3 we are going to show that the problem is in fact NP-hard, even when the contact network is known.

1.1.2 SI model and some generalizations

To quantitatively dive into the description of the framework, we shall now introduce models for epidemic propagation. At this point we are going to introduce the discrete time SI model , the simplest one. Its generalizations are postponed to Chapter 2.

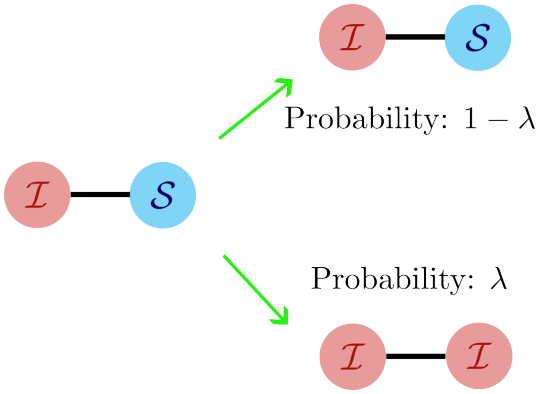

Fig. 1.1 The infection mechanism in the SI model. The circles represent individuals and the black line is a link.

Discrete-time Susceptible-Infectious (SI) model. The simplest way to model the epidemic spread over a population is to use a two-state variables model, in which the individual can either be Susceptible ($\mathcal{S}$) or Infectious ($\mathcal{I}$). Let us consider N individuals interacting at discrete times on a contact graph $\mathcal{G}$. We define $x_i^t \in \{\mathcal{S}, \mathcal{I}\}$ the *state* of an individual $i \in \{1, \dots, N\}$ at time $t = 0, \dots, T$, where T is the total number of time steps, sometimes called *horizon time*. We call patient zero every individual who is infectious at time 0. The set of all the patients zero is therefore: $\{i : x_i^0 = \mathcal{I} \quad \forall i = 1, \dots, N\}$. We define γ the probability for an individual to be the patient zero, independently of the others' state. As a consequence, there are on average $N\gamma$ patients zero. Typically, $\gamma \sim \frac{1}{N}$. The idea is that we want to study a generic case of epidemic spread in a sub-global population, for which it is reasonable to have more than one patient zero. We now model the infection process: when an individual i in the state $\mathcal{I}$ has an isolated contact with an individual j in the state $\mathcal{S}$, this will change its state from $\mathcal{S}$ to $\mathcal{I}$ with probability λ, as in Figure 1.1. In the SI model it is not possible to recover: the transition $\mathcal{I} \to \mathcal{S}$ is forbidden. The infection dynamics described so far is Markov. Defining $x = \{x_i^t, \forall i = 1, \dots, N, \forall t = 1, \dots, T\}$, we thus have:

$$P(x) = P^0(x^0) \prod_{t=0}^{T-1} P^{t+1}(x^{t+1}|x^t),$$

where naturally $x^t = \{x_i^t, \forall t = 0, \dots, T\}$. The patient zero probability, as defined above, implies that each individual is the patient zero with probability

γ independently of the others:

$$P^0(x^0) = \prod_{i=1}^{N} p^0(x_i^0)$$

where simply $p^0(x_i^0) = \gamma\delta_{x_i^0,\mathcal{I}} + (1-\gamma)\,\delta_{x_i^0,\mathcal{S}}$. For what concerns the infection process, each individual i changes state only due to its contacts. Let us define the set $\partial i(t) = \{j = 1, \ldots, N : j \text{ is a contact of } i \text{ at time } t\ \}$, which is simply the set of all the contacts (or *neighbors* in graph theory) of i at time t. In the infection dynamics, only the neighbors[1] of an individual are responsible for its infection. This has two implications:

1. The states of two individuals at time $t+1$ are independent, conditioned to the past:
$$P^{t+1}(x^{t+1}|x^t) = \prod_{i=1}^{N} p^{t+1}(x_i^{t+1}|x^t)$$

2. The $t+1$ state of an individual only depends on the state at time t of itself and its contacts.

$$p^{t+1}(x_i^{t+1}|x^t) = p^{t+1}(x_i^{t+1}|x_i^t, x_{\partial i(t)}^t)$$

To compute the transition functions $p^{t+1}(x_i^{t+1}|x_i^t, x_{\partial i(t)}^t)$ we observe that, if the individual remains $\mathcal{S}$ at time $t+1$ it means that, among its $\mathcal{I}$ contacts, nobody was able to infect it:

$$P(x_i^{t+1} = \mathcal{S}|x_i^t = \mathcal{S}, x_{\partial i(t)}^t) = \prod_{j\in\partial i(t)} (1 - \lambda\delta_{x_j^t,\mathcal{I}}) \tag{1.1}$$

by normalization, it is possible to find the transition $\mathcal{S} \to \mathcal{I}$. The process is therefore:

$$P(x) = \prod_{i=1}^{N} \left(p^0(x_i^0) \prod_{t=0}^{T-1} p(x_i^{t+1}|x_i^t, x_{\partial i(t)}^t) \right). \tag{1.2}$$

There are generalizations to the SI model, which we quickly overview here, postponing their precise treatment to Chapter 2.

- The Susceptible Infectious Recovered (SIR) model, allows for recovery by introducing the $\mathcal{R}$ (Recovered/Removed) state. In this model an $\mathcal{I}$

[1] actually, only the $\mathcal{I}$ neighbors, but it is not necessary now to specify.

individual becomes $\mathcal{R}$ after a random amount of time, which depends on a recovery probability, typically indicated with letter μ. In the SIR model, an $\mathcal{R}$ individual never gets back to S or I state, so no one can ever be infected twice.

- The SIS model, which is, like the SI, a two-state variable model. Here, however, the $\mathcal{I}$ individual recovers after a random time to the state $\mathcal{S}$.
- The SEIR model. It includes the E (Exposed) state, which simply describes the case in which an individual has just been infected, but it is still not infectious i.e. it can not infect other individuals.
- Continuous-time models. Every model defined above can be generalized to its continuous time correspondent.
- Many other model generalizations are possible: the pre-symptomatic state, the vaccinated state, etc... may be introduced. In this thesis, however, we are trying to find some fundamental and general results for epidemic inference. The aim is to identify which features impact the most when doing inference. We will see that these are the network topology, the value of the infection rate and the way in which clinical tests are collected. Some other factors, as the number of states included in the model, do not seem to impact dramatically on the performance, as shown in section 2.4.3. For this reason, our treatment will focus on the simplest models.

Whichever model is chosen for describing an epidemic process, we will call $x = \{x_i^t\}_{i=1,\dots,N}^{t=0,\dots,T}$ the whole epidemic trajectory. Each x_i^t is the epidemic state of the individual $i \in \{1, \dots, N\}$ (with N being the total number of individuals) at time t. If we choose the SI model, then $x_i^t \in \{\mathcal{S}, \mathcal{I}\}$; if we choose the SEIR model, then $x_i^t \in \{\mathcal{S}, E, \mathcal{I}, \mathcal{R}\}$.

1.1.3 A Bayesian formulation: prior and posterior

Once the model is fixed, it is possible to assign a probability to each epidemic trajectory x. This probability is usually called *prior distribution* or more simply *prior.*

$$P(x) = \text{prior distribution}$$

The prior is the probability that a specific epidemic cascade takes place. It depends on the choice of the model and on the contact network. For the SI model the prior is given in equation (1.2). A general formulation of the prior is given later in section 2.1.4. Clinical tests, also called observations, are sources of information that we want to use to reconstruct the epidemic cascade. They carry information about the epidemic state of the tested individual at the specific time in which the individual is tested. Each observation o can be encoded in a quadruplet $o = (o_i, o_t, o_s, o_f)$, where o_i is the tested individual, o_t is the time at which the test is made, o_s is the outcome, o_f is the error rate of the test. The clinical tests, in fact, are in general non perfect. They have a false negative and a false positive rate. For sake of simplicity we are going to assume that the two error rates are equal to each others and both equal to o_f. The set of all the clinical tests collected is the set $\mathcal{O}$. Let us therefore define the *posterior distribution* or *posterior* as:

$$\mathcal{P}(x|\mathcal{O}) = \frac{P(x)P(\mathcal{O}|x)}{P(\mathcal{O})} \tag{1.3}$$

in which the r.h.s. is the Bayes formula. The posterior function assigns a probability to each epidemic cascade x conditioned to the observations $\mathcal{O}$. Let us analyze the three terms of the r.h.s. of the posterior distribution:

- $P(x)$: It is the prior distribution. As shown later, section 2.1.4, the prior can be rewritten as a product of local factors (i.e. each factor depending on an individual and its neighbors).
- $P(\mathcal{O}|x)$: is the probability of observing $\mathcal{O}$ given that the epidemic realization is x. It is called *likelihood*. This is a function that measures the compatibility between observations and the considered epidemic trajectory. Evaluating the likelihood is typically easy. In fact, the distribution $P(\mathcal{O}|x)$ factorizes over the observations:

$$P(\mathcal{O}|x) = \prod_{o \in O} p(o|x)$$

this is because the several observations are independent of each others if the underlying epidemic state from which they are taken is known [2]. Moreover, each observation $o = (o_i, o_t, o_s, o_f)$ only depends on the state of the individual i at the testing time t:

$$P(\mathcal{O}|x) = \prod_{o=(o_i,o_t,o_s,o_f)\in O} p(o|x_i^t)$$

The exact value of each $p(o|x_i(t))$ is also known: indeed, if the false rate is zero, this function is simply 1 when the observed state s is equal to the actual epidemic state $x_i(t)$ and 0 when they differ. Otherwise, if the false rate is nonzero, then:

$$p(o = (o_i, o_t, o_s, o_f)|x_i^t) = \begin{cases} 1 - o_f & \text{if } x_i^t = s \\ f & \text{if } x_i^t \neq s \end{cases}$$

- $P(\mathcal{O})$ is simply the normalization term due to the fact that the posterior is a probability. Thus:

$$P(\mathcal{O}) = \sum_x P(x)P(\mathcal{O}|x).$$

If we consider a continuous-time model, then the sum becomes an integral. This normalization factor is the reason why making inference is difficult. It is in fact the sum over all possible epidemic cascades. This means that this sum has in general a number of addends which grows exponentially with the number of individuals.

Typically, the prior distribution and the likelihood are easy to compute, i.e. there is a polynomial algorithm which is able to assign for each configuration x and observations set $\mathcal{O}$ their likelihood $P(\mathcal{O}|x)$ and prior $P(x)$ values. Computing the posterior, instead, is an NP-hard problem, as discussed in section 1.3. Approximations are therefore needed to address the computation and the marginalization of the posterior.

[2] Actually, this is an approximation. For example, one hospital might receive a batch of flawed clinical tests. In that case the results among individuals which are tested in that same hospital would be correlated. We are going to ignore this possibility.

1.1.4 Inference process

Having introduced the prior, the likelihood and the posterior distributions allows us to describe the inference process now. The goal is to reconstruct an unknown epidemic cascade x^*, called *planted*, which is supposed to be a fair sample of the prior distribution. The ideal goal of epidemic inference is to use the available information, in the form of the observations set $\mathcal{O}$, to guess x^*. Observations are assumed to be sampled from the likelihood. It is not possible, in general, to exactly reconstruct the precise epidemic cascade x^* from $\mathcal{O}$: this is because many epidemic realizations are compatible with a particular set $\mathcal{O}$ of observations. As a consequence, the best possible result inference can provide is formulated by means of probabilities. Epidemic inference, thus, aims at assigning a probability $\mathcal{P}(x|\mathcal{O})$ to each possible epidemic realization x, conditioned to the observations $\mathcal{O}$. Perfect reconstruction is achieved when $\mathcal{P}(x|\mathcal{O})$ is peaked around x^*. However, for realistic scenarios, the distribution might not be such. Nonetheless, studying posterior marginals, i.e. the posterior probability for each individual to be infectious at certain times, can give huge insights on the epidemic cascade and can open to possible studies of policies to adopt in order to contain the epidemic process.

1.1.5 Evaluating Performance: estimators computation

To quantify the predicting power of the posterior, we now define some estimators. We start by defining $p_{i,t}(x_i^{*,t}, x_i^t|\mathcal{O})$ as the marginal probability of having the planted state $x_i^{*,t} \in \{0,1\}$ and the inferred state $x_i^t \in \{0,1\}$ of one individual $i = 1, \ldots, N$ at a given time $t \in \{0, 1, \ldots, T\}$. For simplicity, $x_i^t = 0$ if the state of individual i at time t is $\mathcal{S}$ and $x_i^t = 1$ if the state is $\mathcal{I}$. With this definition we can introduce some estimators.

Maximum Mean Overlap (MMO) At a given time t, the overlap between the planted configuration $x^{*,t}$ and a configuration $\hat{x}^t$ is calculated as follows:

$$O_t(x^{*,t}, \hat{x}^t) = \frac{1}{N} \sum_{i=1}^{N} \delta_{x_i^{*,t}, \hat{x}_i^t}$$

During the inference process, the planted configuration is unknown. The best Bayesian estimator is obtained by assuming that $\underline{x}^{*,t}$ follows the posterior distribution. The best we can do, therefore, is to compute the mean overlap over the posterior:

$$\mathrm{MO}_t(\hat{x}^t) = \sum_{\underline{x}^t} P(x^t|\mathcal{O}) O_t(x^t, \hat{x}^t)$$

We define the Maximum Mean Overlap estimator as:

$$\hat{x}^{t,\mathrm{MMO}} = \mathrm{argmax}_{x^t} \mathrm{MO}_t(\hat{x}^t)$$

which is achieved setting:

$$\hat{x_i}^{t,\mathrm{MMO}} = \mathrm{argmax}_{x_i^t} p_{i,t}(x_i^t|\mathcal{O})$$

The overlap $O_t(x^{*,t}, \hat{x}^{t,\mathrm{MMO}})$ provides a quantitative estimation of the accuracy of the Maximum Mean Overlap estimator.

Minimum Mean Squared Error (MMSE) The squared error (SE) at a given time t between the planted configuration $x^{t,*}$ and an estimator $\hat{x}^t$ is:

$$\mathrm{SE}_t(x^{*,t}, \hat{x}^t) = \frac{1}{N} \sum_{i=1}^{N} (x_i^{*,t} - \hat{x}_i)^2$$

Similarly to the overlap, the best Bayesian estimator for the squared error $\hat{x}^{t,\mathrm{MMSE}}$ is the one that minimizes the squared error averaged over the posterior:

$$\mathrm{MSE}_t(\hat{x}^t) = \sum_{x^t} P(x^t|\mathcal{O}) \mathrm{SE}_t(x^t, \hat{x}^t)$$

This is achieved by setting

$$\hat{x}_i^{t,\mathrm{MMSE}} = \sum_{x_i^t} p_{i,t}(x_i^t|\mathcal{O}) x_i^t.$$

Notice that this is not necessarily a discrete value in $\{0, 1\}$.

Area Under the Curve (AUC) The Area Under the Curve (AUC) is defined as the area under the Receiver Operating Characteristic (ROC) curve [5], which is computed as follows. At a fixed time t, the marginal probability $p_{i,t}(x_i^t = \mathcal{I}|\mathcal{O})$ is computed for each individual. For a given threshold $\rho \in [0, 1]$, the true positive rate TPR(ρ) (respectively the false positive rate FPR(ρ)) is the fraction of positive (respectively negative) individuals i in the planted with $p_{i,t}(x_i^t = \mathcal{I}|\mathcal{O}) \geq \rho$. The ROC is the parametric plot of TPR(ρ) versus FPR(ρ), with ρ as the varying parameter. These three estimators will be used in the following to quantify the performance of the algorithms studied.

1.1.6 Algorithms and thresholds

In Chapter 2, algorithms for epidemic inference are described. Their aim is to approximate the marginal distributions of the posterior $\mathcal{P}(x|\mathcal{O})$. Approximations are needed because computing the exact form of $\mathcal{P}(x|\mathcal{O})$ and its marginals is unfeasible with a computer. The number of operations scales exponentially with the number of individuals. Approximations allow to build algorithms which are polynomial in time w.r.t. the total number of individuals, making it possible to perform inference. Therefore, algorithms are a crucial tool for epidemic containment. However, developing algorithms alone is not enough. It is in fact necessary to understand if the information provided is sufficient. For example, trying to reconstruct an epidemic cascade for a population of one million individuals with only one clinical test is impossible due to lack of information. In other words, there exists an *intrinsic* ignorance due to the limited available information $\mathcal{O}$. In Chapter 3, a method for quantifying such ignorance is introduced and theoretical bounds for inference are discussed.

1.1.7 Inferring the prior distribution

So far, we have described the epidemic inference problem as the process of computing (or at least approximating) the posterior distribution $\mathcal{P}(x|\mathcal{O})$, assuming to know the prior distribution. Actually, this assumption is quite heavy. The prior in fact depends on:

1. The epidemic model: the model fixes the dynamics of the epidemic. For example, in the SI model there exist only two states and the transition $\mathcal{I} \to \mathcal{S}$ is forbidden, while in the SIS model it is allowed. Even the number of states changes depending on the model. The prior probability distribution therefore can only be written once the model is fixed.

2. The contact network: the probability of a particular epidemic cascade is strongly affected by the network structure. An isolated node, for example, is infectious with very low probability (exactly the probability of being the patient zero).

3. The so called *hyper-parameters* of the model: even when the model is fixed, there are some free parameters that describe the patient zero distribution, the infection rate, the recovery rate etc... which drastically change the prior distribution.

Assuming to exactly know the prior, therefore, means to assume the knowledge of the model, the contact network and the hyper-parameters. Nonetheless, inferring the prior from the set of observations $\mathcal{O}$ is a titanic goal. Typically, the epidemic model is *chosen*, trying to capture the fundamental features of the specific epidemic one is interested in studying. For example, if we want to describe a rapid outbreak in a very short time window, an SEI model might be sufficient because the recovery events would be so rare to impact in a negligible way. The choice of the model is therefore gauged according to the epidemic studied. Inferring the contact network, instead, is a very hard task. A typical assumption in fact is to know the network. This hypothesis is quite strong and indeed one important research path to explore is to study performance bounds when decreasing the number of known links in the network, but this topic is not treated in this thesis. For what concerns hyper-parameters, it is possible to infer them from the observations $\mathcal{O}$. Many methods in literature are implemented to reconstruct hyper-parameters. Of course, as for the posterior computation, inference of hyper-parameters is not an easy task and can be affected by lack of information. The procedure for inferring the hyper-parameters is sketched in the next paragraph, where connections with Statistical Physics are described.

1.2 Connections with Statistical Physics

The epidemic inference problem has a huge overlap with Statistical Physics. Let us consider equation (1.3):

$$\mathcal{P}(x|\mathcal{O}) = \frac{P(x)P(\mathcal{O}|x)}{\sum_{x'} P(x')P(\mathcal{O}|x')},$$

in which we have substituted the expression of the denominator as the normalization of the numerator. Defining:

$$H(x, \mathcal{O}) = -\log\left(P(x)P(\mathcal{O}|x)\right), \tag{1.4}$$

the posterior becomes:

$$\mathcal{P}(x|\mathcal{O}) = \frac{e^{-H(x,\mathcal{O})}}{\sum_{x'} e^{-H(x',\mathcal{O})}} = \frac{e^{-H(x,\mathcal{O})}}{Z}. \tag{1.5}$$

This is the form of the canonical probability distribution in statistical physics. Continuing this parallelism, we can also define the free energy, a fundamental quantity in Physics which also plays a key role in inference:

$$F = -\log Z = -\log P(\mathcal{O}). \tag{1.6}$$

The free energy quantifies the information carried by the observations. In fact, if the observations are very selective, then the expression $\sum_{x'} P(x')P(\mathcal{O}|x')$ is small because many terms of $P(\mathcal{O}|x')$ in the sum are small. As a consequence $F = -\log P(\mathcal{O})$ is big. On the other hand, if e.g. no observation is made, then $F = -\log\left(\sum_x P(x)\right) = 0$. We will come back to the interpretation of F in Chapters 2 and 3. For now it is sufficient to stress that the posterior can be re-written as a canonical probability distribution. This allows to import some general methods like mean field [6], Monte Carlo [7], message passing [8, 6] algorithms or variational methods [9, 10] to approximate the posterior distribution. Each method introduces a different approximation with different effects in terms of performance and speed of the algorithm. A more precise treatment of algorithms is given in Chapter 2.

1.2.1 Descending free energy to infer the prior

Typically, the prior and the likelihood distributions depends on some hyper-parameters, as discussed in paragraph 1.1.7. These hyper-parameters are in general not known and must be inferred. We call θ the set of all the hyper-parameters of the prior distribution. The free energy, therefore, becomes:

$$F(\theta) = -\log \sum_x P_\theta(x) P_\theta(\mathcal{O}|x).$$

Inferring hyper-parameters by minimizing F is a typical procedure in practical problems, which is sometimes called *maximization of evidence* [11, 12] or *maximization of type II likelihood* [13, 14]. This intuitively corresponds to find the set of parameters θ such that the observations $\mathcal{O}$ collected are the most typical possible. Indeed, low free energy means that the configuration space of trajectories compatible with the observations set has a high probability. In other words, minimizing $F(\theta)$ means maximizing $P(\mathcal{O}|\theta)$ with respect to θ, namely to find the $\tilde{\theta}$ such that $\mathcal{O}$ has the highest possible probability to be collected.

1.2.2 Planted problems

The canonical distribution formalism of epidemic inference in (1.5) is actually common to many other problems, as: time series analysis [15], machine learning [16–19, 12, 20], error correcting code theory [21, 22, 17, 23, 24], force field inference [25–28]. In all these problems the typical aim is to reconstruct an original configuration x^*, usually denoted as the *planted configuration* or simply *planted*. For example, for the error correcting codes, x^* is a message sent in input through a noisy channel. The corresponding (perturbed) output is $\mathcal{O}$. The aim of the receiver is to guess x^* by computing the posterior distribution $\mathcal{P}(x|\mathcal{O})$. In this context the output $\mathcal{O}$ plays the same role of the observations $\mathcal{O}$ in epidemic inference. In order to compute the posterior, we suppose the receiver to know both the coding scheme and the stochastic distortion process played by the noise in the channel, i.e. the likelihood $P(\mathcal{O}|x)$. The posterior is thus always:

$$\mathcal{P}(x|\mathcal{O}) = \frac{P(x)P(\mathcal{O}|x)}{\sum_{x'} P(x')P(\mathcal{O}|x')},$$

which is identical to (1.3). In general, it is called *Bayesian Inference* the process of reconstructing, by means of the posterior computation, a planted configuration x^* generated from a prior distribution $P(x^*)$, relying on an observation set $\mathcal{O}$ which is correlated to the planted by a likelihood $P(\mathcal{O}|x^*)$.

1.2.3 Bayes optimality and Nishimori conditions

Let us suppose to have some observations $\mathcal{O}$ of the hidden planted x^*. We want to give an estimate $\bar{A}$ of some observable $A(x^*)$ which depends on the stochastic trajectory x^*. A reasonable idea is to average $A(x)$ over the posterior:

$$A(x^*) \simeq \langle A(x) \rangle_{x \sim \mathcal{P}(x|\mathcal{O})} = \bar{A}(\mathcal{O}).$$

The estimate $\bar{A}$ is therefore a random variable depending on the observations $\mathcal{O}$ (which in turn depend on the planted). Its average is:

$$\mathbb{E}\left[\bar{A}(\mathcal{O})\right] = \sum_{x^*,\mathcal{O}} P_{\tilde{\theta}}(\mathcal{O}|x^*) P_{\tilde{\theta}}(x^*) \bar{A}(\mathcal{O}),$$

where we explicitly highlighted the dependence of the prior and likelihood on the hyper-parameters set $\tilde{\theta}$. If the prior and the likelihood are known in the inference process, it is said that inference is made under *Bayes optimal conditions* or under *Bayes optimality*. In this case it is possible to write the posterior as:

$$\mathcal{P}_{\tilde{\theta}}(x|\mathcal{O}) = \frac{P_{\tilde{\theta}}(\mathcal{O}|x) P_{\tilde{\theta}}(x)}{P_{\tilde{\theta}}(\mathcal{O})},$$

where of course also the posterior depends on the (known) hyper-parameters of the prior and likelihood. Therefore, we can rewrite the average of the observable

as:

$$\begin{aligned}
\mathbb{E}\left[\bar{A}(\mathcal{O})\right] &= \sum_{x^*,\mathcal{O}} \mathcal{P}_{\tilde{\theta}}(x^*|\mathcal{O}) P_{\tilde{\theta}}(\mathcal{O}) \bar{A}(\mathcal{O}) = \\
&= \sum_{x^*,\mathcal{O}} \mathcal{P}_{\tilde{\theta}}(x^*|\mathcal{O}) P_{\tilde{\theta}}(\mathcal{O}) \sum_x A(x) \mathcal{P}_{\tilde{\theta}}(x|\mathcal{O}) = \\
&= \sum_{x^*,\mathcal{O},x} \mathcal{P}_{\tilde{\theta}}(x^*|\mathcal{O}) P_{\tilde{\theta}}(\mathcal{O}) A(x) \mathcal{P}_{\tilde{\theta}}(x|\mathcal{O}) = \\
&= \sum_{\mathcal{O},x} P_{\tilde{\theta}}(\mathcal{O}) A(x) \mathcal{P}_{\tilde{\theta}}(x|\mathcal{O}) = \\
&= \sum_{\mathcal{O},x} A(x) P_{\tilde{\theta}}(x, \mathcal{O}) = \\
&= \sum_x A(x) P_{\tilde{\theta}}(x) = \\
&= \sum_{x^*} A(x^*) P_{\tilde{\theta}}(x^*),
\end{aligned}$$

where in the second passage we wrote the definition of $\bar{A}(\mathcal{O})$ and in the last passage we simply changed the name to an index. We end up with:

$$\mathbb{E}_{x^* \sim P(x^*)}\left[A(x^*)\right] = \mathbb{E}_{x \sim \mathcal{P}(x|\mathcal{O}), \mathcal{O} \sim P(\mathcal{O}|x^*), x^* \sim P(x^*)}\left[A(x)\right] \tag{1.7}$$

which means that the average of the estimate and the average of the quantity $A(x^*)$ coincide. In other words, the posterior estimate coincides with the planted observable, on average. If we consider a very large system, thus, it may happen that at least for some observables which benefit of the self-averaging property (the so called intensive quantities), the relation (1.7) becomes:

$$A(x^*) \approx A(x),$$

where x^* is sampled from the prior and x is sampled from the corresponding posterior. At the thermodynamic limit the approximation becomes an equality and, therefore, the planted configuration is a fair sample of the posterior. Equation (1.7) is probably the simplest yet very instructive example of *Nishimori condition.* The concept of Nishimori condition is widely diffused in literature [29–34]. It is due to Nishimori [35, 36], who however introduced it as a result in the disordered Ising model. This concept is in fact much more general. The natural extended framework to define Nishimori conditions is indeed in

Bayesian inference and is due to Iba [33]. The idea behind Nishimori conditions is that, when they hold, averages over the posterior coincide with averages over the prior. Note that in the computation above we made the hypothesis of knowing the prior. What happens if we don't? For example, when the correct hyper-parameters $\tilde{\theta}$ are not known we can only resort to some guessed hyper-parameters (which can be randomly chosen of, better, inferred), which we call θ. We have that the posterior consequently depends on θ:

$$\mathcal{P}_\theta(x|\mathcal{O}) = \frac{P_\theta(x)P_\theta(\mathcal{O}|x)}{P(\mathcal{O})}$$

Thus, the computation changes and the fourth passage of the previous computation becomes:

$$\mathbb{E}\left[\bar{A}(\mathcal{O})\right] = \sum_{\mathcal{O},x} P_{\tilde{\theta}}(\mathcal{O})A(x)\mathcal{P}_\theta(x|\mathcal{O})$$

but we can't move further because now the two hyper-parameters θ (with which inference is done) and $\tilde{\theta}$ (from which the planted was sampled) are different. As a consequence, we understand that the Nishimori condition is satisfied only when the true generation parameters are known, which is the Bayes optimality.

1.2.4 Nishimori line in the Ising spin glass

The concept of *Nishimori condition* actually originated from the study of spin glasses [35]. The so called *Nishimori line* was introduced as a line in the Ising spin glass model's phase diagram. It was an extremely relevant discovery because on this line some thermodynamic quantities (as the internal energy) can be exactly computed ([37], Chapter 12. Spin Glasses: Constraints and Frustration, pag 247). The derivation of these properties was obtained by relying on a gauge invariance, which we are going to sketch here. Let us consider an undirected graph $G = (V, \mathcal{E})$. The Ising spin glass model on graph G is defined by the Hamiltonian:

$$H[\underline{J}](\underline{\sigma}) = -\sum_{(i,j)} J_{ij}\sigma_i\sigma_j$$

where:

- each edge on the graph between site $i \in V$ to site $j \in V$ is denoted with a couple $(i, j) \in \mathcal{E}$;
- $\underline{J} = \{J_{ij}\}$ is the set of the random independent couplings, drawn from a bimodal distribution:

$$P(J_{ij} = 1) = 1 - p;$$
$$P(J_{ij} = -1) = p$$

- the spins in the set $\underline{\sigma} = \{\sigma_1, \ldots, \sigma_N\} \in \{-1, +1\}^N$ are the dynamical variables of the model
- N is the total number of sites.

(for an introduction to the ferromagnetic and disordered Ising models we refer the reader to [37] or to [38]). If the following substitutions are made:

$$\sigma_i \to \sigma_i s_i =: \sigma_i^{(\underline{s})} \qquad\qquad J_{ij} \to J_{ij} s_i s_j =: J_{ij}^{(\underline{s})}$$

where $\underline{s} = (s_1, \ldots, s_N) \in \{-1, 1\}^N$, then the Hamiltonian remains unaltered since $\sigma_i^2 = 1 \ \forall \ i = 1, \ldots, N$. This gauge invariance is crucial to derive the properties of the Nishimori line. We now define the Nishimori temperature as:

$$\beta_N = \frac{1}{2} \log \frac{(1-p)}{p} \tag{1.8}$$

where p is defined above as the probability of sampling a -1 link. The reason of this definition is that the thermodynamic computations becomes easy, as we are going to see now. Due to this definition, the probability of having a link J is:

$$\frac{P(J=1)}{P(J=-1)} = \frac{p}{1-p} = e^{-2\beta_N}.$$

Therefore:

$$\frac{P(J)}{P(-J)} = e^{-2\beta_N J}.$$

Since the ratio of the two probabilities is fixed, we can rewrite them as:

$$P(J) = \frac{e^{\beta_N J}}{n}, \quad \text{for } J = \pm 1$$

where n is a normalization, which we fix now:

$$1 = P(J=1) + P(J=-1) = \frac{e^{\beta_N J} + e^{-\beta_N J}}{n} = \frac{2\cosh(\beta_N)}{n}.$$

So we have that the Nishimori condition is equivalent to:

$$P(J|J \in \{-1,1\}) = \frac{e^{\beta_N J}}{2\cosh(\beta_N)}. \tag{1.9}$$

This, used altogether with the gauge invariance, simplifies the computation of the energy, which is defined as:

$$U(\beta) = \mathbb{E}_{\underline{J}} \left\{ \frac{Z_U[\beta, \underline{J}]}{Z[\beta, \underline{J}]} \right\}$$

where the $\mathbb{E}_{\underline{J}}$ stands for the average over $\underline{J}$ and

$$Z_U[\beta, \underline{J}] = \sum_{\underline{\sigma}} H[\underline{J}](\underline{\sigma}) e^{-\beta H[\underline{J}](\underline{\sigma})}$$
$$Z[\beta, \underline{J}] = \sum_{\underline{\sigma}} e^{-\beta H[\underline{J}](\underline{\sigma})}$$

represent the thermal average over $\underline{\sigma}$ in the canonical ensemble. A crucial observation to make here is that both Z_U and Z are invariant after the gauge transformation $J_{ij} \to J_{ij}^{(\underline{s})}$ and $\sigma_i \to \sigma_i^{(\underline{s})}$ defined above, for every $\underline{s} \in \{-1,1\}^N$. Therefore, by summing all over the possible gauges $\underline{s}$ we have:

$$U(\beta) = \frac{1}{2^N} \sum_{\underline{s}} \mathbb{E}_{\underline{J}^{(\underline{s})}} \left\{ \frac{Z_U[\beta, \underline{J}^{(\underline{s})}]}{Z[\beta, \underline{J}^{(\underline{s})}]} \right\}.$$

This quantity is in general extremely hard to compute for a general value of β. However, at the the Nishimori temperature, we have a formula for the probability distribution of the $\underline{J}^{(\underline{s})}$, using eq (1.9) and the gauge relation:

$$P_{\underline{s}}(J_{ij}) = \frac{e^{\beta_N J_{ij} s_i s_j}}{2\cosh(\beta_N)}.$$

Therefore, the quenched average over the $\underline{J}^{(\underline{s})}$ must be performed w.r.t. this measure. Observe that, if we choose no gauge transformation, i.e. $\underline{s} =$

$(1, 1, \ldots, 1)$ then we get back to (1.9). The internal energy is therefore:

$$\begin{aligned} U(\beta_N) &= \sum_{\underline{s}} \sum_{\underline{J}} \prod_{(ij)} \frac{e^{\beta_N J_{ij} s_i s_j}}{2\cosh(\beta_N)} \frac{Z_U[\beta_N, \underline{J}]}{Z[\beta_N, \underline{J}]} = . \\ &= \frac{1}{[2\cosh(\beta_N)]^{|\mathcal{E}|}} \sum_{\underline{J}} \sum_{\underline{s}} e^{\beta_N \sum_{(ij)} J_{ij} s_i s_j} \frac{Z_U[\beta_N, \underline{J}]}{Z[\beta_N, \underline{J}]} = \\ &= \frac{1}{[2\cosh(\beta_N)]^{|\mathcal{E}|}} \sum_{\underline{J}} Z[\beta_N, \underline{J}] \frac{Z_U[\beta_N, \underline{J}]}{Z[\beta_N, \underline{J}]} = \\ &= \frac{1}{[2\cosh(\beta_N)]^{|\mathcal{E}|}} \sum_{\underline{J}} Z_U[\beta_N, \underline{J}], \end{aligned}$$

where $|\mathcal{E}|$ is the number of edges because $\mathcal{E}$ is the edge set. The final formula can be further simplified to:

$$U(\beta_N) = -|\mathcal{E}| \tanh(\beta_N).$$

(see [37]). The origin on the Nishimori condition is the spin glass theory. However, after the discovery of many connections between physics and inference [22, 24, 23, 21], Iba [33] derived the Nishimori result in the more general framework of statistical inference. The result of Iba is of great importance also because it allows to interpret the sum over the gauge configurations $\underline{s}$, which so far is mathematically rigorous, but a bit obscure.

Annealed VS Quenched average Without being exhaustive, it is worth at least to mention the difference between thermal and quenched averages. We refer the interested reader to [39, 40]. In statistical physics of disordered systems, there are typically two sets of variables:

1. The dynamical variables: they are *fast* variables. In the Ising spin glass model, they are the σ variables. The physical meaning of each σ_i is to be the instantaneous magnetization of site i. These variables fluctuate at nonzero temperature due to thermal noise. When we average over these them, in fact, we are doing the so called *thermal average*.

2. The quenched variables are instead the *slow* ones. In the Ising spin glass model, they are the J's. Their physical meaning is the interaction among sites. These interactions may change in time due to e.g. deformation of

the material. However, these changes should be very slow compared to the thermal fluctuations of the dynamical variables.

The idea is that when averaging over σ we are averaging over the fast fluctuations, so that we can consider the J's fixed. Once thermal averages are evaluated, we can averaged again over the set J. The average over the slow fluctuations is called *quenched.*

1.2.5 Planted spin glass

We now come back to the planted problems, with the aim of discussing one very important implication of Nishimori conditions: replica symmetry. To do so, we introduce the planted spin glass. This will allow us to connect the seemingly distant Nishimori conditions in equation (1.7) to the Nishimori temperature in equation (1.9). We will finally move to describe their implications in terms of replica symmetry. First, we introduce the planted spin glass model, following [34]. It is an inference problem where one wants to reconstruct a planted configuration $\underline{\sigma}^* = (\sigma_1^*, \dots, \sigma_N^*) \in \{-1, 1\}^N$ of N binary values, which is randomly uniformly sampled from the set $\{-1, 1\}^N$. The observations provided to reconstruct the planted are in the form of an edge set built following the scheme:

1. Take M random couples (i, j), $i = 1, \dots, N$ and $j = 1, \dots, N$ among the $\binom{N}{2}$ possible couples.
2. For each couple (i, j) take the product $\sigma_i^* \sigma_j^*$ which is 1 if $\sigma_i^* = \sigma_j^*$ and -1 conversely. Flip it with probability $\rho^* \in [0, 1]$. Call J_{ij} this result.
3. Return the graph $G = (V, \mathcal{E})$, where $V = \{1, \dots, N\}$ and $\mathcal{E}$ is the edge set of the M couples extracted in point 1. Return, also, the set $\underline{J} = \{J_{ij}\}_{(i,j)\in\mathcal{E}}$.

The parameter ρ^* is a sort of noise in the observation because it flips the result of the product. The set $\underline{J}$ plays the role of observations set. Following the usual Bayesian approach of (1.5), we have the posterior:

$$\mathcal{P}(\underline{\sigma}|\underline{J}) = \frac{P(\underline{J}|\underline{\sigma})P(\underline{\sigma})}{P(\underline{J})}. \tag{1.10}$$

In this posterior form, note that the noise ρ^* is a hyper-parameter, which we might not know in the inference process. Therefore, we call ρ^* the true hyper-parameter with which the J's were generated and we introduce ρ as the hyper-parameter used during inference. Due to independent choice of the couples (i, j) we have that the likelihood factorizes:

$$\begin{aligned} P(\underline{J}|\underline{\sigma}) &= \prod_{(i,j)} p_{ij}(J_{ij}|\sigma_i\sigma_j) = \\ &= \prod_{(i,j)} \left[\delta\left(J_{ij} - \sigma_i\sigma_j\right)(1-\rho) \; + \delta\left(J_{ij} + \sigma_i\sigma_j\right)\rho\right]. \end{aligned}$$

The prior is uniformly random:

$$P(\underline{\sigma}) = 2^{-N}\mathbb{I}[\underline{\sigma} \in \{-1, 1\}^N],$$

so the posterior is now well defined. To map this model onto the Ising spin glass in the canonical ensemble, we want to rewrite the likelihood in an exponential form. So we simply rephrase ρ as:

$$\begin{aligned} \rho &= \mathcal{N}(\beta)e^{\beta} \\ 1-\rho &= \mathcal{N}(\beta)e^{-\beta} \end{aligned}$$

then the normalization $\mathcal{N}(\beta) = (2\cosh(\beta))^{-1}$, so

$$\rho = e^{\beta}/(2\cosh(\beta)) \tag{1.11}$$

and the single factor of the likelihood becomes:

$$p_{ij}(J_{ij}|\sigma_i\sigma_j) = e^{\beta J_{ij}\sigma_i\sigma_j}$$

so:

$$\mathcal{P}(\underline{\sigma}|\underline{J}) = \frac{\prod_{(i,j)} e^{\beta J_{ij}\sigma_i\sigma_j}}{2^N\left(2\cosh(\beta)\right)^M P(\underline{J})}$$

We can evaluate $P(\underline{J})$ as the sum of the numerator in equation (1.10) :

$$P(\underline{J}) = \frac{Z[\beta, \underline{J}]}{2^N\left(2\cosh(\beta)\right)^M}$$

where $Z[\beta, \underline{J}]$ is the partition function of the Ising spin glass. The posterior is therefore:

$$\mathcal{P}(\underline{\sigma}|\underline{J}) = \frac{\prod_{(i,j)} e^{\beta J_{ij}\sigma_i\sigma_j}}{Z[\beta, \underline{J}]}$$

which would be identical to the canonical probability distribution of an Ising spin glass model, if it wasn't for the fact that the set $\underline{J}$ is correlated with the planted configuration $\underline{\sigma}^*$ and it is not independently sampled. However, if we now perform the following gauge transformation (which does not alter the posterior distribution):

$$J_{ij} \to J_{ij}\sigma_i^*\sigma_j^* = \tilde{J}_{ij} \qquad\qquad \sigma_i \to \sigma_i\sigma_i^* = \tilde{\sigma}_i$$

for all σ's and J's, we end up with the Ising spin glass model. This is because:

- The planted configuration is mapped to the all-spins-up configuration, i.e. $(1, \ldots, 1)$. As a consequence, when we collect the couplings $\tilde{J}_{ij}$ in this gauge we simply multiply $1 \cdot 1$ and flip it with probability ρ^*.
- Each J_{ij} is therefore $+1$ if the result is not flipped and -1 if it is flipped
- Thus, each J_{ij} is i.i.d. and it is $+1$ with probability $1 - \rho$ and -1 with probability ρ.

Now the model, described by:

$$\mathcal{P}(\underline{\tilde{\sigma}}|\underline{\tilde{J}}) = \frac{\prod_{(i,j)} e^{\beta \tilde{J}_{ij}\tilde{\sigma}_i\tilde{\sigma}_j}}{Z[\beta, \underline{\tilde{J}}]}$$

is exactly an Ising spin glass model. The mapping from planted spin glass to Ising spin glass is completed. To resume the steps of this mapping it was necessary to:

1. take a planted configuration;
2. extract the couplings from the planted with a (unknown) noise ρ^*;
3. introduce the noise parameter ρ used for computing the posterior and map it to an inverse temperature β;
4. gauge transform the resultant model.

Now we have again an Ising spin glass model, with the advantage of interpreting the inverse temperature β as a noise hyper-parameter. We can therefore impose Bayes optimality by setting $\rho = \rho^*$. In other words, we study now the case in which we know the correct hyper-parameter ρ^*. In this case we find the corresponding temperature by using (1.11):

$$\rho^* = \frac{e^{\beta}}{e^{\beta} + e^{-\beta}} = \frac{1}{1 + e^{-2\beta}}.$$

So the inverse temperature in the Bayes optimal condition is:

$$\beta = \frac{1}{2} \log \frac{\rho^*}{1 - \rho^*}$$

This is exactly the Nishimori temperature in (1.8). The statistical inference approach, therefore, has led to a natural derivation of the Nishimori temperature. The Nishimori line, thus, can be interpreted as the zone of the phase diagram in which we know the prior and the likelihood distributions. This concludes the bridge that connects the Nishimori conditions in Bayesian inference and the Nishimori line in spin-glass theory. We can now use this bridge to characterize the replica symmetry on the Nishimori line (some deep studies on this are [41, 34, 42]). To (very shortly and not-at-all thoroughly[3]) introduce what is replica symmetry, we say that it is a property of probability distributions in high dimensions (like the ones we have seen so far, i.e. posterior distributions in inference and canonical distributions in physics). When the distribution is replica symmetric, then two things, among many others, happen. To describe them, let us first define the overlap of two Ising configurations: the overlap q between two configurations $\underline{\sigma}^1$ and $\underline{\sigma}^2$ is:

$$q = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^1 \sigma_i^2$$

It is a scalar product which measures how much the two configurations are aligned. When a distribution $P(\sigma)$ is replica symmetric, then:

1. If we sample two configurations $\underline{\sigma}^1$ and $\underline{\sigma}^2$ from this distribution, then on average they will have a fixed overlap q.

[3] we refer the reader to the extremely clear description provided in [37], Chapter 12: *What is a glass phase?* pag 252.

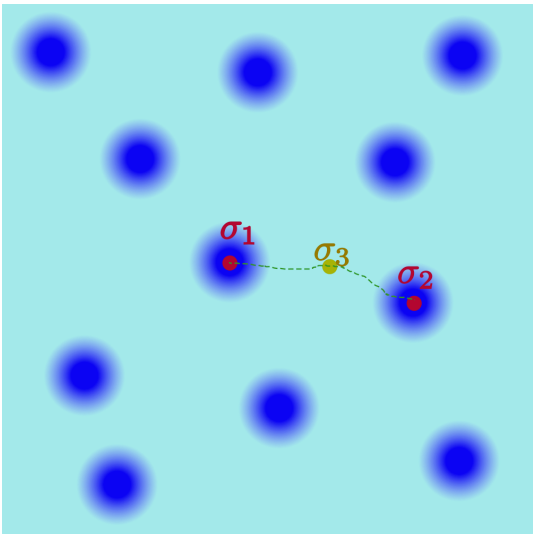

Fig. 1.2 Clustering of high probability zones for a Replica Symmetry Broken distribution. To pass from a configuration σ_1 to another σ_2 located in two disconnected zones of high probability, the dynamics should visit some other configuration σ_3 of low probability. Therefore, this is very difficult to happen and the dynamics gets trapped.

2. In a Monte Carlo simulation the dynamics encounters few (i.e. non exponential in N) attractors.

When the distribution is replica symmetry broken, typically these two points become false, i.e. there is no more a fixed overlap q and the dynamics gets stuck in one of the very many attractors.

These two points might be explained pictorially with a 2 dimensional image (keeping in mind, however, that P is a high dimensional probability distribution), as shown in Figure 1.2. The idea is that the probability distribution has some zones of high probability. These zones might be disconnected, meaning that to progressively (e.g. by single spin-flips) deform one configuration $\underline{\sigma}^1$ to another $\underline{\sigma}^2$ which are on two different zones of high probability, one has to pass for some configuration $\underline{\sigma}^3$ which is not in a high probability zone. As a consequence, the dynamics[4] gets stuck in one island of high probability. When the system becomes replica symmetry broken, the number of such islands become exponential. For what concerns the overlap, two sampled configurations

[4]e.g. Monte Carlo dynamics, which typically relies on some local updates like the Metropolis scheme [43].

are on average in a high probability zone. If we sample two configuration of the same island, than they have high overlap. If we sample two configurations belonging to two different islands, they will have lower overlap. The reader must be aware that we are hiding a lot in this discussion, for example:

- there are several ways (or steps) in which replica symmetry can be broken (1 step, 2 steps, up to infinite replica symmetry breaking) which describe how the zones of high probability are nested;
- there exists the dynamical replica symmetry breaking transition, a phenomenon in which the dynamics behaves as if as the system was Replica Symmetry Broken, but the thermodynamics is replica symmetric, namely there is only one overlap on average.

A clear introduction to replica theory (with detailed computations) is in [40]. For the present aim, we will simply say that if the distribution of overlaps concentrates around a value, then the probability distribution is replica symmetric. We are going to see now that this is exactly what happens in the planted spin glass model.

1.2.6 Nishimori conditions & Replica Symmetry

The important result that we can finally show is that the Ising spin glass is always replica symmetric at the Nishimori temperature, i.e. on the Nishimori line. To do so we use the bridge with Bayesian inference. We first rewrite in a lighter notation equation (1.7):

$$\mathbb{E}[A(x^*)] = \mathbb{E}[A(x)]$$

in which we (with some abuse of notation) now use the symbol $\mathbb{E}[\cdot]$ for the averages in general. The distribution over which the average is computed is fixed by the variable name. The x^* variable is a planted configuration, so is distributed according to the prior. The other variable comes from the posterior. The generalization of the first Nishimori relation to a two-variables observable $A(x, y)$ is also true (see [34, 44]), namely:

$$\mathbb{E}[A(x^*, x)] = \mathbb{E}[A(y, z)],$$

where again, the average is performed respectively w.r.t. the prior for x^* and the posterior for x, y, z. This equation can be proved by direct computation, as eqn. (1.7). Its meaning is also similar: on average, we can estimate an observable by exchanging the planted configuration with a configuration sampled from the posterior. However, this equation is crucial when applied to Ising spin glass:

$$\mathbb{E}[A(\underline{\sigma}^*, \underline{\sigma})] = \mathbb{E}[A(\underline{\sigma}^1, \underline{\sigma}^2)].$$

If we set $A(\underline{\sigma}^1, \underline{\sigma}^2)$ equal to the overlap between $\underline{\sigma}^1, \underline{\sigma}^2$, then the average value of the overlap between the planted and a configuration is equal to the average value q of the overlap between two configurations. Let us compute the first term, remembering that, due to the gauge, the planted is simply $\underline{\sigma}^* = (1, 1, \dots, 1)$:

$$\mathbb{E}[A(\underline{\sigma}^*, \underline{\sigma})] = \mathbb{E}\left[\frac{1}{N}\sum_i \sigma_i\right] = \mathbb{E}\left[\frac{1}{N}\sum_i \sigma_i\right]$$

but this is simply the average over the magnetization m. So we have:

$$\mathbb{E}[m] = \mathbb{E}[q]$$

we can repeat this reasoning to find equations for all the moments of the distributions of m and q, arriving to an equality between the two probability distributions of the magnetization and the overlap:

$$p(m) = p(q).$$

Since the magnetization is a self concentrating quantity (see for example [34]), its distribution is actually peaked on its average. As a consequence, the distribution of overlaps is peaked on a single value. We can finally conclude that the Ising spin glass is replica symmetric on the Nishimori line. This property is considered to be much more general than this. The idea is that Nishimori conditions (i.e. Bayes optimality) should always imply replica symmetry [41].

1.2.7 Nishimori conditions in epidemic inference

The study of Nishimori conditions motivates us to always infer the hyper-parameters of a distribution. Otherwise, if we just guessed them, we risked to

get stuck in one of the very many attractors of the replica symmetry broken posterior. In Chapter 2, therefore, detailed description to hyper-parameter learning is given. In Chapter 3, additionally, a more detailed analysis on the inference phase diagram is provided for the epidemic problem. The aim of Chapter 3 is to do for epidemic inference what we have just done for the planted spin glass: characterizing its phase diagram. We will develop a semi-analytical tool based on the replica symmetric cavity equations.

1.3 Epidemic inference is an NP-hard problem

Here we show that computing the posterior of the epidemic model is a NP-hard problem. Take a population of N individuals on a graph $G = (V, \mathcal{E})$, $|V| = N$. Consider the time discrete SI epidemic process of T temporal steps on graph G, described by the prior $P(x)$, from which the planted configuration $x^* = (x_1^*, \dots, x_N^*)$ is extracted. Each $x_i^* = (x_i^{1,*}, \dots, x_i^{T,*})'$ is the trajectory of individual i, for $i \in V$. Let $\mathcal{O}$ be the set of all the observations on x^*. . Let us call posterior the distribution $\mathcal{P}(x|\mathcal{O})$ which gives to each epidemic trajectory the probability of being the planted, conditioned to the observations. Suppose now that there exists an algorithm $\mathcal{A}$ which computes in polynomial time the posterior. The algorithm takes in input the set of observations $\mathcal{O}$, the prior P, the graph G and it outputs the posterior distribution $\mathcal{P}$. If such an algorithm exists, then we can solve the Unweighted Minimum Steiner Tree (UMST) problem in polynomial time. The UMST problem requires to find, given a graph $G = (V, \mathcal{E})$ and a subset of $L \subset V$ of the vertex set, if it exists a sub-tree H of G which connects all the vertices in L and with number of vertices less or equal to a constant w, i.e. $|H| \leq w$.

1.3.1 Intuitive explanation

It is possible to map UMST into the epidemic inference problem. Here we suggest a non rigorous mapping. It is sufficient to take the time discrete SI model and build the observations set by imposing that the individuals corresponding to the vertices in L are observed $\mathcal{I}$ at time T. The observations set $\mathcal{O}_L$ built this way allows to map the epidemic problem to UMST if we set

infection probability λ and patient zero probability γ to approach zero. In fact, for $\gamma \to 0$, the prior probability for an individual to be the patient zero goes to zero. However, the observations force the posterior to allow for at least one patient zero, in order to explain the $\mathcal{I}$ observations. Let us consider the posterior:

$$\mathcal{P}(x|\mathcal{O}) \propto P(\mathcal{O}|x)P(x)$$

where the proportionality is w.r.t. x. The quantity $P(\mathcal{O}|x)$ in this context is a hard constraint, so it is one if $x_l^T = \mathcal{I}, \forall l \in L$ and it is zero otherwise. This means that, for all the trajectories x satisfying the constraint, the posterior is proportional to the prior: $\mathcal{P}(x|\mathcal{O}) \propto P(x)$. This will be useful to understand which are the trajectories x having the highest posterior probability. Take in fact, among the ones which satisfy all the constraints, the trajectory x which has r patients zero and z infectious individuals at final time. The posterior probability of this trajectory is $\mathcal{P}(x|\mathcal{O}) \approx \gamma^r \lambda^{z-r}$. This is because there are $z - r$ infection events and r patients zero. If we consider the limit in which $\lambda \to 0$, $\gamma \to 0$, $\frac{\gamma}{\lambda} \to 0$, we see that the most probable trajectories are the ones that minimize r and $z - r$. Since γ is infinitely smaller that λ, the most probable trajectories are the ones that minimize r and among them the most probable are the ones that minimize $z - r$. The minimum value for r is 1 because if $r = 0$ then no epidemic happens. So we say that the most probable trajectories have $r = 1$. Any trajectory x' with $r > 1$ is infinitely less probable than a trajectory x with $r = 1$:

$$\frac{\mathcal{P}(x|\mathcal{O})}{\mathcal{P}(x'|\mathcal{O})} = \lambda^c \frac{\gamma}{\gamma^r} = \frac{\lambda^c}{\gamma^{r-1}} \to \infty$$

where $c \in \mathbb{Z}$. Among the trajectories with $r = 1$, the most probable are the ones that minimize z. All the others have infinitely smaller probability to happen. Of course, $z_{min} \geq |L|$ because there are at least $|L|$ individuals observed $\mathcal{I}$. So far, we have concluded that the posterior distribution is nonzero only for trajectories with one patient zero and which minimize z such that all the constraints are satisfied. These trajectories are the solutions of UMST. In fact, the infection chain $(x^0, x^1, \ldots, x^T)$ connects all the individuals in L with the minimum number of infections in between. Using the algorithm $\mathcal{A}$ we could therefore compute the posterior and sampling from it would allow us to obtain the minimum number of solution(s) to UMST, which is an NP-complete

problem.

1.3.2 Proof

A slightly different construction leads us to a rigorous mapping between the UMST and the epidemic marginalization. Let us take one vertex $l \in L$. We build the observation set by stating that at time zero all individuals in V are observed in the state $\mathcal{S}$ except for l, which is observed $\mathcal{I}$. Due to the observations at time zero we have that there is only one patient zero which is the individual l. At time $T > |L|$ all the individuals in L are observed in the state $\mathcal{I}$. This builds the observations set. We now show that this observations set, with a proper choice of the infection parameter λ, results in a posterior which, if marginalized, allows to solve the UMST problem. The posterior of a configuration x which satisfies the observation constraints is:

$$\mathcal{P}(x|\mathcal{O}) = \frac{P(x)P(\mathcal{O}|x)}{P(\mathcal{O})} = \frac{P(x,\mathcal{O})}{P(\mathcal{O})}$$

where again, the quantity $P(\mathcal{O}|x)$ is a hard constraint, so it is one if $x_l^T = \mathcal{I}, \forall l \in L$ and it is zero otherwise. We now want to find an upper and a lower bound to this expression. To do so, we introduce an equivalent formulation of the epidemic spread. Let us associate to each edge $e \in \mathcal{E}$ and for each time $t = 1, \dots, T$ a gate Boolean variable $g_e^t \in \{0, 1\}$ which is 1 if the gate is open and 0 if the gate is closed. If a gate is open, then the infection can pass at that time through that edge: this means that if one $\mathcal{I}$ individual is connected at a certain time to an $\mathcal{S}$ individual, it will infect it if and only if the gate is open. Each gate has independent probability λ to be open. We now define $z(x)$ as the total number of infected individuals in the trajectory x excluding the patient zero. Note that $z(x) \in \{0, \dots, N-1\}$. Thus, there exists a minimum number $z_{\min} = \min_x z(x)$. We now consider the probability of the following event: $z(x)$ specific gates are open and the other channels can be open or closed, such that the observations constraints are satisfied. This event has probability equal to $\lambda^{z(x)}$ and it is a lower bound for $P(x, \mathcal{O})$. In fact, the probability of the event x is the sum of several possible configurations of channels among which there

is the event just considered. Therefore:

$$\lambda^{z(x)} \leq P(x, \mathcal{O})$$

This is valid only for x satisfying the observations constraints, i.e. for $P(\mathcal{O}|x)$. For any x:

$$\lambda^{z(x)} P(\mathcal{O}|x) \leq P(x, \mathcal{O})$$

To find an upper bound, we define $\underline{g} = \{g_e^t\}_{e\in\mathcal{E}}^{t=0,\dots T-1}$ as a specific configuration of the channels and $n(\underline{g})$ as the number of open gates in the configuration $\underline{g}$. We notice that:

$$\begin{aligned}
P(x) &\leq \sum_{\underline{g}:n(\underline{g})\geq z(x)} \lambda^{n(\underline{g})}(1-\lambda)^{|\mathcal{E}T|-n(\underline{g})} \\
&\leq \lambda^{z(x)} \sum_{\underline{g}:n(\underline{g})\geq z(x)} (1-\lambda)^{|\mathcal{E}T|-n(\underline{g})} \\
&\leq \lambda^{z(x)} \sum_{\underline{g}:n(\underline{g})\geq z(x)} 1 \\
&\leq \lambda^{z(x)} \sum_{\underline{g}} 1 \\
&\leq \lambda^{z(x)} T^{|\mathcal{E}|}.
\end{aligned}$$

Observing that $P(\mathcal{O}|x) \in \{0, 1\}$ we have that $P(x, \mathcal{O}) \leq P(x)$, so:

$$\lambda^{z(x)} P(\mathcal{O}|x) \leq P(x, \mathcal{O}) \leq \lambda^{z(x)} T^{|\mathcal{E}|}. \tag{1.12}$$

Now we sum over x:

$$\sum_x P(\mathcal{O}|x)\lambda^{z(x)} \leq P(\mathcal{O}) \leq \sum_x \lambda^{z(x)} T^{|\mathcal{E}|}. \tag{1.13}$$

The lower bound can be minored by choosing only one element of the whole sum. In particular we choose a trajectory x such that $z(x) = z_{\min}$:

$$\sum_x P(\mathcal{O}|x)\lambda^{z(x)} \geq \lambda^{z_{\min}}$$

The upper bound is:

$$\begin{aligned}\sum_x \lambda^{z(x)} T^{|\mathcal{E}|} &= T^{|\mathcal{E}|}\left(\sum_{x:z(x)=z_{\min}} \lambda^{z(x)} + \sum_{x:z(x)>z_{\min}} \lambda^{z(x)}\right) \\ &\leq T^{|\mathcal{E}|}\lambda^{z_{\min}} \#(\text{trajectories}) \\ &= T^{|\mathcal{E}|}\lambda^{z_{\min}} T^N\end{aligned}$$

$$\lambda^{z_{\min}} \leq P(\mathcal{O}) \leq T^{|\mathcal{E}|+N}\lambda^{z_{\min}}.$$

We now take the logarithm:

$$z_{\min}\log\lambda \leq \log P(\mathcal{O}) \leq (|\mathcal{E}|+N)\log T + z_{\min}\log\lambda$$

and we divide by $\log\lambda$:

$$z_{\min} \leq \frac{\log P(\mathcal{O})}{\log\lambda} \leq \frac{(|\mathcal{E}|+N)\log T}{\log\lambda} + z_{\min}$$

If:

$$\left|\frac{(|\mathcal{E}|+N)\log T}{\log\lambda}\right| < \frac{1}{2}$$

which means:

$$\lambda < T^{-2(|\mathcal{E}|+N)}$$

then we have that:

$$z_{\min} \leq \frac{\log P(\mathcal{O})}{\log\lambda} < \frac{1}{2} + z_{\min}$$

so that the integer number closest to $\frac{\log P(\mathcal{O})}{\log\lambda}$ is $z_{\min}$. It is important to stress that we found a bound to λ which is independent of the specific instance to infer. This proves that being able to compute the partition function of the epidemic problem allows to find the minimum size of the Steiner tree. Computing the partition function is therefore at least as hard of an NP-complete problem, thus it is NP-hard. Now we show that not only computing the partition function, but also computing marginals is an NP-hard problem. This is a simple consequence of the computations just shown. Intuitively, we have just proved that the epidemic posterior for $\lambda < T^{-2(|\mathcal{E}|+N)}$ is dominated by the configurations x with $z(x) = z_{\min}$. This implies that if we compute the total number of infectious

individuals (which corresponds to the sum of the marginals of being infectious at final time), we should have a number close to $z_{\min} + 1$, i.e. the patient zero plus the infected individuals. In formulae

$$\sum_{i=1}^{N} p_i(x_i^T = \mathcal{I}|\mathcal{O}) \approx z_{\min} + 1$$

Therefore, computing the marginals allows to solve the UMST. Let us now make this argument rigorous. Using (1.12) and (1.13) we have:

$$\frac{\lambda^{z(x)} P(\mathcal{O}|x)}{T^{|\mathcal{E}|+N} \lambda^{z_{\min}}} \leq P(x|\mathcal{O}) \leq \frac{\lambda^{z(x)} T^{|\mathcal{E}|}}{\lambda^{z_{\min}}}.$$

We focus now only on the upper bound in order to estimate the probability of having $z(x) > z_{\min}$:

$$\begin{aligned}
\sum_{x:z(x)>z_{\min}} P(x|\mathcal{O}) &\leq \sum_{x:z(x)>z_{\min}} \frac{\lambda^{z(x)} T^{|\mathcal{E}|}}{\lambda^{z_{\min}}} \\
&\leq \lambda \sum_{x:z(x)>z_{\min}} T^{|\mathcal{E}|} \\
&\leq \lambda T^{|\mathcal{E}|+N} \\
&< \frac{1}{N}.
\end{aligned}$$

For

$$\lambda T^{|\mathcal{E}|+N} < \frac{1}{N},$$

which corresponds to

$$\lambda < \frac{1}{T^{|\mathcal{E}|+N} N},$$

we have:

$$\sum_{x:z(x)>z_{\min}} P(x|\mathcal{O}) < \frac{1}{N}. \tag{1.14}$$

We use this to estimate $\left|\sum_{i=1}^{N} p_i(x_i^T = \mathcal{I}|\mathcal{O}) - z_{\min}\right|$. First we rewrite the term:

$$\begin{aligned}\sum_{i=1}^{N} p_i(x_i^T = \mathcal{I}|\mathcal{O}) &= \sum_{i=1}^{N}\sum_{x} \mathcal{P}(x|\mathcal{O})\delta_{x_i^T,\mathcal{I}} \\ &= \sum_{x} \mathcal{P}(x|\mathcal{O}) \sum_{i=1}^{N} \delta_{x_i^T,\mathcal{I}} \\ &= \sum_{x} (z(x)+1)\mathcal{P}(x|\mathcal{O}) \\ &= (z_{\min}+1) \sum_{x:z(x)=z_{\min}} \mathcal{P}(x|\mathcal{O}) + \sum_{x:z(x)>z_{\min}} (z(x)+1)\mathcal{P}(x|\mathcal{O}).\end{aligned}$$

Where we used that $z(x)+1 = \sum_{i=1}^{N} \delta_{x_i^T,\mathcal{I}}$. Now we use (1.14) on the second addend of the sum to have:

$$\begin{aligned}\sum_{i=1}^{N} p_i(x_i^T = \mathcal{I}|\mathcal{O}) &< (z_{\min}+1) + (z_{\max}+1)\frac{1}{N} && (1.15)\\ &= (z_{\min}+1) + 1. && (1.16)\end{aligned}$$

Thus, the integer part of $\sum_{i=1}^{N} p_i(x_i^T = \mathcal{I}|\mathcal{O})$ is $z_{\min}$. We have therefore proved that being able to compute the marginals for $\lambda < T^{-|\mathcal{E}|-N}N$ allows to find the UMST. Exactly computing the epidemic marginals is an NP-hard problem. However, we can prove something more: even accurately approximating epidemic marginals is NP-hard. Let us suppose to be able to compute each epidemic marginal with an error $\varepsilon > 0$:

$$\hat{p}_i(x_i^T = \mathcal{I}|\mathcal{O}) = p_i(x_i^T = \mathcal{I}|\mathcal{O}) \pm \varepsilon$$

then, let us choose $\lambda < T^{-|\mathcal{E}|-N}(\delta N)^{-1}$, where $\delta > 1$. Using equation (1.16):

$$\sum_{i=1}^{N} \hat{p}_i(x_i^T = \mathcal{I}|\mathcal{O}) < (z_{\min}+1) + \frac{1}{\delta} \pm N\varepsilon.$$

$$\frac{1}{\delta} + N\varepsilon \leq 1$$

Therefore, approximating the marginals with an error $\varepsilon < \left(1 - \frac{1}{\delta}\right) N^{-1}$ is again an NP-hard problem because it would still allow to find $z_{\min}$.

Chapter 2

Inference of the single instance

This Chapter is devoted to algorithms which approximate the posterior distribution at fixed single instance. This means that we are interested in reconstructing a hidden fixed planted x^*, which represents an epidemic trajectory propagated along a given contact graph G. We hypothesize to have a given set of observations $\mathcal{O}$ from x^*. The aim is to use the information contained in the graph G and in the observations $\mathcal{O}$ to reconstruct the hidden epidemic instance x^*, by approximately computing marginals of the posterior distribution $\mathcal{P}(x|\mathcal{O})$. In the first section of this chapter we introduce the prior distributions of several epidemic models. Then we move to the study of algorithms used to approximate the posterior. We present the Causal Variational Approach (CVA), a method devised to efficiently sample from the posterior and we compare its performance to other existing algorithms, showing its superiority. Finally we discuss, using the Causal Variational Approach, a possible simplification rule for the epidemic inference problem, a result that we call *model reduction*.

2.1 Models

Agent-based models play a pivotal role in epidemic literature, as evidenced by their widespread use in studies such as [45–47]. These models serve as fundamental tools for capturing the epidemic spread within a network. In the introduction, we delved into the discrete-time Susceptible Infectious (SI) model. This section expands on our exploration by presenting more general

models and highlighting their significance in the context of inference. The underlying concept is that the broader and more comprehensive the model, the more finely it describes the dynamics of the epidemic process. However, it is noteworthy that a simpler model may exhibit a similar cascade description compared to a more complex counterpart. In such instances, opting for the simpler model proves advantageous for two compelling reasons: first, the coding of the inference algorithm is more straightforward; second, the parameter space is smaller, rendering it more intuitive for interpretation. In the following section we introduce the prior distribution of SI, SEIR, SIS models for discrete and continuous time dynamics, both in the usual markovian and in the non-markovian case. We subsequently move to the description of network models. The reader who wants to skip the precise mathematical details of each model can directly move to section 2.1.4 in which a general equation which describes priors is introduced.

2.1.1 SEIR discrete-time model

The Susceptible-Exposed-Infectious-Recovered model takes into account four possible states in which an individual can be:

1. $\mathcal{S}$: the individual has never been infected. It is therefore not able to infect anyone. It can be infected, with probability λ, by an $\mathcal{I}$ individual if there is a contact between the two. If an $\mathcal{S}$ individual gets infected, it becomes Exposed (E)

2. E: the individual has been infected recently. In this state the individual is still not infectious, so it can not infect any other. At the same time, it has been infected so it can not be infected again. Every time step it might undergo the transition $E \to \mathcal{I}$ with probability ν.

3. $\mathcal{I}$: the individual is infectious, so it can infect $\mathcal{S}$ individuals (with probability λ) in contact with it. Every time step, it can recover with probability μ.

4. $\mathcal{R}$: the individual is recovered (or removed, i.e. dead). It can not infect nor be infected. It does not undergo any transition: an $\mathcal{R}$ individual always remain $\mathcal{R}$.

While writing this list of states, we have introduced not only the four states characterizing the model, but also some transition probabilities. The infection probability λ was already introduced in the SI model. Here we have introduced the latency probability ν to pass from E to $\mathcal{I}$ and the recovery probability μ to pass from $\mathcal{I}$ to $\mathcal{R}$. Given the nature of the transitions in time, the future state only depends on the present. The probability of the process is therefore markovian:

$$P(x) = P(x^0) \prod_{t=0}^{T-1} P(x^{t+1}|x^t) \tag{2.1}$$

where $x = \{x_i^t, \forall i = 1, \ldots, N, \forall t = 0, \ldots, T\}$ is the complete epidemic cascade and $x^t = \{x_i^t, \forall i = 1, \ldots, N\}$ is the state of all the individuals at fixed time t. The terms in equation (2.1) factorize: the initial time term can be rewritten as:

$$P(x^0) = \prod_{i=1}^{N} p(x_i^0)$$

because every individual is the patient zero independently with the same probability, where $p^0(x_i^0) = \gamma \delta_{x_i^0,\mathcal{I}} + (1-\gamma)\, \delta_{x_i^0,\mathcal{S}}$ (see section 1.1.2 for more details). The transition term $P(x^{t+1}|x^t)$, as in the SI model case, factorizes because the individuals' states at time $t+1$ are independent, conditioned to the state at time t. Moreover, as in the SI model, the transition of an individual i between times t and $t+1$ is only due to its neighbors $\partial i(t)$. This is because the transition $\mathcal{S} \to E$ depends on the contacts and the other transitions only depend on the single individual (i.e. they are independent of the neighbors). This implies that:

$$P(x) = \prod_{i=1}^{N} \left(p^0(x_i^0) \prod_{t=0}^{T-1} p^{t+1}(x_i^{t+1}|x_i^t, x_{\partial i(t)}^t) \right)$$

which is formally identical to (1.2), but in this case the transition rates are different:

$$
\begin{aligned}
p^{t+1}(x_i^{t+1} = \mathcal{S} | x_i^t = \mathcal{S}, x_{\partial i(t)}^t) &= \prod_{j \in \partial i} (1 - \lambda \delta_{x_j^t, \mathcal{I}}) \\
p^{t+1}(x_i^{t+1} = E | x_i^t = \mathcal{S}, x_{\partial i(t)}^t) &= 1 - \prod_{j \in \partial i} (1 - \lambda \delta_{x_j^t, \mathcal{I}}) \\
p^{t+1}(x_i^{t+1} = E | x_i^t = E) &= 1 - \nu \\
p^{t+1}(x_i^{t+1} = \mathcal{I} | x_i^t = E) &= \nu \\
p^{t+1}(x_i^{t+1} = \mathcal{I} | x_i^t = \mathcal{I}) &= 1 - \mu \\
p^{t+1}(x_i^{t+1} = \mathcal{R} | x_i^t = \mathcal{I}) &= \mu.
\end{aligned}
$$

In which we wrote the nonzero probability transition rates. The other transitions are zero because the only allowed ones are $\mathcal{S} \to E \to \mathcal{I} \to \mathcal{R}$.

2.1.2 Recurrence: SIS model

The SI and SEIR models have in common that each individual can only be affected once by the infection phenomenon. This is because there is no possibility for an individual to jump again to the $\mathcal{S}$ state. A model is said to be recurrent if we allow such a possibility. The simplest example is the SIS model. The model is again markovian, with a probability distribution which is always:

$$
P(x) = \prod_{i=1}^{N} \left(p^0(x_i^0) \prod_{t=0}^{T-1} p(x_i^{t+1} | x_i^t, x_{\partial i(t)}^t) \right)
$$

where the transition probabilities are:

$$
\begin{aligned}
p(x_i^{t+1} = \mathcal{S} | x_i^t = \mathcal{S}, x_{\partial i(t)}^t) &= \prod_{j \in \partial i} (1 - \lambda \delta_{x_j^t, \mathcal{I}}) \\
p(x_i^{t+1} = \mathcal{I} | x_i^t = \mathcal{S}, x_{\partial i(t)}^t) &= 1 - \prod_{j \in \partial i} (1 - \lambda \delta_{x_j^t, \mathcal{I}}) \\
p(x_i^{t+1} = \mathcal{I} | x_i^t = \mathcal{I}) &= 1 - \mu \\
p(x_i^{t+1} = \mathcal{S} | x_i^t = \mathcal{I}) &= \mu
\end{aligned}
$$

For the recovery probability $\mu = 0$ we come back to the SI model.

2.1.3 From discrete to continuous time models

The epidemics spread continuously in time. However, if one particular epidemic spreads with an approximately fixed timescale Δt, then it is possible to coarse grain a continuous time model in epochs of time Δt. For example, for the case of COVID-19, the discrete-time epidemic description may be adopted by setting the epoch Δt to one day. For a discrete time epidemic which starts at time 0 and ends at time T:

$$x = \begin{pmatrix} x_1^0 & x_1^1 & \cdots & x_1^T \\ \vdots & \vdots & \ddots & \vdots \\ x_N^0 & x_N^1 & \cdots & x_N^T \end{pmatrix}.$$

For continuous time models, x can no longer be represented as a matrix because it becomes a continuous function along the rows (which might be parametrized depending on the specific model). The representation used so far is therefore not usable in the continuous-time case. However, it is possible to introduce a new notation which works both in the discrete and in the continuous time models. Take for example the SI model. To fully characterize the trajectory of an individual i it is sufficient to specify its infection time $t_i^{\mathcal{I}}$. There is therefore, a map from the representation $(x_i^0, x_i^1, \ldots, x_i^T)$ to a real number $t_i^{\mathcal{I}}$. For example, the trajectory $(\mathcal{S}, \mathcal{S}, \mathcal{S}, \mathcal{I}, \mathcal{I}, \mathcal{I}, \ldots, \mathcal{I})$ is mapped onto $t_i^{\mathcal{I}} = 3$ because the individual gets infectious at time 3. The trajectory in which the individual never gets infected is conventionally mapped at $T + 1$:

$$x_i = (\mathcal{S}, \mathcal{S}, \ldots, \mathcal{S}) \to t_i^{\mathcal{I}} = T + 1$$

For the SEIR model the mapping can be done in the same way. We just need to introduce three transition times, which respectively correspond to the times when the individual gets exposed, infectious and recovered.

$$(x_i^0, x_i^1, \ldots, x_i^T) \to (t_i^E, t_i^{\mathcal{I}}, t_i^{\mathcal{R}}).$$

We can, therefore, define $\underline{t}_i$ as the tuple which encodes the transition times of a single individual i. For the SI model $\underline{t}_i = t_i^{\mathcal{I}}$ and for SEIR $\underline{t}_i = (t_i^E, t_i^{\mathcal{I}}, t_i^{\mathcal{R}})$. We finally define $\underline{t} = \{\underline{t}_i, \forall i = 1, \ldots, N\}$. Now we have a notation to encode a continuous-time epidemic trajectory. To write down the probability distribution

we do a continuous time limit. Notice that for the SIS model the number of possible transition times is infinite. It is therefore not possible to deal with it using this approach. In this thesis, inference in the SIS model is not treated, but we refer the interested reader to [48].

Continuous-time SI model

We provide here all the details for the SI model computation, under the hypothesis of a time-independent contact network. The generalizations follow the very same procedure and will be dealt with later on. More details can be found in [9]. We first introduce a time set $\mathcal{T}_{\delta t} = \{0, \delta t, 2\delta t, 3\delta t, \dots, T\}$, which depends on $\delta t \in \mathbb{R}^+$ (it will be sent to zero). We then rewrite equation (1.2) :

$$P(x) = \prod_{i=1}^{N} \left(p^0(x_i^0) \prod_{t \in \mathcal{T}_{\delta t}} \prod_{j \in \partial i} p(x_i^{t+\delta t} | x_i^t, x_{\partial i}^t) \right)$$

by plugging the transition probabilities of the SI model, eq (1.1), we have:

$$P(x) = \prod_{i=1}^{N} \left[p^0(x_i^0) \left(\prod_{j \in \partial i} \prod_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t} (1-\lambda) \right) \left(1 - \prod_{j \in \partial i} \left(1 - \lambda \delta_{x_j^{t_i^{\mathcal{I}}-\delta t}, \mathcal{I}} \right) \right) \right]$$

where we simply substituted the $\mathcal{S} \to \mathcal{S}$ transition for all the time steps in which i does not get infected (until time $t_i^{\mathcal{I}}$) and the $\mathcal{S} \to \mathcal{I}$ transition at time $t_i^{\mathcal{I}}$. The notation $\prod_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t}$ means that $t \in \mathcal{T}_{\delta t}$ and $t_j^{\mathcal{I}} + \delta t \leq t \leq t_i^{\mathcal{I}} - 2\delta t$. We multiply now the two products and write $p^0(x_i^0)$ explicitly :

$$P(\underline{t}) = \prod_{i=1}^{N} \left[\delta_{t_i^{\mathcal{I}},0} \gamma + (1 - \delta_{t_i^{\mathcal{I}},0})(1-\gamma) \left(\prod_{j \in \partial i} \prod_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t} (1-\lambda) - \prod_{j \in \partial i} \prod_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-\delta t} (1-\lambda) \right) \right] \tag{2.2}$$

so that everything is written as a function of $\underline{t}$ and the dependency on x is eliminated. Notice that the infection probability should scale as the number of time steps; otherwise the epidemic dynamic would be trivial (either all or no individuals infected). Since the number of time-steps is $|\mathcal{T}_{\delta t}| = T/\delta t$, then the infection probability scales as $\lambda = \tilde{\lambda} \delta t$. We now rewrite the products over time

using the identity $a = e^{\log a}$:

$$\prod_{j\in\partial i}\prod_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t}(1-\tilde{\lambda}\delta t) = \prod_{j\in\partial i} e^{\sum_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t}\log(1-\tilde{\lambda}\delta t)} \simeq$$
$$= \prod_{j\in\partial i} e^{-\tilde{\lambda}\sum_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t}\left(\delta t+O(\delta t^2)\right)} = \tag{2.3}$$
$$= \prod_{j\in\partial i} e^{-\tilde{\lambda}\left(t_i^{\mathcal{I}}-\delta t-t_j^{\mathcal{I}}\right)_+ +O(\delta t)}.$$

Where we introduced:

$$(a)_+ = \begin{cases} a & \text{if } a > 0 \\ 0 & \text{if } a \leq 0 \end{cases}$$

to substitute the value of the sum. We now notice that in the distribution $P(\underline{t})$ there is the difference of the two products, which can be manipulated:

$$\left(\prod_{j\in\partial i}\prod_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t}(1-\lambda) - \prod_{j\in\partial i}\prod_{t=t_j^{\mathcal{I}}+\delta t}^{t_i^{\mathcal{I}}-2\delta t}(1-\lambda)\right) =$$
$$= \left(\prod_{j\in\partial i} e^{-\tilde{\lambda}\left(t_i^{\mathcal{I}}-\delta t-t_j^{\mathcal{I}}\right)_+} - \prod_{j\in\partial i} e^{-\tilde{\lambda}\left(t_i^{\mathcal{I}}-t_j^{\mathcal{I}}\right)_+}\right) \simeq$$
$$= -\delta t\frac{\mathrm{d}}{\mathrm{d}(\delta t)}\left(\prod_{j\in\partial i} e^{-\tilde{\lambda}\left(t_i^{\mathcal{I}}+\delta t-t_j^{\mathcal{I}}\right)_+}\right)\Bigg|_{\delta t=0} + O(\delta t^2).$$

We see that the entire probability distribution has a scaling leading term of δt. Plugging the manipulation into (2.2) and evaluating the limit $\delta t \to 0$:

$$F(\underline{t}) = \left(\prod_{i:t_i^{\mathcal{I}}=0}\gamma\right)\left(\prod_{i:t_i^{\mathcal{I}}>0}(1-\gamma)\frac{\mathrm{d}}{\mathrm{d}(\delta t)}\left(-\prod_{j\in\partial i} e^{-\tilde{\lambda}\left(t_i^{\mathcal{I}}+\delta t-t_j^{\mathcal{I}}\right)_+}\right)\Bigg|_{\delta t=0}\right)$$

where we defined the density distribution $F(\underline{t}) = \frac{\mathrm{d}^N(P(\underline{t}))}{\mathrm{d}(\delta t)^N}\Big|_{\delta t=0}$. We also separated the contributions of patients zero from the others. The last passage is to compute

the derivative:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}(\delta t)}\left(\prod_{j\in\partial i} e^{-\tilde{\lambda}\left(t_i^{\mathcal{I}}+\delta t-t_j^{\mathcal{I}}\right)_+}\right) = \\
=&\frac{\mathrm{d}}{\mathrm{d}(\delta t)}\left(e^{-\tilde{\lambda}\sum_{j\in\partial i}\left(t_i^{\mathcal{I}}+\delta t-t_j^{\mathcal{I}}\right)_+}\right) = \\
=&-\tilde{\lambda}\sum_{k\in\partial i}\theta(t_i^{\mathcal{I}}-t_j^{\mathcal{I}})e^{-\sum_{j\in\partial i}\tilde{\lambda}\left(t_i^{\mathcal{I}}+\delta t-t_j^{\mathcal{I}}\right)_+}.
\end{aligned}$$

Where

$$\theta(t) = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t \leq 0 \end{cases}$$

The continuous-time SI model prior density distribution is thus:

$$F(\underline{t}) = \left[\left(\prod_{i:t_i^{\mathcal{I}}=0}\gamma\right)\left(\prod_{i:t_i^{\mathcal{I}}>0}(1-\gamma)\tilde{\lambda}\left(\sum_{k\in\partial i}\theta(t_i^{\mathcal{I}}-t_k^{\mathcal{I}})\right)\prod_{j\in\partial i}e^{-\tilde{\lambda}\left(t_i^{\mathcal{I}}-t_j^{\mathcal{I}}\right)_+}\right)\right]$$

Notice that the term $\sum_{k\in\partial i}\theta(t_i^{\mathcal{I}}-t_k^{\mathcal{I}})$ ensures that there is at least one patient zero. If it was not the case, in fact, there would be an individual i for which $t_i^{\mathcal{I}}$ is the minimum infection time and $t_i^{\mathcal{I}} > 0$. The term $\sum_{k\in\partial i}\theta(t_i^{\mathcal{I}}-t_k^{\mathcal{I}})$ would then be zero, vanishing the product. We now have a physical interpretation to the term $\tilde{\lambda}$, which is the infection *rate.*

Continuous SEIR model

Generalizing to the SEIR model is straightforward. It is sufficient to evaluate the continuous time also for each of the transitions. For example, the additional term due to latency is in the form:

$$\begin{aligned}
\nu\prod_{t=t_i^E}^{t_i^{\mathcal{I}}-\delta t}(1-\nu) &= \tilde{\nu}\delta t\prod_{t=t_i^E}^{t_i^{\mathcal{I}}-\delta t}(1-\delta t\tilde{\nu}) = \\
&= \tilde{\nu}\delta t e^{-\tilde{\nu}\left(t_i^{\mathcal{I}}-t_i^E\right)+O(\delta t)}.
\end{aligned}$$

The density distribution is:

$$F(\underline{t}) = \left(\prod_{i:t_i^{\mathcal{I}}=0} \gamma \right) \left[\prod_{i:t_i^{\mathcal{I}}>0} \left((1-\gamma)\tilde{\lambda} \left(\sum_{k\in\partial i} \theta(t_i^E - t_j^{\mathcal{I}}) \right) \prod_{j\in\partial i} e^{-\tilde{\lambda}\left(t_i^E - t_j^{\mathcal{I}}\right)_+} \right) \times \right.$$
$$\left. \times \ \tilde{\nu} e^{-\tilde{\nu}\left(t_i^{\mathcal{I}} - t_i^E\right)} \tilde{\mu} e^{-\tilde{\nu}\left(t_i^{\mathcal{R}} - t_i^{\mathcal{I}}\right)} \right].$$

2.1.4 Time varying contact network

If the contact network changes in time, it is possible to generalize the above equations by introducing a time varying infection probability λ_{ij}^t, which is equal to λ during the contact time-interval and is equal to zero when for all the times t in which i and j are not in contact with each others. Since we have a generic interaction which is time and edge dependent, we need to keep an integral in equation (2.3):

$$\prod_{j\in\partial i} e^{-\sum_{t=t_j^{\mathcal{I}}}^{t_i^{\mathcal{I}}-\delta t} \tilde{\lambda}_{ji}^t \delta t} \to \prod_{j\in\partial i: t_j^{\mathcal{I}} < t_i^{\mathcal{I}}} e^{-\int_{t_j^{\mathcal{I}}}^{t_i^{\mathcal{I}}-\delta t} \tilde{\lambda}_{ji}(t)\mathrm{d}t} =: \prod_{j\in\partial i: t_j^{\mathcal{I}} < t_i^{\mathcal{I}}} e^{-\left(\Lambda_{ji}(t_i^{\mathcal{I}}-\delta t) - \Lambda_{ji}(t_j^{\mathcal{I}})\right)}$$

where the function Λ_{ji} is a primitive of $\tilde{\lambda}_{ji}$. The quantity $\tilde{\lambda}_{ji}(t)$ is the non constant infection rate. Notice that:

- The infection rate is a non negative function of time: $\tilde{\lambda}_{ji}(t) \geq 0$
- The infection rate is not normalized: $\int \mathrm{d}t \tilde{\lambda}_{ji}(t) \neq 1$. In fact, for a constant rate, this integral is infinity.
- Differently from the time-constant case, the non constant rate gives rise to an unnormalized infection probability. Take the PDF for an individual i to get infected from j at time t:

  $$f(t) = \tilde{\lambda}_{ji}(t) e^{-\int_0^t \mathrm{d}u \tilde{\lambda}_{ji}(u)}$$

  This quantity is not normalized. Which means that :

  $$F(\infty) := \int_0^\infty f(t)dt \leq 1.$$

While for constant rates the equality holds, for non constant rates (which for example decrease strongly over time), the total integral of this improper PDF can be less than 1. This intuitively corresponds to rates which do not guarantee an infection over time. In reality, actually, this can happen: two individuals in contact (one in the $\mathcal{I}$ and the other in the $\mathcal{S}$ state) might never pass the infection one to the other! The infectiousness of the $\mathcal{I}$ individual decreases indeed after some days. If after the first days the infection does not happen, then it becomes extremely unlikely to have it later. This phenomenology is instead impossible in the (unrealistic) case of constant rate (sooner or later the infection will take place!). When normalizing the $f(t)$ for non constant rates, thus, we always have to keep in mind that there is a probability $1 - F(\infty)$ to have no infection.

Now that we have generalized to a non constant rate in time, what follows is identical to the derivation in the case of constant network, but for sake of completeness we repeat here the computations in this more general case. We proceed step by step for the SI model case. The generalized form of equation (2.3) is:

$$
\begin{aligned}
&\prod_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}\prod_{t=t_j^{\mathcal{I}}}^{t_i^{\mathcal{I}}-\delta t}(1-\lambda_{ji}^t)-\prod_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}\prod_{t=t_j^{\mathcal{I}}}^{t_i^{\mathcal{I}}}\left(1-\lambda_{ji}^t\right)=\\
\overset{\delta t\to 0}{\longrightarrow}&\prod_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}e^{-\left(\Lambda_{ji}(t_i^{\mathcal{I}}-\delta t)-\Lambda_{ji}(t_j^{\mathcal{I}})\right)}-\prod_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}e^{-\left(\Lambda_{ji}(t_i^{\mathcal{I}})-\Lambda_{ji}(t_j^{\mathcal{I}})\right)}=\\
=&-\delta t\frac{\mathrm{d}}{\mathrm{d}(\delta t)}\prod_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}e^{-\left(\Lambda_{ji}(t_i^{\mathcal{I}}+\delta t)-\Lambda_{ji}(t_j^{\mathcal{I}})\right)}\Bigg|_{\delta t=0}=\\
=&\delta t\prod_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}e^{-\left(\Lambda_{ji}(t_i^{\mathcal{I}}+\delta t)-\Lambda_{ji}(t_j^{\mathcal{I}})\right)}\sum_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}\frac{\mathrm{d}\Lambda_{ji}(t_i^{\mathcal{I}}+\delta t)}{\mathrm{d}(\delta t)}=\\
=&\delta t e^{-\sum_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}\left(\Lambda_{ji}(t_i^{\mathcal{I}}-\delta t)-\Lambda_{ji}(t_j^{\mathcal{I}})\right)}\sum_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}\tilde{\lambda}_{ji}(t_i^{\mathcal{I}})=\\
=&\delta t e^{-\sum_{j\in\partial i:t_j^{\mathcal{I}}<t_i^{\mathcal{I}}}\left(\Lambda_{ji}(t_i^{\mathcal{I}}-\delta t)-\Lambda_{ji}(t_j^{\mathcal{I}})\right)}\sum_{j\in\partial i}\theta(t_i^{\mathcal{I}}-t_j^{\mathcal{I}})\tilde{\lambda}_{ji}(t_i^{\mathcal{I}}).
\end{aligned}
$$

The final form of the distribution is therefore:

$$F(\underline{t}) = \left[\left(\prod_{i:t_i^{\mathcal{I}}=0} \gamma \right) \left(\prod_{i:t_i^{\mathcal{I}}>0} (1-\gamma) e^{-\sum_{j\in\partial i: t_j^{\mathcal{I}}<t_i^{\mathcal{I}}} \left(\Lambda_{ji}(t_i^{\mathcal{I}})-\Lambda_{ji}(t_j^{\mathcal{I}})\right)} \sum_{j\in\partial i} \theta(t_i^{\mathcal{I}}-t_j^{\mathcal{I}}) \tilde{\lambda}_{ji}(t_i^{\mathcal{I}}) \right) \right]. \tag{2.4}$$

Having introduced a time and edge-dependent function $\tilde{\lambda}_{ij}(t)$ actually allows to treat the general case of non homogeneous and non constant infection rates. This last form is therefore very general and allows to describe, e.g., the non Markov models which are introduced in the next paragraph.

Non-Markov models

So far we have described only Markov processes, in which the infection state of the population only depends on the immediate past state. In other words, there is no memory in the models we have described. The reason is that all the rates (infection, latency, recovery) introduced so far are constant in time. In real epidemic scenarios, actually, it seems more meaningful to introduce an infection rate which changes during time after the infection. More precisely, let $i \in \{1, \dots, N\}$ be an individual who gets infected at time $t_i^{\mathcal{I}}$. In the standard SI model individual i would infect its contacts with probability λ from time $t_i^{\mathcal{I}}$ forever on. This is quite an unrealistic description of the epidemic transmission, which is expected to be very low the first hours after the infection, increase during time until it reaches a peak and finally decrease to zero. The SEIR model is an approximation of this phenomenology. However, it is possible and sometimes convenient to maintain a two variable model and generalize instead the infection rate. We introduce here the non-markovian infection rate, which changes over time after the infection:

$$\lambda_{ij}(t) = \lambda(t - t_i^{\mathcal{I}}) \qquad \text{with} \qquad \lambda(t<0) = 0.$$

This definition is both valid for discrete and continuous time models and means that the infection probability from i to one contact $j \in \partial i$ at time t is equal to a function which depends only on time passed after the infection event. This definition is rather tricky, so we make her some remarks. The outgoing infection rate just defined is:

- asymmetric: $\lambda_{ij}(t) = \lambda(t - t_i^{\mathcal{I}}) \neq \lambda(t - t_j^{\mathcal{I}}) = \lambda_{ji}(t)$.
- homogeneous among the contacts, i.e. $\lambda_{ij}(t)$ does not depend by any means on the contact j;
- explicitly independent of the source individual, i.e. $\lambda_{ij}(t)$ does not depend explicitly on i. The dependence is only in the infection time. So, e.g. the infection rate of an individual i with $t_i^{\mathcal{I}} = 3$ at time $t = 6$ is equal to the infection rate of an individual j with $t_j^{\mathcal{I}} = 23$ at time $\tau = 26$ because:

$$\lambda_{ik}^{6} = \lambda(t - t_i^{\mathcal{I}}) = \lambda(3) = \lambda(\tau - t_j^{\mathcal{I}}) = \lambda_{jm}^{26}$$

  for $j \in \partial i$ and $m \in \partial j$

The infection rate just defined makes the stochastic process non-markovian. To see this, we can think in the discrete time case. If we want to know the transition probabilities of the population infection state from time $t-1$ to time t we need to know all the infection rates $\{\lambda(t - 1 - t_i^{\mathcal{I}}), \forall i = 1, \ldots, N\}$, which however depend on the infection times of all the individuals, which can be determined only knowing the history of the dynamics. Let us write the explicit prior for the non-Markov SI model in the discrete and continuous case. For the discrete case the form is identical to (1.2) with the transitions rates that change from (1.1) to:

$$P(x_i^{t+1} = \mathcal{S} | x_i^t = \mathcal{S}, x_{\partial i(t)}^t) = \prod_{j \in \partial i(t)} (1 - \lambda(t - t_j^{\mathcal{I}}))$$

For the continuous time model we use the general form of equation (2.4) substituting the functional form of the rate. We simply have to compute the

difference of the primitives $\Lambda_{ji}(t_i^{\mathcal{I}} - \delta t) - \Lambda_{ji}(t_j^{\mathcal{I}})$ in this specific case, which is:

$$\Lambda_{ji}(t_i^{\mathcal{I}}) - \Lambda_{ji}(t_j^{\mathcal{I}}) = \int_{t_j^{\mathcal{I}}}^{t_i^{\mathcal{I}}} \tilde{\lambda}_{ji}(t)\mathrm{d}t =$$
$$= \int_{t_j^{\mathcal{I}}}^{t_i^{\mathcal{I}}} \lambda(t - t_j^{\mathcal{I}})\mathrm{d}t = \int_{0}^{t_i^{\mathcal{I}} - t_j^{\mathcal{I}}} \lambda(u)\mathrm{d}u =$$
$$= \Lambda(t_i^{\mathcal{I}} - t_j^{\mathcal{I}}).$$

Where Λ is the primitive of λ. The prior for the continuous time non-markovian SI model is therefore:

$$F(\underline{t}) = \prod_{i=1}^{N} \left[\left(\prod_{i:t_i^{\mathcal{I}}=0} \gamma \right) \left(\prod_{i:t_i^{\mathcal{I}}>0} (1-\gamma) e^{-\sum_{j\in\partial i} \Lambda(t_i^{\mathcal{I}} - t_j^{\mathcal{I}})} \sum_{j\in\partial i} \lambda(t_i^{\mathcal{I}} - t_j^{\mathcal{I}}) \right) \right]$$

This closes the description of the epidemic models used in this thesis. We compact the results into a more general equation in the next paragraph.

Modeling some network ignorance

If the contact network is partially known, there may happen that some individuals who only had registered contacts with $\mathcal{S}$ people, end up being infected by some $\mathcal{I}$ individual which is not registered in the contact graph. To model such an ignorance, a self-infection α is typically introduced. The self infection α is defined as the probability for an isolated $\mathcal{S}$ individual to get infected in one time step. The transition probability in the discrete time models is therefore:

$$P(x_i^{t+1} = \mathcal{S} | x_i^t = \mathcal{S}, x_{\partial i(t)}) = (1-\alpha) \prod_{j\in\partial i(t)} (1 - \lambda \delta_{x_j^t, \mathcal{I}}).$$

Passing to continuous limit is straightforward: it is sufficient to consider the self-infection of i as the infection rate of another individual, always in the $\mathcal{I}$ state, attached to i . For the SI non-markovian model, for example, the final

form is:

$$F(\underline{t}) = \prod_{i=1}^{N} \left[\left(\prod_{i:t_i^{\mathcal{I}}=0} \gamma \right) \left(\prod_{i:t_i^{\mathcal{I}}>0} (1-\gamma) e^{-\alpha t_i^{\mathcal{I}} - \sum_{j\in\partial i} \Lambda(t_i^{\mathcal{I}} - t_j^{\mathcal{I}})} \left(\alpha + \sum_{j\in\partial i} \lambda(t_i^{\mathcal{I}} - t_j^{\mathcal{I}}) \right) \right) \right]$$

A compact notation

It is now useful to write down an equation which summarizes the prior distributions seen so far (time discrete, continuous, SI, SEIR, constant and non-constant network). The best notation relies on the transition times $\underline{t} = (\underline{t}_1, \ldots, \underline{t}_N)$, because they describe both the discrete and the continuous-time models. We recall here that each transition time $\underline{t}_i$ is the set of the times at which individual i changes its state. By direct comparison with each model described so far, we can see that a general form for the priors is:

$$P(\underline{t}) = \prod_{i=1}^{N} \psi(\underline{t}_i, \underline{t}_{\partial i}) \tag{2.5}$$

where ψ differs depending on the model and $\underline{t}_{\partial i} = \{\underline{t}_j, \forall j \in \partial i\}$. We are going to use P to both express a probability distribution for discrete time case and a probability density function for the continuous time case. For example, take the SI model at discrete time and fixed contact network, equation (1.2):

$$\begin{aligned}
P(x) =& \prod_{i=1}^{N} p^0(x_i^0) \prod_{t=0}^{T-1} p(x_i^{t+1} | x_i^t, x_{\partial i}^t) = \\
=& \prod_{i=1}^{N} p^0(x_i^0) \left(\prod_{t=0}^{t_i-2} p(x_i^{t+1} = \mathcal{S} | x_i^t = \mathcal{S}, x_{\partial i}^t) \right) p(x_i^{t_i^{\mathcal{I}}} = \mathcal{I} | x_i^{t_i^{\mathcal{I}}-1} = \mathcal{S}, x_{\partial i}^t) = \\
=& \prod_{i=1}^{N} \left[p^0(x_i^0) \left(\prod_{t=0}^{t_i^{\mathcal{I}}-2} \prod_{j\in\partial i} (1 - \lambda \delta_{x_j^t, \mathcal{I}}) \right) \left(1 - \prod_{j\in\partial i} (1 - \lambda \delta_{x_j^{t_i^{\mathcal{I}}-1}, \mathcal{I}}) \right) \right] \\
=& \prod_{i=1}^{N} \left[p^0(x_i^0) \left(\prod_{t=0}^{t_i^{\mathcal{I}}-2} \prod_{j\in\partial i} \left(1 - \lambda \theta(t - t_j^{\mathcal{I}})\right) \right) \left(1 - \prod_{j\in\partial i} \left(1 - \lambda \delta(t_i^{\mathcal{I}} - 1 - t_j^{\mathcal{I}})\right) \right) \right],
\end{aligned}$$

where in the third passage we have substituted the transition probability in equation (1.1) and in the fourth passage we rephrased the x state with the

infection time notation by introducing the theta function

$$\theta(t) = \begin{cases} 1 & t > 0 \\ 0 & t \leq 0 \end{cases}.$$

We clearly see, thus, that for this model we have, comparing with equation (2.5):

$$\psi(t_i^{\mathcal{I}}, t_{\partial i}^{\mathcal{I}}) = p^0(x_i^0) \left(\prod_{t=0}^{t_i^{\mathcal{I}}-2} \prod_{j \in \partial i} \left(1 - \lambda\theta(t - t_j^{\mathcal{I}})\right) \right) \left(1 - \prod_{j \in \partial i} \left(1 - \lambda\delta(t_i^{\mathcal{I}} - 1 - t_j^{\mathcal{I}})\right) \right). \tag{2.6}$$

Equation (2.5) tells us that the prior distribution factorizes over local functions, each one depending on a central individual and its neighbors. Note that:

- The factor is normalized. In the discrete case: $\sum_{\underline{t}_i} \psi(\underline{t}_i, \underline{t}_{\partial i}) = 1$. In the continuous case there would be an integral. To see this, we notice that $\psi(\underline{t}_i, \underline{t}_{\partial i}) = p(x_i^0) \prod_t p^{t+1}(x_i^{t+1} | x_i^t, x_{\partial i}^t)$. Summing over $\underline{t}_i$ at fixed $\underline{t}_{\partial i}$ means to sum over all possible i trajectories.

- $\psi(\underline{t}_i, \underline{t}_{\partial i}) \neq p(\underline{t}_i | \underline{t}_{\partial i})$. Indeed, the conditioned distribution of $\underline{t}_i$ at fixed $\underline{t}_{\partial i}$ is:

$$\begin{aligned} \frac{p(\underline{t}_i, \underline{t}_{\partial i})}{p(\underline{t}_{\partial i})} &= \frac{1}{p(\underline{t}_{\partial i})} \sum_{\underline{t} \setminus (\underline{t}_i, \underline{t}_{\partial i})} P(\underline{t}) = \\ &= \frac{1}{p(\underline{t}_{\partial i})} \sum_{\underline{t} \setminus (\underline{t}_i, \underline{t}_{\partial i})} \prod_j \psi(\underline{t}_j, \underline{t}_{\partial j}). \end{aligned}$$

The more compact notation just introduced is useful to describe the algorithms in a more general framework. Before doing so, we quickly explain how networks and observations are simulated and modeled.

Observation modeling

A clinical test is the observation in the epidemic inference problem and can be modeled as a 4-uple $o = (o_s, o_i, o_t, o_f)$ where o_s is the state (which depending on the model can be $\mathcal{S}$,E,$\mathcal{I}$,$\mathcal{R}$) found by the clinical test, $o_i = 1, \ldots, N$ is the tested individual, o_t is the time at which test o is done and o_f is the noise (i.e.

false positive/negative rate). The set of all the observations is $\mathcal{O}$. To simulate and observation o, it is sufficient to randomly draw an individual o_i from the planted x^* and evaluate its state $x^*_{o_i,o_t}$at a randomly drawn time o_t. Then this result is corrupted with probability o_f to obtain o_s. The likelihood is the probability of observing $\mathcal{O}$ conditioned on the trajectory $\underline{t}$. The observations $o \in \mathcal{O}$ are independent of each others, conditioned to the epidemic trajectory. In other words:

$$P(\mathcal{O}|\underline{t}) = \prod_{o\in\mathcal{O}} p(o|\underline{t})$$

each observation o, in turn, depends on the whole epidemic trajectory only through the individual o_i corresponding to that observation. The likelihood can be written as:

$$P(\mathcal{O}|\underline{t}) = \prod_{o\in\mathcal{O}} p(o|\underline{t}_{o_i})$$

Each factor of the likelihood is the probability of observing the state o_s at time o_t, given the epidemic state $\underline{t}_{o_i}$:

$$p(o|\underline{t}_{o_i}) = o_f\left(1-\delta_{o_s,x^{o_s}_{o_i}}\right) + (1-o_f)\delta_{o_s,x^{o_s}_{o_i}}$$

which means that the probability of observing the incorrect state o_s for the individual o_i is equal to the false rate o_f, while the probability to observe the correct state, $o_s = x^{o_s}_{o_i}$ of the individual o_i is one minus the false rate. Since the likelihood factorizes over observations, we can group the observations over the same individual:

$$\begin{aligned} P(\mathcal{O}|\underline{t}) &= \prod_{i=1}^{N} \prod_{o\in\mathcal{O}:o_i=i} p(o|\underline{t}_i) = \\ &=: \prod_{i=1}^{N} p(\{o\}_{o_i=i}|\underline{t}_i), \end{aligned}$$

where the last passage is a definition. Since the prior can be written in the form (2.5), we have that:

$$\begin{aligned} P(\underline{t}|\mathcal{O}) &= \frac{P(\underline{t})P(\mathcal{O}|\underline{t})}{P(\mathcal{O})} = \\ &= \frac{\prod_{i=1}^{N}\psi(\underline{t}_i,\underline{t}_{\partial i})p(\{o\}_{o_i=i}|\underline{t}_i)}{\sum_{\underline{t}'}\prod_{i=1}^{N}\psi(\underline{t}'_i,\underline{t}'_{\partial i})p(\{o\}_{o_i=i}|\underline{t}'_i)}. \end{aligned} \tag{2.7}$$

As already mentioned in equation (1.4), the posterior can be rewritten in a canonical ensemble form, by defining $H(\underline{t}) = -\log P(\underline{t})P(\mathcal{O}|\underline{t})$. We thus have that the Hamiltonian is a sum of local terms, namely:

$$H(\underline{t}) = -\sum_i \log\left(\psi(\underline{t}_i, \underline{t}_{\partial i}) p(\{o\}_{o_i=i}|\underline{t}_i)\right).$$

This form is useful to discuss some approximations as the mean field method, section 2.2.1.

Network modeling

Epidemics propagate along networks of individuals who interact with each other in a rather complex way. For example, there are more sociable individuals, jobs that require connecting with many people, and there might be a correlation between a person's job and their interests. Therefore, the typical places where an artist goes during their free time might be different from those visited by a gym instructor. This complexity is studied by network theorists, who also provide models to simulate contact graphs. To give four examples of very simple (though quite unrealistic) network models, we consider the tree, the random regular, the Erdős–Rényi (which are all introduced in section 1.1.1) and the proximity networks. The latter is defined, in 2 dimensions, as follows: individuals are assigned uniformly random an x and y coordinate, $(x, y) \in [0, 1] \times [0, 1]$. A radius $0 \leq R \leq 1$ is fixed and the Euclidean distance between two individuals i of coordinates (x_i, y_i) and j with coordinates (x_j, y_j) is defined as $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$. Each individual $i \in \{1, \ldots, N\}$ is linked to all the individuals j such that $d_{ij} \leq R$. In other words: $\partial i = \{j \in \{1, \ldots, N\} : d_{ij} \leq R\}$. This model accounts for an interaction which spreads in fixed dimensions (2 in this case, in general D if we distribute the points in D dimensions), which is not realistic because individuals might fly among cities and countries, take the car to move throughout the neighborhoods, meet people from other cities. None of the above network models is complex enough to describe a realistic network for epidemic spread. It is important to test the inference algorithms on complicated and realistic graphs. Otherwise, we might overestimate the predictive power of some methods. For example, we are going to introduce later on the Belief Propagation, which works optimally

if the contact network has a tree structure. Therefore, if we tested it on a tree network, it would show extraordinarily good performance. However, realistic contact networks have loops. To have a fair estimate of the performances on real-like scenarios, we have to use the most realistic models available. We are thus going to introduce now two network models which we use later as benchmarks to test the inference algorithms that will be introduced in the next sections of this chapter.

- *OpenABM-Covid19* : it firstly appeared in [49]. The contact graph is the superposition of densely connected graphs representing interactions within households and a sparser network mirroring occupation relationships. A random time-varying network is additionally used to model contacts in public transportation, transient social gatherings, etc. The number of interactions through the random network is extracted from a negative binomial distribution to allow for rare super-spreading events. Memberships to both fixed and dynamic graphs are determined by the age of individuals, e.g. children live with adults, elderly people have fewer interactions than other age groups, etc.

- *Space-temporal model from geolocation data*: this model has been developed in [50]. Individuals are assigned to households that are localized in an urban area according to the actual population density. Using the available geo-location data, other venues as schools, research institutes, bars, bus stops, workplaces and supermarkets are similarly displaced in the map. Individuals can visit a number of locations with a probability that decreases as the household-target distance increases. The duration of contacts between individuals concurrently visiting the same venue is assumed to be known and gathered by contact tracing smartphone applications. Some interesting and realistic features naturally arise from this contact dynamics, such as the presence of super-spreaders.

These two models seem to integrate the difficulties of the Erdős–Rényi and proximity graphs. They take in fact into account the correlation between the probability of having a link between two individuals and their euclidean distance. They include however many long range interactions due to work reasons, schools or recreational activities. These two models allow to generate

networks that are used later to test the performance of the algorithms. It is now the time to explain such algorithms.

2.2 Variational methods

Variational approaches are widely used for several applications including machine learning and computational biology [51–54]. They rely on the minimization (respectively maximization) of a functional which quantifies the error (resp. the quality) of an approximating function. The name comes from the calculus of variations, which is typically used to stabilize this functional. In this context, we are going to study strategies for minimizing the Kullback-Leibler (KL) divergence, introduced in [55]. Given two probability distributions $\mathcal{P}$ and $\mathcal{Q}$, defined on a domain $\mathcal{X}$, the KL divergence between $\mathcal{Q}$ and $\mathcal{P}$ is:

$$D_{KL}(\mathcal{Q}||\mathcal{P}) = \int_{x\in\mathcal{X}} \mathcal{Q}(x) \log \frac{\mathcal{Q}(x)}{\mathcal{P}(x)} \mathrm{d}x = \left\langle \log \frac{\mathcal{Q}(x)}{\mathcal{P}(x)} \right\rangle_{x\sim\mathcal{Q}}.$$

This quantity is always positive and benefits of the triangular inequality. It is zero if and only if $\mathcal{P} = \mathcal{Q}$ (almost everywhere). It is therefore similar to a distance between two probabilities except for the fact that it is non-symmetric with respect to inversion of $\mathcal{P}$ and $\mathcal{Q}$ (see [53]). It is thus a pseudo distance. It can be used in the context of inference to measure how much an approximation $\mathcal{Q}$ is able to reproduce the posterior $\mathcal{P}$. In particular, let us suppose that we want to approximate a posterior distribution $\mathcal{P} : \mathcal{X} \to [0, 1]$. If one defines a family of approximating functions $\mathcal{Q}^\omega$ which depend on a continuous parameter $\omega = (\omega_1, \omega_2, \ldots, \omega_m)$, with $\omega \in \Omega \subset \mathbb{R}^m$, it is possible to optimize the KL divergence $D_{KL}(\mathcal{Q}^\omega||\mathcal{P})$ between $\mathcal{Q}^\omega$ and the posterior $\mathcal{P}$ with respect to the set ω in order to find the optimal approximating function $\mathcal{Q}^{\tilde{\omega}}$, for $\tilde{\omega} = \arg\min_\Omega D_{KL}(\mathcal{Q}^\omega||\mathcal{P})$. If the family $\{\mathcal{Q}^\omega\}_{\omega\in\Omega}$ contains the function $\mathcal{P}$, then the minimization of the KL divergence leads to the exact posterior because the KL divergence is zero if and only if the two functions are equal to each others. In general, the functional form of the posterior is unknown and $\mathcal{P} \notin \{\mathcal{Q}^\omega\}_{\omega\in\Omega}$. Therefore, the optimal $\mathcal{Q}^{\tilde{\omega}}$ is only an approximation. To find $\mathcal{Q}^{\tilde{\omega}}$, one typically sets the gradient $\partial_\omega \mathcal{Q}^\omega = 0$ in order to find some fixed point equations. Sometimes, however, the functional form of $\mathcal{Q}^\omega$ does not allow to

find such equations and one must resort to a gradient descent. To practically see how the variational method works we now show an application to find a mean field approximation of the posterior.

2.2.1 The (too) naive mean field method

In this subsection we try to approximate the discrete time SI model posterior with a variational mean field (MF) approach. This approach produces a bad approximation for the posterior. However, it is instructive to understand the procedure in order to build stronger methods, as the Causal Variational Approach, section 2.3. Let $\mathcal{P}(x|\mathcal{O})$ be the posterior distribution of an epidemic model which is too hard to compute exactly. We can try to approximate its marginals by means of a *mean field* approximation, which means that we choose a family of functions completely factorized over the individuals:

$$\mathcal{Q}^{\omega}(x) = \prod_{i=1}^{N} q_i^{\omega_i}(\underline{t}_i)$$

Each ω_i is in turn a set of parameters. For example, for a time discrete SI model, the parameter $\omega_i = (\omega_i^0, \ldots, \omega_i^T)$ controls each value of the marginal distribution $q_i^{\omega_i}(t_i^{\mathcal{I}})$. Notice that we are using the equivalent notation of infection times $\underline{t}$, defined in section 2.1.3, interchangeably with the trajectory notation x. They are equivalent so this is totally harmless. We now have to optimize over the $N \times T$ matrix:

$$q = \begin{pmatrix} q_1^{\omega_1^0} & \cdots & q_1^{\omega_1^T} \\ \vdots & \ddots & \vdots \\ q_N^{\omega_N^0} & \cdots & q_N^{\omega_N^T} \end{pmatrix},$$

where each value of the matrix is defined as $q_i^{\omega_i^t} := q_i^{\omega_i}(t_i^{\mathcal{I}} = t)$. The notation might appear a cumbersome, but the idea behind is simply that we are optimizing over all the possible values each marginal can take in time[1]. The optimal mean field approximation is then found by minimizing the divergence

[1] A short-hand notation would be to optimize over $q_i(t_i^{\mathcal{I}})$ for all $i = 1, \ldots, N$ and for all $t_i^{\mathcal{I}} = 0, \ldots, T$. It could then be written that the optimal approximating function is given by $\frac{\delta D_{KL}(\mathcal{Q}||\mathcal{P})}{\delta q_i(t_i^{\mathcal{I}})} = 0$. However, this notation may lead to confusion when applied to more complex approximating families that are going to be introduced later on.

between the approximating function and the posterior:

$$
\begin{aligned}
D_{KL}(\mathcal{Q}^{\omega}||\mathcal{P}) &= \left\langle \log \frac{\mathcal{Q}^{\omega}(x)}{\mathcal{P}(x|\mathcal{O})} \right\rangle_{x\sim\mathcal{Q}} = . \\
&= \left\langle \log \frac{\prod_i q_i^{\omega_i}(t_i^{\mathcal{I}})}{\mathcal{P}(\underline{t}|\mathcal{O})} \right\rangle_{\underline{t}\sim\mathcal{Q}} = \\
&= \sum_{i=1}^{N} \left\langle \log q_i^{\omega_i}(t_i^{\mathcal{I}}) \right\rangle_{\underline{t}\sim\mathcal{Q}} - \langle \log \mathcal{P}(\underline{t}|\mathcal{O}) \rangle_{\underline{t}\sim\mathcal{Q}} = \\
&= \sum_{i=1}^{N} \left\langle \log q_i^{\omega_i}(t_i^{\mathcal{I}}) \right\rangle_{\underline{t}\sim\mathcal{Q}} - \langle \log P(\underline{t}) P(\mathcal{O}|\underline{t}) \rangle_{\underline{t}\sim\mathcal{Q}} + \langle \log P(\mathcal{O}) \rangle_{\underline{t}\sim\mathcal{Q}} = \\
&= \sum_{i=1}^{N} \left\langle \log q_i^{\omega_i}(t_i^{\mathcal{I}}) \right\rangle_{\underline{t}\sim\mathcal{Q}} - \langle \log P(\underline{t}) P(\mathcal{O}|\underline{t}) \rangle_{\underline{t}\sim\mathcal{Q}} + \log P(\mathcal{O}) = \\
&= -S[\mathcal{Q}^{\omega}] + U[\mathcal{Q}^{\omega}] + \log P(\mathcal{O}).
\end{aligned}
$$

We defined, in the last passage, the entropy of the distribution S and the energy U. The reason why U is called energy is due to the analogy between Bayesian inference and statistical physics, equation (1.5): the product of likelihood and prior can be seen as the exponential of a Hamiltonian. We now optimize over $\mathcal{Q}^{\omega}$ to find a fixed point equation:

$$
0 = \frac{\partial D_{KL}(\mathcal{Q}^{\omega}||\mathcal{P})}{\partial \omega_j^t} = \log q_j^{\omega_j}(t) + \frac{\partial U[\mathcal{Q}^{\omega}]}{\partial \omega_j^t}
$$

the derivative of $\log P(\mathcal{O})$ vanishes because it does not depend on ω. Since each marginal is normalized, we have to impose the constraint that $\sum_{t_i^{\mathcal{I}}=0}^{T} q_i^{\omega_i}(t_i^{\mathcal{I}}) = 1$, for all individuals i. This simply means that we have to add a Lagrange multiplier α_i for each of the individuals. When we derive this term too we end up with:

$$
\begin{aligned}
0 &= \log q_j^{\omega_j}(t) + \frac{\partial U[\mathcal{Q}^{\omega}]}{\partial \omega_j^t} - \frac{\partial \left(\sum_i \alpha_i \sum_{t_i^{\mathcal{I}}} q_i^{\omega_i}(t_i^{\mathcal{I}}) \right)}{\partial \omega_j^t} = \\
&= \log q_j^{\omega_j}(t) + \frac{\partial U[\mathcal{Q}^{\omega}]}{\partial \omega_j^t} - \alpha_j.
\end{aligned}
$$

The fixed point equation is therefore:

$$
q_j^{\omega_j}(t) = \alpha_j e^{-\frac{\partial U[\mathcal{Q}^{\omega}]}{\partial \omega_j^t}}.
$$

where each α_j is fixed by imposing the normalization condition:

$$\alpha_j = \frac{1}{\sum_t e^{-\frac{\partial U[\mathcal{Q}^\omega]}{\partial \omega_j^t}}}.$$

This is a rather elegant equation that however is hiding a huge fundamental problem which makes this mean field approach useless for the epidemic context. As the equation suggests, in fact, the computation of mean field marginals has been reduced to compute the derivatives of the energy. The problem is that the energy is, for many choices of ω, infinite! The energy is finite only for very non-physical values of ω. This implies that the equation above is only a formal solution to the marginalization problem. To see why the energy diverges, we simply compute it.

$$\begin{aligned} U =& - \langle \log P(\underline{t}) P(\mathcal{O}|\underline{t}) \rangle_{\underline{t} \sim \mathcal{Q}} = \\ =& - \langle \log P(\underline{t}) \rangle_{\underline{t} \sim \mathcal{Q}} - \log \langle P(\mathcal{O}|\underline{t}) \rangle_{\underline{t} \sim \mathcal{Q}} . \end{aligned}$$

The divergent term is due to the prior, so we study that one:

$$\begin{aligned} \langle \log P(\underline{t}) \rangle_{\underline{t} \sim \mathcal{Q}} &= \left\langle \log \prod_i \psi(t_i^{\mathcal{I}}, \underline{t}_{\partial i}^{\mathcal{I}}) \right\rangle_{\underline{t} \sim \mathcal{Q}} = \\ &= \sum_i \left\langle \log \psi(t_i^{\mathcal{I}}, \underline{t}_{\partial i}^{\mathcal{I}}) \right\rangle_{\underline{t} \sim \mathcal{Q}} . \end{aligned}$$

The divergent terms are the $\left\langle \log \psi(t_i^{\mathcal{I}}, \underline{t}_{\partial i}^{\mathcal{I}}) \right\rangle$. This is due to the fact that the function $\psi(t_i^{\mathcal{I}}, \underline{t}_{\partial i}^{\mathcal{I}})$ is zero for some values of $t_i^{\mathcal{I}}, \underline{t}_{\partial i}^{\mathcal{I}}$ due to causal constraints. In fact, ψ must be zero all the times that $0 < t_i^{\mathcal{I}} < \min_{j \in \partial i}\{t_j^{\mathcal{I}}\}$, because i can get infected either because one contact infected it, or because i is the patient zero. In other words, if there is an infection event which can not be justified because all the contacts are susceptible, then the transition is impossible[2]. The problem of the mean field approximation is that, in general, it is not able to capture correlations among individuals (the approximation is in fact fully factorized). As a consequence, the average contains terms of the form

[2]as stated in paragraph 2.1.4, the transition could be explained by introducing some self-infection, which models some network ignorance. In this case the MF method would work. We would like, however to build a method which performs better and better if information increases. For the MF method instead, we have quite the opposite, i.e. the method works only if our ignorance on the network is high!

$q_i(t_i^{\mathcal{I}}) \prod_{j \in \partial i} q_j(t_j^{\mathcal{I}}) \log 0$, which diverge. The only way to avoid these infinities is to impose the probability distribution to be 0 every time the transition is 0. This, however, forces the marginals of each individual $i = 1, \ldots, N$ to be nonzero only for $t_i^{\mathcal{I}} = 0, T+1$, i.e. patient zero or not infected at all. This polarization of the solution is due to the causal constraints and leads to the conclusion that this variational mean field approximation does not work for epidemic inference.

2.3 The Causal Variational Approach

Part of this section is from [9]. When an approximation does not work there is at least one way to go on: trying to build a better approximation! The standard naive mean field procedure fails because of hard causal constraints which introduce strong correlations, impossible to be captured by the MF method. In other words, the prior distribution has constraints due to the fact that every time an individual changes state from $\mathcal{S}$ to $\mathcal{I}$ there must be at least one infectious contact which caused the infection:

$$p(x_i^{t+1} = \mathcal{I} | x_i^t = \mathcal{S}, x_{\partial i} = \{\mathcal{S}, \mathcal{S}, \ldots, \mathcal{S}\}) = 0$$

in the infection time notation, equation (2.5): $\psi(t_i^{\mathcal{I}}, t_{\partial i}^{\mathcal{I}}) = 0$ if $\{t_j \geq t_i, \forall j \in \partial i\}$. One path to follow is to enrich the family of approximating functions: we need an approximation that includes some correlation. In particular, the family of functions $\mathcal{Q}^\omega$ must respect all the causal constraints imposed by the infection dynamics. At the same time, these approximating functions must be easy to compute, otherwise it would be meaningless to approximate the posterior $\mathcal{P}$ with a function $\mathcal{Q}^{\tilde{\omega}}$ which we are not able to compute in polynomial time. Actually, there exists a functional form which respects both the requirements. That is the prior, which naturally respects all the constraints (actually the constraints *come* from the prior) and it is hypothesized to be polynomially computable. An idea is thus to build a family of approximating functions $\mathcal{Q}^\omega$ which are functionally identical to the prior. By minimizing the KL divergence $D_{KL}(\mathcal{Q}^\omega || \mathcal{P})$ w.r.t. $\omega \in \Omega$, we would find the best prior-like function approximating the posterior. A way to build a family $\{\mathcal{Q}^\omega\}$ from the prior P is to allow the transition rates of the prior to be a free parameter to optimize later. For example, if the prior

has a constant infection probability of $\lambda = 0.2$, which enters P through the functions ψ, we can build a family of approximating functions $\mathcal{Q}^{\lambda}$ which are identical to P but for the fact that λ is a free parameter and is not fixed to 0.2. We can even generalize it further, by allowing the infection rate to be heterogeneous in the graph $\lambda \to \{\lambda_{ij}\}_{i,j=1}^{N}$, so to build a much richer family $\mathcal{Q}^{\{\lambda_{ij}\}_{i,j=1}^{N}}$. We can also modify the other rates (latency, recovery, patient zero probability) or introduce new rates to build the approximating $\{\mathcal{Q}^{\omega}\}$ family[3]. In general, we approximate the posterior using a generalized prior distribution. This is exactly the idea of the Causal Variational Approach [9], whose name comes from the fact that is a variational method which respects all the causal constraints of the prior distribution by definition. In the next section we build more organically the approximating family of functions for some specific models. Then we discuss how to minimize the KL divergence (which is a bit more tricky then in the naive MF case). Finally, we test the method against other existing ones.

2.3.1 CVA for epidemic models: the approximating functions

In this section we are going to describe the Causal Variational Approach (CVA) for the epidemic models. Let P_θ be the prior of an epidemic model, in which we explicitly wrote the dependence on the so called *hyper-parameters*, namely the patient zero probability, the infection, latency and recovery probabilities (or rates) depending on which model we choose, by writing the symbol θ. We know from equation (2.5) that the prior can be expressed as:

$$P_\theta(\underline{t}) = \prod_{i=1}^{N} \psi_\theta(\underline{t}_i, \underline{t}_{\partial i})$$

The function ψ_θ contains the hyper-parameters of the chosen model. Let x^* (or $\underline{t}^*$ in the equivalent transition time notation) be the planted configuration (i.e. the unknown epidemic we want to infer) and $\mathcal{O}$ be the set of all the observations taken from x^*. Let the posterior be the probability distribution

$$\mathcal{P}_\theta(\underline{t}|\mathcal{O}) = \frac{P_\theta(\underline{t})P_\theta(\mathcal{O}|\underline{t})}{P_\theta(\mathcal{O})}.$$

[3]the family should be built to contain the prior: there should exist an ω_p such that $\mathcal{Q}^{\omega_p} = P$

The Causal Variational Approach family of approximating functions for the posterior $\mathcal{P}$ is defined as:

$$\mathcal{Q}^{\omega}(\underline{t}) = \prod_i q^{\omega_i}(\underline{t}_i, \underline{t}_{\partial i})$$

where the functional form of q^{ω_i} is identical to ψ_θ except for the fact that each hyper-parameter is substituted by a set of inference parameters which we are going to introduce now for each model.

2.3.2 Markov SI and SEIR models

We now treat the CVA approximation for continuous and discrete-time markovian models. A special paragraph is instead dedicated to non-markovian models, which can also be easily treated with CVA, but are more tricky and need a small generalization, as shown later on. If the model is markovian, the prior's parameters to generalize are:

- Patient zero probability γ. We generalize this by introducing the set of probabilities $\{\gamma_i\}_{i=1}^N$. They approximate each individual's posterior probability of being the zero patient. When decreasing the KL divergence between the CVA approximation and the posterior, thus, we should end up, if the observation set is sufficiently informative, to a set $\{\tilde{\gamma}_i\}_{i=1}^N$ such that:
$$x_i^*(t=0) = \mathcal{I} \iff \tilde{\gamma}_i \text{ is high, } \forall i = 1, \dots, N.$$
- Infection probability (rate) λ. We could generalize it by introducing a set $\{\lambda_{ij}^t\}_{(i,j)\in\mathcal{E}}^{0\le t<T}$, each one representing the posterior probability (or rate in the continuous time models) to have infection along the edge (i,j) at time t. However, a slightly lighter parametrization is used in CVA, which simplifies the computations when minimizing of the KL divergence. The idea is to introduce a set of parameters $\{\lambda_i^t\}_{i=1,\dots,N}^{0\le t<T}$ which represents the incoming infection probability: λ_i^t is the probability (rate for the continuous case) for the individual i to be infected by one of its infectious neighbors. In other words, we are parametrizing the incoming infection probability to be homogeneous along the contacts: $\lambda_{ji}^t = \lambda_i^t, \forall j \in \partial i$. This should work because it allows to satisfy observations. For example,

consider for an SI model an individual i observed susceptible at time τ_1 and infectious at time $\tau_2 > \tau_1$. The CVA must assign $\gamma_i = 0$ and $\lambda_i^{t<\tau_1} = 0$ because the observation at time τ_1 imposed i to be $\mathcal{S}$ until τ_1. The observation at time τ_2, however, witnesses an infection event, which happened between τ_1 and τ_2. CVA thus sets $\lambda_i^{\tau_1<t<\tau_2} > 0$ in order to allow the individual i to be infected by one of its contacts. Notice that λ_i^t can be interpreted as an infection susceptibility of the individual i at time t. For the discrete time case, therefore, we can parametrize the posterior infection probability with the set $\{\lambda_i^t\}_{i=1,\dots,N}^{t=0,\dots,T-1}$ of $N \times T$ elements. For the continuous case, instead, we would have an infinite amount of parameters due to the infinite number of time steps. We therefore parametrize again each λ_i^t with a Gaussian function of 3 parameteres: peak, mean and standard deviation:

$$\lambda_i^t = \lambda_{i,p} e^{-\left(\frac{t-\lambda_{i,m}}{\lambda_{i,s}}\right)^2}$$

where $\lambda_{i,p}, \lambda_{i,m}, \lambda_{i,s}$ control respectively the peak, the mean and the width of the infection rate.

- Self-infection: it plays an important role in inference. It has been introduced as a parameter to compensate for graph ignorance. It is very useful to use it as a CVA parameter also when the network is completely known. In fact, it allows to justify infection events witnessed by observations which can not be easily explained by tuning the infection probabilities $\{\lambda_i^t\}_{i=1,\dots,N}^{0\leq t<T}$. We thus introduce a set of parameters $\{\alpha_i^t\}_{i=1,\dots,N}^{0\leq t<T}$, which can be thought as the posterior self-infection. As for the infection, in the discrete time case we have $N \times T$ parameters. For the continuous time case we instead have the $3N$ parameters $\{\alpha_{i,p}, \alpha_{i,m}, \alpha_{i,s}\}$ which are related to the self infection by:

$$\alpha_i^t = \alpha_{i,p} e^{-\left(\frac{t-\alpha_{i,m}}{\alpha_{i,s}}\right)^2}$$

- Latency: to apply CVA to the SEIR model we also need to generalize the latency probabilities (rates). This is done very similarly to the previous parameters described, by introducing a set $\{\nu_i^t\}_{i=1,\dots,N}^{0\leq t\leq T}$ which represents the posterior latency distributions of each individual. Also in this case

we re-parametrize the time dependency of the rates using a Gaussian distribution of parameters: $\nu_{i,p}, \nu_{i,m}, \nu_{i,s}$

- Recovery: the generalization for the recovery probability is formally identical to the latency. For the continuous time case, we have Gaussian functions with parameters $\mu_{i,p}, \mu_{i,m}, \mu_{i,s}$.

This is how the approximating CVA family of functions $\mathcal{Q}^{\omega}$ for the markovian case are built. Notice that the number of parameters scales linearly with N. The optimization process of those parameters is described later on in this section. We first introduce the approximating functions for the non-markovian case.

Generalization of CVA for non-Markov models

When the infection rate is non constant and depends on the infection time of the individual, see section 2.1.4, the parametrization described above does not work well. Indeed, the CVA infection rate introduced earlier can be thought as the susceptibility to infection, while the non-markovian infection rate, as explained in paragraph 2.1.4, is an outgoing infectiousness. To be as clear as possible, we add a subscript to the CVA infection rates and to the prior non-markovian infection rates, which become respectively $\lambda^t_{i,\mathrm{IN}}$ and $\lambda_{\mathrm{OUT}}(t-t_i^{\mathcal{I}})$ for each individual $i = 1, \dots, N$. The resulting infection rate along the edge $(i \to j)$ is defined as:

$$\lambda^t_{ij} := \lambda_{\mathrm{OUT}}(t - t_i^{\mathcal{I}})\lambda^t_{j,\mathrm{IN}}.$$

The CVA parameters, therefore, are unaltered with respect to the markovian case. Their interpretation changes. For the markovian case they are the approximate posterior infection rates. In the non markovian case, instead, they are multiplied to the prior non constant infection rate to have the total resulting rate λ^t_{ij}. To recap, to each individual i corresponds a set of parameters ω_i which is well defined for each of the models:

- SI, discrete time: $\omega_i = (\gamma_i, \{\alpha_i^t\}_{t=0}^{T-1}, \{\lambda^t_{i,\mathrm{IN}}\}_{t=0}^{T-1})$
- SEIR, discrete time: $\omega_i = (\gamma_i, \{\alpha_i^t\}_{t=0}^{T-1}, \{\lambda^t_{i,\mathrm{IN}}\}_{t=0}^{T-1}, \{\nu_i^t\}_{t=0}^{T-1}, \{\mu_i^t\}_{t=0}^{T-1})$
- SI, continuous time: $\omega_i = (\gamma_i, \alpha_{i,p}, \alpha_{i,m}, \alpha_{i,s}, \lambda_{i,p}, \lambda_{i,m}, \lambda_{i,s})$

- SEIR, continuous time:

$$\omega_i = (\gamma_i, \alpha_{i,p}, \alpha_{i,m}, \alpha_{i,s}, \lambda_{i,p}, \lambda_{i,m}, \lambda_{i,s}, \nu_{i,p}, \nu_{i,m}, \nu_{i,s}, \mu_{i,p}, \mu_{i,m}, \mu_{i,s})$$

both for the Markov and for the non-Markov models.

2.3.3 Minimizing the CVA KL divergence

Now that the family of CVA approximating functions has been built, it is necessary to look for the best approximation belonging to that family. This is achieved by minimizing the KL divergence $D_{KL}(\mathcal{Q}^\omega||\mathcal{P})$ with respect to $\omega = (\omega_1, \ldots, \omega_N)$, where each ω_i is the set of parameters defined above for each model. The Kullback-Leibler divergence is:

$$D_{KL}(\mathcal{Q}^\omega||\mathcal{P}) = \int \mathrm{d}\underline{t}\mathcal{Q}^\omega(\underline{t}) \log \frac{\mathcal{Q}^\omega(\underline{t})}{\mathcal{P}(\underline{t})}.$$

The derivative of the divergence w.r.t. a generic parameter ω_i^r is:

$$\frac{\partial D_{KL}(\mathcal{Q}^\omega||\mathcal{P})}{\partial \omega_i^r} = \int \mathrm{d}\underline{t} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log \frac{\mathcal{Q}^\omega(\underline{t})}{\mathcal{P}(\underline{t}|\mathcal{O})} + \int \mathrm{d}\underline{t}\mathcal{Q}^\omega(\underline{t}) \frac{\partial \log \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r}.$$

Let us focus on the second addend of the l.h.s.

$$\begin{aligned}
\int \mathrm{d}\underline{t}\mathcal{Q}^\omega(\underline{t}) \frac{\partial \log \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} &= \int \mathrm{d}\underline{t}\mathcal{Q}^\omega(\underline{t}) \frac{1}{\mathcal{Q}^\omega(\underline{t})} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} = \\
&= \int \mathrm{d}\underline{t} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} = \\
&= \frac{\partial}{\partial \omega_i^r} \int \mathrm{d}\underline{t}\mathcal{Q}^\omega(\underline{t}) = \\
&= \frac{\partial}{\partial \omega_i^r} 1 = 0.
\end{aligned}$$

This implies that the derivative of the divergence is:

$$\begin{aligned}\frac{\partial D_{KL}(\mathcal{Q}^\omega||\mathcal{P})}{\partial \omega_i^r} &= \int \mathrm{d}\underline{t} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log \frac{\mathcal{Q}^\omega(\underline{t})}{\mathcal{P}(\underline{t}|\mathcal{O})} = \\ &= \int \mathrm{d}\underline{t} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} + \int \mathrm{d}\underline{t} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log P(\mathcal{O}) = \\ &= \int \mathrm{d}\underline{t} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} + \cancel{\log P(\mathcal{O}) \frac{\partial}{\partial \omega_i^r} \int \mathrm{d}\underline{t} \mathcal{Q}^\omega(\underline{t})} = \\ &= \int \mathrm{d}\underline{t} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})},\end{aligned}$$

where we have cancelled the last addend because the integral of the $\mathcal{Q}^\omega$ gives 1 due to normalization. Now the equation of the derivative is tractable. We have in fact eliminated the partition function $P(\mathcal{O})$ which is too hard to compute. To further manipulate this formula we multiply and divide by $\mathcal{Q}^\omega(\underline{t})$:

$$\begin{aligned}\frac{\partial D_{KL}(\mathcal{Q}^\omega||\mathcal{P})}{\partial \omega_i^r} &= \int \mathrm{d}\underline{t} \frac{\mathcal{Q}^\omega(\underline{t})}{\mathcal{Q}^\omega(\underline{t})} \frac{\partial \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} = \\ &= \int \mathrm{d}\underline{t} \mathcal{Q}^\omega(\underline{t}) \frac{\partial \log \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} = \\ &= \left\langle \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} \frac{\partial \log \mathcal{Q}^\omega(\underline{t})}{\partial \omega_i^r} \right\rangle_{\underline{t} \sim \mathcal{Q}^\omega}.\end{aligned}$$

The function $\mathcal{Q}^\omega$ is by definition identical in form to the prior:

$$\mathcal{Q}^\omega(\underline{t}) = \prod_i q^{\omega_i}(\underline{t}_i, \underline{t}_{\partial i}),$$

which implies that:

$$\begin{aligned}\frac{\partial D_{KL}(\mathcal{Q}^\omega||\mathcal{P})}{\partial \omega_i^r} &= \left\langle \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} \frac{\partial}{\partial \omega_i^r} \sum_j \log q^{\omega_j}(\underline{t}_j, \underline{t}_{\partial j}) \right\rangle_{\underline{t} \sim \mathcal{Q}^\omega} = \\ &= \left\langle \log \frac{\mathcal{Q}^\omega(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} \frac{\partial}{\partial \omega_i^r} \log q^{\omega_i}(\underline{t}_i, \underline{t}_{\partial i}) \right\rangle_{\underline{t} \sim \mathcal{Q}^\omega}. \qquad (2.8)\end{aligned}$$

We have therefore a form which, at fixed ω, is the average of polynomially-computable functions. This means that at fixed epidemic trajectory $\underline{t}$ and at fixed values of parameters ω, the quantity inside can be numerically evaluated. To compute the average we need to sample $\underline{t}$ from $\mathcal{Q}^\omega$. This can be efficiently

Algorithm 1 Sampling from time discrete SI model

Input: The parameters set ω, the contact graph G, the horizon time T.

- Initialize a $N \times T$ matrix x, to interpret as the state of each individual at each time. Initialize the matrix completely equal to $\mathcal{S}$
- **for** $i = 1 : N$
  - set $x_i^0 = \mathcal{S}$ with probability $1 - \gamma_i$ and $x_i^0 = \mathcal{I}$ with probability γ_i
- **for** $t = 1 : T$; **for** $i = 1, \dots, N$
  - **if** $x_i^{t-1} = \mathcal{S}$
    - **for** $j \in \partial i$
      - **if** $x_j^{t-1} = \mathcal{I}$ then set $x_i^t = \mathcal{I}$ with probability λ_i^{t-1}.
    - set $x_i^t = \mathcal{I}$ with self-infection probability α_i^{t-1}.
    - **else** set $x_i^t = \mathcal{I}$
- **Return** x

achieved because sampling from $\mathcal{Q}^\omega$ is as easy as sampling from the prior, due to their identical functional forms. We are now going to describe the sampling process. Later, we show how to minimize the KL divergence using a gradient descent procedure.

Sampling the trajectory

The form of the CVA approximating functions allows to efficiently sample from them. They represent indeed the prior stochastic processes, with the only difference of having other hyper-parameters. For example, for the SI time-discrete model, it is sufficient to follow the procedure described in Algorithm 1, in which we first sample the patients zero independently with probabilities $\{\gamma_i\}_{i \in V}$. Then we simulate, in a loop over time, the infection and self infection process for all the $\mathcal{S}$ individuals. Notice that there is a probability of $\prod_i (1 - \gamma_i)$ for the time-zero configuration to be populated only by $\mathcal{S}$ individuals. To avoid this possibility we must sample the zero-time configuration from Algorithm 2, which forces at least one individual to be the patient zero. To see how the algorithm work, call n_{pz} the number of patients zero. The idea is to sample

Algorithm 2 Sampling at least one patient zero.

Input: The parameters set ω, the contact graph $\mathcal{G}$, the configuration x.

- **for** $i = 1 : N$
  - set $x_i^0 = \mathcal{I}$ with probability $\gamma_i / \left(1 - \prod_{j \geq i}(1 - \gamma_j)\right)$; otherwise to $\mathcal{S}$
  - **if** $x_i^0 = \mathcal{I}$ then **break** the loop
- **for** $k = i + 1 : N$ (i.e. for the remaining population)
  - set $x_i^0 = \mathcal{I}$ with probability γ_i; otherwise to $\mathcal{S}$

from $P(x_i^0 = \mathcal{I} | n_{pz} > 0)$. We can compute this function:

$$\begin{aligned} P(x_i^0 = \mathcal{I} | n_{pz} > 0) &= \frac{P(x_i^0 = \mathcal{I}, n_{pz} > 0)}{P(n_{pz} > 0)} = \\ &= \frac{P(x_i^0 = \mathcal{I})}{P(n_{pz} > 0)} = \\ &= \frac{\gamma_i}{1 - \prod_j (1 - \gamma_j)}. \end{aligned}$$

To sample recursively from this probability distribution, we start from individual 1:

$$P(x_1^0 = \mathcal{I} | n_{pz} > 0) = \frac{\gamma_1}{1 - \prod_j (1 - \gamma_j)}$$

Say that we sample the individual 1 to be $\mathcal{S}$ at time zero from the distribution $P(x_1^0 = \mathcal{I} | n_{pz} > 0)$. Now for the individual 2 we want to sample from:

$$\begin{aligned} P(x_2^0 = \mathcal{I} | n_{pz} > 0, x_1^0 = \mathcal{S}) &= \frac{P(x_2^0 = \mathcal{I} | x_1^0 = \mathcal{S})}{P(n_{pz} > 0 | x_1^0 = \mathcal{S})} = \\ &= \frac{P(x_2^0 = \mathcal{I})}{1 - \prod_{i \geq 2}(1 - \gamma_i)} = \\ &= \frac{\gamma_2}{1 - \prod_{i \geq 2}(1 - \gamma_i)}. \end{aligned}$$

Applying recursively this reasoning leads to Algorithm 2. An optimization and generalization of Algorithm 1 is possible by using a Gillespie simulation, described in detail in Algorithm 3 for the more general case of the non-Markov SI model. The idea is to store the tentative infection times of all the individuals in a queue $\underline{t} = (t_1^{\mathcal{I}}, \ldots, t_N^{\mathcal{I}})$, and update them recursively until their value

Algorithm 3 The Gillespie algorithm for continuous time sampling. An example for non-Markov SI model.

Input: The parameters set ω, the contact graph G, the horizon time T.

- Initialize a queue $\underline{t}$
- Sample all the patients zero, $t_i^{\mathcal{I}} = 0$, using Algorithm 2
- **for** $i = 1 : N$ such that $t_i^{\mathcal{I}} > 0$
  - Extract a value of self infection time from the distribution $\{\alpha_i^t\}_{t\in\mathbb{R}^+}$ and set it equal to $t_i^{\mathcal{I}}$
- **Loop** over the increasingly sorted queue $\underline{t}$
  - Take the element $t_i^{\mathcal{I}}$ coming from the sorted queue
  - **for** $j \in \partial i$
    - Extract the tentative infection time $t_{i\to j}$ from the distribution $\{\lambda_{ij}^t\}_{t\in\mathbb{R}^+} = \{\lambda_{\text{OUT}}(t - t_i^{\mathcal{I}})\lambda_{i,\text{IN}}^t\}_{t\in\mathbb{R}^+}$
    - **if** $t_{i\to j} < t_j^{\mathcal{I}}$ then update setting $t_j^{\mathcal{I}} = t_{i\to j}$
    - **Remove** i from the queue and and save $t_i^{\mathcal{I}}$ as the infection time of i.

is correctly sampled. To this end, the first step consists in sampling all the patients zero by means of Algorithm 2. Secondly, self-infection events are sampled from the self-infection distribution $\{\alpha_i^t\}_{t\in\mathbb{R}^+}$. Finally, contagion events are sampled by extracting from the queue the individual j corresponding to the minimum infection time: $j = \arg\min_{i\in\{1,\dots,N\}} t_i^{\mathcal{I}}$. If an individual i tries to infect another individual j at time $t_{i\to j}$, then j updates its infection time by taking the minimum between $t_{i\to j}$ and its current value of $t_j^{\mathcal{I}}$. This way the list is updated recursively. We describe an efficient procedure to sample from continuous distributions (as the self-infection rate $\{\alpha_i^t\}_{t\in\mathbb{R}^+}$, the infection rate, the recovery and latency rates) in Appendix A.

Gradient descent

Once the sampling process is implemented, it is possible to numerically estimate the averages with respect to the CVA function. In particular, the derivative of $D_{KL}(\mathcal{Q}^\omega|\mathcal{P})$ in equation (2.8) can be computed. To optimize the Kullback-Leibler divergence, a gradient descent procedure can be implemented. The standard gradient descent would require to compute the partial derivative w.r.t. all the parameters and change each parameter of a small fraction ε of the opposite of the gradient:

$$\omega_i^{k,r+1} = \omega_i^{k,r} - \varepsilon \frac{\partial D_{KL}(\mathcal{Q}^\omega||\mathcal{P})}{\partial \omega_i^r}$$

where the superscripts r and $r+1$ stand for the iteration of the update. The quantity ε is a real number named learning rate. This procedure in principle leads to a minimum of the KL divergence. It is known, however, that directly descending the gradient might not be the fastest way to optimize a function [56–59]. In this context that is due to the existence of several scales for each derivative. Thus, the scale-free technique of the Sign Descender [59, 58] is preferred:

$$\omega_i^{k,r+1} = \omega_i^{k,r} - \varepsilon \left|\omega_i^{k,r}\right| \text{ sign}\left(\frac{\partial D_{KL}(\mathcal{Q}^\omega||\mathcal{P})}{\partial \omega_i^r}\right)$$

The interpretation of this technique is to change each parameter by a fraction ε of its current value, opposite to the sign of the derivative. Iterating this procedure it is possible to optimize the KL divergence.

CVA Free energy

In order to estimate if the KL divergence is actually decreasing, we can compute the *variational free energy*:

$$\begin{aligned} F[\omega] &= U[\omega] - S[\omega] \\ &= -\int \mathrm{d}\underline{t}\mathcal{Q}^{\omega}(\underline{t}) \log P(\underline{t})P(\mathcal{O}|\underline{t}) + \int \mathrm{d}\underline{t}\mathcal{Q}^{\omega}(\underline{t}) \log \mathcal{Q}^{\omega}(\underline{t}) = \\ &= \int \mathrm{d}\underline{t}\mathcal{Q}^{\omega}(\underline{t}) \log \frac{\mathcal{Q}^{\omega}(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} = \\ &= D_{KL}(\mathcal{Q}^{\omega}(\underline{t})||P(\underline{t})P(\mathcal{O}|\underline{t})). \end{aligned}$$

This quantity is closely related by the divergence with the posterior:

$$F[\omega] = D_{KL}(\mathcal{Q}^{\omega}||\mathcal{P}) - \log P(\mathcal{O}).$$

Notice that the variational free energy can be computed numerically because it does not require to evaluate $\log P(\mathcal{O})$. Once the iterative updating scheme has reached convergence (i.e. the variational free energy does not decrease further and starts oscillating), the optimization process can be stopped. A set of parameters $\tilde{\omega}$ is the approximate minimum point of the KL divergence. If we hypothesize that the CVA approach gives a good estimate of the posterior, namely $\mathcal{Q}^{\tilde{\omega}} \approx \mathcal{P}$, then variational free energy computed at $\tilde{\omega}$ gives an estimate of the free energy $F = -\log P(\mathcal{O})$ defined in equation (1.6):

$$\begin{aligned} F[\tilde{\omega}] &= D_{KL}(\mathcal{Q}^{\tilde{\omega}}||\mathcal{P}) - \log P(\mathcal{O}) \approx \\ &\approx D_{KL}(\mathcal{P}||\mathcal{P}) - \log P(\mathcal{O}) \\ &= -\log P(\mathcal{O}) = F. \end{aligned}$$

Marginalization

Once the KL is optimized, i.e. the parameters set $\tilde{\omega}$ is obtained, marginalizing is simple. It is sufficient to sample a set of M trajectories $(\underline{t}_1, \ldots, \underline{t}_M)$ from the CVA distribution $\mathcal{Q}^{\tilde{\omega}}$. Then we can treat each $\underline{t}_m$ as a sample from the posterior. Computing marginals simply consists to make histograms with these samples. Example: to compute the average infection time of individual i we

take the M samples and sample average their i_th component:

$$\left\langle t_i^{\mathcal{I}} \right\rangle = \frac{1}{M} \sum_{m=1}^{M} (\underline{t}_m)_i^{\mathcal{I}} ,$$

where $(\underline{t}_m)_i^{\mathcal{I}}$ is the infection time of the individual i for the m_th sample.

Hard to soft constraints

When working with hard constraints, logarithms can be infinite due to some zeros. For the MF case, these $\log 0$'s were pathological and related to causal correlations which the mean field method is not able to capture. For CVA, instead, they are only apparent problems which can actually be cured by introducing very small parameters to smooth the constraints. If we look at $D_{KL}(\mathcal{Q}^{\omega}||\mathcal{P})$, indeed, we see that there is a $\log P(\mathcal{O}|x)$, which is $\log 0$ if there is no false rate and x violates some constraints of $\mathcal{O}$. It is sufficient to introduce a small false rate, (order 10^{-5}) to completely cure the problem, without reasonably affecting the performance of CVA.

Parallelization

The minimization of the KL divergence reduces to sampling. This process can be done in parallel. Each CPU can sample a configuration and compute the quantity inside the average of equation (2.8). Then it is sufficient to average over the CPUs to find the final result. As a consequence, the CVA algorithm can be run in parallel.

Variance Reduction

To accelerate the optimization process it is also possible to resort to the variance reduction technique from reinforcement learning [60–62], which simply consists in modifying equation (2.8) by subtracting the term $\left\langle \log \frac{\mathcal{Q}^{\omega}(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} \right\rangle \left\langle \frac{\partial}{\partial \omega_i^r} \log q^{\omega_i}(\underline{t}_i, \underline{t}_{\partial i}) \right\rangle$,

which is zero because:

$$\left\langle \frac{\partial}{\partial \omega_i^r} \log q^{\omega_i}(\underline{t}_i, \underline{t}_{\partial i}) \right\rangle = \left\langle \frac{\partial}{\partial \omega_i^r} \log \mathcal{Q}^{\omega}(\underline{t}) \right\rangle = \\ = \int \mathrm{d}\underline{t} \frac{\mathcal{Q}^{\omega}(\underline{t})}{\mathcal{Q}^{\omega}(\underline{t})} \frac{\partial \mathcal{Q}^{\omega}(\underline{t})}{\partial \omega_i^r} = \\ = \frac{\partial}{\partial \omega_i^r} \int \mathrm{d}\underline{t} \mathcal{Q}^{\omega}(\underline{t}) = 0.$$

The derivative of the KL can thus be rewritten as:

$$\frac{\partial D_{KL}(\mathcal{Q}^{\omega}||\mathcal{P})}{\partial \omega_i^r} = \left\langle \left(\log \frac{\mathcal{Q}^{\omega}(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} - \left\langle \log \frac{\mathcal{Q}^{\omega}(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} \right\rangle \right) \frac{\partial}{\partial \omega_i^r} \log q^{\omega_i}(\underline{t}_i, \underline{t}_{\partial i}) \right\rangle_{\underline{t} \sim \mathcal{Q}^{\omega}} = \\ = \left\langle \left(\log \frac{\mathcal{Q}^{\omega}(\underline{t})}{P(\underline{t})P(\mathcal{O}|\underline{t})} - F[\omega] \right) \frac{\partial}{\partial \omega_i^r} \log q^{\omega_i}(\underline{t}_i, \underline{t}_{\partial i}) \right\rangle_{\underline{t} \sim \mathcal{Q}^{\omega}}.$$

This typically makes convergence easier in the update process.

Hyper-parameters inference

An interesting problem in epidemic inference is to reconstruct the hyper-parameters of the prior distribution, which are typically indicated using the letter θ. To do so, one typically takes the free energy and descends it, as described in section 1.2.1. It is important to notice that the approximate CVA distributions may contain some hyper-parameter of the prior. For example, in the non-markovian case, the approximating posterior rate λ_{ij}^t is the product of a CVA parameter $\lambda_{j,\mathrm{IN}}^t$ and the prior hyper-parameter $\lambda_{\mathrm{OUT}}(t - t_i^{\mathcal{I}})$. It is more correct in this case, therefore, to write the approximating functions as $\mathcal{Q}_{\theta}^{\omega}$, where θ is the set of hyper-parameters. This allows us to understand that both the posterior and its CVA approximation depend on the prior hyper-parameters θ. For this paragraph, therefore, we are going to explicitly write the dependence on the hyper-parameters θ as a subscript of the distributions. What we want to minimize w.r.t. θ and ω is:

$$F[\omega](\theta) = D_{KL}(\mathcal{Q}_{\theta}^{\omega}(\underline{t})||P_{\theta}(\underline{t})P(\mathcal{O}|t)).$$

Taking the derivative w.r.t. a generic CVA parameter ω_i^r brings to eq. (2.8). The derivative w.r.t a generic hyper-parameter θ_k is very similar, but some

contributions from the prior do not vanish:

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}F[\omega](\theta) &= \frac{\partial}{\partial\theta_k}\int \mathrm{d}\underline{t}\mathcal{Q}^{\omega}_{\theta}(\underline{t})\log\frac{\mathcal{Q}^{\omega}{}_{\theta}(\underline{t})}{P_\theta(\underline{t})P_\theta(\mathcal{O}|t)} = \\
&= \int \mathrm{d}\underline{t}\left(\frac{\partial}{\partial\theta_k}\mathcal{Q}^{\omega}_{\theta}(\underline{t})\right)\log\frac{\mathcal{Q}^{\omega}{}_{\theta}(\underline{t})}{P_\theta(\underline{t})P_\theta(\mathcal{O}|t)} + \cancel{\int \mathrm{d}\underline{t}\mathcal{Q}^{\omega}_{\theta}(\underline{t})\frac{\partial}{\partial\theta_k}\log\mathcal{Q}^{\omega}{}_{\theta}(\underline{t})} + \\
&\quad - \int \mathrm{d}\underline{t}\mathcal{Q}^{\omega}_{\theta}(\underline{t})\frac{\partial}{\partial\theta_k}\log P_\theta(\underline{t})P_\theta(\mathcal{O}|t) = \\
&= \int \mathrm{d}\underline{t}\left(\frac{\partial}{\partial\theta_k}\mathcal{Q}^{\omega}_{\theta}(\underline{t})\right)\log\frac{\mathcal{Q}^{\omega}{}_{\theta}(\underline{t})}{P_\theta(\underline{t})P_\theta(\mathcal{O}|t)} - \int \mathrm{d}\underline{t}\mathcal{Q}^{\omega}_{\theta}(\underline{t})\frac{\partial}{\partial\theta_k}\log P_\theta(\underline{t})P_\theta(\mathcal{O}|t) = \\
&= \left\langle \log\frac{\mathcal{Q}^{\omega}_{\theta}(\underline{t})}{P_\theta(\underline{t})P_\theta(\mathcal{O}|\underline{t})}\frac{\partial\log\mathcal{Q}^{\omega}_{\theta}(\underline{t})}{\partial\theta_k} - \frac{\partial\log P_\theta(\underline{t})P_\theta(\mathcal{O}|t)}{\partial\theta_k}\right\rangle,
\end{aligned}$$

where the vanishing term has already been studied for the gradient descent above. Now we have a first term which is identical in form (and treated accordingly) to the derivative w.r.t. the CVA parameters. The second term is new and represents the rate of change of the prior process due to a modification of the hyper-parameters.

2.3.4 Warm up. CVA for Conditioned Random Walk

Before diving into applications of the Causal Variational Approach to epidemic inference, we test it to the conditioned random walk in 1D. This is a toy model which can even be solved exactly. Nonetheless, it is useful to characterize the CVA by applying it to this simpler problem, before moving to epidemics. We define the random walk in 1D as the time-discrete stochastic process characterized by the presence of a walker jumping at each step to the right or to the left with equal probability[4]. Calling $x = (x^0, x^1, \ldots, x^T)$ a trajectory, we have:

$$\begin{aligned}
P(x) = \prod_{t=0}^{T-1}\begin{cases} p(x^{t+1} = x^t + 1) \\ p(x^{t+1} = x^t - 1) \end{cases} \quad &= \\
= \prod_{t=0}^{T-1}\begin{cases} \frac{1}{2} \\ \frac{1}{2} \end{cases} \quad &= \frac{1}{2^T}
\end{aligned}$$

[4]this model can be generalized to multiple dimensions and to a biased version in which the walker has some preferential directions[63].

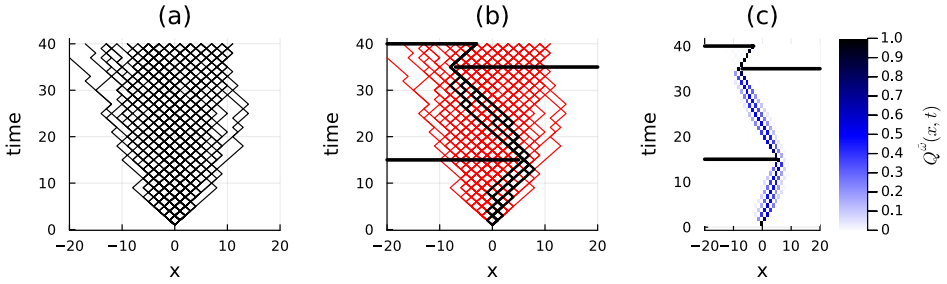

Fig. 2.1 Conditioned Random Walk. On the left (a), the random walk distribution. Each trajectory has the same probability. In the center (b), the walls are added and the effect is to select only some trajectories, plotted in black. All the trajectories that hit the walls (i.e. visit prohibited zones of the space-time) are discarded and plotted in red. On the write (c), the Causal Variational Approach reconstruction of the marginal probability of the walker to visit each particular zone of the space at fixed time. Figure taken from [9]

where x^0 is conventionally set to 0 and going to the left or the right means respectively decreasing or increasing of 1 unit, as in Figure 2.1 (a). Each $x^t \in \{-T, -T+1, \dots, 0, \dots, T-1, T\}$. Every trajectory has the same probability. From this process we can define the *conditioned random walk* by introducing a set of constraints in space and time: we constrain the walker not to visit certain zones of the space in certain periods of time. We can represent these constraints as walls in the 2D space-time, as in Figure 2.1 (b). . The original distribution might assign probability to the set of trajectories which violate the constraints. Let us define $\mathcal{W} = \{w_1, w_2, \dots, w_T\}$ the set of all the constraints (or walls). Each $w_t \subset \mathbb{R}$ is a set in the 1D space that the walker can not visit[5] at time t; in other words: $x^t \notin w_t, \forall t = 0, \dots, T$. We define the conditioned random walk as the distribution probability $\mathcal{P}(x|\mathcal{W})$ to have a trajectory which satisfies the constraints. Using Bayes' law:

$$\mathcal{P}(x|\mathcal{W}) = \frac{P(x)P(\mathcal{W}|x)}{\sum_{x'} P(x')P(\mathcal{W}|x')} = \\ = \frac{P(\mathcal{W}|x)}{\sum_{x'} P(\mathcal{W}|x')},$$

where

$$P(\mathcal{W}|x) = \prod_{t=1}^{T} \mathbb{I}[x^t \notin w_t]$$

[5] in this notation, $w^t = \emptyset$ means that no constraint is present at time t.

and where we simplified the constant $P(x) = 1/2^T$. We can think the walls as the result of observations. In fact, let us suppose that we had a detector which observed the walker with a precision which allows only to discard the zones in $\mathcal{W}$. In that case we can define the set $\mathcal{O} = \{o_1, \ldots, o_T\}$ as the complementary of $\mathcal{W}$, i.e. $o_t = \{-T, -T+1, \ldots, 0, \ldots, T-1, T\} \setminus w_t$. This allows us to interpret the conditioned random walk as an inference problem. Before seeing the CVA approximation of this problem, we describe its exact solution. We will then compare some marginals of CVA to the exact ones.

Exact solution to marginalization of the conditioned RW

We define the probability column:

$$\underline{p} = \begin{pmatrix} p_{-T} \\ \vdots \\ p_0 \\ \vdots \\ p_T \end{pmatrix}$$

where each element p_i represents the probability of having the walker in the position i, for $i \in \{-T, \ldots, T\}$. Notice that $\sum_i p_i = 1$. Now we define the transition probability matrix:

$$A = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \ldots \\ \frac{1}{2} & 0 & \frac{1}{2} & \\ 0 & \frac{1}{2} & 0 & \\ \vdots & & & \ddots \end{pmatrix}$$

which is an off diagonal matrix with elements representing the transition probabilities. The evolution in time is:

$$\underline{p}^{t+1} = A\underline{p}^t = A^2\underline{p}^{t-1} = \ldots A^{t+1}\underline{p}^0$$

where $\underline{p}^0$ is by definition concentrated on the origin, because it is the probability of a walker to be in a certain position at initial time:

$$\underline{p}^0 = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

Notice that the time evolution A preserves the normalization of the probability column. The unconditioned random walk marginals can be therefore computed by evaluating the probability column at fixed time and place:

$$p_i^t = \left(A^t \underline{p}^0\right)_i$$

When walls are added, the transition matrices are substituted by new matrices that are zero on the forbidden transitions. For each wall $w_t \in \mathcal{W}$ at time t we substitute A with a matrix B_{w_t} such that the elements are:

$$\left(B_{w_t}\right)_{ij} = \begin{cases} A_{ij} & \text{if } j \notin w_t \\ 0 & \text{if } j \in w_t \end{cases}$$

Notice that these new transition rates do not preserve the normalization. We have to normalize manually.

$$p_i^t = \frac{\left(\sum_{j=-T}^{T} \left(\prod_{k=t}^{T-1} B_{w_k} \underline{e}_i\right)_j\right) \left(\prod_{k=0}^{t-1} B_{w_k} \underline{p}^0\right)_i}{\sum_{j=-T}^{T} \prod_{k=0}^{T-1} B_{w_k} \underline{p}_i^0}$$

where $\underline{e}_i$ is the column which is zero everywhere except for the i_th element which is 1. Let us interpret this formula: to compute the marginal we evolved from the zero time state until time t, by multiplying the new transition matrices $\prod_{k=0}^{t-1} B_{w_k}$ to the initial time probability column $\underline{p}_0$. We evaluated the result at position i. That quantity is the probability for the walker to arrive at time t to position i. The walker, however, might hit some walls at future times $k > t$. The posterior conditioned probability must take it into account by evaluating the probability for a walker that at time t is in position i to *survive,* i.e. not to

hit any walls. We therefore evolve a walker which starts at time t at position i, represented with the probability column $\underline{e}_i$, until the end of the walk and we sum all over the possible final positions, in order to compute the probability to survive. We normalize everything by the probability for a walker starting at the origin to survive.

Comparison with Causal Variational Approach

The CVA simply treats the conditioned random walk as an inference problem, where the prior is the random walk distribution $P(x)$, the observations $\mathcal{O}$ are the complementary of the walls $\mathcal{W}$ and the approximating family of functions is again a random walk with some parameters generalization. Instead of having all the transitions identical to 1/2, we set them to be dependent on time and space. We have a set $\omega = \{\omega_i^t\}_{i=-T,\dots T}^{t=0,\dots T}$ of parameters which reproduce the random walk by setting them all to 1/2. For simplicity, we define each ω_i^t to be the probability to jump to the right if the walker is on site i at time t. The CVA approximating function is thus defined as:

$$\mathcal{Q}^\omega(x_1, x_2, \dots, x_T) = \prod_{t=0}^{T-1} \begin{cases} \omega_{x^t}^t & \text{if } x^{t+1} = x^t + 1 \\ 1 - \omega_{x^t}^t & \text{if } x^{t+1} = x^t - 1 \end{cases} = \\ =: \prod_{t=0}^{T-1} q(x^{t+1}, x^t; \omega_{x^t}^t).$$

Sampling from $\mathcal{Q}^\omega$ is trivial because it is a heterogeneous random walk. This allows to compute (and descend) the gradient of the divergence $D_{KL}(\mathcal{Q}^\omega|\mathcal{P})$:

$$\frac{\partial D_{KL}(\mathcal{Q}^\omega|\mathcal{P})}{\partial \omega_i^s} = \left\langle \sum_{t=0}^{T-1} \frac{\partial \log\left(q(x^{t+1}, x^t; \omega_{x^t}^t)\right)}{\partial \omega_i^s} \right\rangle_{x\sim\mathcal{Q}^\omega} = \\ = \left\langle \sum_{t=0}^{T-1} \frac{1}{q(x^{t+1}, x^t; \omega_{x^t}^t)} \frac{\partial q(x^{t+1}, x^t; \omega_{x^t}^t)}{\partial \omega_i^s} \right\rangle_{x\sim\mathcal{Q}^\omega} = \\ = \left\langle \sum_{t=0}^{T-1} \delta_{x^t,i}\delta_{t,s} \frac{1}{q(x^{s+1}, x^s; \omega_i^s)} \frac{\partial q(x^{s+1}, i; \omega_i^s)}{\partial \omega_i^s} \right\rangle_{x\sim\mathcal{Q}^\omega}.$$

this allows to find the best $\tilde{\omega}$ and to compute the approximate marginal posteriors, which are compared to the exact ones in Figure 2.2. In this simple 2D case (1D of space + 1D of time) it is possible to visualize the CVA parameters,

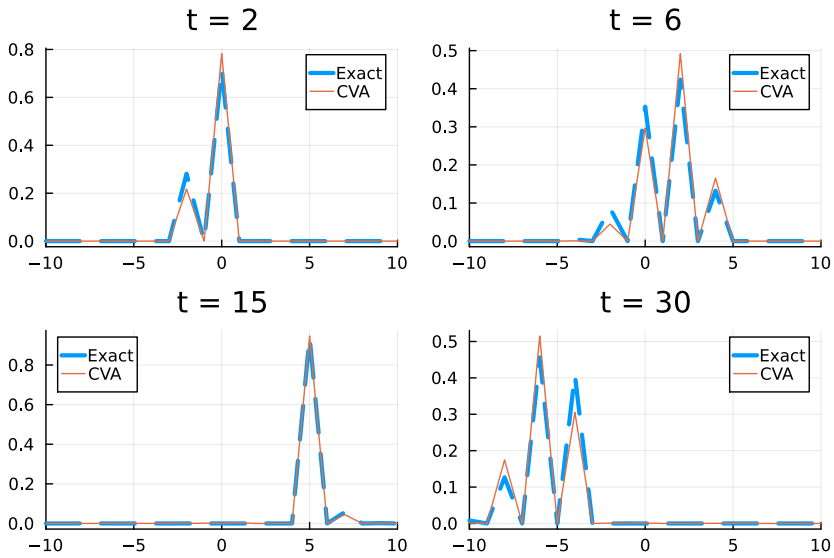

Fig. 2.2 The comparison between exact marginals and CVA approximation shows good agreement. The horizontal axis represents the space and the vertical axis the marginal probability for a walker to visit that zone at fixed time. The total walking time is equal to 40 and the walls are the same of Figure 2.1. Figure taken from [9]

which are represented in Figure 2.3 For the epidemic case the dimension of the space is too high and does not allow such a simple graphical representation. The epidemic case is, moreover, very interesting since it does not admit an exact solution computed in polynomial time.

2.4 Results for the Causal Variational Approach

This results section is partially based on the published paper [9]

We use now the CVA for studying epidemic inference. First, we test the validity of this approximation against other existing methods for the case of an SI model. To do so, we use the network simulators described in paragraph 2.1.4. We will see that CVA is robust and is the top performing method. We then test the ability of CVA to infer the hyper-parameters, by computing and descending the free energy landscape of the posterior distribution as a function of the hyper-parameters. Finally, we use CVA to conjecture a general result on epidemic

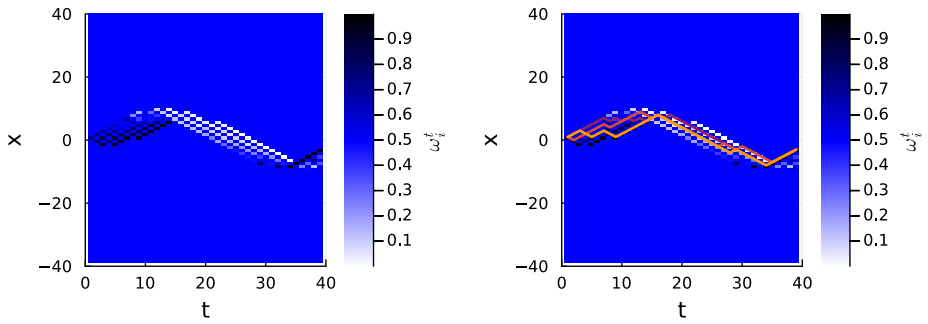

Fig. 2.3 The CVA parameters after the descent phase. *Left:* Each pixel represents the probability to jump up. A bright pixel signals therefore a high rate of going down, a black pixel means that the walker will jump probably up and blue means that CVA did not move the parameter from the initial 1/2 value. *Right:* three trajectories sampled from CVA distribution and superposed to the parameters. The total walking time is equal to 40 and the walls are the same of Figure 2.1. Figure taken from [9]

models, which we name *model reduction*: we show that it is almost equivalent to use SEIR or SI model to infer a hidden planted configuration which was generated with the SEIR model. This result suggests that the complexity of the model should not play a crucial role in inference performance.

2.4.1 Results on synthetic networks

In this first part of the results section we compare CVA to other existing methods in literature.

Other methods for inference

This section provides a list of the other inferential techniques whose performances are compared with those of Causal Variational Approach.

1. *Sib.* This method is based on a Belief Propagation approach to epidemic spreading processes [8, 6]. It is thoroughly introduced in Chapter 3, where its generalization to an ensemble method is introduced and used to derive general properties of the posterior. Sib performs well when epidemic models are on random contact networks, while it may suffer from the presence of loops in the graph.

2. *Backward-time Mean Field heuristic* . This method is based on a heuristic way to deal with observations from clinical tests and on a Mean Field approximation (different to the one introduced in 2.2.1) of the prior distribution, which is considered to be factorized over nodes at fixed time. An advantage of this method relies on its simplicity and small computational cost; however, it typically shows poorer performance with respect to the other methods. We refer to [6] for additional details about the Mean Field approximation. The heuristic scheme is instead discussed in the next point.

3. *Backward-time heuristic (with sampling)* – The MF method developed in [6] and described above relies, in addition to the MF ansatz for the prior distribution, on a heuristic way to deal with observations: it assumes that if an individual is tested $\mathcal{I}$ at time t, this means that it became positive at time $t - \tau$, with τ properly tuned (with the strength that performance seem approximately independent of τ). It is natural to wonder about the performance of this heuristic, regardless of the MF approximation. In [9], therefore, the MF estimation of marginal probabilities is substituted by sampling trajectories forward in time. As shown later on, this method performs almost the same of its MF correspondent. Therefore, the heuristic seems to be the limiting factor in the performance of the method.

4. *Monte Carlo.* Since epidemic trajectories can be fully described in terms of the infection time vector $\underline{t}$, the Markov Chain Monte Carlo (MCMC) in [9] defines dynamics on these continuous variables that eventually converge to a stationary distribution (i.e. the posterior $\mathcal{P}$). At each step of the MCMC, a node i is randomly selected and a new value of its infection time, denoted with $\hat{t}_i^{\mathcal{I}}$, is proposed, by drawing it from a probability $K(\hat{t}_i^{\mathcal{I}}|t_i^{\mathcal{I}})$. The initial condition for the Markov Chain is sampled from the prior distribution P. To diminish the effect of initial equilibration time an initial number of steps is typically required to let the MC forget the initial condition and sample efficiently the posterior distribution.

5. *Soft-Margin.* The Soft-Margin estimator is described in [64]. For the comparison with CVA the method is adapted by sampling from the prior probability distribution $P(x)$ and weighting each sample with the

observation likelihood $P(\mathcal{O}|x)$, for which a small artificial false rate is added with the aim of softening the constraints, resulting in an improvement of the method's performance. The Soft-Margin technique is asymptotically exact (except for the small false rate). However, when the population size grows, the probability to sample a trajectory x which satisfies the observation constraints $P(\mathcal{O}|x)$ dramatically decreases. Therefore the method is too slow for large population sizes.

Two of the aforementioned methods (Monte Carlo, and Soft-Margin) are asymptotically exact. This means that, if run for a sufficient amount of time, they provide the exact posterior. In order to have a fair comparison, the run-time of CVA, MCMC and Soft-Margin is approximately the same. Notice that while CVA and Soft-Margin are parallel algorithms, MCMC can only be used sequentially.

Results on Proximity graphs

We start by describing performance on small proximity graphs (see section 2.1.4 for their definition) of $N = 50$ individuals. To accumulate statistics, several graphs are simulated and for each one a planted epidemic trajectory x^* sampled from the prior $P(x)$. Then from each planted x^* several observation sets $\mathcal{O}_{n_1} \subset \mathcal{O}_{n_2} \subset \mathcal{O}_{n_3} \subset \mathcal{O}_{n_4} \subset \dots$ are built with increasing number of observations: n_k is the number of observations in the k-th observation set. Each method approximates then the posterior marginals of $\mathcal{P}(x|\mathcal{O}_k)$. Observations are noiseless, i.e. no false rate, and performed on a randomly chosen fraction of the population at a fixed time, which for the present case is the horizon time T. The AUC at initial time (which quantifies the ability of a method to reconstruct the patient-zero, see section 1.1.5) and at the final time (the so called *risk assessment*) are shown in Figure 2.4 as functions of the number n_k of observations available. As expected, in both cases the average performance of the methods improve when the number of observations increases. In particular, the soft-margin method is expected to converge to the exact results for this type of experiment when the number N of individuals is small. The results obtained with CVA are very close (and closer than any other technique) to those obtained by means of Soft Margin (denoted with the world *soft*), even in the regime with only a few observations.

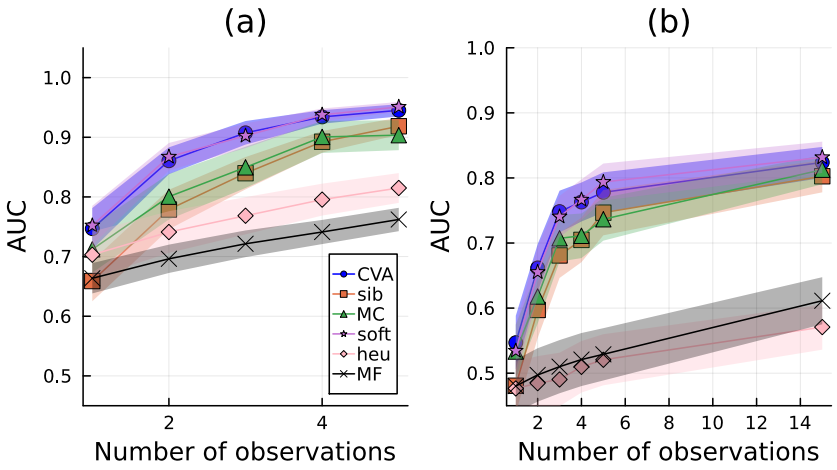

Fig. 2.4 Study of Area under the ROC (AUC) as a function of the number of observations at horizon time T and initial time 0, respectively in panel (a) and panel (b). The simulated contact graph is a proximity network with average connectivity $2.2/N$. For both simulations in panels (a) and (b), the total number of individuals is $N = 50$, the probability of being the patient zero is set to $\gamma = 1/N$, and the infection rate is $\lambda = 0.1$. For each epidemic realization, inference is performed for an increasing number of noiseless observations (here the false rate is 0) at time T . Thick lines and shaded areas indicate the averages and the standard errors computed over 40 instances. Figure taken from [9]

Open ABM and StEM networks

To further investigate the performances of CVA against the other techniques, two realistic dynamic contact network instances are considered, one generated using the *spatio-temporal epidemic model* (StEM) in [50] and the other using the discrete-time OpenABM model in [49] (see section 2.1.4). Epidemic realizations are generated using a continuous-time SI model on these contact graphs. For all different methods (CVA, Sib, Soft-Marg, and MCMC) the corresponding AUCs are shown as functions of time (in days), in Figure 2.5 (a), (c) and (b), (d) for OpenABM and the StEM respectively. Two different observation protocols are adopted in this comparison:

1. Observations times are randomly scattered with uniform distribution in the interval $[1, T]$. Moreover, observations are biased towards tested-positive outcomes to mimic a realistic scenario where symptomatic, which are a subset of the infected individuals, are more likely to be tested than susceptible ones.

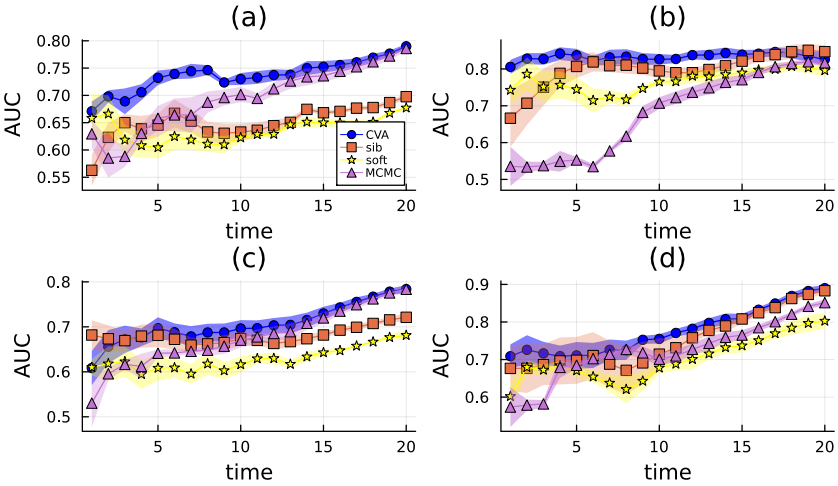

Fig. 2.5 A comparison of performances: Causal Variational Approach (CVA), Belief Propagation (sib) and SoftMargin (soft), and MCMC (MC) are compared by studying the AUC associated with the prediction of the infected individuals as a function of time on several instances of dynamic contact network generated using the OpenABM (in panel (a) $N = 2000$, in (c) $N = 1000$) and the StEM (panels (b) and (d)) for $N = 904$. The infection rate is set to $\lambda = 0.15$ for the latter and $\lambda = 0.02$ for the former. Observations' false rate is zero in both cases. For panels (c) and (d), observations are performed at the last time of the dynamics, i.e. the horizon time T. For the results in panels (a) and (b), observation times are extracted uniformly in the range $[1, T]$; at each observation time, infected nodes are observed with a biased probability equal to $1.1 n_{\mathcal{I}}(t)/N$where $n_{\mathcal{I}}(t)$ is the number of infectious individuals at time t and N is the total number of individuals. The total number of observations is $n_{obs} = 0.1\, N$ for OpenABM and $n = 100$ for the StEM. Figure taken from [9]

2. Observations are performed at the horizon time T with no bias due to sympthoms.

Panels (a) and (b) are associated with the observations scattered in time, while panels (c) and (d) correspond to observations at the last time only. In panel (a) simulations are run for $N = 2000$, while in panel (c) the total number of individuals is $N = 1000$. It is easy to see that, in panels (a) and (c), CVA (blue dots) is the best-performing method in terms of AUC. Only MCMC (pink triangles) reaches comparable AUC around $t = T$. The results achieved by Belief Propagation (Sib) are similar to those produced by CVA when the size of the graph is $N = 1000$ (panel (c)). However, for $N = 2000$, they significantly deteriorate (panel (a)). For the instances generated according to StEM, the comparison reveals that CVA achieves the largest values of the AUC at all times

and Belief Propagation (sib, orange squares) performs comparably only at the horizon time. MCMC approaches CVA performances in the last time epochs, while it is not able to predict the zero patient. Indeed, the AUC associated with MCMC predictions for all parametrizations is slightly larger than 0.5 (which correspond to random guess) for $t < 5$ when observations are performed at the horizon time.

2.4.2 Results on hyper-parameter inference

A crucial task in epidemic inference is to estimate the hyper-parameters. It is in fact usual to have no prior information on the e.g. infection rate and recovery rate. In this paragraph we aim to visualize the process of parameters inference made by CVA (see section 2.3.3). To do so, we take an SI model with patient zero and infection probability respectively fixed to γ^* and λ^* and generate a planted trajectory x^*, from which we take some observations $\mathcal{O}$. We then explore the entire space of hyper-parameters $\Theta = \{(\gamma, \lambda) \in [0,1] \times [0,1]\}$: for each couple γ, λ we find the CVA approximation of the posterior and its corresponding free energy. We end up with a free energy landscape as in Figure 2.6. If the approximation of the free energy is good, then the minimum of the landscape should coincide, or at least be near to the correct hyper-parameter couple γ^*, λ^*. In formulae, for each (γ, λ) we minimize the approximating family of CVA functions $\{\mathcal{Q}^{\omega}_{(\gamma,\lambda)}\}$, obtaining $\mathcal{Q}^{\tilde{\omega}}_{(\gamma,\lambda)}$, then we find the corresponding free energy $F[\tilde{\omega}](\gamma, \lambda) = D_{KL}(\mathcal{Q}^{\tilde{\omega}}_{(\gamma,\lambda)}(\underline{t})||P_{(\gamma,\lambda)}(\underline{t})P(\mathcal{O}|t))$ and we plot it. On the same plot we pin the hyper-parameter couple (γ^*, λ^*) from which $\mathcal{O}$ has been generated. Of course, the actual method for CVA to infer the hyper-parameters is much faster than reconstructing the whole free energy landscape, as described in paragraph 2.3.3: CVA simply descends the KL simultaneously w.r.t. its parameters ω and the hyper-parameters $\theta = (\gamma, \lambda)$. To visualize and test the efficacy of this procedure, we start from some different initial conditions (γ_0, λ_0) and for each initial condition we perform and plot the descent: in few steps it reaches the minimum zone, which corresponds to the correct hyper-parameters couple (γ^*, λ^*). We conclude that CVA simultaneous descent of the parameters and hyper-parameters is effective. Being able to infer hyper-parameters is a crucial task, not only for practical applications, but also for theoretical developments, as described in the next paragraph.

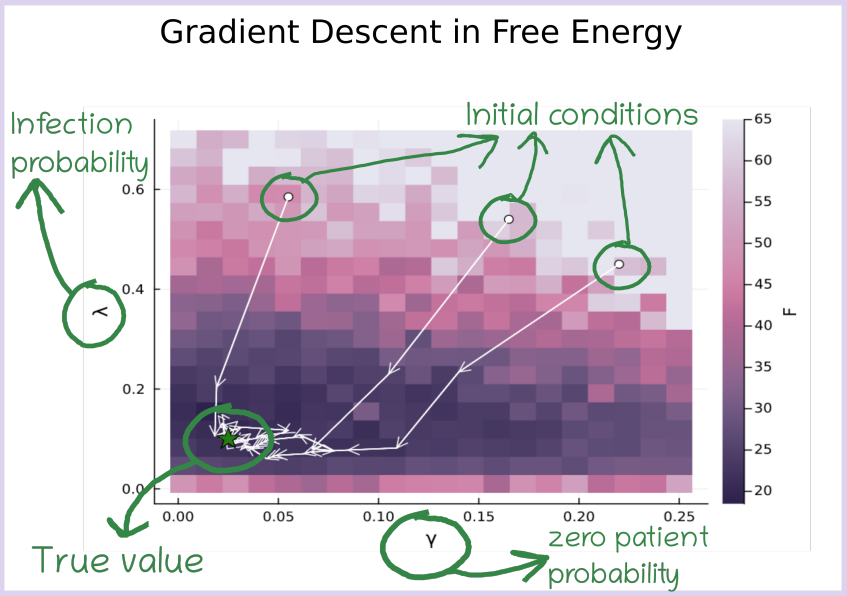

Fig. 2.6 Heat map of the free energy landscape as a function of the hyper-parameters of the generative SI model. The experiment is performed on a proximity graph with $N = 50$ individuals and density $\rho = 2/N$; the epidemic model is characterized by patient zero probability $\gamma^* = 1/N$ and infection rate $\lambda^* = 0.1$, shown here as a green star. We perform a large number of observations ($n_{obs} = 2N$) at uniformly randomly distributed times. The lowest values of this free energy landscape are concentrated around the exact hyper-parameters couple (γ^*, λ^*). The oriented paths (white arrows) represent the gradient descent dynamics of the algorithm in the hyper-parameters space, starting from three different initial conditions (γ_0, λ_0). Figure taken from [9]

2.4.3 Model reduction

Inference of hyper-parameters is now used to show a result which hints at a possible simplification in the epidemic inference problem: the model reduction, i.e. the possibility to use simple models to infer dynamics generated by more complicated models. In particular, the planted trajectories in this numerical experiment are generated with a SEIR model and they are inferred using two different prior models:

- a non-Markov SI model;
- a SEIR model.

If model reduction holds, then the results obtained by means of SI and SEIR inference should be similar to each others. The results in Figure 2.7 represent four different kinds of possible tests of the model reduction conjecture. All of them have in common the idea to use a SEIR model with a fixed set θ^* of hyper-parameters to generate the planted x^*, from a set $\mathcal{O}$ is extracted. Then two estimations are made, one based on the SI model CVA and one with the SEIR model CVA. None of them have access to θ^*. In both cases, thus, the hyper-parameters have to be inferred. The SI model CVA infers some effective hyper-parameters $\tilde{\theta}_{\mathrm{SI}}$. The SEIR model CVA infers a set $\tilde{\theta}_{\mathrm{SEIR}}$. The model reduction conjecture claims that the effective hyper-parameters $\tilde{\theta}_{\mathrm{SI}}$ and $\tilde{\theta}_{\mathrm{SEIR}}$ should reproduce approximately the same dynamics. To test this, four experiments are reported here:

1. The number of observations n_{obs} is a small fraction of the population, $n_{obs} = N/10$. In this regime, represented in panel (a), the *SEIR posterior*, namely the CVA approximate posterior $\mathcal{Q}^{\tilde{\omega}}_{\theta^*}$, corresponding to the hyper-parameters θ^* is strictly the best performing, as expected. This means that the number of observations is not sufficient to fully reconstruct the prior. Interestingly, the performance (measured with AUC over time) of the CVA approximations which do not have access to the hyper-parameters, namely $\mathcal{Q}^{\tilde{\omega}_{\mathrm{SI}}}_{\theta_{\mathrm{SI}}}$ for the SI and $\mathcal{Q}^{\tilde{\omega}_{\mathrm{SEIR}}}_{\theta_{\mathrm{SI}}}$ for the SEIR, are similar! This experiment goes therefore in the direction of the model reduction.
2. The number of observation is high: $n_{obs} = N/2$. The hyper-parameter-inferring approximations perform equally to the CVA SEIR distribution

$\mathcal{Q}^{\tilde{\omega}}_{\theta^*}$. Also in this case the SI and SEIR solutions perform the same. See panel (b)

3. The average number of infectious individuals over time can also be studied. This number is a quantity which depends on the prior distribution P^θ. Therefore, it strongly depends on the hyper-parameters. The correct average number of infectious individuals corresponds to θ^*, which is labeled in the plot as the *SEIR prior*. In panel (c) an observation set is taken in order to be typical, i.e. the number of observed infectious individuals over time is proportional to the number infectious individuals of the planted x^*. In this case, both the SI and the SEIR hyper-parameters-inferring methods reproduce results comparable with the correct SEIR prior.

4. In panel (d), the previous experiment is repeated. This time, however, the observations set is biased, namely in this set there is a much higher fraction of infectious individuals than in the planted. The two prior-inferring methods both fail to reproduce the correct result for the average number of infectious individuals. However, also this result is in the direction of the model reduction because they both reproduce the same incorrect number of infected individuals over time.

Of course, this is only a preliminary result which should be refined by e.g. analyzing a comparison between SI model and very complex models as the open ABM in [49]. Moreover, the tests performed so far are on single instances with CVA, which is an approximate algorithm. A more solid result would require to analyze the exact posterior for each possible graph, planted configuration and observation. This is a bit far away from the technical possibilities we have at the moment. However, a first step into classifying general posterior properties is the topic of Chapter 3.

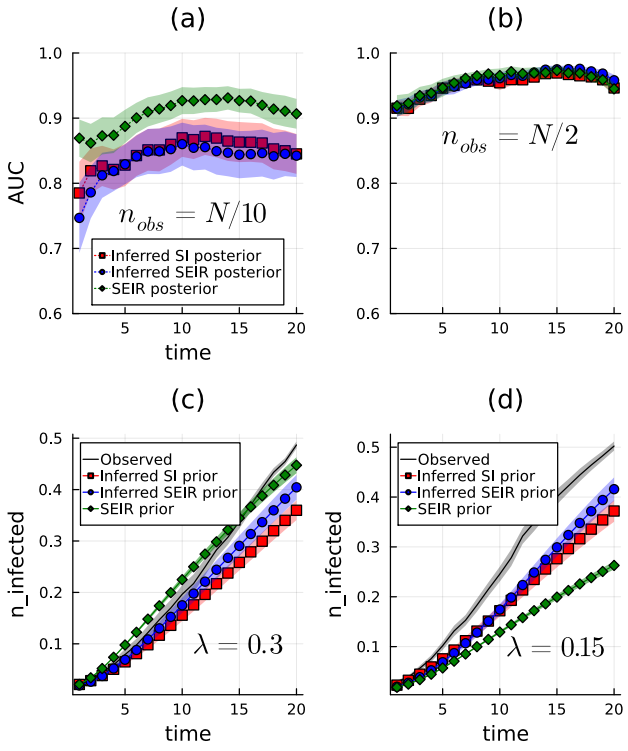

Fig. 2.7 Test of model reduction from a SEIR to an SI model. The numerical experiments are made on a proximity graph with $N = 100$ individuals and density $\rho = 2.2/N$. The observed epidemic realizations are generated using a SEIR model with $\gamma^* = 1/N$, $\lambda^* = 0.3$ (panels (a) and (b) and (c)) and $\lambda^* = 0.15$ (panel (d)), latency delay $\nu^* = 0.5$ and recovery delay $\mu^* = 0.1$. In panel (a) the AUC is plotted as a function of time and the number of observations is $n_{obs} = N/10$. In panel (b) the number is $n_{obs} = N/2$. In both panels (a) and (b) the three different inferred posterior CVA distributions are compared: the SEIR CVA posterior with correct hyper-parameters (green diamonds), the SEIR CVA posterior with inferred hyper-parameters (blue circles) and a SI CVA posterior with inferred hyper-parameters (red squares). Shaded areas represent the error around the average value, computed using 22 instances. In panels (c) and (d) the average fraction of infectious individuals as a function of time estimated using the correct SEIR prior model (green diamonds), a SEIR prior with the inferred hyper-parameters (blue circles), and a SI prior model with the inferred hyper-parameters (red squares). The regimes shown correspond to unbiased observations (panel (c) $\lambda = 0.3$), and to observations preferentially sampled from large outbreaks (panel (d), $\lambda = 0.15$). The black curves represent the average number of infectious of the planted trajectory. Shaded areas represent the standard error computed from 40 realizations of the dynamics. Figure taken from [9]

Chapter 3

Thermodynamic ensemble results

So far, we have studied the inference problem at a single instance level. This means that we had a fixed contact graph G, a fixed planted configuration $\underline{t}^*$ and a fixed observations set $\mathcal{O}$. The aim was to reconstruct $\underline{t}^*$ by approximating the posterior distribution. This chapter provides theoretical information bounds to reconstruction, characterizing the epidemic regimes which are harder/easier to infer. It is obvious that increasing the number (or decreasing the noise) of observations always improves the quality of inference, but it is less clear is the relation between inference quality and e.g. infection rate, patient zero probability, network density. To answer these questions we need to find average results (we don't want our claims to be dependent on a particular network or a specific epidemic trajectory). However, naively averaging over all the (infinite!) set of graphs, planted trajectories and observations is unfeasible. We need to:

- average over specific sets or *ensembles* of graphs, e.g. random regular graphs, Erdős–Rényi, etc...
- resort to finer techniques which allow to average over infinite sets.

The technique used is the *replica symmetric cavity metho*d [65, 66], introduced later on (section 3.1.6) as an ensemble version of the *belief propagation* (BP) algorithm (see for example [37], *Belief Propagation*, page 291). This chapter starts introducing BP in general, as a method for (approximately) compute

marginals of high-dimensional probability distributions. We illustrate an application of BP to single instance inference, introducing *sib* [8], an algorithm for risk assessment and patient zero reconstruction based on BP. The aim of this Chapter is to present an ensemble generalization of sib, which we call here *epidemble* (epidemic ensemble) [44], an algorithm based on the replica symmetric cavity method and aimed at studying ensemble properties of the posterior marginals.

3.1 Belief Propagation

We now give a brief overview of Belief Propagation (BP), mainly following [37], [34] and [67]. Consider the problem of computing marginals of a probability distribution $P : \mathcal{X} \to [0, 1]$ of the form:

$$P(x) = \frac{1}{Z} \prod_{a=1}^{M} \psi_a(\boldsymbol{x}_a), \tag{3.1}$$

where $x = (x_1, \ldots, x_N) \in \mathcal{X}$ and $\boldsymbol{x}_a$ is the subset of $(x_1, \ldots, x_N)$ which the function ψ_a depends on. This functional form includes equation (2.7), by setting $x = \underline{t}$, $\boldsymbol{x}_a = \underline{t}_i, \underline{t}_{\partial i}$, $M = N$. This is thus a more general form w.r.t. epidemic posterior and can be interpreted graphically as a factor graph.

3.1.1 Factor Graph

It is possible to associate a graph to probability distributions as the one in equation (3.1): it is called *factor graph* and it is a bipartite graph, i.e. its vertices belong to two disjoint sets and the edges only connect vertices of one set with vertices of the other. To build the factor graph corresponding to equation (3.1), take all the functions $\{\psi_a, a = 1, \ldots, M\}$ and draw a square vertex for each one. Label the square which corresponds to the function ψ_a with the letter a. Then, draw a circular vertex for each of the coordinates $\{x_i, i = 1, \ldots, N\}$ of the domain variable x. We have now M squares and N circles. Now, connect each square a to the circles corresponding to the arguments of ψ_a. The factor graph is built, see Figure 3.1. The square vertices are called *function nodes,* while the circular vertices are the *variable nodes.*

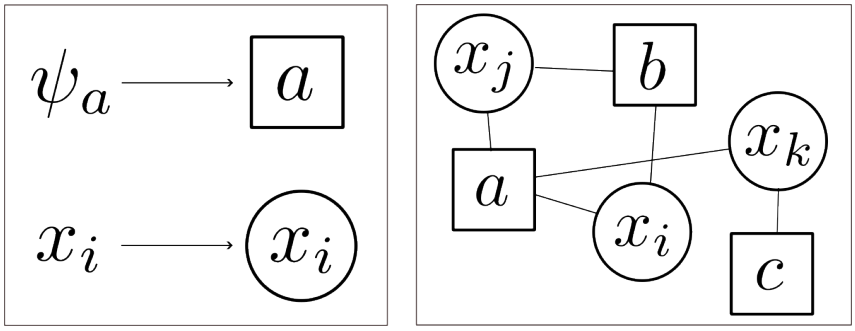

Fig. 3.1 Construction of the factor graph. *Left:* for each function ψ_a a square vertex (function node) a is drawn. At each variable x_i is associated a circular vertex (variable node). *Right:* a small example of factor graph, in which $\psi_a = \psi_a(x_i, x_j, x_k), \psi_b = \psi_b(x_i, x_j)$ and $\psi_c = \psi_c(x_k)$.

Finally, we define the neighborhood ∂a of the function node a as the set of variable nodes attached to a, i.e.

$$\partial a = \{i \in \{1, \dots, N\} : x_i \text{ is an argument of } \psi_a\}.$$

Similarly, we define:

$$\partial i = \{a \in \{1, \dots, M\} : \psi_a \text{ contains } i \text{ as an argument}\}$$

3.1.2 BP update equations

Suppose we want to compute marginals of the distribution in equation (3.1). If the functions $\{\psi_a, a = 1, \dots, M\}$ are such that the associated factor graph has no loops, then it is possible to build an algorithm which exactly marginalizes the distribution. The idea is that, if no loop is present, then the process of cutting one edge separates the graph in two disjoints sub-graphs. This implies that two variable nodes i, j attached to the same function node a become independent if we *remove* the factor a, as shown in Figure 3.2. Identically, removing a variable node separates the function nodes attached to it. We define, for a couple of nodes node η, ζ, regardless of being function or variable nodes, the sub-graph :

$$S_\eta^{(\zeta)} = \{(i, a) \text{ connected to } \eta, \forall\, i \in \{1, \dots, N\}, \forall\, a \in \{1, \dots, M\} \text{ if } \zeta \text{ is removed}\}$$

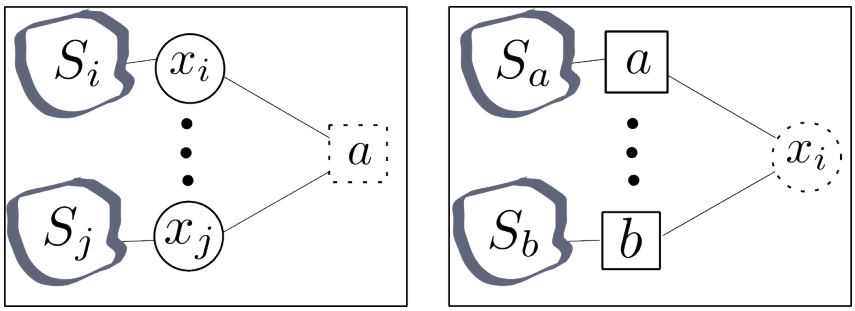

Fig. 3.2 Removing a node from the tree factor graph separates it in disjoint sub-graphs. *Left:* removing a function node separates the nodes of ∂a. In this illustration we represented i, j and their attached sub-graphs S_i and S_j. The black dots represent the other nodes in ∂a. The sub-graphs attached to the nodes in ∂a are separated if a is removed. *Right:* if a node x_i is removed, then the factors attached to it separate and their attached sub-graphs become disjoint.

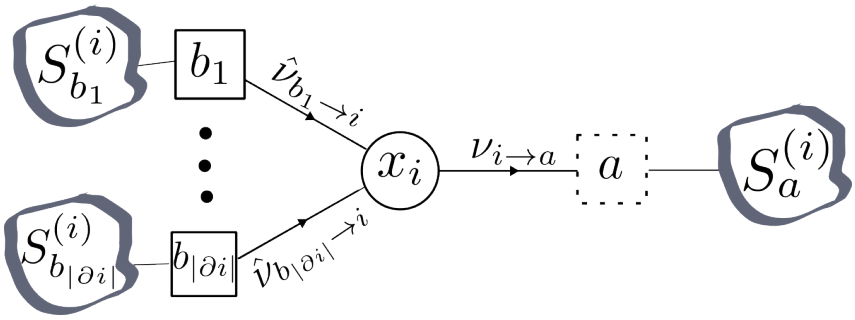

Fig. 3.3 BP message from i to a. The cavity is made by removing factor a.

We write $S^{(\zeta)}_\eta = (V^{(\zeta)}_\eta, F^{(\zeta)}_\eta)$, where $V^{(\zeta)}_\eta$ and $F^{(\zeta)}_\eta$ are respectively the set of variable nodes and the set of function nodes in the sub-graph $S^{(\zeta)}_\eta$. Now we can derive the Belief propagation equations. We define for each edge (i, a) the quantity $\nu_{i\to a}(x_i)$ as the marginal distribution of x_i in the factor graph from which a is removed. The quantity $\nu_{i\to a}(x_i)$ is called *cavity marginal* because it has been defined by removing a node (namely by generating a cavity). To have an expression for the cavity marginal, we marginalize the distribution obtained by removing the function node a from equation (3.1). The cavity marginal $\nu_{i\to a}(x_i)$ is thus:

$$\nu_{i\to a}(x_i) \propto \sum_{x\setminus x_i} \prod_{b\in\{1,\dots M\}\setminus a} \psi_b(\boldsymbol{x}_b).$$

Where the proportionality symbol is w.r.t. x_i. Looking at Figure 3.3, we understand that removing a separates the sub-graph $S^{(i)}_a$ attached to a from

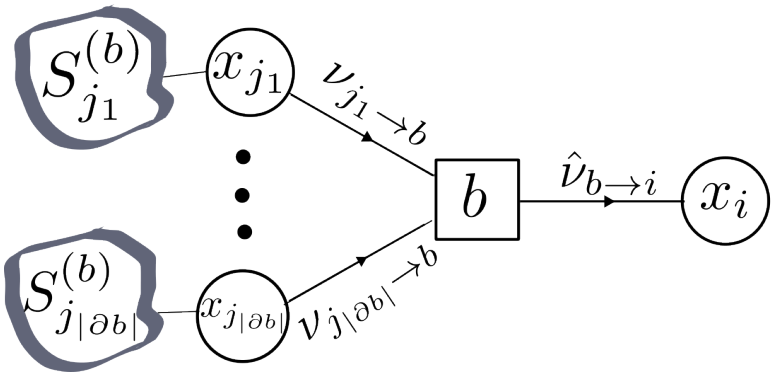

Fig. 3.4 The BP function-to-variable message from b to i. The cavity is made by removing all the factors but b attached to x_i.

the rest of the graph. Therefore, the sum over all the variable nodes in $S_a^{(i)}$ does not depend on x_i and it is just a proportionality term. We have:

$$\nu_{i \to a}(x_i) \propto \prod_{b \in \partial i \setminus a} \sum_{\{x_j\}_{j \in V_b^{(i)}}} \psi_b(\boldsymbol{x}_b) \prod_{c \in F_b^{(i)}} \psi_c(\boldsymbol{x}_c)$$

in which we switched the sum with the product because the sets $S_b^{(i)}$ for $b \in \partial i \setminus a$ are disjoint due to the absence of loops. Defining:

$$\hat{\nu}_{b \to i}(x_i) \propto \sum_{\{x_j\}_{j \in V_b^{(i)}}} \psi_b(\boldsymbol{x}_b) \prod_{c \in F_b^{(i)}} \psi_c(\boldsymbol{x}_c), \tag{3.2}$$

we have:

$$\nu_{i \to a}(x_i) = \frac{1}{z_{i \to a}} \prod_{b \in \partial i \setminus a} \hat{\nu}_{b \to i}(x_i). \tag{3.3}$$

This is the first BP equation. The distribution $\hat{\nu}_{b \to i}(x_i)$, as written in equation (3.2), is exactly equal to the marginal distribution of x_i if the node x_i was attached only to b. Also the quantity $\hat{\nu}_{b \to i}(x_i)$, therefore, can be interpreted as a cavity marginal: this time the cavity is made by removing all the function nodes but b attached to x_i. To distinguish between the sets of cavity marginals $\{\nu_{i \to a}\}_{i=1,\dots,N}^{a \in \partial i}$ and $\{\hat{\nu}_{a \to i}\}_{i=1,\dots,N}^{a \in \partial i}$ we respectively call them *variable-to-function messages* and *function-to-variable messages*. To rewrite the expression of $\hat{\nu}_{b \to i}(x_i)$ only in terms of local [1] functions, we take its definition we rearrange

[1] i.e. involving only nodes attached to b

products and sums keeping in mind the graphical interpretation in Figure 3.4:

$$\hat{\nu}_{b\to i}(x_i) \propto \sum_{\boldsymbol{x}_b \setminus x_i} \psi_b(\boldsymbol{x}_b) \prod_{j\in\partial b} \sum_{\{x_j\}_{j\in S_j^{(b)}}} \prod_{c\in F_j^{(b)}} \psi_c(\boldsymbol{x}_c).$$

The equation might appear messy, but if we compare it with Figure 3.4 we see that the first sum in the equation is over the first layer of variables $\boldsymbol{x}_b \setminus x_i$; then, for each variable j in this layer we sum over the sub-graph attached to it. This sum is the marginal of each x_j (for $j \in \partial b$) if we remove the factor b from the graph. Thus, by definition of variable-to-function messages:

$$\hat{\nu}_{b\to i}(x_i) = \frac{1}{z_{b\to i}} \sum_{\boldsymbol{x}_b \setminus x_i} \psi_b(\boldsymbol{x}_b) \prod_{j\in\partial b} \nu_{j\to b}(x_j). \tag{3.4}$$

This is the second BP equation. In equations (3.3) and (3.4) we have eliminated the proportionality symbol by defining the normalization constants:

$$z_{i\to a} = \sum_{x_i} \prod_{b\in\partial i\setminus a} \hat{\nu}_{b\to i}(x_i)$$
$$z_{b\to i} = \sum_{\boldsymbol{x}_b} \psi_b(\boldsymbol{x}_b) \prod_{j\in\partial b} \nu_{j\to b}(x_j),$$

which are simply obtained by summing over x_i the numerator. The system of equations (3.3) and (3.4) defines the update rule for BP: we can initialize the messages $\{\nu_{i\to a}\}_{i=1,\dots,N}^{a\in\partial i}$ and $\{\hat{\nu}_{a\to i}\}_{i=1,\dots,N}^{a\in\partial i}$, e.g. to constant distributions and then iterate equations (3.3) and (3.4) until a fixed point is found. From the fixed point messages, as shown in the net paragraph, it is possible to compute marginals of the distribution (3.1). For a tree, this fixed point is always reached and it is exact [37], i.e. all the marginals computed with BP are the exact marginals of equation (3.1).

3.1.3 Marginals

The BP scheme allows to compute marginals as functions of the messages. For example, to compute the marginal distribution of the node x_i, i.e. $p_i(x_i) = \sum_{x\setminus x_i} P(x)$, we can add a foo function node $\tilde{a}_i$ to the original factor graph by attaching it to the only variable x_i. The marginal of $p_i(x_i)$ is then equal to the

variable-to-function message $\nu_{i\to\tilde{a}_i}(x_i)$. In fact, this is the marginal of x_i after the removal of the foo node $\tilde{a}_i$, which is exactly what we are looking for. In conclusion:

$$p_i(x_i) = \sum_{x\setminus x_i} P(x) = \\ = \frac{1}{z_i} \prod_{a\in\partial i} \hat{\nu}_{a\to i}(x_i)$$

and we can forget about the foo function node $\tilde{a}_i$. Note that we defined

$$z_i = \sum_{x_i} \prod_{a\in\partial i} \hat{\nu}_{a\to i}(x_i)$$

Similarly, it is possible to compute the joint marginal of several variables [37]. An important example, which we use for the computation of energy, is the joint marginal of the variables ∂a attached to a function node a:

$$p_{\partial a}(\boldsymbol{x}_a) = \frac{1}{z_a}\psi_a(\boldsymbol{x}_a) \prod_{i\in\partial a} \nu_{i\to a}(x_i),$$

where:

$$z_a = \sum_{\boldsymbol{x}_a} \psi_a(\boldsymbol{x}_a) \prod_{i\in\partial a} \nu_{i\to a}(x_i).$$

The Belief Propagation scheme is a useful tool to compute marginals. In the next paragraph we are going to estimate thermodynamic quantities using the BP algorithm. Being able to evaluate the free energy, for example, is of extreme importance for inferring the hyper-parameters.

3.1.4 BP estimation of thermodynamic quantities

We are now going to compute the BP equations for energy, entropy and free energy.

Energy The energy of a probability distribution of the form of equation (3.1) can be expressed by first defining an Hamiltonian. We rewrite equation (3.1):

$$P(x) = \frac{1}{Z} \exp\left(\sum_{a=1}^{M} \log \psi_a(\boldsymbol{x}_a)\right)$$

then, in analogy with canonical formalism in statistical physics (see equation (1.5)), we define the Hamiltonian as:

$$H(x) = -\sum_{a=1}^{M} \log \psi_a(\boldsymbol{x}_a)$$

The internal energy can be defined as the average of the Hamiltonian:

$$\begin{aligned} U &= -\sum_{a=1}^{M} \sum_{\boldsymbol{x}_a} p_{\partial a}(\boldsymbol{x}_a) \log \psi_a(\boldsymbol{x}_a) = \\ &= -\sum_{a=1}^{M} \frac{1}{z_a} \sum_{\boldsymbol{x}_a} \psi_a(\boldsymbol{x}_a) \prod_{i \in \partial a} \nu_{i \to a}(x_i) \log \psi_a(\boldsymbol{x}_a). \end{aligned}$$

This is the BP estimate of the internal energy, which for a tree is exact.

Entropy The entropy of the probability distribution in eq. (3.1) is:

$$S = -\sum_{x} P(x) \log P(x).$$

For a tree, there exists an exact formula for the distribution $P(x)$ in terms of the local marginals, which is very useful for computing the entropy:

$$P(x) = \prod_{a=1}^{M} p_{\partial a}(\boldsymbol{x}_a) \prod_{i=1}^{N} p_i(x_i)^{1-|\partial i|}. \tag{3.5}$$

This can be proved by induction on the number M of function nodes. If $M = 1$, then there is only one function node, so $x = \boldsymbol{x}_a$:

$$P(\boldsymbol{x}_a) = p_{\partial a}(\boldsymbol{x}_a) \prod_{i \in \partial a} p_i(x_i)^{1-1} = p_{\partial a}(\boldsymbol{x}_a)$$

which is a tautology. Assuming the formula to be correct for M function nodes, we try to compute it for $M + 1$ function nodes. The idea is that, since the

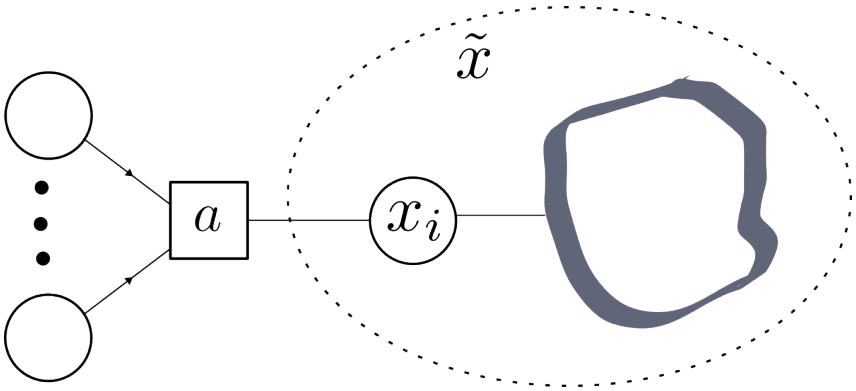

Fig. 3.5 The node a which is connected to only one node x_i which has degree greater than one. $\tilde{x}$ is the set of all the variable nodes except for $\partial a \setminus i$ (i.e. the variable nodes in the dashed ellipse).

graph is a tree, there must exist at least one function node a (located at the border of the tree, see Figure 3.5) which is connected at most to one variable node i which has degree greater than 1. We call $\tilde{x} = x \setminus \{x_j\}_{\partial a \setminus i}$ the vector of all the variable nodes in the sub-graph $i \cup S_i^{(a)}$ induced by eliminating a. We can write the probability of the whole graph by separating the terms coming from $\partial a \setminus i$ from the others:

$$P(x) = \mathbb{P}(\tilde{x})\mathbb{P}(x \setminus \tilde{x}|\tilde{x}) = \mathbb{P}(\tilde{x})\mathbb{P}(\{x_j\}_{\partial a \setminus i}|\tilde{x}).$$

Where we are using the symbol $\mathbb{P}$ to indistinctly express marginals or conditioned probabilities of P. The meaning of each term is determined by the variables' names. Now we notice that the variables in $\partial a \setminus i$ depend on $\tilde{x}$ only through x_i:

$$\begin{aligned} P(x) &= \mathbb{P}(\tilde{x})\mathbb{P}(\{x_j\}_{\partial a \setminus i}|x_i) = \\ &= \mathbb{P}(\tilde{x})\frac{\mathbb{P}(\{x_j\}_{\partial a \setminus i}, x_i)}{\mathbb{P}(x_i)} = \\ &= \mathbb{P}(\tilde{x})\frac{\mathbb{P}(\boldsymbol{x}_a)}{\mathbb{P}(x_i)} \\ &= \mathbb{P}(\tilde{x})\frac{p_{\partial a}(\boldsymbol{x}_a)}{p_i(x_i)}. \end{aligned} \tag{3.6}$$

We now compute $\mathbb{P}(\tilde{x})$. Notice that $\mathbb{P}(\tilde{x}) = \sum_{\boldsymbol{x}_a \setminus x_i} P(x)$. It is possible to represent $\mathbb{P}(\tilde{x})$ as a factor graph which is very similar to the one of $P(x)$. It is sufficient to define $\phi_a(x_i) = \sum_{\boldsymbol{x}_a \setminus x_i} \psi_a(\boldsymbol{x}_a)$. This is a function node attached

to x_i only. Therefore, we can then merge $\phi_a(x_i)$ to another function node $c \in \partial i \setminus a$, by simply defining $\phi_c(\boldsymbol{x}_c) = \phi_a(x_i)\psi_c(\boldsymbol{x}_c)$. The factor graph of $\mathbb{P}(\tilde{x})$ is therefore obtained from the one of P by:

- erasing the function node a and the variable nodes $\partial a \setminus i$;
- substituting a function node ψ_c (c chosen from $\partial i \setminus a$) with the function node ϕ_c.

The function $\mathbb{P}(\tilde{x})$ is therefore now expressed with M function nodes, so we can use the induction hypothesis:

$$\mathbb{P}(\tilde{x}) = p_i(x_i)^{1-(|\partial i|-1)} \left(\prod_{b=1,\dots M+1}^{b\neq a} p_{\partial b}(\boldsymbol{x}_b) \prod_{j=1,\dots,N}^{j\neq i} p_j(x_j)^{1-|\partial j|} \right) =$$
$$= p_i(x_i)^{2-|\partial i|} \left(\prod_{b=1,\dots M+1}^{b\neq a} p_{\partial b}(\boldsymbol{x}_b) \prod_{j=1,\dots,N}^{j\neq i} p_j(x_j)^{1-|\partial j|} \right).$$

Using equation (3.6) we prove the thesis. Finally, we can write the equation for entropy:

$$S = -\sum_{a=1}^{M} \sum_{\boldsymbol{x}_a} p_{\partial a}(\boldsymbol{x}_a) \log p_{\partial a}(\boldsymbol{x}_a) - \sum_{i=1}^{N} (1 - |\partial i|) \sum_{x_i} p_i(x_i) \log p_i(x_i).$$

3.1.5 Loopy BP

All the results provided so far for the Belief Propagation algorithm are obtained under the hypothesis that the factor graph is a tree. However, it is possible to use the BP scheme for all the factor graphs: it is sufficient to define a couple of messages $\nu_{i\to a}$ and $\hat{\nu}_{a\to i}$ for each edge (i, a) in the factor graph and update the sets $\{\nu_{i\to a}\}_{(i,a)}$ and $\{\hat{\nu}_{a\to i}\}_{(i,a)}$ using equations (3.3) and (3.4). This scheme is called *loopy* belief propagation, due to the fact that non-tree factor graphs contain loops. In general, loopy BP is an approximate scheme which might not provide the exact solution. However, in some cases the approximation is very good. An example is in the epidemic case, where even when loops are present BP is among the top performing methods (see Figures 2.4 and 2.5). For a generic factor graph, we can mimic equation (3.5) and define the Bethe

un-normalized probability:

$$\hat{P}(x) = \prod_{a=1}^{M} b_{\partial a}(\boldsymbol{x}_a) \prod_{i=1}^{N} b_i(x_i)^{1-|\partial i|}$$

where the letter b stands for the *beliefs,* i.e. the BP approximate marginals.

$$b_{\partial a}(\boldsymbol{x}_a) = \frac{1}{z_a} \psi_a(\boldsymbol{x}_a) \prod_{i \in \partial a} \nu_{i \to a}(x_i)$$
$$b_i(x_i) = \frac{1}{z_i} \prod_{a \in \partial i} \hat{\nu}_{a \to i}(x_i)$$

Beliefs are the approximations of the exact marginals of $P(x)$, respectively $p_{\partial a}$ and p_i. For tree factor graphs $p_{\partial a} = b_{\partial a}$ and $p_i = b_i$, but when loops are present, the beliefs are only approximations of the marginals. It is remarkable, however, that the Bethe probability distribution just defined is related to the exact joint [67] by a proportionality relation:

$$\begin{aligned}
\hat{P}(x) &= \prod_{a=1}^{M} \left(\frac{b_{\partial a}(\boldsymbol{x}_a)}{\prod_{i \in \partial a} b_i(x_i)} \right) \prod_{i=1}^{N} b_i(x_i) = \\
&= \prod_{a=1}^{M} \left(\frac{\frac{1}{z_a} \psi_a(\boldsymbol{x}_a) \prod_{i \in \partial a} \nu_{i \to a}(x_i)}{\prod_{i \in \partial a} \frac{1}{z_i} \prod_{b \in \partial i} \hat{\nu}_{b \to i}(x_i)} \right) \prod_{i=1}^{N} \frac{1}{z_i} \prod_{a \in \partial i} \hat{\nu}_{a \to i}(x_i) = \\
&= ZP(x) \prod_{a=1}^{M} \left(\frac{\frac{1}{z_a} \prod_{i \in \partial a} \nu_{i \to a}(x_i)}{\prod_{i \in \partial a} \frac{1}{z_i} \prod_{b \in \partial i \setminus a} \hat{\nu}_{b \to i}(x_i)} \right) \prod_{i=1}^{N} \frac{1}{z_i} = \\
&= ZP(x) \prod_{a=1}^{M} \left(\frac{\frac{1}{z_a} \cancel{\prod_{i \in \partial a} \nu_{i \to a}(x_i)}}{\prod_{i \in \partial a} \frac{z_{i \to a}}{z_i} \cancel{\nu_{i \to a}(x_i)}} \right) \prod_{i=1}^{N} \frac{1}{z_i} = \\
&= ZP(x) \left(\prod_{a=1}^{M} \frac{1}{z_a} \right) \left(\prod_{a=1}^{M} \prod_{i \in \partial a} \frac{z_i}{z_{i \to a}} \right) \left(\prod_{i=1}^{N} \frac{1}{z_i} \right) = \\
&= \frac{Z}{\prod_a z_a \prod_{(a,i)} \frac{z_{i \to a}}{z_i} \prod_i z_i} P(x) = \\
&= \frac{Z}{Z_{\text{Bethe}}} P(x).
\end{aligned}$$

This results is valid for a generic factor graph and states that the Bethe distribution is:

- non normalized;

- proportional the exact joint distribution.

The quantity:

$$F_{\text{Bethe}} = -\log Z_{\text{Bethe}} = -\sum_a \log z_a - \sum_{(a,i)} \log \frac{z_{i\to a}}{z_i} - \sum_i \log z_i \tag{3.7}$$

is called *Bethe free energy*. For a tree, the Bethe free energy coincides with the free energy $F = -\log Z$. Notice that the term $\frac{z_{i\to a}}{z_i}$ can be rewritten. In fact:

$$\begin{aligned} z_i &= \sum_{x_i} \prod_{b\in\partial i} \hat{\nu}_{a\to i}(x_i) = \sum_{x_i} \hat{\nu}_{a\to i}(x_i) \prod_{b\in\partial i\setminus a} \hat{\nu}_{b\to i}(x_i) = \\ &= \sum_{x_i} \hat{\nu}_{a\to i}(x_i) z_{i\to a} \nu_{i\to a}(x_i) \end{aligned}$$

So we have that:

$$\frac{z_i}{z_{i\to a}} = \sum_{x_i} \hat{\nu}_{a\to i}(x_i)\nu_{i\to a}(x_i) =: z_{ia}$$

Notice that these passages are only valid at fixed points because in these manipulations we are using BP iterations as if they were identities. We can therefore rewrite the Bethe free energy in a form which is more useful for the epidemic problem:

$$F_{\text{Bethe}} = -\log Z_{\text{Bethe}} = -\sum_a \log z_a + \sum_{(a,i)} \log z_{ia} - \sum_i \log z_i. \tag{3.8}$$

3.1.6 Generalization to the Ensemble: Cavity Method

The BP equations described so far consist in an iterative scheme aimed at marginalizing a fixed probability distribution $P(x)$. However, in physics and inference, one is sometimes interested in studying averages over an *ensemble* of distributions $\{P_{\underline{J}}(x)\}_{\underline{J}\in\mathcal{J}}$, in which $\underline{J}$ is a random vector. In physics, for example, $\underline{J}$ may be the set of the ferromagnetic and antiferromagnetic couplings of an amorphous material. Each piece of material corresponds to a certain fixed configuration $\underline{J}$ of couplings. Averaging over $\underline{J}$, therefore, enables to find general properties of the material, regardless of the specific piece considered. The single piece of material in physics corresponds to the single instance in inference. For example, the quantity corresponding to $\underline{J}$ in inference is the triplet $G, \underline{\tau}, \mathcal{O}$ of contact graph, planted configuration and observations set.

The variable $\underline{J}$ is called *disorder* in the physicists' language and the average over $\underline{J}$ is called average over *disorder* or ensemble average. It is possible to extend the Belief Propagation algorithm to compute ensemble averages with the so called *replica symmetric cavity method* ([37], Probabilistic analysis, pag. 321). This scheme is computationally similar to BP, but conceptually different: at fixed $\underline{J}$, the BP equations can be iterated until convergence of messages $\{\nu_{i\to a}\}_{(i,a)}$, $\{\hat{\nu}_{a\to i}\}_{(i,a)}$. Those messages depend on the value of $\underline{J}$ (which enters the BP equations through the function nodes), so we now explicitly call them $\{\nu^{(\underline{J})}_{i\to a}\}_{(i,a)}$, $\{\hat{\nu}^{(\underline{J})}_{a\to i}\}_{(i,a)}$. However, we observe that each message $\nu_{i\to a}$ (or $\hat{\nu}_{a\to i}$) at fixed $\underline{J}$ explicitly depends only on a sub-vector $\underline{J}_{i\to a}$ (respectively $\underline{J}_{a\to i}$) of the disorder: we can therefore rewrite the set of messages as $\{\nu^{(\underline{J}_{i\to a})}_{i\to a}\}_{(i,a)}$, $\{\hat{\nu}^{(\underline{J}_{a\to i})}_{a\to i}\}_{(i,a)}$. The idea of the replica symmetric cavity method for averaging over $\underline{J}$ is to extract a disorder instance $\underline{J}_{i\to a}$ (respectively $\underline{J}_{a\to i}$) altogether with every update of the BP message $\nu_{i\to a}$ (respectively $\hat{\nu}_{a\to i}$). The central hypothesis of the method is that this process converges to a fixed distribution $\Psi(\nu,\hat{\nu})$ of messages. To understand where this hypothesis comes from, we can think of running BP on an infinite graph. At the fixed point the distribution of messages $\Psi(\nu,\hat{\nu})$ remains unaltered if we keep BP running. Of course, it is impossible to effectively run BP on an infinite graph, so the cavity method resorts on finding a fixed point of the distribution of messages $\Psi(\nu,\hat{\nu})$. We can imagine one iteration of the cavity method as a single BP update of one message in the infinite graph. A way to implement this method numerically is to use the so called *population dynamics* technique. This consists in approximating the distribution $\Psi(\nu,\hat{\nu})$ with a histogram of messages. Typically, a number n is fixed, which is called the *population size*; n function-to-variable $\{\hat{\nu}_i\}_{i=1}^n$ and n variable-to-function $\{\nu_i\}_{i=1}^n$ messages are initialized. Then the BP scheme is used altogether with the ensemble sampling. For example, to update the variable-to-function message ν_k, the replica symmetric cavity method requires to:

1. Extract the number d_{res} of entering messages. This number is called *residual degree* and is the number of incoming function-to-node messages needed to update ν_k. If we call (i,a) the edge on which the variable-to-function message ν_k lays, then the residual degree is $|\partial i \setminus a|$. If the degree distribution of a graph is known, then the residual degree distribution is also known, see Appendix B.

2. Take d_{res} function-to-variable messages in $\{\hat{\nu}_i\}_{i=1}^n$
3. Extract the sub-vector $\underline{J}_k$ which enters the BP equation for ν_k.
4. Use the BP equation to update ν_k.

When the empirical distribution of messages converges, the messages can be used to evaluate ensemble properties. For example, to compute the one-point marginal b, it is sufficient to average among a population of believes, namely:

1. extract the degree d of a randomly chosen node from the degree distribution;
2. extract d function-to-variable messages from the converged set $\{\hat{\nu}_i\}_{i=1}^n$;
3. extract the sub-vector $\underline{J}_1$ of disorder needed to compute a single belief ;
4. compute the corresponding BP belief $b^{\underline{J}_1}$
5. repeat for $m \propto n$ times
6. average among the m beliefs: $b(x) = \frac{1}{M}\sum_{k=1}^m b^{\underline{J}_k}(x)$.

Notice that the cavity method relies on two crucial hypothesis: the independence of the entries of each $\underline{J}$ and the replica symmetric assumption. The first hypothesis is used during the update: we should, for every update, sample an entire single instance of disorder $\underline{J}$. This is however unfeasible, due to the fact that when the system is at thermodynamic limit $\underline{J}$ has infinite coordinates. However, hypothesizing independence among the entries of $\underline{J}$, we can extract only the sub-vector $\underline{J}_{i\to a}$ which enters the BP update equation. If the disorder is correlated, this procedure is no longer valid and the cavity method can not be used. In epidemics, the role of $\underline{J}$ is played by the graph and planted trajectory $\underline{t}^*$, which is a vector of random infection times. These times are correlated by the underlying stochastic dynamics which generated them. It is therefore not possible to apply the cavity method immediately and it is necessary to build a machinery to map the disorder onto a larger space of independent entries, as explained in section 3.3.1. The second hypothesis is the replica symmetry. For updating, e.g., a function-to-variable message, we first have to extract the incoming variable-to-function messages in order to run BP. This extraction is made randomly uniform on the set $\{\nu_i\}_{i=1}^n$. This means that each message

is equivalent to the others and no correlation among messages is considered. This equivalence among messages is the replica symmetric hypothesis. There are methods that generalize the cavity to the 1RSB phase, but we refer the reader to [37], *The 1RSB cavity method.* pag. 429. The replica symmetric cavity method, together with population dynamics, is the tool which we are going to use in the epidemic problem. In replica symmetric regimes, we expect the method to provide exact results (except for the numerical approximation due to the finite size of the population) for all the graph ensembles having no short loops, i.e. the only cycles allowed must diverge with the graph size. Before entering the details, it is useful and pedagogical to talk about Sib, the application of the BP scheme to epidemic inference at fixed instance.

3.2 Sib: the BP Application to Epidemics

The algorithm which applies BP to epidemic inference is called Sib [8, 6]. Some results of Sib are presented, in comparison with the Causal Variational Approach, in section 2.4. Sib has proved to be one of the top performing methods in epidemic Bayesian inference. In this thesis, we are only going to give a brief introduction to the method, referring the reader to [8, 6] for the results and the details .

3.2.1 Effective loops in factor graph

The idea behind Sib is simple. We take the epidemic posterior and we rewrite it in the form of a factor graph, then we run BP. However, there are some difficulties which arise when we build the graph and some refinements must be

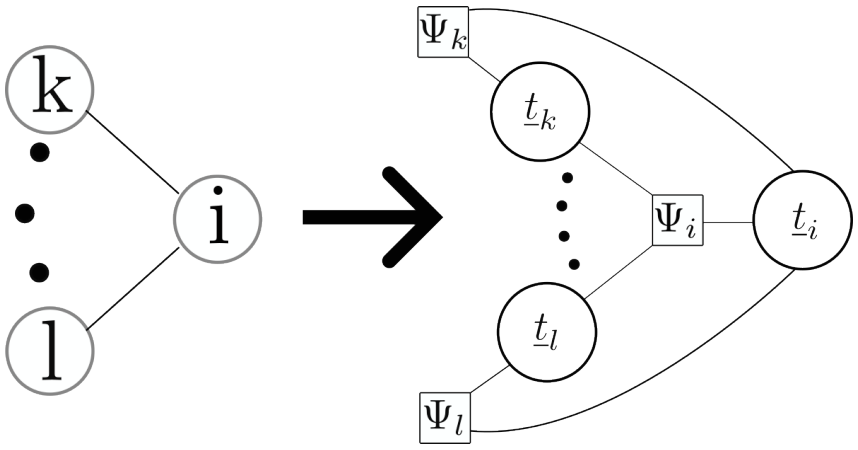

Fig. 3.6 The naive factor graph construction in paragraph 3.2.1 introduces loops in the factor graph corresponding to a tree contact network.

done. Let us start by writing the epidemic posterior (see section 2.1.4):

$$\begin{aligned}\mathcal{P}(\underline{t}|\mathcal{O}) &= \frac{1}{P(\mathcal{O})}P(\underline{t})P(\mathcal{O}|\underline{t}) = \\ &= \frac{1}{P(\mathcal{O})}\left(\prod_{i=1}^{N}\psi(\underline{t}_i,\underline{t}_{\partial i})\right)\left(\prod_{o\in\mathcal{O}}p(o|\underline{t}_{o_i})\right) = \\ &= \frac{1}{P(\mathcal{O})}\prod_{i=1}^{N}\left(\psi(\underline{t}_i,\underline{t}_{\partial i})\prod_{o\in\mathcal{O}:o_i=i}p(o|\underline{t}_i)\right) = \\ &= \frac{1}{P(\mathcal{O})}\prod_{i=1}^{N}\Psi_i(\underline{t}_i,\underline{t}_{\partial i}),\end{aligned} \tag{3.9}$$

where we defined

$$\Psi_i(\underline{t}_i,\underline{t}_{\partial i}) = \psi(\underline{t}_i,\underline{t}_{\partial i})\prod_{o\in\mathcal{O}:o_i=i}p(o|\underline{t}_i) \tag{3.10}$$

and $\{o\}_i = \{o : o_i = i\}$. We have a distribution which is the product of local factors, as in eqn (3.1). In this case, the number of variable nodes N coincides with the number of function nodes. One is tempted to take this distribution and immediately use BP. To do so, it is sufficient to associate a function node for each of the factors Ψ_i and to associate a variable node for each $\underline{t}_i$, with $i = 1,\dots,N$. Then, one simply connects each function to its variables. However, this is not a good strategy. In fact, as shown in Figure 3.6, a tree contact network is mapped onto a loopy factor graph. This implies

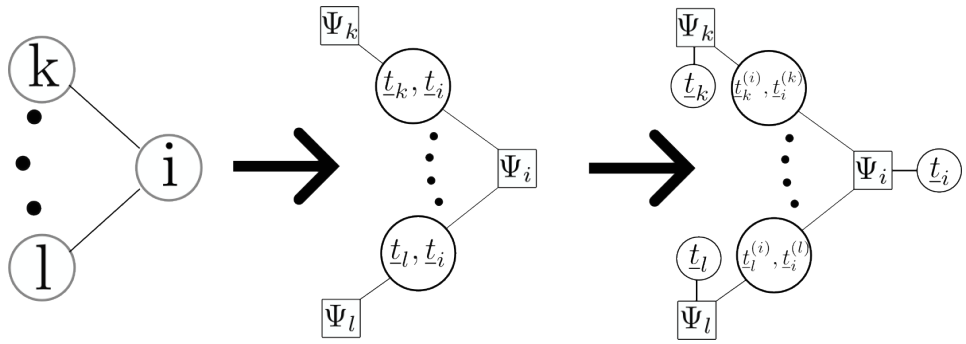

Fig. 3.7 The construction of tree factor graphs as in Sib. *Left:* the original contact network is a tree. *Center:* the central idea for building a tree factor graph is to associate to each link (i, j) a variable node $\underline{t}_i, \underline{t}_j$. However, this construction introduces the variable $\underline{t}_i$ in more nodes. This makes no sense, since different variable nodes should represent different variables. In general, a variable $\underline{t}_j$ in this scheme is introduced $|\partial j|$ times. *Right:* the definitive construction of the factor graph. The problem of introducing the same variable in more nodes is solved by the definition of copies: for each individual i are defined $|\partial i|$ copies of its dynamic state, one for each link. This way different nodes correspond to different variables. However, since we want all the copies $\underline{t}_i^{(j)}$ to have the same value for each $j \in \partial i$, we also introduce the variable $\underline{t}_i$ which is attached only to Ψ_i, which we redefined by including the hard constraint $\prod_{j \in \partial i} \delta_{\underline{t}_i, \underline{t}_i^{(j)}}$.

that BP would provide inexact results even if the contact graph is a tree. This limitation can be overcome by a more clever construction, described in [8]: the idea is to build the factor graph by associating to each individual i the function node Ψ_i and at each edge (i, j) the variable node made by the couple $(\underline{t}_i, \underline{t}_j)$. In other words, we have enlarged the variable nodes' domain, as in Figure 3.7, *center*. Having associated a variable node to each edge allows to eliminate the loops of the previous construction. This way, a tree contact network is mapped onto a tree factor graph. As a consequence, the BP algorithm is at least exact for tree contact networks. However, we are still not able to run BP for this factor graph. In fact, we developed the method in the previous sections under the hypothesis that each variable node represented a different variable. In this case, instead, the variable nodes $(\underline{t}_i, \underline{t}_k)$ and $(\underline{t}_i, \underline{t}_l)$ have the trajectory variable $\underline{t}_i$ in common. Therefore, we still have to modify a bit the construction of the factor graph. We need to define, for each individual i (i.e. for each function node Ψ_i) a set of *copies* $\{\underline{t}_i^{(j)}, j \in \partial i\}$. Every copy represents the trajectory of i. For the link (i, j), the node is defined as the couple $(\underline{t}_i^{(j)}, \underline{t}_j^{(i)})$. This way, we have that different nodes are made by different variables. BP can be run. Still,

there is a problem: we want all the copies to have the same value. We need to introduce a constraint for each of the Ψ_i. To do so, we attach to each Ψ_i a new variable node $\underline{t}_i$. Now Ψ_i is a function of $(\underline{t}_i, \{\underline{t}_i^{(j)}\}_{j\in\partial i}, \{\underline{t}_j^{(i)}\}_{j\in\partial i})$. The first arguments, $\underline{t}_i, \{\underline{t}_i^{(j)}\}_{j\in\partial i}$, all represent the same trajectory. We therefore multiply each Ψ_i for the constraint that all the copies must be equal to $\underline{t}_i$:

$$\phi_i(\underline{t}_i, \{\underline{t}_i^{(j)}\}_{j\in\partial i}, \{\underline{t}_j^{(i)}\}_{j\in\partial i}) = \Psi_i(\underline{t}_i, \{\underline{t}_i^{(j)}\}_{j\in\partial i}, \{\underline{t}_j^{(i)}\}_{j\in\partial i}) \prod_{j\in\partial i} \delta_{\underline{t}_i, \underline{t}_i^{(j)}}. \tag{3.11}$$

There is a little abuse of notation here. In fact, as defined in equation (3.9), the function Ψ_i does depend only on $\underline{t}_i, \underline{t}_{\partial i}$. We instead used the same symbol Ψ_i for a function of the copies $\underline{t}_i, \{\underline{t}_i^{(j)}\}_{j\in\partial i}, \{\underline{t}_j^{(i)}\}_{j\in\partial i}$. However, the effect of the copies is simply to substitute the variable $\underline{t}_i^{(j)}$ to the variable $\underline{t}_i$ when the link i, j is taken into account. We temporally changed symbol to ϕ_i when we added the constraints. However, it is easy to get lost only looking at equations. To effectively run BP, it is necessary to have the graphical counterpart in mind (Figure 3.7, *Right*). When looking at figures, it becomes clear what are exactly the arguments of each function node. To avoid introducing new symbols, therefore, we are going to keep calling Ψ_i the function nodes with the constraints on the copies.

3.2.2 BP equations for epidemic model

The factor graph for the epidemic model has two different kinds of variable nodes: the couple-trajectory and the single-trajectory nodes. The single-trajectory nodes all have degree one because they are attached only to their corresponding function node, while the couple-trajectory ones all have degree 2. Distinguishing the two nodes types allows to reduce the set of message passing BP equations. Looking at Figure 3.7, *Right*, we write down the BP equations. Let's start with variable-to-function messages, equation (3.3).

$$\nu_{(k,i)\to\Psi_i}(\underline{t}_k^{(i)}, \underline{t}_i^{(k)}) = \hat{\nu}_{\Psi_k\to(k,i)}(\underline{t}_k^{(i)}, \underline{t}_i^{(k)}) \tag{3.12}$$

$$\nu_{k\to\Psi_k}(\underline{t}_k) \propto 1 \tag{3.13}$$

these first two equations are trivial due to the degree of the variable nodes. The function-to-variable messages are nontrivial:

$$\hat{\nu}_{\Psi_i\to(k,i)}(\underline{t}_i^{(k)},\underline{t}_k^{(i)}) \propto \sum_{\{\underline{t}_j^{(i)},\underline{t}_i^{(j)}\}_{j\in\partial i\setminus k}} \Psi_i(\underline{t}_i,\{\underline{t}_i^{(j)}\}_{j\in\partial i},\underline{t}_{\partial i}^{(i)}) \prod_{j\in\partial i\setminus k} \nu_{(j,i)\to\Psi_i}(\underline{t}_j^{(i)},\underline{t}_i^{(k)}) \tag{3.14}$$

$$\hat{\nu}_{\Psi_i\to i}(\underline{t}_i^{(k)}) \propto \sum_{\{\underline{t}_j^{(i)},\underline{t}_i^{(j)}\}_{j\in\partial i}} \Psi_i(\underline{t}_i,\{\underline{t}_i^{(j)}\}_{j\in\partial i},\underline{t}_{\partial i}^{(i)}) \prod_{j\in\partial i} \nu_{(j,i)\to\Psi_i}(\underline{t}_j^{(i)},\underline{t}_i^{(k)}). \tag{3.15}$$

Notice that in equation (3.14) there is also the message $\nu_{i\to\Psi_i}(\underline{t}_i)$ in the product, but it was not written because it is a constant of proportionality, as stated in equation (3.13). In principle we have now four equations for updating the messages and we want to find a fixed point. In fact, we can reduce the number of update equations to one. Indeed, equation (3.13) does not update the messages. Moreover, equation (3.12) can be plugged into equations (3.14) and (3.15):

$$\hat{\nu}_{\Psi_i\to(k,i)}(\underline{t}_i^{(k)},\underline{t}_k^{(i)}) \propto \sum_{\{\underline{t}_j^{(i)},\underline{t}_i^{(j)}\}_{j\in\partial i\setminus k}} \Psi_i(\underline{t}_i,\{\underline{t}_i^{(j)}\}_{j\in\partial i},\underline{t}_{\partial i}^{(i)}) \prod_{j\in\partial i\setminus k} \hat{\nu}_{\Psi_j\to(j,i)}(\underline{t}_j^{(i)},\underline{t}_i^{(k)}) \tag{3.16}$$

$$\hat{\nu}_{\Psi_i\to i}(\underline{t}_i^{(k)}) \propto \sum_{\{\underline{t}_j^{(i)},\underline{t}_i^{(j)}\}_{j\in\partial i}} \Psi_i(\underline{t}_i,\{\underline{t}_i^{(j)}\}_{j\in\partial i},\underline{t}_{\partial i}^{(i)}) \prod_{j\in\partial i} \hat{\nu}_{\Psi_j\to(j,i)}(\underline{t}_j^{(i)},\underline{t}_i^{(k)}). \tag{3.17}$$

So we have only two update equations, but we don't really need to solve them both: equation (3.16) is a relation among the set of messages $\{\hat{\nu}_{\Psi_i\to(k,i)}\}_{k\in\partial i}^{i=1,\dots,N}$ and does not depend on the set $\{\hat{\nu}_{\Psi_i\to i}\}_{i=1,\dots,N}$. Iterating equation (3.16) until convergence leads to fixed point cavity marginals. To compute marginals, we observe that the messages in enq (3.17) are the marginals, because there is no cavity in considering that Ψ_i is the only factor of i. The Sib algorithm works by finding the fixed point of eqn (3.16). In the original paper [8], some optimizations are made to increase the velocity of computation of the updates. We refer the interested reader to the paper. Moreover, some similar optimizations are discussed in the next section for the cavity method.

3.3 Cavity method application to Epidemic

Being able to reconstruct the epidemic posterior is a crucial task. However, sometimes the available information (e.g. the number or the quality of the clinical tests) might not be enough to make reasonable predictions. In that case, it is meaningless to run any inference algorithm (as CVA or Sib). It is therefore desirable to develop a method which might predict the feasibility of inference for each epidemic regime. In this section we study *Epidemble,* a method that quantifies information bounds, identifying the regimes where inference is possible. The method computes the expected values of the most used statistical estimators (e.g. AUC, MME, MMO) as functions of the epidemic hyper-parameters. In order to evaluate such expected values, the method works at the thermodynamic limit, namely for the total number N of individuals that goes to infinity. This is achieved by means of the replica symmetric cavity method. Epidemble is therefore the ensemble version of the Sib algorithm, section 3.2. However, passing from the single instance to the ensemble algorithm is nontrivial, due to correlations in the planted times, which make impossible to directly apply the population dynamics technique to the problem. This issue and its solution are described in the next section, where Epidemble method is presented.

3.3.1 The disorder is correlated

This entire section is devoted to the SI model[2], so we are going to lighten a bit the notation: to express the infection time of an individual i we will write t_i instead of $t_i^{\mathcal{I}}$, forgetting about the superscript $\mathcal{I}$. An epidemic trajectory is therefore $\underline{t} = (t_1, \dots, t_N)$ and this notation is adopted interchangeably with $x = \{x_i^t, \forall i = 1, \dots, N, \forall t = 1, \dots, T\}$. The cavity method assumes independence of the entries of disorder: to update a message, in fact, the sub-vector of the disorder needed to compute the BP update is sampled. Disorder, however, is correlated in the epidemic problem. To see this, let us distinguish among the two sources of random disorder: the graph G and the planted

[2]because all the difficulties are already present in the SI model. It is in fact quite simple to generalize to SIR and SEIR model the cavity method described in this section, see subsection 3.4.9

time, which we indicate now with the letter $\underline{\tau}$. The disorder due to graph is independent: to update the message, we sample each degree independently of the other according to the degree distribution, as described in section 3.1.6 . The entries of the planted, instead, are not independent of each others. This impedes a possible direct application of the cavity method to the Sib equation (3.16).

3.3.2 Failure of cavity method for correlated disorder

To see where the cavity method breaks down, let us simplify the treatment by supposing that every individual is observed at time T without noise (false rate). We start from equation (3.16) for the SI model and try to build the (naive) ensemble version from it:

$$\hat{\nu}_{\Psi_i\to(k,i)}(\underline{t}_i^{(k)},\underline{t}_k^{(i)}) = \frac{1}{z_{\Psi_i\to k}} \sum_{\{\underline{t}_j^{(i)},\underline{t}_i^{(j)}\}_{j\in\partial i\setminus k}} \Psi_i(\underline{t}_i,\{\underline{t}_i^{(j)}\}_{j\in\partial i},\underline{t}_{\partial i}^{(i)}) \prod_{j\in\partial i\setminus k} \hat{\nu}_{\Psi_j\to(j,i)}(\underline{t}_j^{(i)},\underline{t}_i^{(k)})$$

Notice that $\Psi_i(t_i,\{t_i^{(j)}\}_{j\in\partial i},t_{\partial i}^{(i)})$ by definition in equation (3.10) depends on the observation on the individual i, which is for hypothesis noiseless and at final time. To implement the cavity method we would be tempted to initialize a population of messages $\{\hat{\nu}_k\}_{k=1}^n$ and to design the following update algorithm:

1. Extract an entry $j\in\{1,\dots,n\}$ with the aim of updating $\hat{\nu}_j$.
2. Extract the number $d-1$ from the residual degree distribution.
3. Extract $d-1$ random messages from $\{\hat{\nu}_k\}_{k=1}^n$
4. Extract the infection time τ_j and the observation from it: $o=(o_s,o_i,o_t,o_f)$, where $o_s=\mathcal{S}\delta_{\tau_j,T+1}+\mathcal{I}(1-\delta_{\tau_j,T+1})$, $o_i=j$, $o_t=T$, $o_f=0$.
5. Update ν_j with the Sib BP equation (3.16).

This algorithm is not possible to run due to passage 4. The extraction of τ_j can be done only extracting the whole stochastic process of epidemic spread. In other words, the components of $\underline{\tau}$ are correlated by the dynamic. This simple version of the cavity method is therefore not doable.

3.3.3 Enlarging the disorder space to make it independent

We need to represent the planted trajectory (planted disorder) by means of a set of independent random variables. To do so we have to look at a deeper level in the stochastic prior process. There is a way to generate an epidemic cascade by sampling a set of independent variables. We describe it now for a finite single instance, so to write new BP equations which admit a cavity extension. To represent the planted $\underline{\tau}$ with independent disorder, it is sufficient to sample for each node $i = 1, \dots, N$ its time-zero state $x_i^0 \in \{0, 1\}$, a boolean variable which states if the individual is a patient zero ($0 = \mathcal{S}, 1 = \mathcal{I}$), and for each edge $(i, j) \in \mathcal{E}$, a couple s_{ij} and s_{ji} of infection delays. Each s_{ij} is a number representing the time needed for the individual i, from the moment it is infectious, to infect j in the absence of any other neighbor. For example, if the infection time of i is τ_i, then if $x_j^0 = 0$ and $\partial j = \{i\}$, it is true that $\tau_j = \tau_i + s_{ij}$. In general, an individual k is either a patient zero or has an infection time which is the minimum among the tentative infection times from its neighbors:

$$\tau_k = (1 - x_k^0) \min_{j \in \partial k} \{\tau_j + s_{jk}\}. \tag{3.18}$$

If we extract all the $\{x_i^0\}_{i=1,\dots,N}, \{s_{ij}\}_{i=1\dots,N}^{j \in \partial i}$ then there exists one unique trajectory associated to them[3], which satisfies equation (3.18) for all $k = 1 \dots, N$ and can be numerically obtained by Algorithm 4. Notice now that all the variables $\{x_i^0\}_{i=1,\dots,N}, \{s_{ij}\}_{i=1\dots,N}^{j \in \partial i}$ are independent from each others: each x_i^0 is sampled from $p(x) = \delta_{x,0}(1 - \gamma) + \delta_{x,1}\gamma$; each s_{ij} is sampled from the probability that an infectious individual i infects a susceptible individual j: $p(s) = \lambda(1 - \lambda)^s$. We therefore have a set of random independent variables which can be used to represent the planted $\underline{\tau}$. We can therefore set the planted disorder to $\mathcal{D} = \{x_i^0\}_{i=1,\dots,N}, \{s_{ij}\}_{i=1\dots,N}^{j \in \partial i}$. The probability of having a trajectory $\underline{\tau}$ given the disorder set $\mathcal{D} = \{x_i^0\}_{i=1,\dots,N}, \{s_{ij}\}_{i=1\dots,N}^{j \in \partial i}$ is a deterministic function

[3] the converse is not true: we might have that the same epidemic trajectory corresponds to different extractions of delays.

Algorithm 4 Building the trajectory from the patient zero states and the infection delays

Input: $\{x_i^0\}_{i=1,\dots,N}, \{s_{ij}\}_{i=1\dots,N}^{j\in\partial i}$ and the contact graph G.
Output: the planted $\underline{\tau}$.

- Initialize a queue $\underline{\tau}^q = (\tau_1^q, \dots, \tau_N^q)$ and a vector of times $\underline{\tau} = (\tau_1, \dots, \tau_N)$.
- Set $\tau_i^q = 0$ if $x_i^0 = 1$ for each $i = 1, \dots, N$.
- **Loop** over time $t = 0, \dots, T$
  - **Loop** over the queue $\underline{t}^q$ by entering from its minimum value τ_k^q
    - Set $\tau_k = \tau_k^q$
    - Infect the neighbors of k: set $\tau_j^q = \min(\tau_j^q, \tau_k^q + s_{kj})$ for all $j \in \partial k$
    - Remove τ_k
- **Return** $\underline{\tau}$.

which factorizes over local contributions:

$$P(\underline{\tau}|\mathcal{D}) = \prod_{i=1}^{N} \mathbb{I}\left[\tau_i = x_i^0 \min_{j\in\partial i}\{\tau_j + s_{ji}\}\right] = \tag{3.19}$$
$$= \prod_{i=1}^{N} \psi^*(\tau_i, \tau_{\partial i}|x_i^0, \{s_{ji}\}_{j\in\partial i}).$$

Where we defined

$$\psi^*(\tau_i, \tau_{\partial i}|x_i^0, \{s_{ji}\}_{j\in\partial i}) = \mathbb{I}\left[\tau_i = x_i^0 \min_{j\in\partial i}\{\tau_j + s_{ji}\}\right]. \tag{3.20}$$

We now write the BP equations for the joint distribution $\mathcal{P}(\underline{\tau}, \mathcal{O}, \underline{t}|\mathcal{D})$. Since this joint only depends on independent disorder (graph disorder and planted disorder $\mathcal{D}$), it will be possible to extend BP to replica symmetric cavity method.

3.3.4 Factor graph for the enlarged distribution

To summarize, even though the posterior at fixed disorder (network and planted) is a product of local factors (eqn.(3.9)), the cavity method is not implementable

because the entries of the planted are correlated. A way to circumvent this problem is to pass to the time-zero and delays representation of the planted trajectory, $\mathcal{D} = \{x_i^0\}_{i=1,\dots,N}, \{s_{ij}\}_{i=1\dots,N}^{j\in\partial i}$. This allows to work with independent disorder. The price to pay is to enlarge the domain of the distribution studied: instead of the posterior, we have to study the joint $\mathcal{P}(\underline{\tau}, \mathcal{O}, \underline{t}|\mathcal{D})$, which is:

$$P(\underline{\tau}, \mathcal{O}, \underline{t}|\mathcal{D}) = P(\underline{\tau}|\mathcal{D})P(\mathcal{O}|\underline{\tau}, \mathcal{D})\mathcal{P}(\underline{t}|\mathcal{O}, \underline{\tau}, \mathcal{D}) \tag{3.21}$$

$$= P(\underline{\tau}|\mathcal{D})P(\mathcal{O}|\underline{\tau}, \mathcal{D})\mathcal{P}(\underline{t}|\mathcal{O}) = \tag{3.22}$$

$$= \frac{1}{P(\mathcal{O})}P(\underline{\tau}|\mathcal{D})P(\mathcal{O}|\underline{\tau}, \mathcal{D})P(\mathcal{O}|\underline{t})P(\underline{t}). \tag{3.23}$$

Where equation (3.21) is the definition of conditional probability and the passage to eqn.(3.22) is due to the fact that the posterior does not depend on the disorder: during the inference process, the information on the planted is all contained in the observation set $\mathcal{O}$. This passage is a definition: we define the inference process by claiming that the posterior has no access to the information on the trajectory except from the observations. Finally, equation (3.23) is obtained from the previous line by using Bayes' law on the posterior. Each factor in equation (3.23) corresponds to one of the three steps of the process:

1. sampling the planted from the disorder using $P(\underline{\tau}|\mathcal{D})$, which is a deterministic process since the disorder fixes the trajectory, equation (3.19);
2. sampling the observations from the planted and the disorder via $P(\mathcal{O}|\underline{\tau}, \mathcal{D})$.
3. sampling a configuration from the posterior $\mathcal{P}(\underline{t}|\mathcal{O})$.

Let us discuss the observation term $P(\mathcal{O}|\underline{\tau}, \mathcal{D})$. It is a function of the only planted trajectory in the case of all noiseless observations at final time:

$$P(\mathcal{O}|\underline{\tau}, \mathcal{D}) = \prod_{o\in\mathcal{O}} \mathbb{I}\left[o_s = \left(\mathcal{S}\delta_{\tau_{o_i},T+1} + \mathcal{I}(1-\delta_{\tau_{o_i},T+1})\right)\right] = P(\mathcal{O}|\underline{\tau}). \tag{3.24}$$

We recall here that each observation o is represented by a 4uple $o = (o_i, o_s, o_t, o_f)$, where o_i is the individual tested, o_s is the observed state, o_t is the time at which the test is made, o_f is the false rate (see subsection 2.1.4. If all the

observations are noiseless and made at time T, then we have $o_t = T$ and $o_f = 0$ for all $o \in \mathcal{O}$. To deal with a generic number of observations, with nonzero false rate (which however we suppose to be constant: $o_f = f, \forall o \in \mathcal{O}$) and random observation times, it is necessary to introduce new disorder variables:

- the total number of observations per individual $\{n_o^i\}_{i=1}^N$. Each individual i is observed n_o^i times. We therefore have that the observation set $\mathcal{O}$ has $\sum_{i=1}^N n_o^i$ elements.
- the error bits set $\{\varepsilon_o\}_{o\in\mathcal{O}}$. Each $\varepsilon_o \in \{0, 1\}$ is a boolean variable that states if the observation o is corrupted or not. Each ε_o is independent of the other sources of disorder and it is drawn from the bimodal probability $p(\varepsilon) = f\delta_{\varepsilon,1} + (1-f)\delta_{\varepsilon,0}$, where f is the false rate.
- the observation times $\{t_o\}_{o\in\mathcal{O}}$. Each observation time t_o simply represents the time at which observation o is made.

Notice that these new variables are independent of each others. By defining the observation disorder set as

$$\mathcal{D}_o = \left\{\{\varepsilon_o, t_o\}_{o\in\mathcal{O}}, \{n_o^i\}_{i=1}^N\right\}$$

The probability of observations $\mathcal{O}$ given the planted $\underline{\tau}$ and the observation disorder $\mathcal{D}_o$ is again a deterministic function:

$$\begin{aligned} P(\mathcal{O}|\underline{\tau}, \mathcal{D}_o) &= \prod_{o\in\mathcal{O}} \mathbb{I}\left[o_s = \text{flip}\left(\mathcal{S}\mathbb{I}[\tau_{o_i} > t_o] + \mathcal{I}\mathbb{I}[\tau_{o_i} \leq t_o], \varepsilon_o\right)\right] = \\ &= \prod_{o\in\mathcal{O}} p^*(o|\tau_{o_i}, \varepsilon_o, t_o) \end{aligned}$$

where:

$$\text{flip}(\mathcal{S}, \varepsilon) = \begin{cases} \mathcal{I} & \varepsilon = 1 \\ \mathcal{S} & \varepsilon = 0 \end{cases} \qquad \text{flip}(\mathcal{I}, \varepsilon) = \begin{cases} \mathcal{S} & \varepsilon = 1 \\ \mathcal{I} & \varepsilon = 0 \end{cases}.$$

The equation simply generalizes eqn.(3.24) by considering generic observation times and flipping the result of the observations according to the disorder bits.

Also the joint in equation (3.23) depends now on the observation disorder:

$$P(\underline{\tau}, \mathcal{O}, \underline{t}|\mathcal{D}, \mathcal{D}_o) = \frac{1}{P(\mathcal{O})} P(\underline{\tau}|\mathcal{D}) P(\mathcal{O}|\underline{\tau}, \mathcal{D}_o) P(\mathcal{O}|\underline{t}) P(\underline{t}). \tag{3.25}$$

Notice that at fixed $\mathcal{D}_o$ and $\underline{\tau}$, the observation set is fixed. But it is also true that at fixed $\mathcal{D}$ the planted $\underline{\tau}$ is fixed. At fixed $\mathcal{D}$ and $\mathcal{D}_o$, therefore, the observation set $\mathcal{O}$ is fixed. We can therefore sum the joint distribution in eqn. (3.23) over $\mathcal{O}$ obtaining:

$$P(\underline{\tau}, \underline{t}|\mathcal{D}, \mathcal{D}_o) = \frac{1}{P(\mathcal{O}^*)} P(\underline{\tau}|\mathcal{D}) P(\mathcal{O}^*|\underline{t}) P(\underline{t}), \tag{3.26}$$

where $\mathcal{O}^*$ is the only observation set such that $P(\mathcal{O}^*|\underline{\tau}, \mathcal{D}_o) = 1$. Let us define $Z(\mathcal{D}, \mathcal{D}_o) := P(\mathcal{O}^*)$ and

$$\xi(\underline{t}, \underline{\tau}; \mathcal{D}_o) := P(\mathcal{O}^*|\underline{t}) = \prod_{i=1}^{N} \prod_{o \in \mathcal{O}^*: o_i = i} p(o|t_i) p^*(o|\tau_i, \varepsilon_o, t_o)$$

Where we simply wrote the likelihood expression for the observation set $\mathcal{O}^*$,which is fixed by the planted $\underline{\tau}$ and the observation disorder $\mathcal{D}_o$. The first factor $p^*(o|\tau_i, \varepsilon_o, t_o)$ ensures that the observation is sampled consistently with the prior and the disorder. The second term comes from the likelihood and weights each infection time with the observation. With these definitions, equation (3.26) becomes:

$$P(\underline{\tau}, \underline{t}|\mathcal{D}, \mathcal{D}_o) = \frac{1}{Z(\mathcal{D}, \mathcal{D}_o)} P(\underline{\tau}|\mathcal{D}) \xi(\underline{t}, \underline{\tau}; \mathcal{D}_o) P(\underline{t})$$

Notice that the observation term ξ couples the planted and inferred times. Substituting the definition of each term inside the joint, we obtain the factorized form:

$$P(\underline{\tau}, \underline{t}|\mathcal{D}, \mathcal{D}_o) = \frac{1}{Z(\mathcal{D}, \mathcal{D}_o)} \prod_{i=1}^{N} \psi^*(\tau_i, \tau_{\partial i}|x_i^0, \{s_{ji}\}_{j \in \partial i}) \xi_i(\tau_i, t_i|\{\varepsilon_o, t_o\}_{o_i = i}) \psi_i(t_i, t_{\partial i}). \tag{3.27}$$

Where we defined the observation factor ξ_i as:

$$\xi_i(\tau_i, t_i|\{\varepsilon_o, t_o\}_{o_i = i}) = \prod_{o \in \mathcal{O}: o_i = i} p^*(o|\tau_i, \varepsilon_o, t_o) p(o|t_i)$$

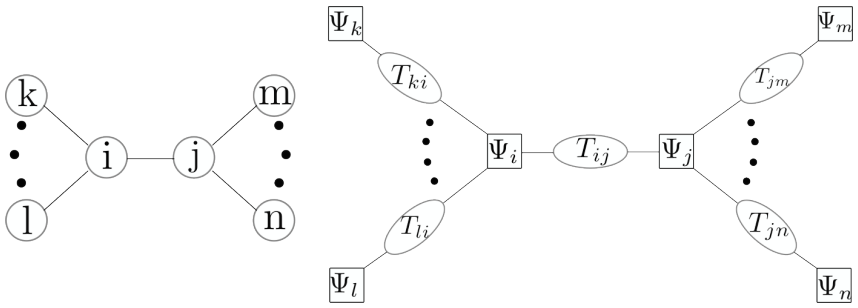

Fig. 3.8 The factor graph representation of the joint distribution in equation (3.28), which corresponds to the joint (3.27) in which loops have been removed by introducing copies. *Left:* a tree-like network of individuals i, j, k, l, m, n. *Right:* the corresponding factor graph

We have a factor graph form, as in equation (3.1), which can be treated similarly to equation (3.9) in Sib. The main difference is that we now have that the dynamical variables are both the inferred $\underline{t}$ and the planted $\underline{\tau}$. We therefore have to define super-variable nodes containing couples of planted and inferred times. Similarly to the single-instance case, the factor graph associated to equation (3.27) contains loops even if the underlying contact network is acyclic. Therefore, copies of the infection times (both planted and inferred) must be introduced. We therefore use the same idea of Sib by placing a function node Ψ_i for each individual $i \in \{1, \dots, N\}$; for each edge (i, j) we place instead a super-variable node $T_{ij} := (\tau_i^{(j)}, \tau_j^{(i)}, t_i^{(j)}, t_j^{(i)})$, as shown in Figure 3.8. Adding the constraints that the copies of the same infection time must be equal to each others, the joint distribution in equation (3.27) becomes:

$$P(\{T_{ij}\}_{(ij)\in\mathcal{E}}|\mathcal{D}, \mathcal{D}_o) = \frac{1}{Z(\mathcal{D}, \mathcal{D}_o)} = \prod_{i=1}^{N} \Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i), \tag{3.28}$$

where $\mathcal{D}_i = \{\{\varepsilon_o, t_o\}_{o:o_i=i}, \{s_{li}\}_{l\in\partial i}, x_i^0\}$ and each factor is:

$$\begin{aligned}\Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i) = \xi(\tau_i^{(j)}, t_i^{(j)}|\{\varepsilon_o, t_o\}_{o_i=i})\psi^*(\tau_i^{(j)}, \tau_{\partial i}^{(i)}|\{s_{li}\}_{l\in\partial i}, x_i^0)\times \\ \times\, \psi(t_i^{(j)}, t_{\partial i}^{(i)}) \prod_{l\in\partial i} \delta_{t_i^{(j)}, t_i^{(l)}} \delta_{\tau_i^{(j)}, \tau_i^{(l)}}\end{aligned} \tag{3.29}$$

Now we finally have a factor graph on which BP equations can run.

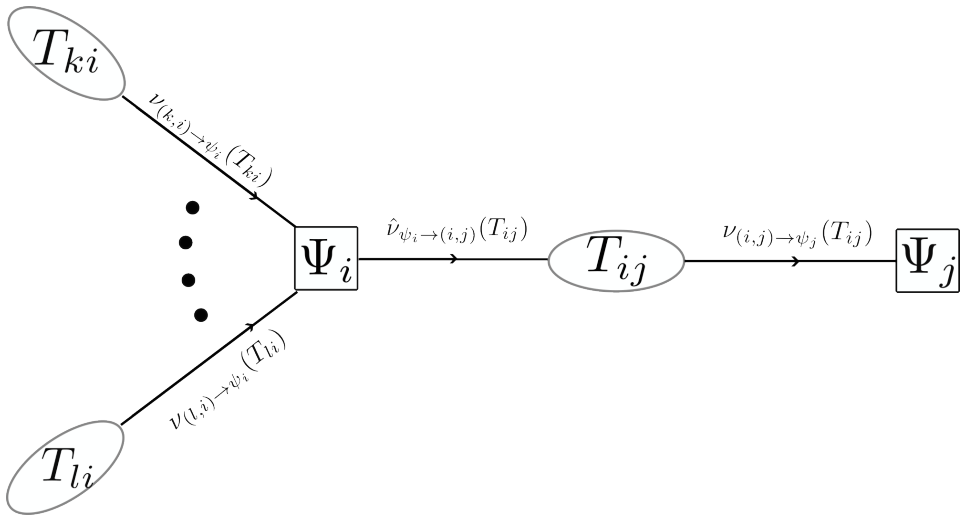

Fig. 3.9 BP messages for the joint factor graph.

3.3.5 Cavity messages for the enlarged distribution

With the help of Figure 3.9, the BP update equations are:

$$\hat{\nu}_{\Psi_i\to(i,j)}(T_{ij}) = \frac{1}{z_{i\to j}} \sum_{\{T_{ki}\}_{k\in\partial i\setminus j}} \Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i) \prod_{k\in\partial i\setminus j} \nu_{(k,i)\to\Psi_i}(T_{ki})$$
$$\nu_{(i,j)\to\Psi_j}(T_{ij}) = \hat{\nu}_{\Psi_i\to(i,j)}(T_{ij})$$

which reduce to:

$$\hat{\nu}_{\Psi_i\to(i,j)}(T_{ij}) = \frac{1}{z_{i\to j}} \sum_{\{T_{ki}\}_{k\in\partial i\setminus j}} \Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i) \prod_{k\in\partial i\setminus j} \hat{\nu}_{\Psi_k\to(k,i)}(T_{ki}). \tag{3.30}$$

If we call T the horizon time (the total number of time-steps in the epidemic process), then each message is defined over a domain of $O(T^4)$ values. We expect, therefore, the BP algorithm to scale with a fourth power w.r.t. total time. This performance can be optimized to $O(T^2)$, as shown in the Appendix D and in [44]. Equation (3.30) is the analogous of the Sib equation (3.14), with the remarkable difference that this allows to build the ensemble counterpart by means of cavity equations. It is sufficient to extract a sub-vector of disorder (in particular: a message to update in the population, its residual degree d_{res}, the time zero state x_i^0, the set of incoming delays $\{s_{li}\}_{l\in\partial i}$, the corruption bits and

Algorithm 5 Replica symmetric cavity method update using population dynamics.

Input: degree distribution $p(d)$, infection probability λ, patient zero probability γ, population $\{\hat{\nu}_i\}_{i=1}^n$ of size n, false rate f.
Output: nothing, the Algorithm updates one message of the population.

- Extract $i \in \{1, \dots, n\}$ uniformly random, this is the index of the message $\hat{\nu}_i$ that will be updated.
- Extract, from the residual distribution of $p(d)$, the residual degree $d_{res} \sim p_{res}(d_{res})$, see Appendix B
- Extract, from the population, d_{res} incoming messages and call them $\{\nu_k\}_{k \in \partial i \setminus j}$
- Set the initial state $x_i^0 = 1$ with probability γ, otherwise set $x_i^0 = 0$.
- Extract d_{res} independent incoming delays $s_1, \dots, s_{d_{res}}$, sampled from $p(s) = \lambda(1-\lambda)^{s-1}$.
- Extract a random number of observations n_{obs} on individual i.
- **for** $o \in n_{obs}$:
  - extract a uniform random observation time $t_o \in \{0, \dots, T\}$.
  - extract a corruption bit $\varepsilon_o = 1$ with probability f.
- Now equation (3.30) can be computed. Use it to update $\hat{\nu}_i$.

the observation times $\{\varepsilon_o, t_o\}_{o \in \mathcal{O}}$, which we remark here are all independent of each others), as explained in Algorithm 5.

3.3.6 Marginal computation

The computation of marginals from the messages is a crucial task. For example, it allows to check the convergence of the algorithm: suppose we have, after k sweeps[4], that the population is $\{\hat{\nu}_i^{(k)}\}_{i=1}^n$. To compute the marginals, we just

[4]a *sweep* is defined as an update of the population, so n message updates form a sweep.

Algorithm 6 Computing the average marginal

Input: degree distribution $p(d)$, infection probability λ, patient zero probability γ, population $\{\hat{\nu}_i\}_{i=1}^n$ of size n, false rate f.
Output: the average marginal.

- Initialize the average marginal m as a matrix of zeros with the size of $T \times T$.
- **for** $i = 1, \ldots, n$
  - Extract the degree d from the degree distribution $p(d)$,
  - take d messages from the population
  - extract one incoming infection delays for each of the d messages
  - extract one time-zero state x_i^0
  - extract a number of observations, the observation times, the error bit
  - use equation (3.31) to compute the belief $b(t_i, \tau_i)$ for each t_i and τ_i
  - $m(t_i, \tau_i) += b(t_i, \tau_i)/n$
- **Return** m

have to extract the disorder and use the BP expression for the beliefs:

$$b(t_i, \tau_i) = \frac{1}{z_i} \sum_{\{T_{il}\}_{l \in \partial i}} \Psi_i(\{T_{il}\}_{l \in \partial i} | \mathcal{D}_i) \prod_{l \in \partial i} \delta_{t_i, t_i^{(l)}} \delta_{\tau_i, \tau_i^{(l)}} \hat{\nu}_{\Psi_l \to (l,i)}(T_{li}). \tag{3.31}$$

By averaging over the disorder, the average belief is found, as described in Algorithm 6. Once the average marginal $m(t_i, \tau_i)$ is found, we can compute for example the fraction of prior infectious individuals:

$$n_{\mathcal{I}}^*(t) = \sum_{\tau_i \leq t} \sum_{t_i} m(t_i, \tau_i) = \sum_{\tau_i \leq t} m^*(\tau_i).$$

This function should converge (apart from fluctuations of the order $1/\sqrt{n}$ due to the finite population size n) to a fixed function when the cavity method converges in population. Computing marginals is therefore a good way to build a convergence criterion. Notice that the posterior number of infectious

individuals,

$$n_{\mathcal{I}}^{\mathcal{P}}(t) = \sum_{\tau_i} \sum_{t_i \leq t} m(t_i, \tau_i) = \sum_{t_i \leq t} m(t_i),$$

must be equal to the prior number $n_{\mathcal{I}}^{*}(t)$ under the Bayes optimal conditions. This is a direct consequence of the Nishimori conditions, section 1.2.6.

3.3.7 Bethe free energy computation

Using equation (3.8):

$$\begin{aligned} F_{\text{Bethe}} = -\log Z_{\text{Bethe}} \; &= -\sum_a \log z_a + \sum_{(a,i)} \log z_{ia} - \sum_i \log z_i = \\ &= -\sum_i \log z_{\Psi_i} + \frac{1}{2} \sum_{i,j \in \partial i} \log z_{ij}, \end{aligned}$$

we can compute the Bethe free energy. Recall the definitions of each term:

$$\begin{aligned} z_a &= \sum_{\boldsymbol{x}_a} \psi_a(\boldsymbol{x}_a) \prod_{i \in \partial a} \nu_{i \to a}(x_i) \\ z_{ia} &= \sum_{x_i} \hat{\nu}_{a \to i}(x_i) \nu_{i \to a}(x_i) \\ z_i &= \sum_{x_i} \prod_{a \in \partial i} \hat{\nu}_{a \to i}(x_i) \end{aligned}$$

In our framework, each variable node is only attached to two function nodes by construction. Moreover, each function node corresponds to one individual and each variable node corresponds to an edge in the contact network. The three quantities above become:

$$\begin{aligned} z_{\Psi_i} &= \sum_{\{T_{ji}\}_{j \in \partial i}} \Psi_i(\{T_{ij}\}_{j \in \partial i}) \prod_{j \in \partial i} \hat{\nu}_{\Psi_j \to (i,j)}(T_{ij}) \\ z_{(i,j)\Psi_i} &= \sum_{T_{ij}} \hat{\nu}_{\Psi_i \to (i,j)}(T_{ij}) \nu_{(i,j) \to \Psi_i}(T_{ij}) \\ z_{(i,j)} &= \sum_{T_{ij}} \hat{\nu}_{\Psi_j \to (i,j)}(T_{ij}) \hat{\nu}_{\Psi_i \to (i,j)}(T_{ij}) = \end{aligned}$$

Notice that the second and the third term for our problem become identical: using the BP equation for the third term:

$$\begin{aligned} z_{(i,j)} &= \sum_{T_{ij}} \hat{\nu}_{\Psi_j \to (i,j)}(T_{ij}) \hat{\nu}_{\Psi_i \to (i,j)}(T_{ij}) = \\ &= \sum_{T_{ij}} \nu_{(i,j) \to \Psi_i}(T_{ij}) \hat{\nu}_{\Psi_i \to (i,j)}(T_{ij}) = \\ &= z_{(i,j)\Psi_i} \end{aligned}$$

The Bethe free energy is therefore transformed to:

$$\begin{aligned} F_{\text{Bethe}} &= -\sum_i \log z_{\Psi_i} + 2 \sum_{(i,j)} \log z_{(i,j)} - \sum_{(i,j)} \log z_{(i,j)} \\ &= -\sum_i \log z_{\Psi_i} + \sum_{(i,j)} \log z_{(i,j)} = \\ &= -\sum_i \log z_{\Psi_i} + \frac{1}{2} \sum_{i,j \in \partial i} \log z_{(i,j)}. \end{aligned}$$

This formula is used to compute the Bethe free energy in the Epidemble algorithm. The free energy is useful for several reasons:

1. it quantifies how informative observations are. In fact, the Bethe free energy is an approximation[5] of the free energy, $F = -\log P(\mathcal{O})$, which is big when $P(\mathcal{O}) = \sum_{\underline{t}} P(\underline{t}, \mathcal{O}) = \sum_{\underline{t}} P(\underline{t}) P(\mathcal{O}|\underline{t})$ is small. Thus, the quantity $P(\mathcal{O})$ is the sum of trajectories (weighted with their prior probability) which are compatible with the observation constraints $\mathcal{O}$. The smaller this sum, the more the observation set is reducing the space of possible trajectories.

2. it allows to perform hyper-parameters inference. The Bethe free energy is in fact a function of the hyper parameters, so it can be descended to infer them. In the results section, we are going to show some results obtained by inferring the patient zero probability γ, the infection probability λ and also the false rate o_f of observations. To infer the parameter γ, the Expectation Maximization method is used, see Appendix C. The other

[5] exact for the graphs studied in the following results section if the posterior is replica symmetric.

parameters are inferred by means of a standard gradient descent on the Bethe free energy.

3. it can be used, in addition to marginals, to check convergence.

3.4 Results

This results section is partially based on the published paper [44].

In this section we discuss results obtained with the Epidemble algorithm. Initially, we characterize estimators under Bayes optimal conditions. Quantifying hardness of inference is, in fact, a nontrivial problem: some estimators might give discordant predictions about hardness. For example, we will see that AUC is low in some regimes where maximum mean overlap (MMO) is high. It is therefore very useful to study the behavior of several estimators, which give us a more complete description of hardness in inference. Analyzing the results, in fact, we will be able to comment about the origin of such differences. Another point analyzed in this section is the consistency between the ensemble and the single instance algorithm (Sib). We are going to compare the cavity method predictions on the estimators with the Sib results obtained on large graphs with the same degree distribution. Subsequently, we deal with the analysis of a particular regime (high infection λ and low patient zero probability γ), for which the posterior shows symptoms of replica symmetry breaking. We present a conjecture about possible replica symmetry breaking under Bayes optimal conditions. Then we move out of Bayes optimality, we study what happens if we make inference with mismatched hyper-parameters (i.e. different from the prior's). In this case replica symmetry breaking is found, as expected. Learning the hyper-parameters by means of the expectation maximization method (Appendix C), allows to recover replica symmetry. The convergence criterion of Epidemble is on the marginals, which must not fluctuate more than the square root of size of the population. If convergence criterion is not reached, the algorithm stops after a fixed number of sweeps (typically the population size is set to $n \sim 10^4$ and the total number of sweeps to 200). Convergence is almost always reached (unless in the special case in which there is suspected RSB) when the prior is known (or inferred), except for the rather interesting

and unexpected regime which shows symptoms of replica symmetry breaking, which is discussed later on. The algorithm shows non-convergence zones, as expected, also when the prior hyper-parameters are not known and we enter the replica symmetry broken phase.

3.4.1 Statistical estimators in Bayes-Optimal case

In the Bayes optimal case the hyper-parameters of the prior are known. Under this condition we start by characterizing the Minimum Mean Squared Error (MMSE), the Maximum Mean Overlap (MMO), the Area Under the ROC (AUC) (see section 1.1.5 for their definition) and the Bethe free energy (Fe) associated with the posterior distribution, computed in section 3.3.7. Notice that, in the ensemble case, computing the ROC is impossible because it would require to evaluate and order the (infinite) list of individual probabilities of being infectious. However, the area under the curve (AUC) can be interpreted as the probability that, given one positive individual i and one negative individual j, their posterior marginal probabilities allow distinguishing which is positive and which is negative. In other words,

$$\mathrm{AUC}(t) = \mathbb{P}\left[P_i(x_i^t = \mathcal{I}|\mathcal{O}) > P_j(x_j^t = \mathcal{I}|\mathcal{O})\middle| x_i^{*,t} = \mathcal{I}, x_j^{*,t} = \mathcal{S}\right].$$

This allows us to compute the AUC in the ensemble case. In Figure 3.10, the fraction of unobserved individuals (dilution) was fixed to dil $= 0.5$ (half of the individuals are observed). All the observations were made at final time $T = 8$. The 2D space explored is that of the patient zero probability and infection, (γ, λ). MMSE, MMO and AUC are shown at three different times (initial time $t = 0$, intermediate time $t = 4$ and final time $t = T = 8$). We can see that MMSE and MMO show the same behavior at all times. For very low infection probability λ, and patient zero probability γ, we see that MMSE is low while MMO and AUC are high, meaning that information contained in the inferred posterior distribution allows to recover the planted configuration with good accuracy. In this regime, in fact, patients zero are few and infect rarely. Thus, they are on average surrounded by a small neighborhood of infectious individuals and well-separated from the other patients zero, making inference

task easy. For high values of patient zero probability γ and infection λ, instead, the population becomes completely infectious in few time steps. Also in this regime, all the estimators show great performance for intermediate ($t = 4$) and final time ($t = T = 8$), because the posterior marginals assign to every individual a probability 1 of being infectious. The hard task is to retrieve the patient zero. At $t = 0$, indeed, MMSE (respectively MMO) is low (resp. high) for high values of γ. However, this does not mean that inference performance is good. Indeed for large γ, the majority of individuals are patients zero, and the other individuals are likely to be infected before the observation time T. Therefore, the observations are (almost) all positive, making it impossible to distinguish the patients zero from the ones infected at later time. Thus, MMSE at time $t = 0$ is low because the marginal posteriors give high probability of being infectious at $t = 0$, independently of the transmission rate λ. However, the (few) non-patients zero remain undetected. A quantity that is sensible to this problem is the AUC, which at time $t = 0$ has in fact a different behavior to the other estimators, signaling hardness of inference. Another (slightly) different quantity is the AUC evaluated only on non observed individuals. When many observations are made, the AUC is dominated by the observed individuals. Thus, evaluating AUC only on non observed individuals (AUCNO) can be a useful tool to quantify the prediction power of the algorithm on unobserved individuals. To see the difference between AUC and AUCNO, we fix the patient zero probability $\gamma = 0.1$, and show in Figure 3.11 these two estimators as functions of the infection probability λ and the observations dilution *dil* (i.e. the fraction of unobserved individuals). We see that the two estimators behave differently, for example at the intermediate time $t = 4$, for low dilution (i.e. many observations) and low transmission rate λ. In this regime there are only few infectious individuals observed (because γ and λ are low). While AUC is close to 1, AUCNO is low, indicating that it is actually hard to find who are the unobserved infectious individuals.

3.4.2 Check of consistency with large single instances

It is natural to wonder whether ensemble results are consistent with large finite-size single instances. To check this point, we see Figure 3.12, where results on large sized ($N = 30000$) instances, obtained by means of Sib algorithm, are

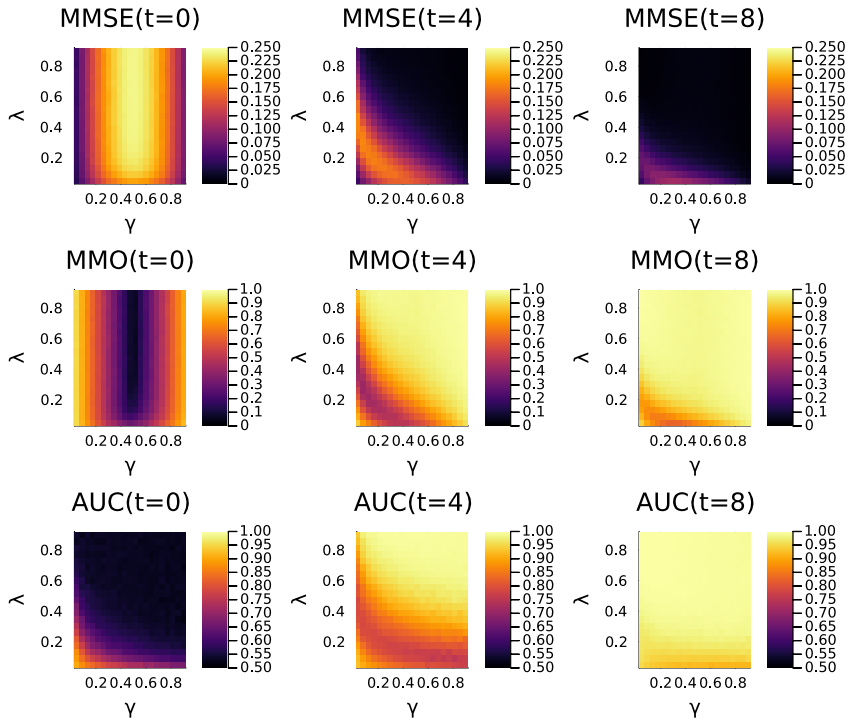

Fig. 3.10 Several statistical estimators (first row: MMSE, second row: MMO, third row: AUC) quantifying the hardness of epidemic inference, as a function of patient zero probability γ and infection probability λ. Each column corresponds to three different times at which the quantities are computed (from left to right: initial time $t = 0$, intermediate time $t = 4$, and final time $t = T = 8$). The three estimators display the same behavior, except for the initial time, when AUC is able to capture for high values of λ and γ that observations are not informative enough. Notice that MMSE quantifies the error in inferring individual's states, so it has a flipped behavior with respect to the other quantities (MMO and AUC are high when inference performance is good). These results were obtained for ER graph ensemble with average degree 3. Figure taken from [44].

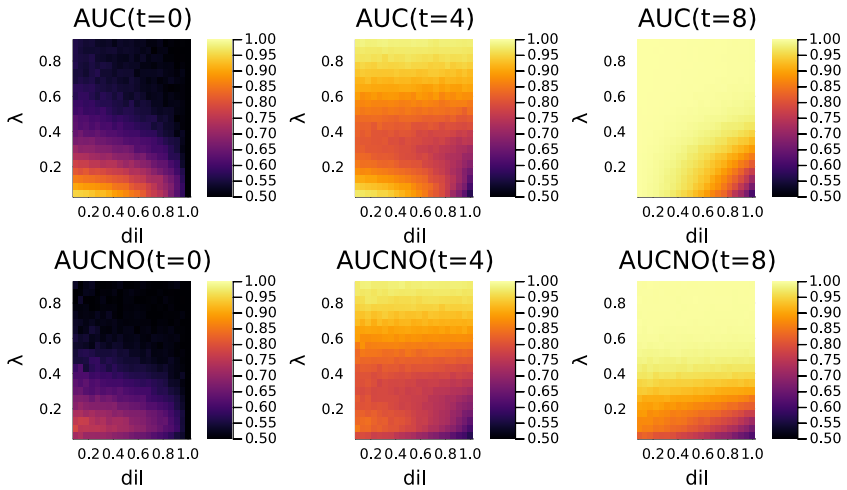

Fig. 3.11 A comparison between AUC evaluated on all individuals (AUC, first row) VS AUC evaluated only on unobserved individuals (AUCNO, second row). The two estimators have a very similar but not identical behavior. In particular, at low dilution (many observations) the AUCNO is systematically smaller than the AUC. In fact, the AUC in this regime is dominated by observed individuals. These results are for ER graph ensemble, with average degree 3. We remark here that AUC is not 0.5 for dilution equal to 1. In fact, ER graphs are heterogeneous (with a Poisson law degree distribution). This implies that some information about the infection probability of each node is contained in the graph itself. For example, the most connected nodes have highest probability of being infected. This allows to achieve some reconstruction also without any observation (dil = 1). Figure taken from [44]

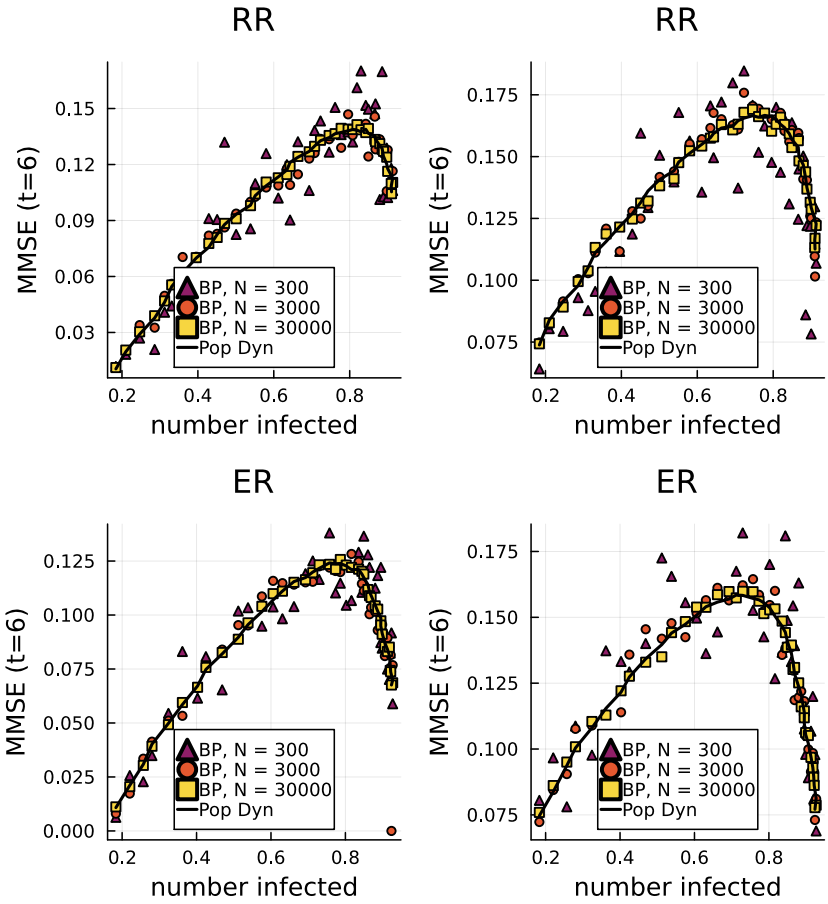

Fig. 3.12 The comparison between Sibyl algorithm (BP) for a single instance of $N = 300$ (triangles), $N = 3000$ (dots), and $N = 30000$ (squares) individuals, and the replica symmetric cavity method results obtained in the thermodynamic limit by means of population dynamics (black solid line). The plots show the MMSE at intermediate time $t = 6$, as function of the number of infectious at final time $T = 8$, which is a function of infection parameters γ and λ. For this plot the patient zero probability is fixed at $\gamma = 0.15$. The first row represents the MMSE for Random Regular graphs (degree 3) while the second row is for Erdős–Rényi (ER) with average degree 3. Each column, instead, is associated with a value of observations dilution *dil*: the first column is for *dil*=0 (all observed) while the second is for *dil*=0.5. There is very good agreement, that increases with the size of the the single instance contact graph.Figure taken from [44]

compared with Epidemble. The MMSE is computed both for the ensemble and the large single instance case and the comparison shows almost perfect consistency.

3.4.3 Ensemble Bethe free energy

All the estimators described so far are time-dependent. The Bethe Free Energy, section 3.3.7, quantifies how much information is in the observations set $\mathcal{O}$ and it is a single number at fixed hyper-paramters. The plots in Figure 3.13 show the Bethe free energy for two different frameworks (analyzed in the previous results sections):

- at fixed observation dilution *dil* and varying γ, λ
- at fixed γ varying (dil, λ) .

The free energy is 0 for *dil*=1 (no observation), as expected. However, it is close to 0 in other cases too, e.g. for high values of infection probability λ. For those values, the infection indeed spreads very fast. As a consequence, at final time all the individuals are infectious. Thus, since the observations are only collected at final time T, they do not carry any valuable information on the planted trajectory: they will always find individuals in the state $\mathcal{I}$. In other words, all the trajectories of the prior end up at time T in the state for which all individuals are infectious. Note, however, that inference is easy in this regime, as it can be checked comparing Figure 3.13 with Figures 3.10 and 3.11. In this regime, although observations are not informative, the prior is concentrated on few trajectories (the ones compatible with all individuals being $\mathcal{I}$ at times $t = 4$ and $t = 8$), making inference trivial. The interesting (and hard) regimes are at intermediate values of γ and λ and for non-zero dilution. In this regime the prior is not concentrated on few trajectories, making inference a non-trivial task.

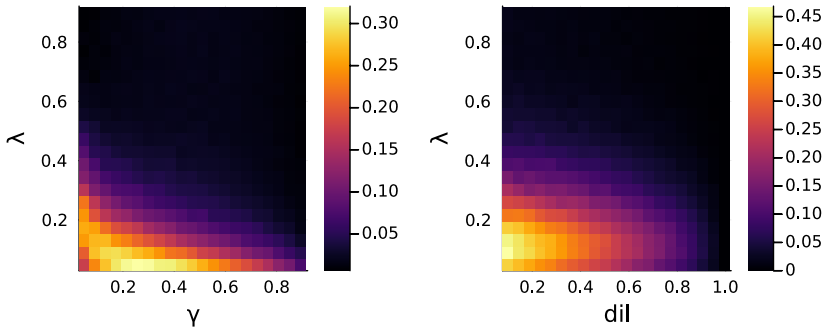

Fig. 3.13 Free energy profile for two different regimes: on the *left panel* as a function of patient zero probability γ and infection probability λ, at fixed dilution $dil = 0.5$; on the *right panel* as a function of observations dilution dil and λ, at fixed patient zero probability $\gamma = 0.1$. The black part of the plot corresponds to the regimes in which observations bring no information, i.e. $F \simeq 0$. This happens obviously at dil=1 because no observation is collected. However, the free energy can be zero also for $dil < 1$. Indeed, when the infection probability λ is high enough, all the individuals are almost surely $\mathcal{I}$ at final time. Observations, which are only made at final time, carry no information in those cases. Only in the intermediate regimes, i.e. when the number of $\mathcal{I}$ and $\mathcal{S}$ individuals are comparable with each others, observations are informative. In this regime the free energy is non-zero and inference is non trivial. The graph ensemble analyzed here is Erdős–Rényi with average degree 3. Figure taken from [44].

3.4.4 More on graph ensembles

The analysis shown so far has been conducted on Erdős–Rényi graphs. To study how inference tasks are affected by the graph structure, we now compare results on three ensembles:

1. the Random Regular (RR) ensemble;
2. the Erdős–Rényi (ER) ensemble;
3. a truncated *fat tailed* (FT) ensemble of graphs, with the degree distribution $p(d) = \frac{1}{Z}\frac{1}{d^2+a}$ for $d \in [d_{min}, d_{max}]$ and $p(d) = 0$ if $d \notin [d_{min}, d_{max}]$. The quantity Z is the normalization of the distribution and the parameter a can be fixed by fixing the average degree.

The third graph ensemble is interesting because it presents highly connected nodes, while still being handled by Belief Propagation (BP), since the distribution of the degree is truncated to a finite maximum value d_{max}. In Figure 3.14 (first row), the Minimum Mean Squared Error (MMSE) at time $t = 6$ is shown for the three graph ensembles. The average degree is fixed to 3 in all three cases. This allows to compare the effects of changing the ensemble. The predictions are thus sensible to the ensemble of graphs chosen. Highly connected nodes, in fact, play a crucial role in changing inference hardness in some regimes. This pushes us to look at applications of the method to a more *realistic* family of graphs. Before doing so, we briefly discuss the effect of noise in the observations.

Noise in observations. When noise affects observations, the inference results get typically worse. This can be seen in Figure 3.14(second row), where the AUC is shown as a function of observations dilution and noise (false rate, *fr*). For false rate equal to 0.5, observations carry no information, since they are wrong half of the time. This is identical to have no observations, i.e. dilution $dil = 1$. For intermediate values, we see that increasing false rate and/or dilution always leads to worse inference, as expected.

Application to a real network. We show here how Epidemble predictions on the infinite size limit can be applied to more realistic random ensembles.

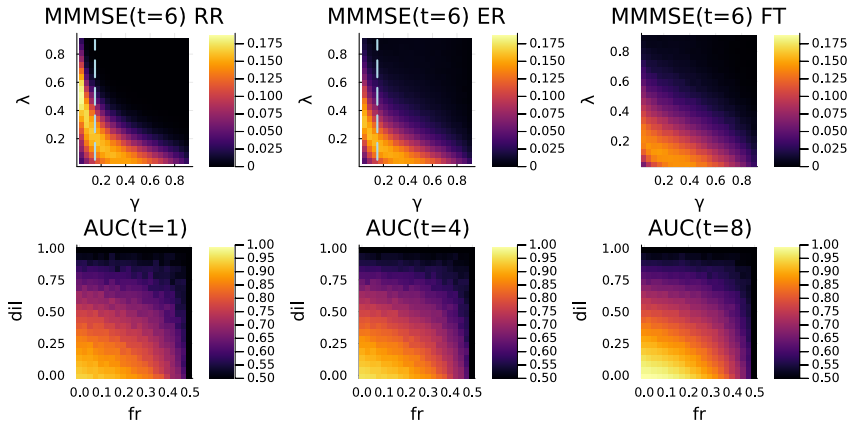

Fig. 3.14 Comparing feasibility of inference for several graph ensembles and for nonzero observation noise. *First row*: the plots show the MMSE at time $t = 6$, with observations collected at final time $T = 8$, as functions of the patient zero probability γ and the infection probability λ. The three plots are (from left to right) for Random Regular (RR), Erdős–Rényi (ER) and Fat Tailed (FT) graph ensembles. The average degree is fixed to 3 for these three ensembles examined. It can be seen that the profiles share similarities, but the more the degree distribution widens (from RR to ER to FT), the flatter is the MMSE. This is due to high-degree nodes: for example, in RR ensemble all nodes have the same degree, so we see that inference is more difficult at low values of γ and high values of λ. In this region, the presence of highly connected nodes simplifies inference because they (and their neighbors) will probably be infectious at time $t = 6$. The dashed lines correspond to the cases studied in Figure 3.12 at dilution 0.5. The only difference is on the y-axis, which represents λ in this and the fraction of infectious for Figure 3.12. *Second row*: the AUC as a function of observations' dilution *dil* and false rate (*fr*). The AUC decreases with *fr* and *dil*. The false positive and negative rates are always assumed to be the same. The patient zero probability is fixed to $\gamma = 0.03$ and the infection probability is $\lambda = 0.03$. The ensemble graph is Random Regular with degree 11. Figure taken from [44].}

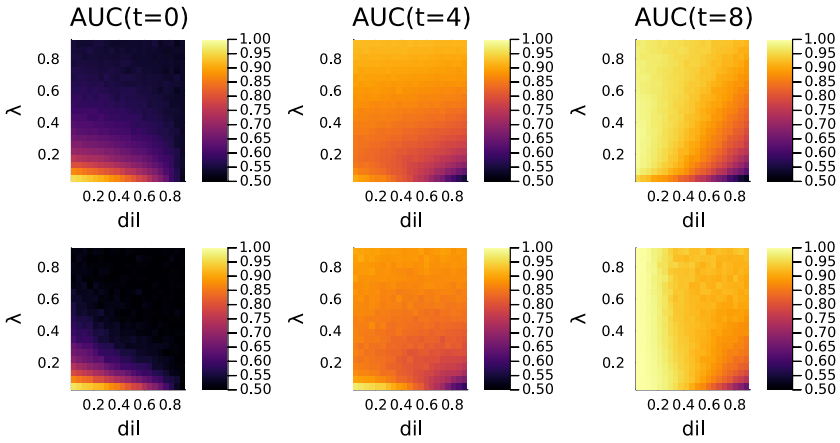

Fig. 3.15 An application of the cavity method to a real network. First row: ensemble predictions for AUC at time 0,4,8. Second row: results obtained on the real network using Sib. Ensemble predictions and Sib (BP) results are near to each others. Some differences can be however noticed for low values of the infection probability λ, where Sib shows higher performance. An explanation should be related to the fact that at a single instance level the observations $\mathcal{O}$ are more informative than in the ensemble case: if one individual is observed in the state $\mathcal{S}$, then the infection cascade did not pass through it. Therefore, correlations among its neighboring individuals drop down due to the $\mathcal{S}$ observation, which factually cuts the network. This reasoning does not subsist in the ensemble case, where the network is not fixed. Therefore, we expect lower AUC in the regimes where $\mathcal{S}$ observations are more frequent, thus at low infection probability. For this plot, the value of the patient zero probability is fixed at $\gamma = 0.2$ and there is no noise in observations, which are all at time $T = 8$. Figure taken from [44].

Given a real finite network, we consider the configuration model ensemble with the same degree distribution of the network. Using cavity method, it is thus possible to approximately find the expected value of each estimator in the thermodynamic limit. This procedure is now applied to a network of sexual contacts from the data repository [68]. The ensemble results are compared with results on the single instance obtained with Sib in Figure 3.15. The first row shows the ensemble predictions, while the second shows single-instance results obtained with Sib. Phase diagrams show the AUC at initial, intermediate and final time as function of the infection probability and the observation dilution in the Bayes Optimal regime. The agreement between the two results is good, implying that the quality of inference performance in this setup depends mostly on the degree distribution, which is a global characteristic of the network.

3.4.5 RSB under Bayes optimality?

A surprising behavior of the Belief Propagation equations was firstly observed in [69] for single instances at small values of γ. In fact, even in the Bayes optimal conditions, BP stops to converge. This breakdown is present also in the thermodynamic limit for the cavity method. This lack of convergence, therefore, seems to be related to a rather profound reason. To understand what is happening, we simplify now the framework by setting the infection probability λ to 1 and by observing all the individuals at final time. This regime is the one studied in Figure 3.16. The black dots represent the number of iterations needed for the cavity method (implemented using population dynamics) to converge. Around $\gamma = \tilde{\gamma} = 0.013$ the algorithm stops converging. An intuitive explanation is the following: for γ around $\tilde{\gamma}$ at final time many individuals are observed infectious ($\mathcal{I}$) and a small (but extensive) part is observed susceptible ($\mathcal{S}$). Since $\lambda = 1$, the sole non-deterministic part of the process is the initial state, so the inference problem reduces to guess the position of the patients zero. The $\mathcal{S}$ individuals not only signal that they were not infected during the epidemic process, but also that any patients zero must be at distance $> T$. For example, for a RR graph, a $\mathcal{S}$ observation excludes the sphere centered in the $\mathcal{S}$-observed individual with $d(d-1)^{T-1}$ individuals. For γ around $\tilde{\gamma}$ these spheres touch and intersect, so that the group of individuals eligible to be the patients zero gets separated in clusters. In Figure 3.17 a 2D plot possible

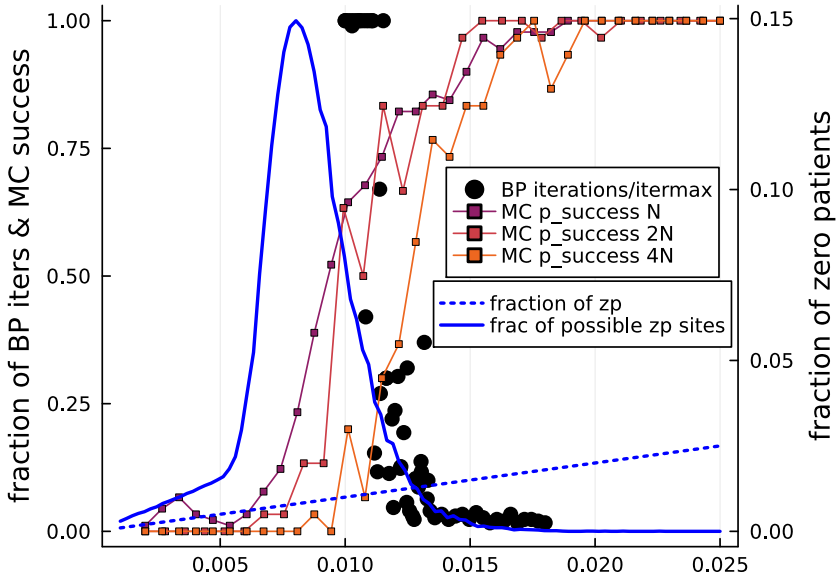

Fig. 3.16 Convergence time for replica symmetric cavity, belief propagation and Monte Carlo methods compared with the number of clusters of possible patients zero in a RR graph with degree 3. Convergence of the three methods breaks down at around $\tilde{\gamma} \simeq 0.013$. The black dots represent the number of iterations for the cavity method (implemented by means of population dynamics) to reach convergence, normalized by the total number of iterations allowed. The continuous squared-marked lines represent the fraction of successful Monte Carlo runs, for several sizes ($N, 2N, 4N$, with $N = 5000$). Convergence is conjectured to be lost due to a Replica Symmetry Breaking transition: the space in which a patient zero can be placed in the posterior becomes clustered (as explained in Figure 3.17). To support this conjecture, it is plotted the fraction of connected components (number of connected components divided by total number of individuals) of individuals that could be the patients zero. This number, as expected, grows sharply in the interval in which BP ceases to converge. The failure of convergence arises when the number of disconnected zones to place the patient zero (continue, blue line) becomes higher to the actual fraction of patients zero (dotted, blue line). This suggests that when the number of zones in which a patient zero might be placed becomes larger than the number of patients zero, then the problem gets hard, as illustrated in Figure 3.17. Figure taken from [44].

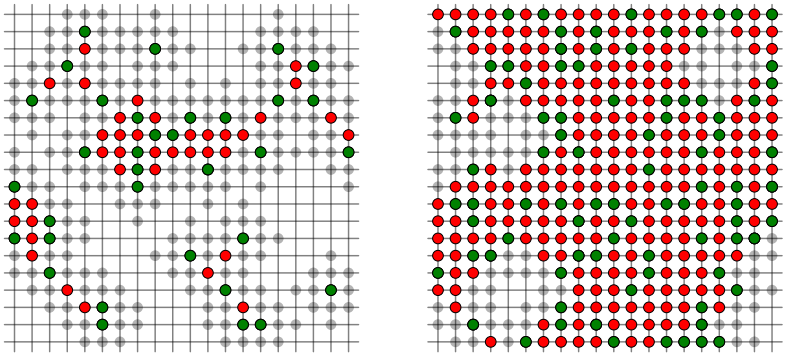

Fig. 3.17 A 2D plot to visualize the geometric change undergone by the configuration space that could explain why cavity method and Monte Carlo schemes stop to converge. In this plot, obtained by simulating epidemic spreading in a 2 dimensional lattice, two scenarios are compared. To the right, γ is higher, namely there are more patients zero (green dots). This implies that the number of infectious (green, red and gray dots) is higher, so the number of $\mathcal{S}$-observed individuals (no dots) is smaller. The patients zero can not be too close to the $\mathcal{S}$-observed individuals because the infection probability is 1, so the observation constraint would be violated. The red dots represent all the individuals which might be the patient zero according to the observations (i.e. individuals tested $\mathcal{I}$ and not too close to $\mathcal{S}$-observed individuals). When the number of patients zero is lower, (*left*) the number of $\mathcal{S}$-observed individuals increases. So the number of possible zones to accommodate patients zero (green plus red dots) reduces and gets clustered. This could result in several separated states of the posterior, each one corresponding to a possible combination of placements for patients zero. Figure taken from [44].

explanation of the phenomenon is given. When the number of $\mathcal{S}$ observations is sufficiently high, due to the fact that there are few patients zero, the space in which patients zero could physically be gets fragmented. To check that this is what actually happens in a Random Regular graph, a single instance was initialized and an epidemic was simulated. The number of connected components in which a patient zero could be present without violating the S observations is plotted in Figure 3.16. This number sharply increases in the decreasing γ direction, around $\tilde{\gamma}$, i.e. when the algorithm stops converging. Further evidence of a phase transition is given by Monte-Carlo dynamics, whose convergence time is also shown in Figure 3.16. The MC was designed in [44] as follows:

1. The planted $\underline{\tau}$ (ground truth) configuration is sampled from the single instance prior.

2. Observations $\mathcal{O}$ are collected from $\underline{\tau}$. The observation protocol is set to observe all the individuals at the final time T (without observation noise).

3. A Metropolis-Hasting Monte Carlo simulation was started in order to sample a configuration satisfying all the observations. To do so, a configuration $\underline{t} \neq \underline{\tau}$ (the planted $\underline{\tau}$ is unknown in the inference process) is sampled from the prior distribution. The initialization configuration typically does not satisfy the observations, which are taken from $\underline{\tau}$. So the following MC move is made:

   (a) An individual is randomly selected and change its time-zero state is changed by sampling the $\mathcal{I}$ state with probability γ (and the $\mathcal{S}$ state with probability $(1-\gamma)$).

   (b) The initial state configuration is evolved (deterministically, since the infection probability is $\lambda = 1$). The configuration at final time is then checked to be consistent with the observations. If that happens, that MC has converged to an acceptable configuration...

   (c) ...otherwise the energy is computed as: $U = -\sum_{i=1}^{N} \log p(o_i|t_i)$ where each o_i is the observation on the ith individual. In principle $p(o_i|t_i)$ should be either 0 (when the configuration does not satisfy the observations) or 1 (when the observation is satisfied). In order to

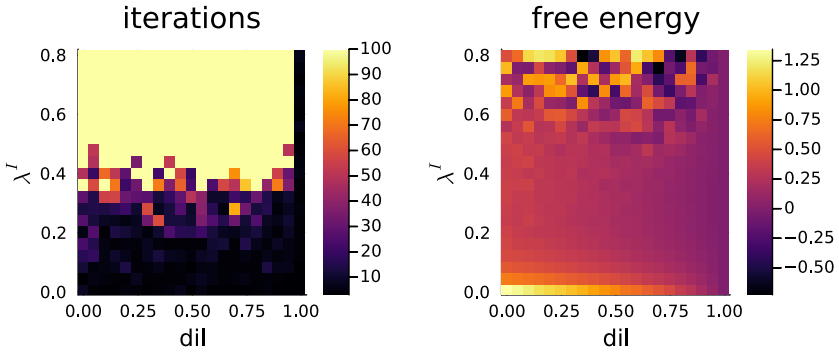

Fig. 3.18 Replica symmetric cavity method's convergence breakdown outside Nishimori conditions. Keeping fixed $\lambda^* = 0.3$ and the parameters $\gamma^* = \gamma^I = 0.03$, while instead moving the inference parameter $0 < \lambda^I < 0.8$ and the dilution between 0 and 1 we see that for high values of λ^I the algorithm does not converge and provides nonphysical results for the Bethe free energy. Figure taken from [44].

avoid infinite energy barriers, a small artificial noise in observations is dded, which is reduced during the Monte Carlo simulation by means of an annealing procedure. In other words, the energy is just a penalization for each broken constraint.

4. At each step, the move in the space of initial states described above is made. The move is accepted by following a standard Metropolis scheme. The MC stops when the configuration satisfies all constraints.

For each value of γ the MC scheme was repeated 60 times. In Figure 3.16 the fraction of runs in which the MC algorithm was able to reach convergence is shown. We clearly see that this quantity drops down around $\tilde{\gamma}$. Due to the failure of BP equations (for finite and infinite graph), the explosion of possible patient zero zones and the failure of the Monte Carlo scheme, Replica Symmetry Breaking transition is conjectured to arise around $\tilde{\gamma}$.

3.4.6 Departing from Bayes-optimal conditions

It is established that when inference is performed without knowing the prior distribution hyper-parameters, it is possible to observe a Replica Symmetry Breaking (RSB) phase transition, which can manifest with a convergence failure of the replica symmetric cavity method algorithm. This is exactly what we see in Figure 3.18. For this plot, the star was used (*) to label the prior

hyper-parameters (with which the planted was generated), e.g. λ^* is the true infection probability, while the superscript ^{I}, which stands for *inference*, was used to denote the hyper-parameters used in the posterior distribution. For example, λ^I is the infection probability used by the algorithm in the inference process. For this plot $\gamma^* = \gamma^I$ (so the algorithm knows the exact value of patient zero probability) and only the infection probability is studied outside Bayes optimality. The free energy landscape is shown as a function of the inference hyper-parameter infection probability λ^I and the observations dilution *dil*. There exists a hyper-parameters zone in which the number of sweeps reaches the maximum allowed number (which was set to 100). In that zone, the estimators show an oscillating behavior. This suggests a breakdown of the algorithm validity, which may be caused by a RSB phase transition. Thus, when the prior hyper-parameters are not known, some difficulties arise in epidemic inference due to breakdown of convergence. A good strategy to avoid this is to infer the prior parameters, as shown in the next paragraph.

3.4.7 Inferring prior hyper-parameters

By approximately minimizing the Bethe free energy, the prior parameters are inferred: for the patient zero probability γ the Expectation Maximization (EM) method is used, Appendix C. For the infection probability λ, instead, a gradient descent (GD) on the free energy is numerically performed using auto derivation. Inference is studied in the same conditions of Figure 3.10. In Figure 3.19 the results computed by inferring the prior parameters are compared with their correspondent in the Bayes optimal case, i.e. the ones in Figure 3.10 (first row) and Figure 3.13(left). The prior parameters are learned by minimizing the Bethe free energy, which agrees almost perfectly with the optimal one. There is a strong agreement also for other estimators, as the MMSE, which are plotted at time $t = 4$. To actually see how well the prior hyper-parameters are inferred, we can see them plotted as functions of their planted correspondent quantities in Figure 3.20. It is again important to compare the results of prior parameters inference with the single instance results on finite graphs. Indeed, the inference results shown so far are for infinitely large graphs. The number of observations is therefore infinite too. It is then crucial to see whether for finite size graphs (and finite information) it is possible to achieve comparable results

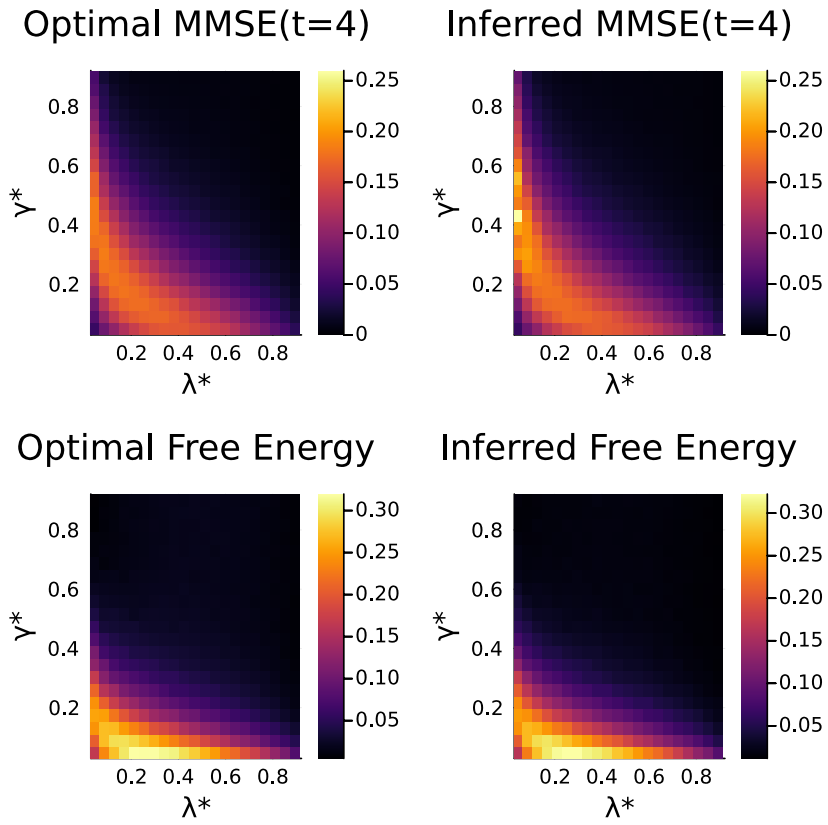

Fig. 3.19 A comparison between estimators (MMMSE and free energy) when prior's hyper-parameters are known (*first column*) VS when they are not known and learned (*second column*). The quantities are represented as functions of the planted parameters γ^* and λ^*. In the first row the MMSE at intermediate time ($t = 4$) is shown: on the left there is the optimal Bayes result, the same of Figure 3.10, while on the right there is the result obtained when λ^I and γ^I are learned. On the second row the same comparison (i.e. Bayes optimality on the left and hyper-parameters' learning on the right) is made for the Bethe free energy. In both cases (MMSE and free energy) the initial conditions for the hyper-parameters were set to $\lambda^I = 0.5$ and $\gamma^I = 0.5$. The results are for the Erdős–Rényi ensemble with average degree of 3. Observations are all collected at final time $T = 8$. Figure taken from [44].

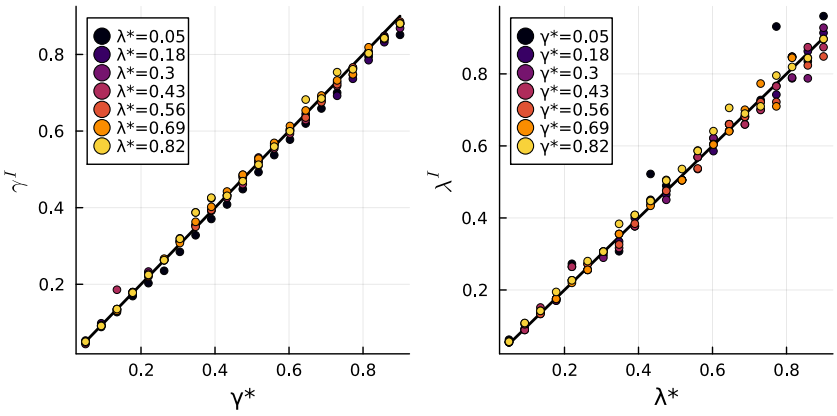

Fig. 3.20 The inferred prior parameters an function of their respective planted quantities. The plot is obtained at zero dilution (all individual observed) and for (uniformly) scattered observations in time. In the left panel, patient zero parameter γ^I is plotted as a function of γ^* for several values of λ^*. The right panel's lines are instead the values of the infection λ^I as function of λ^* for different values of γ^*. Figure taken from [44].

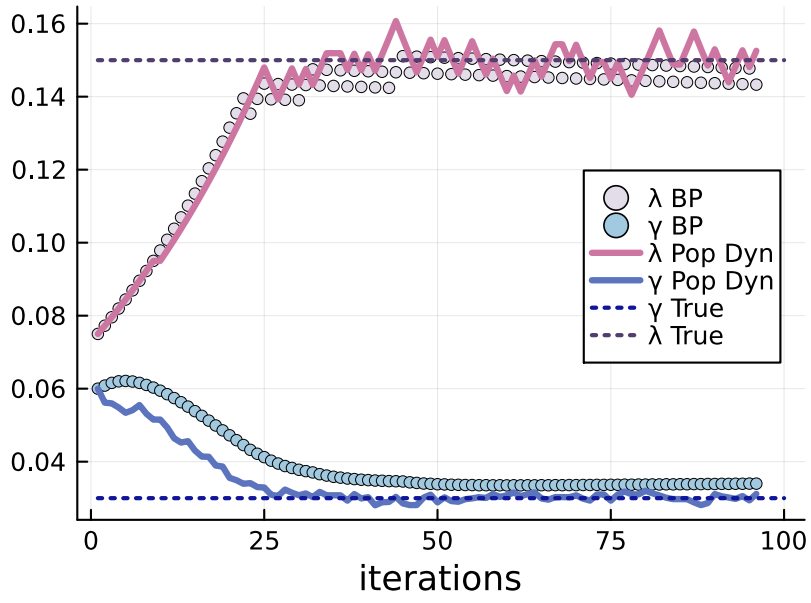

Fig. 3.21 Inference dynamics of ensemble code (cavity method implemented by means of population dynamics) compared with the single instance result, obtained running the Belief Propagation (BP) algorithm (Sib) on a contact network of $N = 10000$ nodes. The plot shows the gradient descent in free energy with respect to the two parameters γ^I and λ^I which respectively represent the patient zero and the infection probabilities. . The results are for Erdős–Rényi (ER) graphs with average degree 3. All the individuals are observed at final time. Figure taken from [44].

to the ensemble. In Figure 3.21 we see that this is the case. A comparison between cavity method and the single instance code is made by analyzing the gradient descent steps on the hyper-parameters. The plot shows that the values inferred by the single instance algorithm are very close to the true ones.

3.4.8 The role of symptoms in inference

In realistic contexts, observations are not collected uniformly random from the population. This is because infectious individuals might manifest some symptoms, which push them to test themselves. The probability of being observed is therefore typically higher for $\mathcal{I}$ than for $\mathcal{S}$. A first consequence is that the real fraction of infectious individuals is lower than the fraction of observed $\mathcal{I}$ individuals. This must be taken into account in inference. To do so, let us call p_+, the probability for an infected individual to be symptomatic. We assume that all infectious symptomatic individuals are tested. All the other individuals are instead tested at random with probability p_r. From this, the probability for an infectious individual to be tested $\mathcal{I}$ is:

$$P(\text{tested}, \text{positive}|\mathcal{I}) = (1 - \text{fr})(p_+ + p_r(1 - p_+))$$

and similarly:

$$P(\text{tested}, \text{negative}|\mathcal{I}) = \text{fr}(p_+ + p_r(1 - p_+))$$

For susceptible states $\mathcal{S}$:

$$P(\text{tested}, \text{positive}|\mathcal{S}) = p_r\text{fr}$$
$$P(\text{tested}, \text{negative}|\mathcal{S}) = p_r(1 - \text{fr})$$

The no-symptoms case is recovered for $p_+ = 0$. We want to see what happens if we neglect the existence of symptoms, i.e. if we infer setting $p_+^I = 0$, where p_+^I is the inference parameter used to account for symptoms. In Figure 3.22 (left panel), we see a substantial overestimation of the infection when the observation bias due to symptoms is neglected. On the right panel, we see that the AUC is systematically higher when the bias is included. We finally see that when the bias p_+ is inferred by minimizing the free energy (by means of a numerical

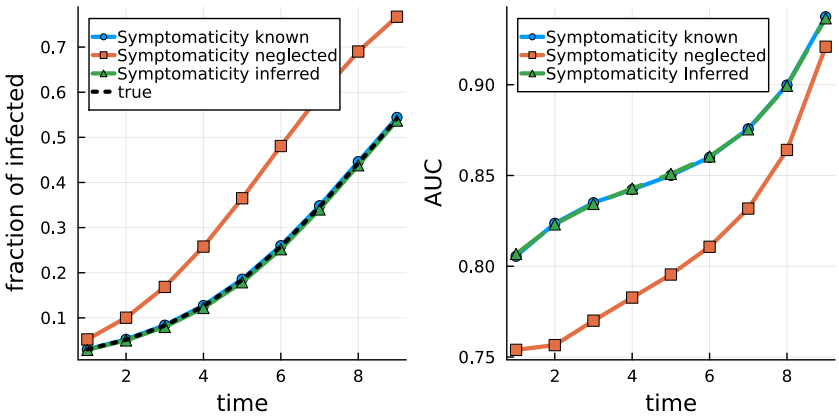

Fig. 3.22 Considering (and inferring) bias in observation due to symptoms allows to recover Nishimori conditions and improves inference performance. All the symptomatic individuals are assumed to be tested. The probability for an infectious individual of being symptomatic was set to $p_+ = 0.5$. Asymptomatic individuals can also be tested. For this plot, the probability for a general individual to be randomly selected for a test was set to $p_r = 0.04$. The *left* plot shows the estimated fraction of infected individuals over time. Considering the bias in the inference process allows to reconstruct this function. On the right plot there is the comparison of AUC when the bias is considered VS when it is neglected. Considering (or inferring) the bias systematically makes the AUC higher. The patient zero probability was set to $\gamma = 0.03$ and the infection probability to $\lambda = 0.25$. The observations are all performed at time $T = 8$. Figure taken from [44].

gradient descent), the results are very close to the optimal ones. This process allows to include the unknown bias without affecting performance.

3.4.9 Generalization to SIR and SEIR

The importance of Epidemble is that it allows to use replica symmetric cavity equations for a problem which shows correlated quenched disorder. The method is therefore presented for the simplest epidemic model (SI), in order to facilitate the explanation of the idea. We describe now, in the framework of the SIR model, a general strategy which can be used to generalize Epidemble to richer epidemic models (SIR and SEIR). After that, we introduce a parametrization that, for the SIR model, allows to maintain the same number of variables of the SI case in the Belief Propagation algorithm.

General strategy

The idea consists to simply increase the number of trajectory variables. While the SI model trajectory needs only one number per individual to be described, the infection time, the SIR trajectory needs two numbers: the infection and the recovery times. Then, we can rewrite for the SIR case equation (3.19). Defining $\underline{\tau} = \{\tau_i^{\mathcal{I}}, \tau_i^{\mathcal{R}}\}_{i=1,\dots,N}$ and $\mathcal{D} = \{x_i^0\}, \{s_{ij}, s_{ji}\}, \{r_i\}$:

$$P(\underline{\tau}|\mathcal{D}) = \prod_{i \in V} \psi^*(\tau_i^I, \tau_i^R, \underline{\tau}_{\partial i}^I, \underline{\tau}_{\partial i}^R, x_i^0, \{s_{ji}\}_{j \in \partial i}, r_i)$$

where the set of recovery delays $\{r_i\}_{i=1,\dots,N}$ represent the time interval in which the individual is in the $\mathcal{I}$ state. Each ψ^* in the SIR model has to take into account that an individual $j \in \partial i$ can only infect the individual i before recovering.

$$\psi^* = \mathbb{I}\left[\tau_i^{\mathcal{I}} = \delta_{x_i^0,\mathcal{S}} \min_{j \in \partial i} f(\tau_j^{\mathcal{I}} + s_{ji}, \tau_j^{\mathcal{R}})\right] \mathbb{I}\left[\tau_i^{\mathcal{R}} = \tau_i^{\mathcal{I}} + r_i\right]$$

where

$$f(\tau, \tau^{\mathcal{R}}) = \begin{cases} \tau & \text{if } \tau < \tau^{\mathcal{R}} \\ T & \text{if } \tau > \tau^{\mathcal{R}} \end{cases}$$

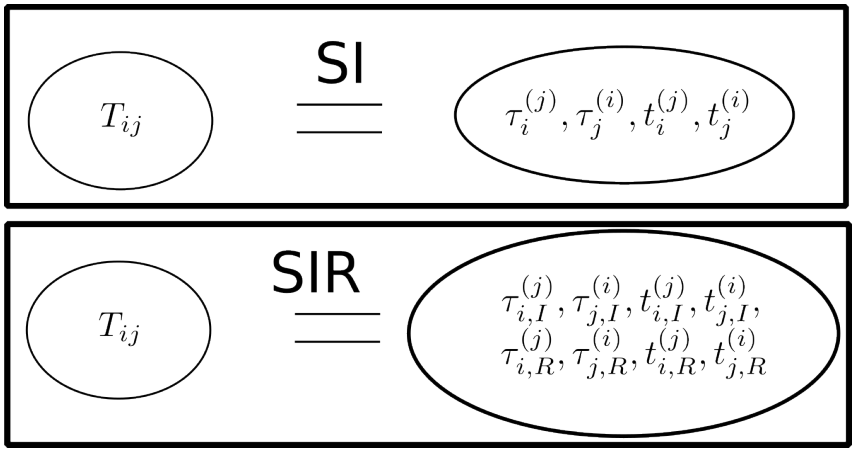

Fig. 3.23 The super-variable for the SIR model includes copies of the planted and inferred recovery time. Figure taken from [44].

The function f ensures that j can not infect i if j has already recovered. The planted trajectory at fixed quenched disorder is now described. The other terms of equation (3.26) are almost identical to the SI case and are discussed for SIR and SEIR in section 2.1.4. We have that equation (3.26), is formally unaltered:

$$P(\underline{\tau}, \mathcal{O}, \underline{t}|\mathcal{D}, \mathcal{D}_o) = \frac{1}{P(\mathcal{O})} P(\underline{\tau}|\mathcal{D}) P(\mathcal{O}|\underline{\tau}, \mathcal{D}_o) P(\mathcal{O}|\underline{t}) P(\underline{t})$$

The variables' and the disorder's domains are however enlarged. The factor graph associated to this equation, as in the SI case, contains loops. Therefore, it is necessary to introduce the usual copies of the infection and the recovery times, exactly as for the SI model. To implement the cavity method it is therefore necessary to define each super-variable T_{ij} by including the recovery (planted and inferred) times, as shown in Figure 3.23. The BP equations can now be run, with messages that however depend on more variables with respect to the SI case. This generalization can be made also for the SEIR model, simply introducing an the exposure time t_i^E (as discussed in section 2.1.4), an exposure delay e_i distributed according to $\zeta(e_i)$ for each $i = 1, \ldots, N$ and generalizing ψ^* by imposing that the infection can not happen during the exposure time interval.

An efficient parametrization for the SIR model

For the SIR model it is possible to obtain an optimized extension of the Epidemble method, which has messages of the same size w.r.t. the SI case. It is sufficient to switch from the infection times to transmission times Instead of using copies of the infection time $\{\tau_i^{\mathcal{I},(j)}\}_{i\in V}^{j\in\partial i}$, it is convenient to describe the epidemic trajectory with the infection transmission times $\{\tau_{ij}\}_{i=1,\dots,N}^{j\in\partial i}$. Each t_{ij} is the time at which i tries to infect its contact j. To actually know the infection time of j it is therefore necessary to take the minimum of the transmissions: \begin{equation}

$$\tau_j^I = \min_{i\in\partial j}\{\tau_{ij}\}. \tag{3.32}$$

The scheme consists therefore to work with $\{\tau_{ij}\}_{i=1,\dots,N}^{j\in\partial i}$, $\{\tau_i^{\mathcal{I}}\}_{i=1,\dots,N}$ and $\{\tau_i^{\mathcal{R}}\}_{i=1,\dots,N}$. After the new messages converge, the infection variables are found with equation 3.32. We now show that with this parametrization no copies of the recovery time have to be introduced, allowing to have a lighter representation for the planted SIR trajectory. We define the new super-variable as $\underline{\tau} = \{\tau_{ij}\}_{i=1,\dots,N}^{j\in\partial i}, \{\tau_i^I\}_{i=1,\dots,N}, \{\tau_i^R\}_{i=1,\dots,N}$ and we have:

$$P(\underline{\tau}|\{x_i^0\},\{s_{ij},s_{ji}\},\{r_i\}) = \prod_{(i,j)} \psi^*(\tau_{ij},\tau_i^{\mathcal{I}},\tau_i^{\mathcal{R}},\{t_{ki}\}_{k\in\partial i},\underline{\tau}_{\partial i}^{\mathcal{R}},x_i^0,s_{ij},r_i)$$

and each factor is:

$$\psi^* = \mathbb{I}\left[\tau_{ij} = f\left(\delta_{x_i^0,S}\min_{k\in\partial i\setminus j}\{\tau_{ki}\}+s_{ij},\tau_i^{\mathcal{R}}\right)\right]\mathbb{I}\left[\tau_i^{\mathcal{R}}=\tau_i^{\mathcal{I}}+r_i\right]\mathbb{I}\left[\tau_i^{\mathcal{I}}=\min_{k\in\partial i}\{\tau_{ki}\}\right]. \tag{3.33}$$

The factor graph equation associated to this construction is therefore:

$$P(\underline{t},\underline{\tau}|\mathcal{D},\mathcal{D}_o) = \prod_i \xi(t_i^{\mathcal{I}},t_i^{\mathcal{R}},\tau_i^{\mathcal{I}},\tau_i^{\mathcal{R}};\{\varepsilon_o,t_o\}_{o_i=i})\times$$

$$\times\prod_{j\in\partial i}\left(\psi^*(\tau_{ij},\tau_i^{\mathcal{I}},\tau_i^{\mathcal{R}},\{\tau_{ki}\}_{k\in\partial i\setminus j},\underline{\tau}_{\partial i}^{\mathcal{R}},x_i^0,s_{ij},r_i)\psi(t_{ij},t_i^{\mathcal{I}},t_i^{\mathcal{R}},\{t_{ki}\}_{k\in\partial i\setminus j},\underline{t}_{\partial i}^{\mathcal{R}})\right)$$

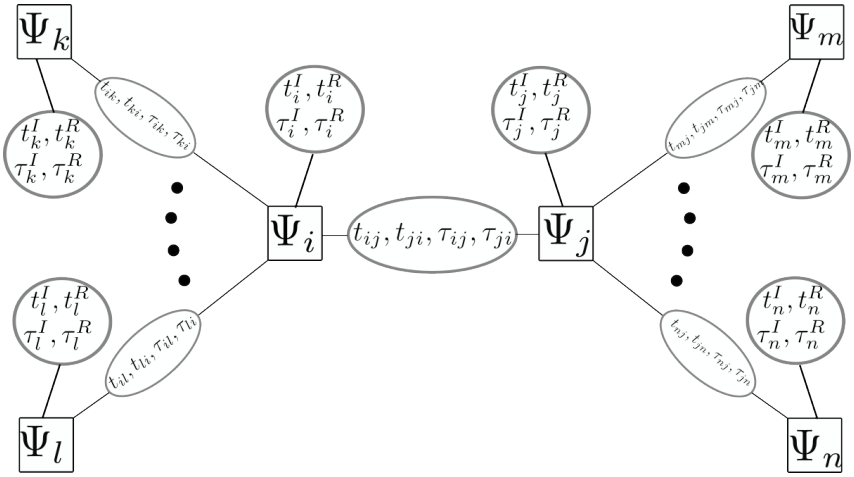

Fig. 3.24 The factor graph for the SIR model using the transmission delay representation. Figure taken from [44].

Where the factors ψ can be obtained by tracing the ψ^* with respect to the disorder:

$$\psi(t_{ij}, t_i^{\mathcal{I}}, t_i^{\mathcal{R}}, \{t_{ki}\}_{k\in\partial i\setminus j}, \underline{t}_{\partial i}^{\mathcal{R}}) = \sum_{x_i^0, s_{ij}, r_i} \omega(s_{ij})(\gamma\delta_{x_i^0,\mathcal{I}} + (1-\gamma)\delta_{x_i^0,\mathcal{S}})\mu(r_i)\times$$
$$\times\, \psi^*(t_{ij}, t_i^{\mathcal{I}}, t_i^{\mathcal{R}}, \{t_{ki}\}_{k\in\partial i\setminus j}, \underline{t}_{\partial i}^{\mathcal{R}}, x_i^0, s_{ij}, r_i),$$

where $\omega(s)$ and $\mu(r)$ are the two (geometric) distributions of the infection delay and recovery delay. The factor graph associated to this equation, differently from equation (3.27), does not contain any loop (see Figure (3.24)) when the underlying contact graph is acyclic, so it is straightforward to implement BP equations on it. Moreover, since the recovery times and the infection times are on leaves (nodes attached only to one factor), they can be traced out, so that the BP equations only involve a number of variables which is the same of the SI case. Note that still the BP equations will be a bit slower than the SI case due to the different nature of the factors (equation (3.33)), which are slower to compute.

Chapter 4

Conclusions and Future perspectives

This work provides an introduction to algorithms and thresholds in epidemic inference on networks. The aim of such studies is to clarify whether and how the available information can be used to reconstruct (and hopefully contain) the epidemic outbreak. Probably, the most immediate message that can be learned from this thesis is that such an aim is far from being reached: even when the contact network is fully known there are huge difficulties in reconstructing the epidemic history. The posterior is hard to compute and can even show signals of replica symmetry breaking out of Bayes optimality. We have conducted an analysis of single-instance and ensemble algorithms. Both research directions present numerous unsolved questions.

4.1 Single Instance Algorithms

Regarding single-instance algorithms, a significant challenge lies in finding methods that are fast, robust, privacy-preserving, parallel, distributed and high-performing. The methods discussed in Chapter 2 all exhibit some limitations in these aspects. For instance, the Causal Variational Approach fails to maintain privacy. Specifically, to execute CVA, the complete observation set must be supplied. An ideal privacy-preserving algorithm, on the other hand, should reconstruct the infection state of an individual solely based on information

about the individual and their direct contacts. Mean field methods, on this hand, are promising (see the MF heuristic at page 80). A possible approach to find new algorithms is to apply techniques from inference, as Expectation Propagation, which could be used to develop new mean field-like methods. This is more quantitatively explained in the next paragraph.

4.1.1 A possible remedy for the infinities in KL divergence

One issue when dealing with the Mean Field approach for epidemics (see section 2.2.1) is that the KL divergence between the mean field approximation and the posterior can be infinite. The KL between the approximation $\mathcal{Q}^{\mathrm{MF}}$ and the posterior $\mathcal{P}$ is indeed:

$$D_{\mathrm{KL}}(\mathcal{Q}^{\mathrm{MF}}||\mathcal{P}) = \sum_{\underline{t}} \mathcal{Q}^{\mathrm{MF}}(\underline{t}) \log \frac{\mathcal{Q}^{\mathrm{MF}}(\underline{t})}{\mathcal{P}(\underline{t})}$$

And this quantity is infinite if it exists a $\underline{t}$ such that $\mathcal{P}(\underline{t}) = 0$ and $\mathcal{Q}^{\mathrm{MF}}(\underline{t}) > 0$. Being $\mathcal{Q}^{\mathrm{MF}}$ factorized over individuals, the KL is finite only if $\mathcal{Q}^{\mathrm{MF}}$ excludes every infection phenomenon, namely each individual is either a patient zero or is never infected. We solved this issue by changing the family of approximating functions to $\mathcal{Q}^{\mathrm{CVA}}$, introducing the Causal Variational Approach (section 2.3). However, there is another way to get around the infinities of the MF approximation. If we change the optimization function to:

$$D_{\mathrm{KL}}(\mathcal{P}||\mathcal{Q}^{\mathrm{MF}}) = \sum_{\underline{t}} \mathcal{P}(\underline{t}) \log \frac{\mathcal{P}(\underline{t})}{\mathcal{Q}^{\mathrm{MF}}(\underline{t})}$$

we can see that the MF approximation exactly reproduces the marginals. If we indeed write $\mathcal{Q}^{\mathrm{MF}}(\underline{t}) = \prod_i q_i(\underline{t}_i)$, then:

$$D_{\mathrm{KL}}(\mathcal{P}||\mathcal{Q}^{\mathrm{MF}}) = \sum_{\underline{t}} \sum_{i} \mathcal{P}(\underline{t}) \log \frac{\mathcal{P}(\underline{t})}{q_i(\underline{t}_i)}.$$

Optimizing the KL gives:

$$\alpha_j = \frac{\delta D_{\mathrm{KL}}(\mathcal{P}||\mathcal{Q}^{\mathrm{MF}})}{\delta q_j(\hat{\underline{t}}_j)} = \sum_{\underline{t} \setminus \underline{t}_j} \frac{\mathcal{P}(\underline{t})}{q_j(\hat{\underline{t}}_j)}$$

where α_j is the Lagrange multiplier due to the normalization of the probability q_j. We therefore have:

$$\alpha_j q_j(\hat{\underline{t}}_j) = \sum_{\underline{t}\setminus\underline{t}_j} \mathcal{P}(\underline{t})$$

and the Lagrange multiplier is easily fixed by imposing normalization:

$$\alpha_j \sum_{\hat{\underline{t}}_j} q_j(\hat{\underline{t}}_j) = \sum_{\underline{t}} \mathcal{P}(\underline{t}) = 1$$

which implies:

$$\alpha_j = 1.$$

We end up having:

$$q_j(\underline{t}_j) = \sum_{\underline{t}\setminus\underline{t}_j} \mathcal{P}(\underline{t})$$

which means that each marginal is the exact one! Optimizing the $D_{\mathrm{KL}}(\mathcal{P}||\mathcal{Q}^{\mathrm{MF}})$ is, however, unfeasible because we should be able to sample from $\mathcal{P}$, which is exactly the starting point. However, methods like Expectation Propagation allow to studying some approximate optimization functions of $D_{\mathrm{KL}}(\mathcal{P}||\mathcal{Q}^{\mathrm{MF}})$ and might turn out to be useful for developing new Mean Field schemes.

4.2 Ensemble study

The exploration of the ensemble problem has only just begun with Epidemble, offering ample opportunities for advancements. For instance, one key aspect is addressing graph ignorance, which might be studied by separating the planted graph, used to extract the planted configuration, from a graph used to make inference. The graphs would coincide under Bayes optimality. To model graph ignorance, we might remove random links from the original true contact network and introduce a self-infection parameter (see section 2.1.4). Additionally, studying in detail the convergence breakdown of the BP equations under Bayes optimality can be an interesting path which might even lead to a better understanding of the Nishimori conditions.

References

[1] Tom Britton and Etienne Pardoux, editors. *Stochastic Epidemic Models with Inference*, volume 2255 of *Lecture Notes in Mathematics*. Springer International Publishing, Cham, 2019.

[2] Shweta Bansal, Bryan T Grenfell, and Lauren Ancel Meyers. When individual behaviour matters: homogeneous and network models in epidemiology. *Journal of The Royal Society Interface*, 4(16):879–891, July 2007. Publisher: Royal Society.

[3] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic Spreading in Scale-Free Networks. *Physical Review Letters*, 86(14):3200–3203, April 2001. Publisher: American Physical Society.

[4] Håkan Andersson and Tom Britton. *Stochastic Epidemic Models and Their Statistical Analysis*, volume 151 of *Lecture Notes in Statistics*. Springer, New York, NY, 2000.

[5] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[6] Antoine Baker, Indaco Biazzo, Alfredo Braunstein, Giovanni Catania, Luca Dall'Asta, Alessandro Ingrosso, Florent Krzakala, Fabio Mazza, Marc Mézard, and Anna Paola Muntoni. Epidemic mitigation by statistical inference from contact tracing data. *Proceedings of the National Academy of Sciences*, 118(32):e2106548118, 2021. Publisher: National Acad Sciences.

[7] Ralf Herbrich, Rajeev Rastogi, and Roland Vollgraf. CRISP: A Probabilistic Model for Individual-Level COVID-19 Infection Risk Estimation Based on Contact Data, June 2022. arXiv:2006.04942 [cs, stat].

[8] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alejandro Lage-Castellanos, and Riccardo Zecchina. Bayesian Inference of Epidemics on Networks via Belief Propagation. *Physical Review Letters*, 112(11):118701, March 2014. Publisher: American Physical Society.

[9] Alfredo Braunstein, Giovanni Catania, Luca Dall'Asta, Matteo Mariani, and Anna Paola Muntoni. Inference in conditioned dynamics through

causality restoration. *Scientific Reports*, 13(1):7350, May 2023. Number: 1 Publisher: Nature Publishing Group.

[10] Indaco Biazzo, Alfredo Braunstein, Luca Dall'Asta, and Fabio Mazza. A Bayesian generative neural network framework for epidemic inference problems. *Scientific Reports*, 12(1):19673, November 2022. Number: 1 Publisher: Nature Publishing Group.

[11] David J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, May 1992.

[12] David J. C. MacKay. Bayesian Methods for Backpropagation Networks. In Eytan Domany, J. Leo van Hemmen, and Klaus Schulten, editors, *Models of Neural Networks III: Association, Generalization, and Representation*, Physics of Neural Networks, pages 211–254. Springer, New York, NY, 1996.

[13] I. J. Good. *The estimation of probabilities: an essay on modern Bayesian methods*. MIT Press, Cambridge, Mass, 1965. Open Library ID: OL18974443M.

[14] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York, New York, NY, 1985.

[15] Genshiro Kitagawa and Will Gersch. *Smoothness Priors Analysis of Time Series*, volume 116 of *Lecture Notes in Statistics*. Springer, New York, NY, 1996.

[16] Gerhard Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer, Berlin, Heidelberg, 2003.

[17] J M Pryce and A D Bruce. Statistical mechanics of image restoration. *Journal of Physics A: Mathematical and General*, 28(3):511–532, February 1995.

[18] Stuart Geman and Donald Geman. Geman, D.: Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. PAMI-6(6), 721-741. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6:721–741, November 1984.

[19] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, March 1991.

[20] Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, New York, NY, 1996.

[21] Nicolas Sourlas. Spin-glass models as error-correcting codes. *Nature*, 339(6227):693–695, June 1989. Number: 6227 Publisher: Nature Publishing Group.

[22] Y. Iba. Bayesian Statistics and Statistical Mechanics. In Hajime Takayama, editor, *Cooperative Dynamics in Complex Physical Systems*, Springer Series in Synergetics, pages 235–236, Berlin, Heidelberg, 1989. Springer.

[23] Manfred Opper and Ole Winther. Mean Field Approach to Bayes Learning in Feed-Forward Neural Networks. *Physical Review Letters*, 76(11):1964–1967, March 1996. Publisher: American Physical Society.

[24] Y. Kabashima and D. Saad. Statistical mechanics of error-correcting codes. *Europhysics Letters*, 45(1):97–103, January 1999. Number: 1 Publisher: EDP Sciences.

[25] Anna Frishman and Pierre Ronceray. Learning Force Fields from Stochastic Trajectories. *Physical Review X*, 10(2):021009, April 2020. Publisher: American Physical Society.

[26] Raphael Sarfati, Jerzy Bławzdziewicz, and Eric R. Dufresne. Maximum likelihood estimations of force and mobility from single short Brownian trajectories. *Soft Matter*, 13(11):2174–2180, March 2017. Publisher: The Royal Society of Chemistry.

[27] Silvan Türkcan, Antigoni Alexandrou, and Jean-Baptiste Masson. A Bayesian Inference Scheme to Extract Diffusivity and Potential Fields from Confined Single-Molecule Trajectories. *Biophysical Journal*, 102(10):2288–2298, May 2012.

[28] Mohamed El Beheiry, Maxime Dahan, and Jean-Baptiste Masson. InferenceMAP: mapping of single-molecule dynamics with Bayesian inference. *Nature Methods*, 12(7):594–595, July 2015. Number: 7 Publisher: Nature Publishing Group.

[29] Diego Alberici, Francesco Camilli, Pierluigi Contucci, and Emanuele Mingione. The Solution of the Deep Boltzmann Machine on the Nishimori Line. *Communications in Mathematical Physics*, 387(2):1191–1214, October 2021.

[30] Pierluigi Contucci, Satoshi Morita, and Hidetoshi Nishimori. Surface Terms on the Nishimori Line of the Gaussian Edwards-Anderson Model. *Journal of Statistical Physics*, 122(2):303–312, January 2006.

[31] Christophe Garban and Thomas Spencer. Continuous symmetry breaking along the Nishimori line. *Journal of Mathematical Physics*, 63(9):093302, September 2022.

[32] Ilya A. Gruzberg, N. Read, and Andreas W. W. Ludwig. Random-bond Ising model in two dimensions: The Nishimori line and supersymmetry. *Physical Review B*, 63(10):104422, February 2001. Publisher: American Physical Society.

[33] Yukito Iba. The Nishimori line and Bayesian statistics. *Journal of Physics A: Mathematical and General*, 32(21):3875, May 1999. Publisher: IOP Publishing.

[34] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, September 2016. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00018732.2016.1211393.

[35] H. Nishimori. Exact results and critical properties of the Ising model with competing interactions. *Journal of Physics C Solid State Physics*, 13:4071–4076, July 1980. ADS Bibcode: 1980JPhC...13.4071N.

[36] Hidetoshi Nishimori. Internal Energy, Specific Heat and Correlation Function of the Bond-Random Ising Model. *Progress of Theoretical Physics*, 66(4):1169–1181, October 1981.

[37] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[38] Giorgio Parisi. *Statistical Field Theory*. Basic Books, January 1988. Google-Books-ID: OF8sAAAAYAAJ.

[39] Francesco Zamponi. Mean field theory of spin glasses, September 2014. arXiv:1008.4844 [cond-mat].

[40] Tommaso Castellani and Andrea Cavagna. Spin-glass theory for pedestrians. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05012, May 2005.

[41] Jean Barbier and Dmitry Panchenko. Strong replica symmetry in high-dimensional optimal Bayesian inference, February 2022. arXiv:2005.03115 [cond-mat, physics:math-ph].

[42] Montanari Andrea. Estimating random variables from random sparse observations. *European Transactions on Telecommunications*, 19(4):385–403, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ett.1289.

[43] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Clarendon Press, February 1999. Google-Books-ID: J5aLdDN4uFwC.

[44] Alfredo Braunstein, Louise Budzynski, and Matteo Mariani. Statistical mechanics of inference in epidemic spreading. *Phys. Rev. E*, 108:064302, Dec 2023.

[45] Fred Brauer. Compartmental Models in Epidemiology. In Fred Brauer, Pauline van den Driessche, and Jianhong Wu, editors, *Mathematical Epidemiology*, Lecture Notes in Mathematics, pages 19–79. Springer, Berlin, Heidelberg, 2008.

[46] Juliana Tolles and ThaiBinh Luong. Modeling Epidemics With Compartmental Models. *JAMA*, 323(24):2515–2516, June 2020.

[47] Linda J. S. Allen. Some discrete-time SI, SIR, and SIS epidemic models. *Mathematical Biosciences*, 124(1):83–105, November 1994.

[48] Stefano Crotti and Alfredo Braunstein. Matrix Product Belief Propagation for reweighted stochastic dynamics over graphs. *Proceedings of the National Academy of Sciences*, 120(47):e2307935120, November 2023. Publisher: Proceedings of the National Academy of Sciences.

[49] Robert Hinch, William J. M. Probert, Anel Nurtay, Michelle Kendall, Chris Wymant, Matthew Hall, Katrina Lythgoe, Ana Bulas Cruz, Lele Zhao, Andrea Stewart, Luca Ferretti, Daniel Montero, James Warren, Nicole Mather, Matthew Abueg, Neo Wu, Olivier Legat, Katie Bentley, Thomas Mead, Kelvin Van-Vuuren, Dylan Feldner-Busztin, Tommaso Ristori, Anthony Finkelstein, David G. Bonsall, Lucie Abeler-Dörner, and Christophe Fraser. OpenABM-Covid19. An agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. *PLOS Computational Biology*, 17(7):e1009146, July 2021. Publisher: Public Library of Science.

[50] Lars Lorch, Heiner Kremer, William Trouleau, Stratis Tsirtsis, Aron Szanto, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. Quantifying the Effects of Contact Tracing, Testing, and Containment Measures in the Presence of Infection Hotspots, November 2022. arXiv:2004.07641 [physics, q-bio, stat].

[51] Florent Krzakala, Andre Manoel, Eric W. Tramel, and Lenka Zdeborová. Variational free energies for compressed sensing. In *2014 IEEE International Symposium on Information Theory*, pages 1499–1503, June 2014. ISSN: 2157-8117.

[52] John R. Hershey, Peder A. Olsen, and Steven J. Rennie. Variational Kullback-Leibler divergence for Hidden Markov models. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 323–328, December 2007.

[53] Ghassen Jerfel, Serena Wang, Clara Wong-Fannjiang, Katherine A. Heller, Yian Ma, and Michael I. Jordan. Variational refinement for importance sampling using the forward Kullback-Leibler divergence. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 1819–1829. PMLR, December 2021. ISSN: 2640-3498.

[54] Friston Karl. A Free Energy Principle for Biological Systems. *Entropy*, 14(11):2100–2121, November 2012. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

[55] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. Publisher: Institute of Mathematical Statistics.

[56] Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated Gradient Descent Escapes Saddle Points Faster than Gradient Descent. In *Proceedings of the 31st Conference On Learning Theory*, pages 1042–1085. PMLR, July 2018. ISSN: 2640-3498.

[57] Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Aarti Singh, and Barnabas Poczos. Gradient Descent Can Take Exponential Time to Escape Saddle Points. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[58] Emmanuel Moulay, Vincent Léchappé, and Franck Plestan. Properties of the sign gradient descent algorithms. *Information Sciences*, 492:29–39, August 2019.

[59] Xiuxian Li, Kuo-Yi Lin, Li Li, Yiguang Hong, and Jie Chen. On Faster Convergence of Scaled Sign Gradient Descent. *IEEE Transactions on Industrial Informatics*, pages 1–10, 2023. Conference Name: IEEE Transactions on Industrial Informatics.

[60] Dian Wu, Lei Wang, and Pan Zhang. Solving Statistical Mechanics Using Variational Autoregressive Networks. *Physical Review Letters*, 122(8):080602, February 2019. Publisher: American Physical Society.

[61] Zdravko Botev and Ad Ridder. Variance Reduction. In *Wiley StatsRef: Statistics Reference Online*, pages 1–6. John Wiley & Sons, Ltd, 2017. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat07975.

[62] Robert M. Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-Reduced Methods for Machine Learning. *Proceedings of the IEEE*, 108(11):1968–1983, November 2020. Conference Name: Proceedings of the IEEE.

[63] Gregory F. Lawler and Vlada Limic. *Random Walk: A Modern Introduction*. Cambridge University Press, June 2010. Google-Books-ID: UBQdwAZDeOEC.

[64] Nino Antulov-Fantulin, Alen Lančič, Tomislav Šmuc, Hrvoje Štefančič, and Mile Šikiš. Identification of Patient Zero in Static and Temporal Networks: Robustness and Limitations. *Physical Review Letters*, 114(24):248701, June 2015. Publisher: American Physical Society.

[65] M. Mézard, G. Parisi, and M. A. Virasoro. SK Model: The Replica Solution without Replicas. *Europhysics Letters (EPL)*, 1(2):77–82, January 1986. Publisher: IOP Publishing.

[66] Marc Mézard and Giorgio Parisi. The Cavity Method at Zero Temperature. *Journal of Statistical Physics*, 111(1):1–34, April 2003.

[67] Alfredo Braunstein. *Algorithms for Optimization, Inference & Learning*. Lecture Notes, 2021.

[68] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.

[69] D. Ghio, A. L. M. Aragon, I. Biazzo, and L. Zdeborova. Bayes-optimal inference for spreading processes on random networks, March 2023. arXiv:2303.17704 [cond-mat].

Appendix A

Extracting a random variable from a continuous distribution

When dealing with sampling, it is almost always necessary to sample a variable from a 1D probability distribution $f : \mathbb{R} \to [0, 1]$. The inverse cumulative method consists in computing the cumulative of f:

$$\mathbb{P}(t < \tau) = F(\tau) = \int_{-\infty}^{\tau} f(s)\mathrm{d}s$$

and to sample a number r uniformly from 0 to 1. This number r can be thought as a random value of the cumulative function. We then look what is the value τ^* for which the cumulative assumes the value of r and that is a sample from f.

$$\begin{aligned} r &\sim \mathrm{Unif}(0, 1) \\ r &= F(\tau^*) \\ \tau^* &= F^{-1}(r). \end{aligned}$$

Now we can apply this to extract e.g. the tentative time at which an individual tries to infect another. If the infection rate from i to j at time t is $\lambda_{ij}(t)$, then the density probability for i to infect j is:

$$\rho(t|t_i) = \lambda_{ij}(t) e^{-\int_{t_i}^{t} \lambda_{ij}(s)\mathrm{d}s}$$

We want to sample from ρ, so we compute the cumulative:

$$\begin{aligned}
R(t|t_i) &= \int_{t_i}^{t} \rho(u)\mathrm{d}u = \\
&= \int_{t_i}^{t} \lambda_{ij}(u) e^{-\int_{t_i}^{u} \lambda_{ij}(s)\mathrm{d}s} \mathrm{d}u = \\
&= \int_{t_i}^{t} -\frac{\partial e^{-\int_{t_i}^{u} \lambda_{ij}(s)\mathrm{d}s}}{\partial u} \mathrm{d}u = \\
&= e^{-\int_{t_i}^{u} \lambda_{ij}(s)\mathrm{d}s}\Big|_{t}^{t_i} = 1 - e^{-\int_{t_i}^{t} \lambda_{ij}(s)\mathrm{d}s}
\end{aligned}$$

and we set it to a random uniform number from 0 to 1:

$$\begin{aligned}
r &= 1 - e^{-\int_{t_i}^{t} \lambda_{ij}(s)\mathrm{d}s} \\
\log r &= -\int_{t_i}^{t} \lambda_{ij}(s)\mathrm{d}s = -\Lambda_{ij}(t) + \Lambda_{ij}(t_i) \\
\Lambda_{ij}(t) &= \Lambda_{ij}(t_i) - \log r \\
t &= \Lambda_{ij}^{-1}(\Lambda_{ij}(t_i) - \log r)
\end{aligned}$$

and t is now a fair sample of the infection distribution ρ.

Appendix B

Sampling from the residual distribution

If $p(d)$ is the degree distribution of the graph, we want to compute the residual degree distribution p_{res}. The residual degree of an edge $(i, j) \in \mathcal{E}$, where $i, j \in V$, is the number of edges of i except for (i, j):

$$d_{res}(i, j) = |\partial i \setminus j|$$

The residual degree is distributed differently to the degree. In fact, while the latter is the probability of picking a node with d edges attached to it, the former is the probability to randomly pick an edge attached to a node with degree $d_{res} + 1$. A high-degree node is more likely to be picked if we uniformly pick an edge. In particular, the probability to pick a node with degree d attached to a randomly picked edge is proportional to $dp(d)$. The residual degree distribution is therefore:

$$p_{res}(d_{res}) = \frac{(d_{res} + 1)p(d_{res} + 1)}{Z}$$

Where $Z = \sum_{d>1} dp(d)$ is the normalization. Thus, for random regular graph with degree $\overline{d}$:

$$\begin{aligned} p_{res}(d_{res}) &= \frac{(d_{res}+1)\delta_{d_{res}+1,\overline{d}}}{Z} = \\ &= \frac{\overline{d}}{Z}\delta_{d_{res},\overline{d}-1} = \\ &= \frac{1}{Z'}\delta_{d_{res},\overline{d}-1} = \\ &= p(\overline{d}-1) \end{aligned}$$

where we canceled the constants with the normalization, obtaining Z'. For the RR graph we actually have that the residual degree distribution coincides with the degree (minus one) distribution. This is true because randomly picking an edge coincides with randomly picking a node. This is not true in general. For the Erdős–Rényi graph, with average degree $\hat{d}$:

$$\begin{aligned} p_{res}(d_{res}) &= \frac{(d_{res}+1)\hat{d}^{d_{res}+1}e^{-\hat{d}}}{Z(d_{res}+1)!} = \\ &= \frac{\hat{d}^{d_{res}}e^{-\hat{d}}}{(d_{res})!}\frac{\hat{d}}{Z} = \\ &= \frac{\hat{d}^{d_{res}}e^{-\hat{d}}}{(d_{res})!}\frac{1}{Z'} = \\ &= p(d_{res}). \end{aligned}$$

It is remarkably true that for the ER graph the residual degree distribution and the degree distribution coincide.

Appendix C

Expectation Maximization

Expectation Maximization (EM) method is an iterative scheme which allows to approximately descend the free energy. Each iteration is separated in two steps:

1. At fixed BP messages, the update for γ at k_{th} iteration is:

$$\gamma_k = \arg\max_{\gamma} \langle \log P\left(\underline{t}, O|\gamma\right)\rangle_{\{\nu\}_k}, \tag{C.1}$$

   where $\{\nu\}_k$ is a shorthand notation for the set of all BP messages at k_{th}iteration.

2. At fixed γ_k, the messages are updated with BP equations.

Equation (C.1) can be explained from the definition of the variational free energy:

$$F[Q](\gamma) := -\langle \log P(\underline{t}, \mathcal{O}|\gamma)\rangle_Q + \langle \log Q(\underline{t})\rangle_Q$$

The posterior distribution $\mathcal{P}(\underline{t}|\mathcal{O};\gamma)$ is the distribution Q which minimizes F. If we evaluate averages with fixed BP messages, then the dependency of F on γ is only on the first addend (the log-likelihood) of the right hand side. Then the optimization on γ reduces to equation (C.1). To actually optimize the log-likelihood term we set to zero the first derivative of equation (C.1) w.r.t. γ. We have, at the k_{th} iteration: \begin{equation}

$$\gamma_k = \frac{1}{N}\sum_{i\in V} p_i^{\mathcal{I},k}(t_i = 0|\mathcal{O}) \tag{C.2}$$

Where $p_i^{I,k}(t_i = 0|\mathcal{O})$ is the posterior probability at k_{th} iteration of individual i to be the patient zero. Expectation Maximization for γ therefore reduces to simply update γ^I with equation (C.2) at every sweep of BP update on the population.

Appendix D

Optimization of Epidemble message passing scheme

Let us now work on the optimization of the BP equations for Epidemble. The BP equations are:

$$\hat{\nu}_{\Psi_i \to (i,j)}(T_{ij}) = \frac{1}{z_{i\to j}} \sum_{\{T_{ki}\}_{k\in\partial i\setminus j}} \Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i) \prod_{k\in\partial i\setminus j} \nu_{(k,i)\to\Psi_i}(T_{ki})$$

$$\nu_{(i,j)\to\Psi_j}(T_{ij}) = \hat{\nu}_{\Psi_i\to(i,j)}(T_{ij})$$

The second equation is due to the fact that node (i, j) only has two functions attached, Ψ_i and Ψ_j. Now we notice that the messages are functions defined over a domain of $O\left(T^4\right)$ variables. In fact recall that $T_{ij} = (\tau_i, \tau_j, t_i, t_j)$. So for the population dynamics scheme we need to store $O(nT^4)$ numbers to implement the scheme and and perform the same number of computations to actually do one sweep. We can optimize this. It is sufficient to express the factor Ψ_i explicitly in the message. The factor Ψ_i is a product of a contribution coming from the planted trajectory distribution, i.e. $P(\underline{\tau}|\mathcal{D})$ a contribution coming from the inferred trajectory posterior distribution, i.e. $P(\underline{t})P(\mathcal{O}|\underline{t})$ and a contribution which couples planted and inferred coming from the observations, $P(\mathcal{O}|\underline{\tau}, \mathcal{D}_o)$. In particular, the exact definition of Ψ_i is given in equation (3.29),

that we write here:

$$\Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i) = \xi(\tau_i^{(j)}, t_i^{(j)}|\{\varepsilon_o, t_o\}_{o_i=i})\psi^*(\tau_i^{(j)}, \tau_{\partial i}^{(i)}|\{s_{li}\}_{l\in\partial i}, x_i^0)\times \\ \times\, \psi(t_i^{(j)}, t_{\partial i}^{(i)}) \prod_{l\in\partial i} \delta_{t_i^{(j)},t_i^{(l)}}\delta_{\tau_i^{(j)},\tau_i^{(l)}}$$

Let us start with the constraint that all the copies of the times must be equal to each others, $\prod_{l\in\partial i}\delta_{t_i^{(j)},t_i^{(l)}}\delta_{\tau_i^{(j)},\tau_i^{(l)}}$. We can simply eliminate the superscript and remove the constraint. We then recall $\xi(\tau_i^{(j)}, t_i^{(j)}|\{\varepsilon_o, t_o\}_{o_i=i}) =: \xi_{t_i}^{\tau_i}$ to shorten notation. Now we have to work on the product $\psi^*\psi$. This can be rewritten as a sum of six simpler terms. The reason is that ψ^* is the sum of three terms and ψ is the sum of two terms. Starting from ψ^*, namely equation (3.20):

$$\psi^*(\tau_i, \tau_{\partial i}|x_i^0, \{s_{ji}\}_{j\in\partial i}) = \mathbb{I}\left[\tau_i = x_i^0 \min_{j\in\partial i}\{\tau_j + s_{ji}\}\right] = \\ = \delta_{x_i^0,\mathcal{I}}\delta_{\tau_i,0} + \delta_{x_i^0,\mathcal{S}}\prod_{j\in\partial i}\mathbb{I}[\tau_i \le \tau_j + s_{ji}] - \delta_{x_i^0,\mathcal{S}}\prod_{j\in\partial i}\mathbb{I}[\tau_i < \tau_j + s_{ji}],$$

where we simply rewrote the minimum (to be the minimum it means that at least there is a j such that $\tau_i = \tau_j + s_{ji}$). Now we write the formula for ψ, see equation (2.6):

$$\psi(t_i, t_{\partial i}) = \prod_{j\in\partial i}(1-\lambda)^{(t_j - t_i - 1)_+} - \prod_{j\in\partial i}(1-\lambda)^{(t_j - t_i)_+}$$

where $(a)_+ = a\theta(a)$ and θ is the Heaviside theta. We can see therefore that the product $\psi^*\xi_{t_i}^{\tau_i}\psi$ is a sum of 6 terms, quite similar among each others:

$$\Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i) = \left(\delta_{x_i^0,\mathcal{I}}\delta_{\tau_i,0} + \delta_{x_i^0,\mathcal{S}}\prod_{j\in\partial i}\mathbb{I}[\tau_i \le \tau_j + s_{ji}] - \delta_{x_i^0,\mathcal{S}}\prod_{j\in\partial i}\mathbb{I}[\tau_i < \tau_j + s_{ji}]\right) \\ \times\, \xi_{t_i}^{\tau_i}\left(\prod_{j\in\partial i}(1-\lambda)^{(t_j - t_i - 1)_+} - \prod_{j\in\partial i}(1-\lambda)^{(t_j - t_i)_+}\right) = \\ = \xi_{t_i}^{\tau_i}\sum_{\alpha=1}^{6} C_\alpha(x_i^0, \tau_i, \tau_{\partial i}, t_i, t_{\partial i})$$

the good thing about this decomposition is that each of the six terms C_α can be written as a product over the neighbors ∂i:

$$\Psi_i(\{T_{il}\}_{l\in\partial i}|\mathcal{D}_i) = \xi_{t_i}^{\tau_i} \sum_{\alpha=1}^{6} \prod_{k\in\partial i} C_\alpha^k(x_i^0, \tau_i, \tau_k, t_i, t_k)$$

now we can rewrite BP equations:

$$\begin{aligned}\hat{\nu}_{\Psi_i\to(i,j)}(T_{ij}) &\propto \sum_{\{T_{ki}\}_{k\in\partial i\setminus j}} \xi_{t_i}^{\tau_i} \sum_{\alpha=1}^{6} \prod_{k\in\partial i} C_\alpha^k(x_i^0, \tau_i, \tau_k, t_i, t_k) \prod_{k\in\partial i\setminus j} \nu_{(k,i)\to\Psi_i}(T_{ki}) \\ &= \xi_{t_i}^{\tau_i} \sum_{\alpha=1}^{6} \Bigg(C_\alpha^j(x_i^0, \tau_i, \tau_j, t_i, t_j)\times \\ &\qquad \times \prod_{k\in\partial i\setminus j} \sum_{T_{ki}} C_\alpha^k(x_i^0, \tau_i, \tau_k, t_i, t_k)\nu_{(k,i)\to\Psi_i}(T_{ki}) \Bigg).\end{aligned}$$

So now the sum over $\{T_{ki}\}_{k\in\partial i\setminus j}$ enters the product. We can further optimize the computation by analyzing $C_\alpha^k(x_i^0, \tau_i, \tau_k, t_i, t_k)$. The dependence on τ_k and t_k are the ones on which we are interested in. Notice in fact that actually C_α^k does not depend on τ_k but only on $\text{sign}(\tau_k - \tau_i + s_{ki})$. Indeed, all of the three addends in the ψ^* depend on τ_k only through $\text{sign}(\tau_k - \tau_i + s_{ki})$. We now call $\sigma_{ki} = 1 + \text{sign}(\tau_k - \tau_i + s_{ki})$ and we have a variable which takes values in $\{0, 1, 2\}$. We now redefine the quantity $C_\alpha^k(x_i^0, \tau_i, \sigma_{ki}, t_i, t_k)$ as a function of σ_{ki}, so that the BP message is:

$$\begin{aligned}\hat{\nu}_{\Psi_i\to(i,j)}(T_{ij}) \propto &\xi_{t_i}^{\tau_i} \sum_{\alpha=1}^{6} \Bigg(C_\alpha^j(x_i^0, \tau_i, \sigma_{ji}, t_i, t_j)\times \\ &\qquad \times \prod_{k\in\partial i\setminus j} \sum_{T_{ki}} C_\alpha^k(x_i^0, \tau_i, \sigma_{ki}, t_i, t_k)\nu_{(k,i)\to\Psi_i}(T_{ki}) \Bigg),\end{aligned}$$

Looking at the BP equation, we see that it is possible to compress the messages. Since the variable τ_k only appears in the messages, we can actually substitute it with the variable σ_{ki}. Defining:

$$m_{\Psi_i\to(i,j)}(\tau_i, \sigma, t_i, t_j) := \hat{\nu}_{\Psi_i\to(i,j)}(\tau_i, 1 + \text{sign}(\tau_k - \tau_i + s_{ki}), t_i, t_j) \tag{D.1}$$

$$m_{(k,i)\to\Psi_i}(\tau_k, \sigma, t_i, t_k) := \sum_{\tau_k} \nu_{(k,i)\to\Psi_i}(T_{ki})\mathbb{I}[\sigma = 1 + \text{sign}(\tau_k - \tau_i + s_{ki})] \tag{D.2}$$

we have:

$$m_{\Psi_i\to(i,j)}(\tau_i,\sigma,t_i,t_j) \propto \xi_{t_i}^{\tau_i}\sum_{\alpha=1}^{6}\Bigg(C_\alpha^j(x_i^0,\tau_i,\sigma_{ji},t_i,t_j)\times \\ \times\prod_{k\in\partial i\setminus j}\sum_{\tau_i,t_i,t_k}\sum_{\sigma'=0}^{2}C_\alpha^k(x_i^0,\tau_i,\sigma',t_i,t_k)m_{(k,i)\to\Psi_i}(\tau_k,\sigma,t_i,t_k)\Bigg).$$

We can simplify further by noticing that the factor ψ only depends on t_k only through $(1-\lambda)^{(t_k-t_i-1)_+}$ or $(1-\lambda)^{(t_k-t_i)_+}$. So, each C_α^k can be rewritten with their explicit dependence on t_k:

$$C_\alpha^k(x_i^0,\tau_i,\sigma',t_i,t_k) =: (1-\lambda)^{(t_k-t_i-b_\alpha)_+}c_\alpha^k(x_i^0,\tau_i,\sigma',t_i)$$

where $b_\alpha\in\{0,1\}$. Defining:

$$\mu_{(k,i)\to\Psi_i}(\tau_k,\sigma,t_i,b_\alpha) := \sum_{t_k}(1-\lambda)^{(t_k-t_i-b_\alpha)_+}m_{(k,i)\to\Psi_i}(\tau_k,\sigma,t_i,t_k), \tag{D.3}$$

we have the BP equation:

$$m_{\Psi_i\to(i,j)}(\tau_i,\sigma,t_i,t_j) \propto \xi_{t_i}^{\tau_i}\sum_{\alpha=1}^{6}\Bigg(C_\alpha^j(x_i^0,\tau_i,\sigma_{ji},t_i,t_j)\times \\ \times\prod_{k\in\partial i\setminus j}\sum_{\tau_i,t_i}\sum_{\sigma'=0}^{2}c_\alpha^k(x_i^0,\tau_i,\sigma',t_i)\mu_{(k,i)\to\Psi_i}(\tau_k,\sigma',t_i,b_\alpha)\Bigg). \tag{D.4}$$

We have compressed the messages. The message μ has $O(T^2)$ variables. So we choose to save this in the population dynamics scheme in order to have $O(nT^2)$ numbers to store, which is a great improvement from $O(nT^4)$. Now we have to relate the message μ with the message m to close the scheme. We simply have to use definitions. From eqn. (D.3):

$$\mu_{(k,i)\to\Psi_i}(\tau_k,\sigma,t_i,b_\alpha) = \sum_{t_k}(1-\lambda)^{(t_k-t_i-b_\alpha)_+}m_{(k,i)\to\Psi_i}(\tau_k,\sigma,t_i,t_k)$$

we use eqn (D.2):

$$\begin{aligned}\mu_{(k,i)\to\Psi_i}(\tau_k,\sigma,t_i,b_\alpha) &= \sum_{t_k}(1-\lambda)^{(t_k-t_i-b_\alpha)_+}\times \\ &\quad\times\sum_{\tau_k}\nu_{(k,i)\to\Psi_i}(T_{ki})\mathbb{I}[1+\mathrm{sign}(\tau_k-\tau_i+s_{ki})=\sigma] \\ &= \sum_{t_k,\tau_k}(1-\lambda)^{(t_k-t_i-b_\alpha)_+}\nu_{\Psi_k\to(k,i)}(\tau_k,\tau_i,t_k,t_i)\mathbb{I}[\sigma_{ki}=\sigma]\end{aligned}$$

where we used for the second passage the original BP equation and the short notation $\sigma_{ki}=1+\mathrm{sign}(\tau_k-\tau_i+s_{ki})$. We also explicitly wrote $T_{ki}=\tau_k,\tau_i,t_k,t_i$. Now we finally use eqn. (D.1):

$$\mu_{(k,i)\to\Psi_i}(\tau_k,\sigma,t_i,b_\alpha) = \sum_{t_k,\tau_k}(1-\lambda)^{(t_k-t_i-b_\alpha)_+}m_{\Psi_k\to(k,i)}(\tau_k,\sigma_{ik},t_k,t_i)\mathbb{I}[\sigma_{ki}=\sigma] \tag{D.5}$$

where $\sigma_{ik}=1+\mathrm{sign}(\tau_i-\tau_k+s_{ik})$. Now we have two equations that allow to relate the sets of messages μ and m. It is convenient to keep in memory only the μ messages, because each μ message only needs $O(T^2)$ numbers to be represented. Each iteration one computes the $O(T^3)$ numbers which represent the message m by extracting random messages μ from the population and using equation (D.4). Once the m is computed, equation (D.5) allows to compute the new μ to put in the population.