Statistical physics methods for machine learning and traffic forecasting

Cyril Furtlehner

Université Paris Saclay

École doctoral de Physique en Île de France

HABILITATION À DIRIGER DES RECHERCHES

Présentée par:

Cyril Furtlehner

le 18/06/2020

Statistical physics methods for machine learning and traffic forecasting

Jury

Cécile Appert-Rolland (*Présidente*)
Adriano Barra (*Rapporteur*)
Alexander Hartmann (*Rapporteur*)
Jean-Pierre Nadal (*Examinateur*)
Kirone Mallick (*Rapporteur*)
Martin Weigt (*Examinateur*)

Chargé de Recherches INRIA Saclay
Équipe Inria TAU
LRI,Université Paris-Saclay
Bât Shannon 660
91405 Orsay Cedex France

Contents

Preamble

Remerciements

Je voudrais tout d'abord remercier Kirone Mallick, Adriano Barra et Alexander Hartmann qui ont eu la gentillesse de prendre sur leur temps pour rapporter ce document ainsi que l'ensemble des membres du jury pour leurs participation, leurs nombreuses questions et remarques encourageantes.

Ce mémoir est le fruit d'influences multiples et diverses. Il me donne l'occasion de saluer et remercier mes collaborateurs de longue date, collègues et amis. En premier lieu Guy Fayolle qui m'a offert la possibilité de me tourner vers le domaine des probabilités et modèles de files d'attente à l'Inria Rocquencourt et de donner une nouvelle impulsion à ma carrière scientifique, Jean-Marc Lasgouttes dont la collaboration datant de cette époque est reflétée dans les parties II et III, ainsi que Arnaud de la Fortelle également rencontré à Rocquencourt dans l'équipe de Guy et responsable de mon intérêt pour les problèmes de prédiction de traffic; Michèle Sebag et Marc Schoenauer dont j'ai rejoint l'équipe TAO de façon permanente à l'Inria, qui m'ont ouvert la porte sur le monde alors inconnu pour moi et fascinant de l'apprentissage machine et avec lesquels (en particulier Michèle) j'ai la chance de collaborer activement; Marc Mézard avec lequel j'ai eu également la chance de travailler au LPTMS et auprès duquel je me suis familiarisé aux algorithmes de propagations de croyances dont il est question dans ce manuscrit et enfin Aurélien Decelle avec lequel je collabore activement sur les propriétés de modèles probabilistes d'apprentissage en particulier les machines de Boltzmann restreintes dont il est question dans la dernière partie.

Organization of the Manuscript

This document traces back my research work done over the last 15 years or so. It is dealing with random walks, exclusion processes, queueing processes, irreversibility, all sort of cycles, belief propagation, traffic congestion, inverse Ising problem, sparse Gaussian copula, clustering and restricted Boltzmann machines. I attempt to unify this into a single document with the expectation of finding some guidelines for future work which will be discussed in the conclusion. For the moment just remark that the document has 4 parts, one to introduce material and subjects relevant to the next three parts dealing with quite distinct and not obviously directly related subjects.

Part II corresponds to research done mainly at Inria Rocquencourt during a period which extend from 2002 to 2012. This concerns the study of stochastic processes like exclusion or queuing processes introduced in Chapter 1 and their application to microscopic road traffic (i.e. at the level of one segment) discussed in Chapter 5. The main questions of interest discussed firstly in Chapter 6 are relative to the emergence of macroscopic phenomena resulting from simple local stochastic dynamical rules, the way to relate these two and possible ways to deal with non-reversibility in absence of integrability. In Chapter 7 these models are used in an applied perspective in order to study the fluctuations of the fundamental diagram of traffic flow.

Part III is overall concerned with the belief propagation algorithm and generalizations introduced in Chapter 3 and how to make it operational for traffic prediction at the level of a conurbation. This line of search started in 2007, and was developed mainly during 2009-2012 thanks to an ANR project I coordinated. An important question in this context is the inverse model problem introduced more specifically in its inverse Ising formulation in Chapter 4 which can be looked at in various ways as discussed in Chapter 9 and 10, in order to find a good trade-off between expressiveness of the model and computational tractability. My interest in this application was revived recently after getting in touch with the Sistema company who found our approach interesting and proposed us to test our method on their data as described at the end of Chapter 10.

Finally part IV is dedicated to the analysis of simple non-supervised learning algorithms, which is a sub-field of Machine learning briefly introduced in Chapter 2. In Chapter 11 by considering a version of Clustering related to belief-propagation, I discuss the question of the “true" number of clusters, which while being ill posed in principle, can actually be addressed in some cases by renormalization group considerations. In Chapter 12 we discuss the restricted Boltzmann machine, which played some time ago a central role in deep learning, in order to study the dynamics of learning under the angle of pattern formation.

Part I

Introductory part

Chapter 1

A few properties of stochastic particle processes

The goal of statistical physics is to understand macroscopic laws from microscopic rules of interactions, like for example the behaviour of a gas or a fluid at equilibrium emerging from Van der Waals interactions. In the last 20 years or so some progress has been made in the understanding of out of equilibrium phenomena and macroscopic transport phenomena thanks to the resolution of simple but insightful processes, for which continuous limits can be obtained. Considering the vastness of the subject, in the following short introduction we only collect a tiny set of elements useful for the reading of Part II.

1.1 Graph theory of nonequilibrium steady states

Consider a continuous time Markov process over a finite state space $\mathcal{S}$. The evolution of the distribution $P_t(\eta)$ of being in a given state $\eta \in \mathcal{S}$ at time t is governed by the master equation

$$\begin{aligned} \frac{dP_t(\eta)}{dt} &= \sum_{\eta'} W_{\eta'\eta} P_t(\eta') - W_{\eta\eta'} P_t(\eta), \\ &= \sum_{\eta'} Q_{\eta\eta'} P(\eta'), \end{aligned} \tag{1.1.1}$$

where $W_{\eta\eta'}$ is the transition rate from state η to η'. Even if this is a linear equation, solving the dynamics of such processes is in general not tractable in practice for large state space. Exact solutions can however be found sometimes, in particular for integrable systems, i.e. relying on the existence of conserved quantities in equal number as there are degrees of freedom. The steady state measure P_∞ (also called invariant measure) is usually easier to determine, because it requires in principle to find the only eigenstate (if the system is ergodic) associated to the eigenvalue zero of the transition operator Q. There are in fact two situations to distinguish depending on whether the process is reversible or not. We denote by

$$J_{\eta\eta'} = W_{\eta\eta'} P_\infty(\eta) - W_{\eta'\eta} P_\infty(\eta'),$$

the probabilistic steady state current between two arbitrary states η and η'. In the first case $J_{\eta\eta'} = 0 \ \forall(\eta, \eta')$ (detailed balance) and we are in the usual thermodynamical equilibrium situation. There is indeed a potential $E(\eta)$ allowing one to express the invariant measure as a Boltzmann distribution

$$P_\infty(\eta) = \frac{e^{-E(\eta)}}{Z},$$

where after choosing some reference state η_r and an arbitrary path $P = \{\eta_0 = \eta_r, \eta_1, \dots \eta_n = \eta\}$ between η_r and η formed of allowed transitions, we can express $E(\eta)$ as

$$E(\eta) = \sum_{i=0}^{n-1} \log \frac{W_{\eta_i \eta_{i+1}}}{W_{\eta_{i+1} \eta_i}}.$$

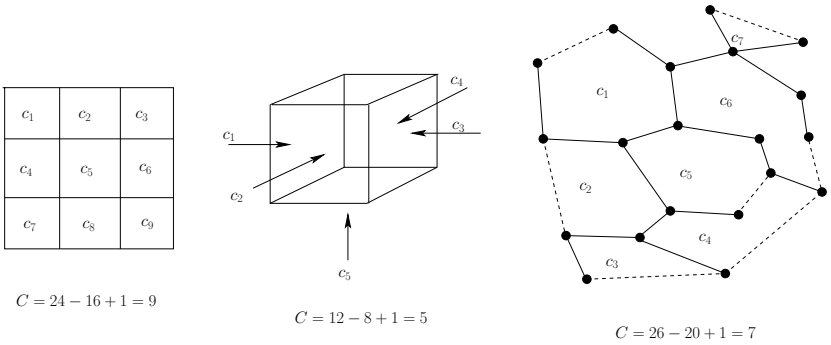


Figure 1.1.1: Example of cycle basis on 2-D and 3-D lattices and a fundamental cycle basis on an arbitrary graph.

The Kolmogorov criterion for reversibility states that the product of $\frac{W_{\eta_i \eta_{i+1}}}{W_{\eta_{i+1} \eta_i}}$ along a closed path is equal to one. This insures that this expression of $E(\eta)$ is independent of the chosen path.

Non-reversibility instead leads to a so-called non-reversible steady-state. There are pairs of states (η, η') for which non-vanishing currents $J_{\eta\eta'} \neq 0$ exist, materialized at macroscopic scale by net currents of particles for instance as illustrated in the next Sections.

To discuss irreversibility in more general terms we make use of the Schnakenberg network theory of irreversible processes [202]. For this let us recall some basic notions of graph theory which will serve also in other parts of the manuscript. In general an (un)oriented graph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of vertices and $\mathcal{E}$ a set of (un)ordered pairs of vertices representing edges. A *spanning tree* of an unoriented connected graph $\mathcal{G}$ is a subgraph of $\mathcal{G}$ which is a tree and which contains all vertices of $\mathcal{G}$. A *rooted tree* is a tree with a given orientation with respect to a specific node called the root, such that all links of a path from any node to the root are oriented toward the root. By definition a *cycle* of $\mathcal{G}$ is an unoriented subgraph where each node has an even degree. A tree has therefore no cycle as a subgraph. The set of cycles is a vector space over $\mathbb{Z}_2$ of dimension $|\mathcal{E}| - |\mathcal{V}| + 1$ for a connected graph. This means that when two cycles are combined, edges are counted modulo 2 and the resulting graph is also a cycle. Examples of cycle basis are shown on Figure 1.1.1. For heterogeneous graphs, a simple way to generate a basis consists first in selecting a spanning tree of the graph and then associate a cycle with each of the $|\mathcal{E}| - |\mathcal{V}| + 1$ remaining links of the graph not contained in the spanning tree (the *chords*), by adding to each one the path on the spanning tree joining the two ends of the link. This yields by definition a *fundamental cycle basis*, associated with the considered spanning tree.

Consider now the state graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E}$ is the set of edges between pair of nodes $(\eta, \eta') \in \mathcal{V} \times \mathcal{V}$ corresponding to non-vanishing $W_{\eta\eta'}$ or $W_{\eta'\eta}$. With a cycle basis at hand, the Kolmogorov criterion for reversibility has to be checked

only on each element of the cycle basis to be satisfied. If not we have still a general statement about the invariant measure that makes use of a variant of the matrix tree theorem also sometime referred to as the Kirchhoff theorem. The invariant measure can be expressed as

$$P_\infty(\eta) = \frac{\sum_{t\in\mathcal{T}_\eta} w(t)}{\sum_{\eta'}\sum_{t\in\mathcal{T}_{\eta'}} w(t)} \tag{1.1.2}$$

where $\mathcal{T}_\eta$ is the set of spanning trees over $\mathcal{G}$ rooted in η and $w(t)$ is the weight associated to a rooted spanning tree t, making use of its orientation:

$$w(t) = \prod_{(\eta,\eta')\in t} W_{\eta\eta'}.$$

This follows from re-expressing the solution to the steady-state equation

$$QP_\infty = 0,$$

based on the fact that

$$\sum_\eta Q_{\eta\eta'} = 0.$$

Indeed, using Cramer's rule to express the relation between $P_\infty(\eta)$ and $P_\infty(\eta_r)$ of a reference state η_r, leads to write $P_\infty(\eta)/P_\infty(\eta_r)$ as the ratio of two determinants, namely the cofactor $\tilde{Q}_{\eta_r\eta}$ of $Q_{\eta_r\eta}$ and the determinant $\tilde{Q}$ of the matrix obtained from Q by replacing $Q_{\eta_r\eta}$ by 1 for all $\eta \in \mathcal{V}$. Then Q having the structure of an admittance-matrix, it is a simple combinatorial fact that in expanding $\tilde{Q}_{\eta_r\eta}$ and $\tilde{Q}$, all cycle contribution cancel, and only contributions from positive diagonal terms contribute, resulting into a sum over spanning trees

$$Q_{\eta_r\eta} = \sum_{t\in\mathcal{T}_\eta} w(t), \qquad \tilde{Q} = \sum_{\eta\in\mathcal{V}}\sum_{t\in\mathcal{T}_\eta} w(t),$$

leading to formula (1.1.2).

Note that the knowledge of a cycle basis is also convenient to formulate thermodynamical properties of the system [202]. Indeed the entropy production $\bar{P}$ [117, 192] of a non-equilibrium steady state P_∞ is given at the microscopic level by

$$\bar{P} = \frac{1}{2}\sum_{\eta\eta'} J_{\eta\eta'}\mathcal{A}_{\eta\eta'}$$

where $\mathcal{A}_{\eta\eta'}$ is the conjugate thermodynamical force to the transition between η and η' also called affinity:

$$\mathcal{A}_{\eta\eta'} = \log\frac{W_{\eta\eta'}P_\infty(\eta)}{W_{\eta'\eta}P_\infty(\eta')}.$$

Expressed with help of a cycle basis $\mathcal{C}$, the entropy production then reads

$$\bar{P} = \frac{1}{2}\sum_{c\in\mathcal{C}} \Phi_c\mathcal{A}_c \tag{1.1.3}$$

where Φ_c are uniquely defined thermodynamical fluxes associated to each element c of the cycle basis and A_c are the corresponding cycle affinity with

$$J_{\eta\eta'} = \sum_{c \ni (\eta,\eta')} \Phi_c, \tag{1.1.4}$$

$$\mathcal{A}_c = \sum_{(\eta,\eta') \in c} \log \frac{W_{\eta\eta'}}{W_{\eta'\eta}}, \tag{1.1.5}$$

where the orientation of c is taken accordingly to the edge $(\eta, \eta') \in c$ in the first relation, and edges $(\eta, \eta') \in c$ are taken with the orientation given by c in the second one. The first relation is easily inverted in particular when the cycle basis is a fundamental cycle basis, because in that case the flux coincide with the current on the chord associated to the cycle. One virtue of this formalism in particular is to make explicit the contributions of cycles for which the Kolmogorov criterion fails to the entropy production (1.1.3). Note also that the $\mathcal{A}_c$ are independent of P_∞. This formalism has been used in [202] to make general statements about non-equilibrium steady states of stochastic processes. In particular the entropy production can decompose as a sum over (non-trivial) cycles and a fluctuation theorem can express the large deviation of the currents associated to these cycles [7].

1.2 Exclusion processes

Introduced originally by Spitzer [213] to understand anomalous diffusion the exclusion process has been subject to a vast amount of studies since then [148, 48, 23] and has become a paradigm microscopic model for transport phenomena. It is a driven lattice gas model which we consider here in one dimension. Sites of the lattice can be either empty (0) or occupied by a particle (1), and each particle can jump to the next [resp. previous] site, each jump being a Poisson process with rate q [resp. 1] if it is empty (see Figure 1.2.1), yielding the following possible transitions between two consecutive sites:

$$10 \quad \underset{1}{\overset{q}{\rightleftarrows}} \quad 01$$

The dynamics of the joint probability distribution $P_t(\eta)$ where $\eta = \{\tau_i, i = 1, \dots N\}$ is a sequence of binary variables $\tau_i \in \{0, 1\}$ encoding the presence or absence of a particle at a given site i, is governed by equation (1.1.1) with $W_{\eta\eta'}$ equal q or 1 for pairs of states (η, η') related respectively by a single forward or backward particle jump. The process can be defined on the ring geometry with periodic boundary conditions or with open boundary conditions. In the latter case rates have to be introduced for particles leaving/entering the system from/to both end sites of the system. These rate represents chemical potentials of reservoirs connected to these end sites.

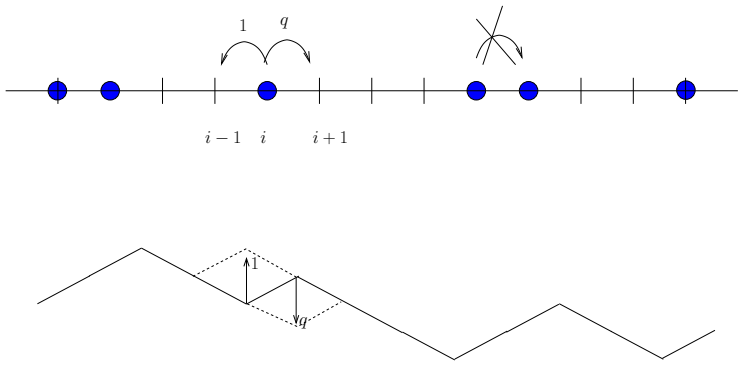


Figure 1.2.1: Asymmetric exclusion process (top). Corresponding fluctuating interface (bottom)

The process is referred to the asymmetric exclusion process (ASEP) when $q \neq 1$ and to the totally asymmetric exclusion process (TASEP) when $q = 0$. This process is non-reversible as soon as $q \neq 1$ or when boundary edge conditions represented by reservoir with distinct chemical potential are present. On the ring geometry, even if the process is non-reversible ($q \neq 1$), the invariant measure is uniform based on a simple partial balance argument. In fact on this geometry many things can be computed thanks to the integrability of the model (see e.g. [206]) through a mapping of the Markov matrix to an Heisenberg spin chain, like the time dependent joint distribution [191], the spectral gap giving the relaxation time to equilibrium [94, 92] or large deviation functionals of particles displacements [51, 49].

On open systems the situation regarding the invariant measure is more complex. Consider for instance a finite system with boundary conditions specified by an incoming rate of particle α on the left most site and an escaping rate β on the right most site of the system. In this case, for $\alpha \neq \beta$ the process is non-reversible and the steady state has not a Gibbs form. Remarkably a closed form expression has been discovered [50] to obtain the steady-state probabilities of each individual state with help of a matrix ansatz. In this representation, a given sequence $\eta = 1010\ldots00$ is represented by a product of matrices D (for 1) and E (for 0), and the corresponding probability measure is obtained by taking the trace

$$\pi_\eta = \frac{1}{Z}\mathrm{Tr}\big(WDEDE\ldots EE\big),$$

where W is an additional matrix which takes into account the boundary property. When inserting this form into the master equation it is immediate to verify

that it formally solves the stationary regime if D, E, W satisfy

$$\begin{aligned} qDE - ED &= D + E \\ DW &= \frac{1}{\beta} W \\ WE &= \frac{1}{\alpha} W. \end{aligned} \tag{1.2.1}$$

Infinite dimensional representation of these quadratic algebra can be obtained with help of q-deformed oscillator algebra [23], and in fact all irreducible finite dimensional representations have been determined [155], thereby implying the consistency of the solution.

When the size L of the system goes to infinity there is a deterministic limit of the dynamics of the particle density $\rho(x,t)$, a so-called hydrodynamic limit represented by the Burger equation, obtained after assuming the behaviour of the rate $q = 1 + v/L$ and doing the rescaling $x = i/L$ and $t' = Lt$ and

$$\frac{\partial \rho}{\partial t} = \frac{\partial^2 \rho}{\partial x^2} - v \frac{\partial}{\partial x}[\rho(1-\rho)],$$

which can actually be rigorously established only when the steady state measure has a product form [21, 42, 214, 133]. The non-linear term in this equation allows the onset of travelling wave, which become shockwaves separating high density phases from low ones, when the diffusion term is absent.

The exclusion process can be equivalently formulated as a fluctuating interface process (see Figure 1.2.1). In the latter representation a configuration is represented by a set of up and down links corresponding respectively to empty and occupied sites. Possible transitions are represented then by moving one point of the path up or down which corresponds respectively to a particle jumping to the left or to the right. The continuous limit of this process obtained by considering the height $h(x)$ of the path at position $x = i/L$ when L becomes large has been extensively studied. This is the famous KPZ equation [123]:

$$\frac{\partial h(x,t)}{\partial t} = \nu \bigtriangledown^2 h(x,t) + \frac{\lambda}{2} \big(\bigtriangledown h(x,t) \big)^2 + \eta(x,t)$$

which constitutes a universality class distinct from the Brownian process, playing a central role in statistical physics.

There are various multi-type particle generalizations of ASEP, where each type has its own hopping rate and overtaking between different species is possible. For instance the multi-species ASEP of Karimipour [124] is integrable on the ring and its steady state measure can be expressed in matrix form. The generalized (hierarchical) multi-species TASEP [58] enjoys as well a matrix form representation for its steady state measure. The possibility of having a closed form expression in terms of a matrix ansatz for a given model has been severely constrained in [113]. In fact the consistency of quadratic diffusive algebra of the form

$$g_{\alpha\beta} D_\alpha D_\beta - g_{\beta\alpha} D_\beta D_\alpha = x_\alpha D_\beta - x_\beta D_\alpha,$$

where α and β denote particles types, and $g_{\alpha\beta}$ and x_α are assumed to be scalars is indeed shown to be absent in general, at the exception of very specific tuning of the parameters, including the aforementioned ones. The link between integrability and existence of matrix forms if not fully elucidated yet seems to hold in general [227].

Another model of interest for the present manuscript is the so-called ABC model [33] considered on the ring geometry. It has three types of particles with the following possible transitions:

$$AB \underset{1}{\overset{q}{\rightleftarrows}} BA$$

$$BC \underset{1}{\overset{q}{\rightleftarrows}} CB$$

$$CA \underset{1}{\overset{q}{\rightleftarrows}} AC$$

This model is reversible for $q \neq 1$ only when the various types of particles have equal densities. If not there are no known consistent quadratic algebra able to express the invariant measure in closed form. Still, this model is interesting in various respects: in the reversible case there is a second order phase transition corresponding to coalescence phenomenon in the diffusive regime where we have the scaling $q = \exp(-\beta/L)$ w.r.t. system size L, the critical point being given by $\beta_c = 2\pi\sqrt{3}$. For sufficiently unbalanced densities the transitions become first order with a coexistence phase showing up in the phase diagram [33, 36] between the ordered and disordered phases. This shows that long range order can take place in one dimension from local transitions with interesting phenomena like e.g. anomalous behaviour of current fluctuations at the transition point [90].

1.3 Queuing processes

Another important class of stochastic particle process is the so-called zero range process [59], which is in fact a special case of queuing network processes [88] well studied in probability theory also sometimes referred to as urn models in statistical physics [91] which generalize the Ehrenfest model to an extensive number of urns. Consider first a station ticket office: travellers arrive at random, say with Poisson arrival rate λ, and get served at the counter with a random service time τ of hopefully finite mean $1/\mu$. A common assumption is that the service time is exponentially distributed with rate μ. Then the number n_t of travellers in the queue is also called a birth and death process. It is ergodic provided that $\lambda < \mu$, and the steady state measure is geometric

$$P_\infty(n_t = n) = (1 - \rho)\rho^n,$$

with $\rho = \lambda/\mu$ representing the mean number of clients in the queue, as a special case of the (general) Little law $\rho = \lambda W$ where W is the mean waiting time in

the queue. More generally if the service rate depends on n and provided that $\lambda < \lim_{n\to\infty} \mu_n$ we have

$$P_\infty(n_t = n) = \frac{1}{Z(\lambda)} \prod_{k=1}^{n} \frac{\lambda}{\mu_k},$$

and $Z(\lambda)$ takes the form of a generating function

$$Z(\lambda) = \sum_{n=0}^{\infty} a_n \lambda^n, \qquad \text{with} \qquad a_n^{-1} = \prod_{k=1}^{n} \mu_k.$$

Its spectral radius corresponds to the limit of ergodicity. Suppose now that there are many servers, assembled into a network, like e.g. a telecommunication network where packets of information have to pass through many servers before reaching a destination. This means that after being served at a server i, the client (or packet) is redirected to another one j with some routing probability p_{ij}. The 2-servers problem for instance represented as random walks in the quarter plane as been studied extensively in [66], using algebraic methods to solve functional equations of bi-variate generating functions. In the zero-range process assumption, the network is usually represented by a regular lattice and particles hop from one site to a neighbouring site. The hopping process is commonly taken as a Poisson process with a rate usually dependent only on the occupation of the departure site. In the queuing theory, general class of queuing networks have been identified which have simple explicit steady state measures. The first one is the so-called Jackson network with exponential service rates [114] possibly open or closed, which typically includes zero-range processes considered in statistical physics. Then BCMP networks were introduced later with more general service policy [18] including multiple types of clients [128]. All have in common that the invariant measure has a product form

$$P_\infty(n_1, \ldots, n_N) = \frac{1}{Z} \prod_{i=1}^{N} p_i(n_i), \tag{1.3.1}$$

for a network composed of N queues. There are two key points for this to occur:

- each server considered in isolation, with in an incoming arrival rate λ is reversible [1].
- there is a (unique up to a multiplicative constant) set of incoming rates $\{\lambda_i, i = 1, \ldots N\}$, associated to each queue i satisfying the so-called traffic equations:
$$\sum_{j \in V(i)} p_{ji}\lambda_j = \lambda_i, \qquad \forall i \in \{1, \ldots N\}$$
(written here for a closed system).

[1] In fact quasi-reversibility is enough. This notion corresponds to having partial balance instead of detailed balanced equations in a context of multi-class customers for instance (see [129])

Then in the form (1.3.1) each factor p_i is the equilibrium distribution of server i taken in isolation with arrival rate λ_i, solution of the traffic equations. Remarkably one should insist on the fact that as whole the system is not reversible in general. Non-vanishing net flows of customers between servers may occur, such that this constitutes a general class of non-reversible steady states expressible in closed form. To illustrate this, let us come back to the exclusion process on the ring of the previous Section. There is a mapping onto a Jackson network as follows: to each empty site we associate a queue, and consider by convention that the clients in the queue are the particles on the left interval between the corresponding empty site and the previous one (see Figure 1.3.1). The traffic

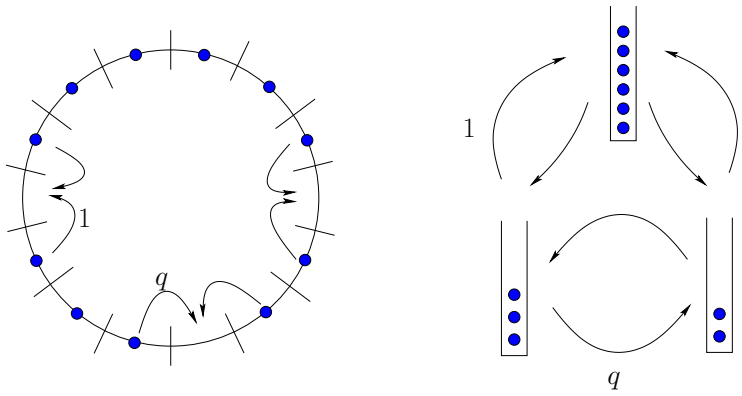


Figure 1.3.1: Mapping between the asymmetric exclusion process on the ring geometry (left) with a simple cyclic Jackson network (right)

equations involve two probabilities $p_{ii+1} = q/(1+q)$ and $p_{i+1\ i} = 1/(1+q)$ and admits a uniform solution $\lambda_i = \lambda$. In the end we get the steady-state joint probability measure describing the clusters of particles as

$$p_\infty(\{n_i, i = 1, \dots M\}) = \frac{1}{\rho^N Z} \prod_{i=1}^{M} \rho^{n_i} \delta\Big(\sum_{i=1}^{M} n_i - N\Big),$$

$$= \frac{1}{Z} \delta\Big(\sum_{i=1}^{M} n_i - N\Big)$$

with $\rho = \lambda/(1+q)$ and λ arbitrary, N being the total number of particles which is constrained, the system being closed, and M the total number of empty sites of the ASEP system. We end up with a uniform distribution over partitions of N into M parts, Z being the overall number of such partitions. This is consistent with the fact previously stated that the invariant measure of ASEP on the ring is the uniform distribution. This correspondence between exclusion process and queuing processes can be used in more complex problems to establish the invariant measure in closed form [67].

Chapter 2

Machine learning: a focus on unsupervised learning

2.1 Machine learning and statistical physics

In the last decade, the field of machine learning (ML) became the center of attention of both the public domain and of scientific research, thanks to spectacular breakthroughs in the training of artificial neural networks [142]. ML is a branch of artificial intelligence (AI) which purpose is to extract information automatically from data. There are three main fields of ML, called *supervised*, *unsupervised* and *reinforcement learning*, which respective goals are schematically dealing with the development of algorithms able to learn complex functions, distributions and policies automatically from data. The first one has to do mainly with classification or regression problems where from any input x like an image one is willing to be able to determine a label about its content or some quantitative information $y = f(x)$, by learning f based on a training dataset of pairs (x, y). In the second one the goal is to learn or model the distribution $p(x)$ itself of the training data in order to extract meaningful features from the data or to be able to generate new realistic ones. The last one is aiming at learning a conditional probability $\pi(a, s)$ where a represents an action (e.g. like a move in a maze) given a state s (the position in the maze) regarding the final goal (escape from the maze).

With the development of deep neural networks taking advantage of the GPU technology, the performance on classification tasks (supervised learning) started to outperform human level at image recognition, and more recently generative models(unsupervised learning) such as generative adversarial networks [93] (GAN) have been able to generate images that cannot be distinguished from true ones [125]. In the context of strategic games like Go, deep neural networks integrated as modules of reinforcement learning algorithms, have also been instrumental in reaching superhuman performances [210].

On the theory side, despite recent significant advances [154, 209, 19, 115, 87] the theoretical understanding of deep learning lag behind these achievements, in various respects like for instance on questions regarding the link between adequate choice of network architecture and complexity of the data. Statistical foundations of learning [228, 226] addresses the question of consistency of machine learning models in very general and elegant terms: given an underlying distribution $F(z)$ from which a set $\mathbf{z} = (z_1, \ldots, z_\ell)$ of training data is sampled, a set Λ of models, a cost function $Q(z, \theta)$ which for a given model $\theta \in \Lambda$ associates to a sample z e.g. a classification score (0 or 1 respectively for correct or incorrect classification) or the negative log likelihood $-\log(p(z, \theta))$, one looks for the optimal model $\theta^\star$ minimizing the following quantity (the risk)

$$\mathcal{R}(\theta) = \int Q(z, \theta) dF(z), \qquad \theta \in \Lambda.$$

This is done by approximating the true risk by the empirical one estimated from the training data

$$\mathcal{R}_{\text{emp}}(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} Q(z_i, \theta),$$

according to the so-called *empirical risk minimization* inductive principle. Consistency means informally that the true risk $R(\theta_\ell)$ of the model θ_ℓ minimizing $\mathcal{R}_{\text{emp}}$ is close to the optimal one $\mathcal{R}(\theta^\star)$. If this is not satisfied we end up in a situation of *overfitting* the data. Intuitively the model will be consistent if its resolution is adapted to the number of data, i.e. not too high, so that some form of law of large numbers applies on the functional space Λ. This is made precise by associating to Λ the Vapnik-Chervonenkis (VC) entropy

$$H^\Lambda(\ell) \stackrel{\text{def}}{=} \mathbb{E}_{z_i\sim F(z)}\big[\log \Omega^\Lambda(z_1,\dots,z_\ell)\big], \tag{2.1.1}$$

with $\Omega^\Lambda(z_1,\dots,z_\ell)$ schematically measuring the number of different realizations of the functional space Λ on the set of points $(z_1,\dots,z_\ell)$. Consistency is insured asymptotically when

$$\lim_{\ell\to\infty} \frac{H^\Lambda(\ell)}{\ell} = 0,$$

while a too high resolution would correspond to having $H^\Lambda(\ell) = \mathcal{O}(\ell\log(2))$ (for a binary classification problem). More practical statements can be obtained by considering the growth function

$$G^\Lambda(\ell) \stackrel{\text{def}}{=} \sup_{z_i\sim F(z)}\big[\log \Omega^\Lambda(z_1,\dots,z_\ell)\big],$$

instead of the VC entropy. As can be proved, there exist a quantity h called the VC dimension, which for a binary classification, is such that $G^\Lambda(h) = h\log(2)$ and beyond which $G^\Lambda(\ell) < h\log(2\frac{\ell}{h})$, for $\ell > h$. It has a simple geometrical interpretation in terms of the maximal number of points that can be shattered by the set of functions Λ, hence characterizing its level of resolution. Typically in practice a ratio of $\frac{\ell}{h} \geq 20$ will be avoiding overfitting.

ML can be reformulated in a language more familiar to physicists [159, 150], to make the relevance of statistical physics more apparent in this context. A common way to relate ML to statistical physics consists in a Bayesian setting to consider the empirical risk as an energy function of a Gibbs measure on the candidate models [207, 178]

$$\begin{aligned} P_{\text{posterior}}(\theta) &\propto P_{\text{prior}}(\theta)\exp\big(-\beta\mathcal{R}_{\text{emp}}(\theta)\big), \qquad \theta\in\Lambda, \\ &\propto \exp\big(-\beta E(\theta)\big) \end{aligned}$$

where P_{prior} and $P_{\text{posterior}}$ represent respectively the prior and the posterior distribution of the models after taking into account the training data z_i, β being an inverse temperature related to output noise of the model. This distribution can be considered as the equilibrium Gibbs distribution associated to a stochastic learning dynamics of the parameters expressed as a Langevin equation [207]

$$d\theta_t = -\nabla_\theta E(\theta)dt + d\eta_t,$$

with $d\eta_t$ a white noise term of inverse variance 2β. In this form the training data appears as a quenched disorder to be averaged over.

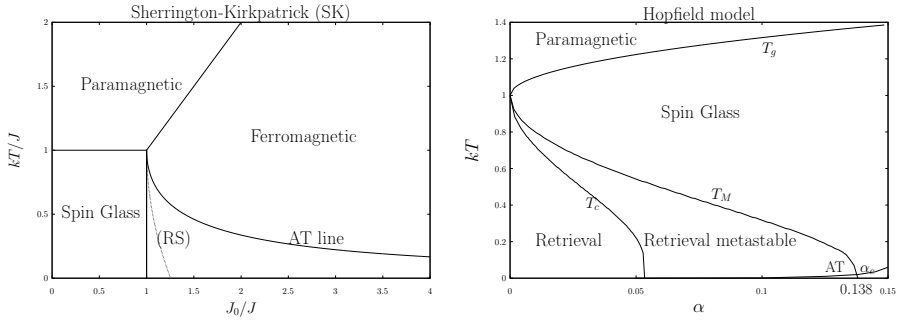


Figure 2.1.1: Phase diagram of the SK model (left) and the Hopfield model (right) obtained with the replica trick formalism and the spontaneous replica symmetry breaking mechanism. For the SK model the couplings are distributed as $J_{ij} \sim \mathcal{N}(\frac{J_0}{N}, \frac{J^2}{N})$ while for the Hopfield model they are the result of superposing $P = \alpha N$ independent patterns, N being the number of variables. The "structured" part of the coupling is of rank 1 in the SK and of rank P in the Hopfield model. In both cases the AT line signals the breakdown of the replica symmetry. The dotted line (RS) on the left indicates the boundary of the SG phase obtained with the replica symmetric saddle point.

As a matter of fact, historically, statistical physics played an influential role in the development of neural networks. In particular, during the 1980s the Hopfield model of associative memory [107] triggered a lot of theoretical studies following the parallel development of spin-glass theory [163]. Its steady state properties are well described in terms of the Gibbs distribution based on the Hamiltonian

$$H[\mathbf{s}] = -\frac{1}{NP} \sum_{i=1,k=1}^{N,P} \xi_i^{(k)} \xi_j^{(k)} s_i s_j,$$

where $\{\xi_i^{(k)} \in \{-1, 1\}, i = 1, \dots N, k = 1, \dots P\}$ represents a set of P patterns to be memorized and encoded into a random spin model of N variables. A natural set of order parameters based on the overlaps

$$m^{(k)} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{S,\xi}\big(\xi_i^{(k)} s_i\big), \qquad \forall k = 1, \dots, P, \tag{2.1.2}$$

can be used to characterize the retrieval phase of the model, where the expectation comprises both thermal averages over the spin configurations $\mathbf{s}$ and expectations with respect to the quenched disorder variables $\xi_i^{(k)}$. The mean-field theory of the Hopfield model has been solved by Amit, Gutfreund and Sompolinsky in [5, 6] using replica's techniques, results which were soon confirmed with help of the cavity method [163], and put later on even firmer mathematical grounds in [219]. There are 3 phases, separated by the transition lines T_g, between the paramagnetic phase and the spin glass phase, and T_c, between the

spin glass phase and the ferromagnetic phase (see Figure 2.1.1). The latter corresponds to the so-called *Mattis states*, i.e. to spin configurations correlated with one of the mixture components and is also called *retrieval phase*. The number of patterns that can be retrieved [6, 165] is an important outcome of such analysis. A different question deals with neural network learning capacity, i.e. the number of pattern that can be stored as a function of its number N of neurons by e.g. the perceptron under various hypothesis [84, 85, 136, 135, 188], which has a close relationship [95, 179] with the computation of the VC entropy (2.1.1). While statistics machine learning theory provides bounds on the true risk giving rise to learning algorithms like support vector machines (SVM) [24], methods from statistical physics can be used in some cases to deliver asymptotic estimates of the learning curve, i.e. the risk as a function of the number of examples [207, 180, 52]. Layered networks were also considered in thermodynamic limits [53, 166] with help of mean-field techniques, in order to clarify mechanisms of information storage and the efficiency of learning algorithms based on associative memory.

More recently, mean-field techniques developed originally in the context of complex systems like as TAP equations [220] and the cavity approximation [163] and their variants related to belief propagation [160] have found a great variety of new playgrounds in Machine learning, such as compressed sensing [138], community detection [46] to mention only a few of them. In these problems, a Bayesian setting is considered with a so-called teacher-student scenario originally introduced in [86]. A simplification can be nicely exploited in the Bayes-optimal case (see [246] and references herein), the notion of Bayes-optimality being actually viewed [112] in the context of error correcting codes as equivalent to the Nishimori line [178] introduced in statistical physics.

2.2 Clustering

In the domain of unsupervised learning, clustering techniques are old standard but widely used tools of Machine learning. It consists in to partitioning an ensemble of objects such that similar ones pertain to the same classes. A precise statement of the problem requires the definition of a similarity measure between objects and of a cost function. As such, it turns out to be an optimization problem, which is generally NP-Hard. Many algorithms have been proposed, ranging from expectation-maximization (EM) types approaches [47] like k-centers and k-means [1] to percolation-like methods for building hierarchies. Some other methods generically called spectral clustering [3, 175] are based on the spectral properties of the affinity matrix or closely related diffusion operators, like a non-backtracking matrix derived from the belief propagation stability [139], such that clusters get associated to separated components of a diffusion process.

From the statistical physics viewpoint depending on the form of the cost

[1] in these algorithms, k is the number of centers to be obtained by alternatively assigning datapoints to candidate centers (expectation step) and then taking the mean of each newly defined cluster as new candidate centers (optimization step)

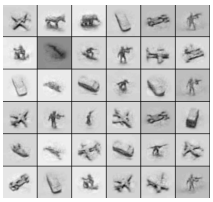
function, the clustering solution may be reformulated as the ground sate of a q-states Potts model which can be solved by Monte-Carlo based methods [237]. This type of models are suitable for Bethe-Peierls approximations, which algorithmic counterpart is known to be belief propagation algorithm to be detailed in the next chapter. Considering a relaxed version of the cost function were clusters are identified by exemplars, and only the similarity of data to their exemplars are taken into account, a clustering algorithm called *affinity propagation* (AP) [70] has been proposed as an instance of the min-sum algorithm to solve the clustering problem. This algorithm turns out to be very efficient compared to other center-based methods like k-centers and k-means, the price to pay for this stability property being a quadratic computational complexity. A basic assumption behind AP, is that each cluster is of spherical shape. This limiting assumption has actually been relaxed by Leone and co-authors in [145, 146], by softening a hard constraint present in AP, which impose that any exemplar has first to point to itself as oneself exemplar. More details on AP will be given in Chapter 3.

Another point common to most clustering techniques, is to fix the free parameter which determines the number of clusters. Some methods based on EM [69] or on information-theoretic consideration have been proposed [215], but mainly use a precise parametrization of the cluster model. There exists also a different strategy based on similarity statistics [56], that have been combined with AP [231], at the expense of a quadratic price. Determining the number of clusters in a given dataset might be considered an ill-defined problem: there is no natural scale, enabling to compare e.g., the k-means solutions obtained for different values of k. How to set the number of clusters is actually part of a broader theoretical question, regarding the consistency of a given clustering algorithm associated to a dataset [134]. The mainstream stability-based approach to consistency (see e.g. [56]) proceeds by comparing the clustering obtained on independent samples of the dataset, computing some stability measure. The parameters of the clustering algorithm are selected so as to optimize the stability criterion. While this approach is empirically efficient, it has some shortfalls [20], and an optimal stability does not necessarily translate into a "best clustering".

Scale properties take an important place in the axiomatization of clustering [134, 1], with the requirement that the cost function and the clustering quality measure be scale invariant. This requirement is commonly faced in statistical physics, where consistent models near phase transition points are expected to be insensitive at large scale to microscopic irrelevant details.

2.3 Restricted Boltzmann machines

Generative models is a generic denomination to designate unsupervised ML systems which purpose is to represent faithfully the density distribution of data embedded into a high dimensional space. These models have been demonstrated on image first by their ability to produce highly realistic purely synthetic faces or objects such as the ones shown on Figure 2.3.1. Various architectures exists,



3D objects - DBM

Salakhutdinov et al. 2009

Faces VAE

Klys et al. 2018

Faces GAN

Chen et al. 2016

Figure 2.3.1: Examples of synthetic images automatically generated by a deep Boltzmann machine (left) a variational auto-encoder (center) and a generative adversarial network (right).

among which the variational auto-encoder (VAE) [132] (a sophisticated version of previously proposed auto-encoders [102, 101]), the generative adversarial network (GAN) [93] and the deep Boltzmann machines (DBM) [196] are the most popular. They all inherit developments made in the context of deep learning and can be used potentially in many different area.

Originally called Harmonium [211], the Restricted Boltzmann Machine (RBM) is also a generative model albeit much simpler than previous ones. It played an important role in deep learning as a way to pre-train deep auto-encoders layer wise [101]. In order to build more powerful models, RBMs can indeed be stacked to form "deep" architectures like DNN or DBM. It is a simple 2-layers undirected neural network which represents the data in the form of a Gibbs distribution of visible and latent variables (see Figure 2.3.2). The former noted $\mathbf{s} = \{s_i, i = 1 \dots N_v\}$ correspond to explicit representations of the data while the latter noted $\sigma = \{\sigma_j, j = 1 \dots N_h\}$ are there to build arbitrary dependencies among the visible units. They play the role of an interacting field among visible nodes. Usually the nodes are Bernoulli distributed, but Gaussian distributions or other distributions on real-valued bounded support are also used [224], ultimately making RBMs adapted to more heterogeneous data sets. Assuming binary variables $s_i, \sigma_j \in \{-1, 1\}$, an energy function is defined for a configuration of nodes

$$E(\mathbf{s}, \sigma) = -\sum_{i,j} s_i W_{ij} \sigma_j + \sum_{i=1}^{N_v} \eta_i s_i + \sum_{j=1}^{N_h} \theta_j \sigma_j \tag{2.3.1}$$

defining a Gibbs distribution between visible and hidden units,

$$p(\mathbf{s}, \sigma) = \frac{e^{-E(\mathbf{s},\sigma)}}{Z} \tag{2.3.2}$$

where W is the weight matrix and η and θ are biases, or external fields on the variables. Each weight vector associated to a given hidden unit and its corresponding bias defines an hyperplan partitioning the visible space into two regions corresponding to the hidden unit being activated or not (see Figure 2.3.2).

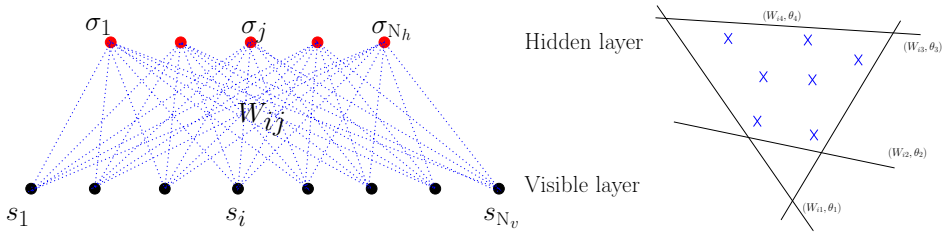


Figure 2.3.2: Bipartite structure of the RBM (left). Hyperplans defined by the weight vectors and bias associated to each hidden variable can delimit fixed density regions in input space (right).

$Z = \sum_{\mathbf{s},\sigma} e^{-E(\mathbf{s},\sigma)}$ is the partition function of the system. The joint distribution between visible variables is then obtained by summing over hidden ones. In this context, learning the parameters of the RBM means that, given a dataset of M samples composed of N_v variables, we ought to infer values to W, η and θ such that new generated data obtained by sampling this distribution should be similar to the input data. The standard method to infer the parameters is to maximize the log likelihood of the model, where the pdf (2.3.2) has first been summed over the hidden variables

$$\mathcal{L} = \sum_j \langle \log(2\cosh(\sum_i W_{ij} s_i - \theta_j)) \rangle_{\text{Data}} - \sum_i \eta_i \langle s_i \rangle_{\text{Data}} - \log(Z). \tag{2.3.3}$$

Over time since its introduction, the RBM has continuously attracted the interest of the research community, firstly because it can be easily used for both continuous and discrete variables [137, 240, 32, 238] and the activation can be tuned to be either binary of relu [174]; secondly because for datasets of modest size it is able to deliver good results [105, 109] comparable to the ones obtain from more elaborated network such as GAN (see for instance [243]).

Considered as a special case of a product of experts, a learning algorithms called contrastive divergence [103] (CD) as been proposed and subsequently refined to Persistence CD [221] (PCD). Efficient and well documented [100] these algorithms are based on a quick Monte Carlo estimation of the response function of the RBM, exploiting the conditional independence of the visible or hidden variables conditionally to the complementary ones.

They all correspond to expressing the gradient ascent on the likelihood as

$$\Delta W_{ij} = \gamma \left(\langle s_i \sigma_j p(\sigma_j|\mathbf{s}) \rangle_{\text{Data}} - \langle s_i \sigma_j \rangle_{p_{\text{RBM}}} \right) \tag{2.3.4}$$

$$\Delta \eta_i = \gamma \left(\langle s_i \rangle_{p_{\text{RBM}}} - \langle s_i \rangle_{\text{Data}} \right) \tag{2.3.5}$$

$$\Delta \theta_j = \gamma \left(\langle \sigma_j \rangle_{p_{\text{RBM}}} - \langle \sigma_j p(\sigma_j|\mathbf{s}) \rangle_{\text{Data}} \right) \tag{2.3.6}$$

where γ is the learning rate. The main problem are the $\langle \cdots \rangle_{p_{\text{RBM}}}$ terms on the right hand side of (2.3.4-2.3.6). These are not tractable and the various methods basically differ in their way of estimating those terms (Monte-Carlo Markov

chains, naive mean-field, TAP...). For an efficient learning the $\langle \cdots \rangle_{\text{Data}}$ terms must also be approximated by making use of random mini-batches of data at each step.

Nevertheless, despite some interesting interpretations of CD in terms of non-equilibrium statistical physics [197], the learning of RBMs remains a set of obscure recipes from the statistical physics point of view: hyperparameters (like the size of the hidden layer) are supposed to be set empirically without much theoretical guidelines. Even for practical purpose, it is intrinsically difficult to efficiently estimate numerically the gradient w.r.t. the parameters of the model, as soon as the network has learned non trivial modes.

The very definition of the RBM allows one to study it in a way similar to the SK or to the Hopfield model. The analogy is actually strengthened by the observation that an RBM with Bernoulli-Gaussian variables is mapped exactly to the Hopfield model [15], the number of patterns of the Hopfield model corresponding to the number of hidden units. Based on that, recent works [17, 16] characterize the retrieval capacity of RBMs. Mean-field based algorithms based on TAP equations have also been proposed [82, 111, 218, 160] in addition to Gibbs sampling based methods. None of these being fully satisfactory (see e.g. [225] for a more detailed discussion), especially if one is willing to learn an empirical distribution with good accuracy. RBM with sparse weight matrix have been considered to analyze compositional mechanisms [2, 168] of features to create complex patterns. From the analysis of related linear models [222, 25], it is already a well established fact that a selection of the most important modes of the singular values decomposition (SVD) of the data is performed in the linear case. In fact in the simpler context of linear feed-forward models the learning dynamics can be fully characterized by means of the SVD of the data matrix [200], showing in particular the emergence of each mode by order of importance with respect to the corresponding singular values.

Chapter 3

Belief propagation and generalizations

3.1 Markov random fields

Markov random fields [141] (MRF) are widely used probabilistic models, able to represent multivariate structured data in order to perform inference tasks. They are at the confluence of probability, statistical physics and machine learning [230]. From the formal probabilistic viewpoint they express the conditional independence properties of a collection of n random variables $\mathbf{x} = \{x_1, \ldots, x_n\}$, in the form of a factorized probability measure, where each factor involves a subset of $\mathbf{x}$. In statistical mechanics the Gibbs measure takes the form of an MRF, to express the thermodynamic equilibrium probability of a system of n interacting degrees of freedom. The practical use of MRF appears also in various applied fields, like image processing, bioinformatics, spatial statistics or information and coding theory, as well as in the context of deep learning as explained in the previous chapter. There are two main generic problems that have to be commonly dealt with when using MRF in practical applications:

Direct inference problems:

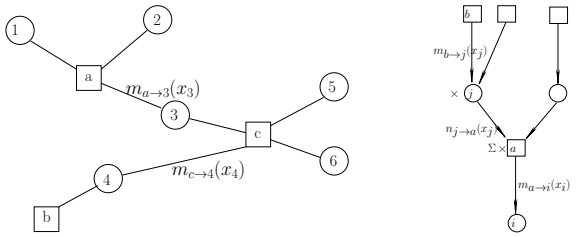
- computation of marginal probabilities, also called marginalization problem:
$$p_i(x_i) = \sum_{\mathbf{x} \setminus x_i} P(x_i),$$
which involves in general an exponential cost with respect to N to be done exactly;

- computing the mode, also referred to as the maximum a posteriori probability (MAP)
$$\mathbf{x}^\star = \operatorname*{argmax}_{\mathbf{x}} P(\mathbf{x}),$$
which is generally an NP hard problem [37, 208].

These two problems are of different nature and involve generally distinct techniques which can share sometimes some similarities. The former can be addressed e.g. by Monte-Carlo sampling or by mean-field methods which boils down to some approximation of the entropy contribution to the free energy; the latter is a combinatorial optimization problem which corresponds to the search for the ground state of a system at zero temperature.

3.2 Belief Propagation

The "belief propagation" algorithm (BP), appeared in the artificial intelligence community for inference problems on Bayesian networks [185]. It is a non-linear iterative map which propagates information on a dependency graph of variables in the form of messages between variables. It has been recognized to be a generic procedure, instantiated in various domains like error correcting codes, signal processing or constraints satisfaction problems with various names



(a) (b)

Figure 3.2.1: Example of factor graph (a) and message propagation rules (b).

depending on the context [140]: the forward-backward algorithm for hidden Markov model selection; the Viterbi algorithm; Gallager's sum-product algorithm in information theory. It has also a nice statistical physics interpretation in the context of mean-field theories, as a minimizer of a Bethe free energy [242], a solver of the cavity equations [163] and its relation to the TAP equations in the spin-glass context [119]. As a noticeable development in the recent years, related to the connection with statistical physics [161], is the emergence of a new generation of algorithms for solving difficult combinatorial problems, like the survey propagation algorithm [164] for constraint satisfaction problems or the affinity propagation for clustering [70]. We consider a set of discrete random variables $\mathbf{x} = \{x_i, i \in \mathcal{V}\} \in \{1, \ldots, q\}^{|\mathcal{V}|}$ obeying a joint probability distribution of the form

$$\mathcal{P}(\mathbf{x}) = \prod_{a \in \mathcal{F}} \psi_a(x_a) \prod_{i \in \mathcal{V}} \phi_i(x_i), \tag{3.2.1}$$

where ϕ_i and ψ_a are factors associated respectively to a single variable x_i and to a subset $a \in \mathcal{F}$ of variables, $\mathcal{F}$ representing a set of cliques and $x_a \stackrel{\text{def}}{=} \{x_i, i \in a\}$. The ψ_a are called the "factors" while the ϕ_i are there by convenience and could be reabsorbed in the definition of the factors. This distribution can be conveniently represented with a bi-bipartite graph, called the factor graph [140]; $\mathcal{F}$ together with $\mathcal{V}$ define the factor graph $\mathcal{G}$, which will be assumed to be connected. The set $\mathcal{E}$ of edges contains all the couples $(a, i) \in \mathcal{F} \times \mathcal{V}$ such that $i \in a$. We denote d_a (resp. d_i) the degree of the factor node a (resp. to the variable node i). The factor graph on the Figure 3.2.1.a corresponds for example to the following measure

$$p(x_1, \ldots, x_6) = \frac{1}{Z} \psi_a(x_1, x_2, x_3) \psi_b(x_4) \psi_c(x_3, x_4, x_5, x_6)$$

with the following factor nodes $a = \{1, 2, 3\}$, $b = \{4\}$ and $c = \{3, 5, 6\}$. Assuming that the factor graph is a tree, computing the set of marginal distributions, called the belief $b(x_i = x)$ associated to each variable i can be done efficiently. The BP algorithm does this effectively for all variables in one single procedure, by remarking that the computation of each of these marginals involves intermediates quantities called the messages $m_{a\to i}(x_i)$ [resp. $n_{i\to a}(x_i)$] "sent" by factor

node a to variable node i [resp. variable node i to factor node a], and which are necessary to compute other marginals. The idea of BP is to compute at once all these messages, using the relation among them as a fixed point equation. Iterating the following message update rules sketched on Figure 3.2.1.b:

$$m_{a\to i}(x_i) \longleftarrow \sum_{\mathbf{x}_a\setminus x_i} \psi_a(\mathbf{x}_a) \prod_{j\in a\setminus i} n_{j\to a}(x_j), \tag{3.2.2}$$

$$n_{j\to a}(x_j) \longleftarrow \phi_j(x_j) \prod_{b\ni j\setminus a} m_{b\to j}(x_j). \tag{3.2.3}$$

yields, when a fixed point is reached, the following result for the beliefs,

$$b(x_i) = \frac{1}{Z_i}\phi_i(x_i)\prod_{a\ni i} m_{a\to i}(x_i),$$

$$b(x_a) = \frac{1}{Z_a}\psi_a(X_a)\prod_{i\in a} n_{i\to a}(x_i).$$

This turns out to be exact if the factor graph is a tree, but only approximate on multiply connected factor graphs. As mentioned before, this set of beliefs corresponds to a stationary point of a variational problem [242]. Indeed, consider the Kullback-Leibler divergence between a test joint distribution $b(\mathbf{x})$ and the reference $p(\mathbf{x})$. The Bethe approximation leads to the following functional of the beliefs, including the joint beliefs $b_a(x_a)$ corresponding to each factor:

$$\begin{aligned} D_{KL}(b\|p) &= \sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{b(\mathbf{x})}{p(\mathbf{x})} \\ &\approx \sum_{a,x_a} b_a(x_a) \log \frac{b_a(x_a)}{\psi(x_a)\prod_{i\in a} b_i(x_i)} + \sum_{i,x_i} \log \frac{b_i(x_i)}{\phi_i(x_i)} \\ &\stackrel{\text{def}}{=} F_{Bethe} = E - S_{Bethe}. \end{aligned}$$

This is equivalent to say that we look for a minimizer of $D_{KL}(b\|p)$ in the following class of joint probabilities:

$$b(\mathbf{x}) = \prod_a \frac{b_a(x_a)}{\prod_{i\in a} b_i(x_i)} \prod_i b_i(x_i), \tag{3.2.4}$$

under the constraint that

$$\sum_{x_a\setminus x_i} b_a(x_a) = b_i(x_i) \qquad \forall a\in\mathcal{F}, \forall i\in a,$$

and that

$$\sum_{\mathbf{x}\setminus x_a} b(\mathbf{x}) \approx b_a(x_a), \qquad \forall a\in\mathcal{F}, \tag{3.2.5}$$

is valid, at least approximately. For a multi-connected factor graph, the beliefs b_i and b_a are then interpreted as pseudo-marginal distribution. It is only when $\mathcal{G}$ is simply connected that these are genuine marginal probabilities of the reference distribution p.

There are a few properties of BP that are worth mentioning at this point. Firstly, BP is a fast converging algorithm:

- Two sweeps over all edges are needed if the factor-graph is a tree.
- The complexity scales heuristically like $KN\log(N)$ on a sparse factor-graph with connectivity $K \ll N$.
- It is N^2 for a complete graph.

However, when the graph is multiply connected, there is little guarantee on the convergence [169] even though in practice it works well for sufficiently sparse graphs. Another limit in this case, is that the fixed point may not correspond to a true measure, simply because (3.2.4) is not normalized and (3.2.5) is approximate. In this sense, the obtained beliefs, albeit compatible with each other are considered only as pseudo-marginals. Finally, for such graphs, the uniqueness of fixed points is not guaranteed, but it has been shown that:

- stable BP fixed points are local minima of the Bethe free energy [97];
- the converse is not necessarily true [232].

There are two important special cases, where the BP equations simplify: (i) For binary variables: $x_i \in \{0,1\}$. Upon normalization, the messages are parameterized as:

$$m_{a\to i}(x_i) = m_{a\to i}x_i + (1-m_{a\to i})(1-x_i),$$

which is stable w.r.t. the message update rule. The propagation of information reduces then to the scalar quantity $m_{a\to i}$.
(ii) For Gaussian variables, the factors are necessarily pairwise, of the form

$$\psi_{ij}(x_i,x_j) = \exp\big(-A_{ij}x_ix_j\big),$$

$$\phi_i(x_i) = \exp\big(-\frac{1}{2}A_{ii}x_i^2 + h_ix_i\big).$$

Since factors are pairwise, messages can be seen as sent directly from one variable node i to another j with a Gaussian form:

$$m_{i\to j}(x_j) = \exp\Big(-\frac{(x_j-\mu_{i\to j})^2}{2\sigma_{i\to j}}\Big).$$

This expression is also stable w.r.t. the message update rules. Information is then propagated via the 2-component real vector $(\mu_{i\to j},\sigma_{i\to j})$ with the following

update rules:

$$\mu_{i\to j} \longleftarrow \frac{1}{A_{ij}}\Big(h_i + \sum_{k\in\partial i\setminus j} \frac{\mu_{k\to i}}{\sigma_{k\to i}}\Big),$$

$$\sigma_{i\to j} \longleftarrow -\frac{1}{A_{ij}^2}\big[A_{ii} + \sum_{k\in\partial i\setminus j} \sigma_{k\to i}^{-1}\big].$$

At convergence the belief takes the form:

$$b_i(x) = \sqrt{\frac{\sigma_i}{2\pi}}\exp\big(-\frac{(x-\mu_i)^2}{2\sigma_i}\big)$$

with

$$\mu_i = \sigma_i\big(h_i + \sum_{j\in\partial i} \frac{\mu_{j\to i}}{\sigma_{j\to i}}\big) \tag{3.2.6}$$

$$\sigma_i^{-1} = A_{ii} + \sum_{j\in\partial i} \sigma_{j\to i}^{-1} \tag{3.2.7}$$

and the estimated covariance between x_i and x_j reads

$$\sigma_{ij} = \frac{1}{A_{ij}(1 - A_{ij}^2\sigma_{i\to j}\sigma_{j\to i})}.$$

In this case, there is only one fixed point even on a loopy graph, not necessarily stable, but if convergence occurs, the single variable beliefs provide the exact marginals [233]. In fact, for continuous variables, the Gaussian distribution is the only one compatible with the BP rules. Expectation propagation [167] is a way to address more general distributions in an approximate manner.

3.3 Generalized BP

In this Section we give all the necessary material concerning the relation between BP, generalized BP (GBP) and mean-field approximations in statistical physics. Further details and references can be found e.g. in [186].

In fact as observed in [131, 171], the Bethe approximation is only the first stage of a systematic entropy cumulant expansion over a poset $\{\alpha\}$ of clusters

$$S = \sum_\alpha \Delta S_\alpha,$$

where ΔS_α is the entropy correction delivered by the cluster α with respect to the entropy of all its subclusters. The decomposition is actually valid at the level of each cluster, such that with help of some Möbius inversion formulae, the corrections

$$\Delta S_\beta = \sum_{\alpha\subseteq\beta} \mu(\alpha,\beta)\ S_\alpha$$

and subsequently the full entropy can be expressed as a weighted sum

$$S = \sum_{\alpha} \kappa_\alpha S_\alpha$$

of individual clusters entropy

$$S_\alpha = -\sum_{\mathbf{x}_\alpha} b_\alpha(\mathbf{x}_\alpha) \log b_\alpha(\mathbf{x}_\alpha),$$

where $\kappa_\alpha \in \mathbb{Z}$ are a set of counting numbers. For example on the 2D square lattice, the Kikuchi approximation amounts to retain as cluster the set of nodes $v \in \mathcal{V}$, of links $\ell \in \mathcal{E}$ and of square plaquettes $c \in \mathcal{C}$ such that on a periodic lattice the corresponding approximate entropy reads

$$S = \sum_{c} S_c - \sum_{\ell} S_\ell + \sum_{v} S_v.$$

In the cluster variational method (CVM), the choice of constraints may be arbitrary, as long as the clusters hierarchy is closed under intersection.

Once identified, the connection between the Bethe approximation and BP led Yedidia et al. to propose in [242] a generalization to BP as an algorithmic counterpart to CVM. In fact they introduced a notion of region, relaxing the notion of cluster used in CVM. In their formulation, any region R containing a factor a should contain all variable nodes attached to a in order to be valid. The approximate free energy functional associated with a set of regions is given by

$$\mathcal{F}(b) = \sum_{R \in \mathcal{R}} \kappa_R \mathcal{F}_R(b_R) + \sum_{R' \subseteq R} \sum_{\mathbf{x}_{R'}} \lambda_{RR'}(\mathbf{x}_{R'}) \big(b_{R'}(\mathbf{x}_{R'}) - \sum_{\mathbf{x}_R \setminus \mathbf{x}_{R'}} b_R(\mathbf{x}_R) \big),$$

where $b_R(\mathbf{x}_R)$ and κ_R are respectively the marginal probability and counting number associated with region R. The $\lambda_{RR'}$ are again Lagrange multipliers enforcing the constraints among regions beliefs. The only constraint for the counting numbers is that for any variable i or node a

$$\sum_{R \ni i} \kappa_R = \sum_{R \ni a} \kappa_R = 1.$$

This ensures the exactness of the mean energy contribution $E(b)$ to the free energy in general as well as the entropy term for uniform distributions in particular. By comparison, there is no freedom in the CVM on the choice of the counting numbers once the set of cluster is given. Additional desirable constraints on the counting numbers are (i) the maxent-normal constraint and (ii) a global unit sum rule for counting numbers,

$$\sum_{R \in \mathcal{R}} \kappa_R = 1. \tag{3.3.1}$$

Condition (i) means that the approximate region based entropy reaches its maximum for the uniform distribution. Condition (ii) insures exactness of the entropy estimate for perfectly correlated distributions. As for belief propagation, a set of compatibility constraints among beliefs are introduced with help of Lagrange multipliers and generalized belief propagation again amounts to solving the dual problem after a suitable linear transformation of Lagrange multipliers hereby defining the messages. Once a fixed point is found a reparameterization property of the joint measure holds:

$$P(\mathbf{x}) \propto \prod_{R \in \mathcal{R}} b_R(\mathbf{x}_R)^{\kappa_R}.$$

When the region graph has no cycle, this factorization involves the true marginals probabilities of each region and is exact.

There is some degree of freedom both in the initial choice of Lagrange multipliers and messages leading to different algorithms without changing the free energy and associated variational solutions. A canonical choice is to connect regions only to their direct ancestor or direct child regions leading to the parent-to-child algorithm. With this choice the constraints are however redundant, some linear dependencies are present and this can potentially affect the convergence of the algorithm by adding unnecessary loops in the factor graph. This problem has been addressed in [183] where for a given region set a construction for a minimal factor graph is proposed.

GBP is a framework corresponding to a wide class of algorithms, which upon a good choice of regions can lead to much accurate results than basic BP. Its systematic use is however made delicate by the following unsolved issues as far as large scale inference is concerned for the marginalization problem:

- there is no automatic and efficient procedure of choosing the regions able to scale with large scale problems for non-regular factor graphs, despite proposals like the region pursuit algorithm [234] whose potential use seems however limited to small size systems.

- without special care the computational cost grows exponentially with respect to region size.

- there are difficult convergence problems associated with GBP which have led some to consider double loop algorithms [245, 98] at the price of additional computational burden.

3.4 Affinity propagation

In the large coupling limits the belief propagation can be straightforwardly adapted to the optimization context in the form of the min-sum algorithm also called belief revision [185], by simply replacing "$\sum$" by "min" (see e.g. [195]). The AP algorithm mentioned in the previous chapter performs a clustering by

identifying exemplars in a min-sum setting. It solves the following optimization problem

$$\mathbf{c}^* = \operatorname{argmin}\big(E[\mathbf{c}]\big),$$

with

$$E[\mathbf{c}] \stackrel{\text{def}}{=} -\sum_{i=1}^{N} S(i, c_i) - \sum_{\mu=1}^{N} \log \chi_\mu[\mathbf{c}] \tag{3.4.1}$$

where $\mathbf{c} = (c_1, \ldots, c_N)$ is the mapping between data and exemplars, $S(i, c_i)$ is the similarity function between i and its exemplar. For datapoints embedded in an Euclidean space, the common choice for S is the negative squared Euclidean distance. A free positive parameter is given by

$$s \stackrel{\text{def}}{=} -S(i, i), \qquad \forall i,$$

the penalty for being oneself exemplar. $\chi_\mu^{(p)}[\mathbf{c}]$ is a set of constraints. They read

$$\chi_\mu[\mathbf{c}] = \begin{cases} p, & \text{if } c_\mu \neq \mu, \exists i \text{ s.t. } c_i = \mu, \\ 1, & \text{otherwise.} \end{cases}$$

$p = 0$ is the constraint of the model of Frey-Dueck. Note that this strong constraint is well adapted to well-balanced clusters, but probably not to ring-shape ones. For this reason Leone et. al. [145, 146] have introduced the smoothing parameter p. Introducing the inverse temperature β,

$$P[\mathbf{c}] \stackrel{\text{def}}{=} \frac{1}{Z} \exp(-\beta E[\mathbf{c}])$$

represents a probability distribution over clustering assignments c. At finite β the classification problem reads

$$\mathbf{c}^* = \operatorname{argmax}\big(P[\mathbf{c}]\big).$$

The AP or SCAP equations can be obtained from the standard BP equation [70, 145] as an instance of the Max-Product algorithm. For self-containedness, let us sketch the derivation here. The BP algorithm provides an approximate procedure to the evaluation of the set of single marginal probabilities $\{P_i(c_i = \mu)\}$ while the min-sum version obtained after taking $\beta \to \infty$ yields the affinity propagation algorithm of Frey and Dueck. The factor-graph involves variable nodes $\{i, i = 1 \ldots N\}$ with corresponding variable c_i and factor nodes $\{\mu, \mu = 1 \ldots N\}$ corresponding to the energy terms and to the constraints (see Figure 3.4.1). Let $A_{\mu \to i}(c_i)$ the message sent by factor μ to variable i and $B_{i \to \mu}(c_i)$ the message sent by variable i to node μ. The belief propagation fixed point equations read:

$$A_{\mu \to i}(c_i = c) = \frac{1}{Z_{\mu \to i}} \sum_{\{c_j\}} \prod_{j \neq i} B_{j \to \mu}(c_j) \chi_\mu^\beta[\{c_j\}, c] \tag{3.4.2}$$

$$B_{i \to \mu}(c_i = c) = \frac{1}{Z_{i \to \mu}} \prod_{\nu \neq \mu} A_{\nu \to i}(c) e^{\beta S(i,c)} \tag{3.4.3}$$

In (3.4.2) we observe first that $\hat{A}_{\mu\to i} \stackrel{\text{def}}{=} A_{\mu\to i}(c_i = \nu \neq \mu)$ is independent of ν and secondly that $A_{\mu\to i}(c_i = c)$ depends only on $B_{j\to\mu}(c_j = \mu)$ and on $\sum_{\nu\neq\mu} B_{j\to\mu}(c_j = \nu)$. As a consequence, the schema can be reduced to the propagation of two quantities

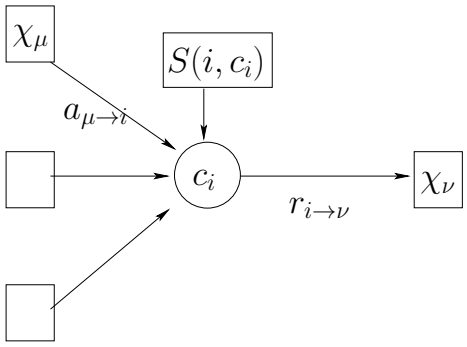


Figure 3.4.1: Factor graph corresponding to AP. Small squares represents the constraints while large ones are associated to pairwise contributions in $E(\mathbf{c})$.

$$a_{\mu\to i} \stackrel{\text{def}}{=} \frac{1}{\beta}\log\Big((N-1)\frac{A_{\mu\to i}(c_i=\mu)}{1-A_{\mu\to i}(c_i=\mu)}\Big),$$

$$r_{i\to\mu} \stackrel{\text{def}}{=} \frac{1}{\beta}\log\Big(\frac{B_{i\to\mu}(c_i=\mu)}{1-B_{i\to\mu}(c_i=\mu)}\Big),$$

called respectively the "availability" and "responsibility" messages, with $q \stackrel{\text{def}}{=} -\frac{1}{\beta}\log p$. Taking the limit $\beta\to\infty$ at fixed q yields

$$a_{\mu\to i} = \min\Big(0, \max\big(-q, \min(0, r_{\mu\to\mu})\big) + \sum_{j\neq i}\max(0, r_{j\to\mu})\Big),\ \mu\neq i, \quad (3.4.4)$$

$$a_{i\to i} = \min\Big(q, \sum_{j\neq i}\max(0, r_{j\to i})\Big), \quad (3.4.5)$$

$$r_{i\to\mu} = S(i,\mu) - \max_{\nu\neq\mu}\big(a_{\nu\to i} + S(i,\nu)\big). \quad (3.4.6)$$

After reaching a fixed point, exemplars are obtained according to

$$c_i^* = \operatorname*{argmax}_\mu\big(S(i,\mu) + a_{\mu\to i}\big) = \operatorname*{argmax}_\mu\big(r_{i\to\mu} + a_{\mu\to i}\big). \quad (3.4.7)$$

Altogether, 3.4.4,3.4.5,3.4.6 and 3.4.7 constitute the equations of SCAP which reduce to the equations of AP when q tends to $+\infty$.

Chapter 4

Inverse problems

4.1 The inverse Ising problem and mean-field methods

Finding the couplings and external fields of an Ising model is a relevant problem in many different areas. Originally considered in the context of neural networks [107] it has been since identified as a key problem - the Boltzmann machine learning problem - in statistical machine learning [104]. The huge production of biological data has led to reconsider this problem and to realize its relevance for the analysis of many biological networks [203, 12]. In the context of social networks it could as well become an important tool for analyzing data to identify influence links and trendsetters in information networks for example, or community detection. From the statistics perspective, the IIP is basically a model selection problem, in the Markov random fields (MRF) family where N binary variables are observed at least pair by pair so that a covariance matrix is given as input data. The optimal solution is then the MRF model with maximal entropy obeying moment constraints, which happens to be the Ising model with highest log-likelihood. It is a difficult problem, where both the graph structure and the values of the fields and couplings have to be found.

Existing approaches fall mainly in the following categories:

- Purely computational efficient approaches rely on various optimization schemes of the log likelihood [144] or on pseudo-likelihood [106] along with sparsity constraints to select the only relevant features.
- Common analytical approaches are based on the Plefka expansion [189] of the Gibbs free-energy by making the assumption that the coupling constants J_{ij} are small. The picture is then of a weakly correlated unimodal probability measure. For example, approaches used in [34, 170] are based on this assumption.
- Another possibility is to assume that relevant coupling J_{ij} have locally a tree like structure. The Bethe approximation [242] can then be used with possibly loop corrections. Again this corresponds to having a weakly correlated unimodal probability measure and these kinds of approaches are referred to as pseudo-moment matching methods in the literature. For example the approaches proposed in [121, 236, 162, 239] are based on this assumption.
- In the case where a multimodal distribution is expected, then a model with many attraction basins is to be found and Hopfield-like models [107, 35] are likely to be more relevant. To be mentioned also is a recent mean-field methods [177] which allows one to find in some simple cases the Ising couplings of a low temperature model, i.e. displaying multiple probabilistic modes.

Consider an Ising model, i.e. a MRF of binary variables $\{s_i \in \{-1, 1\}, i \in \mathcal{V}\}$, where $\mathcal{V}$ is a set of vertices of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{E}$ a set of edges corresponding to interactions between variables (s_i, s_j), associated to some coupling

$J_{ij} \in \mathbb{R}$. We assume that from a set of historical observations, the empirical mean $\hat{m}_i$ [resp. covariance $\hat{\chi}_{ij}$] is given for each variable s_i [resp. each pair of variable (s_i, s_j)]. In this case, from Jayne's maximum entropy principle [116], imposing these moments to the joint distribution leads to a model pertaining to the exponential family, the Ising model in the present case:

$$\mathcal{P}(\mathbf{s}) = \frac{1}{Z[\mathbf{J}, \mathbf{h}]} \exp\Big(\sum_i h_i s_i + \sum_{i,j} J_{ij} s_i s_j\Big) \tag{4.1.1}$$

where the external fields $\mathbf{h} = \{h_i\}$ and the coupling constants $\mathbf{J} = \{J_{ij}\}$ are the Lagrange multipliers associated respectively to mean and covariance constraints when maximizing the entropy of $\mathcal{P}$. They are obtained as minimizers of the dual optimization problem:

$$(\mathbf{h}^\star, \mathbf{J}^\star) = \underset{(\mathbf{h},\mathbf{J})}{\operatorname{argmin}}\, \mathcal{L}[\mathbf{h}, \mathbf{J}], \tag{4.1.2}$$

where

$$\mathcal{L}[\mathbf{h}, \mathbf{J}] \stackrel{\text{def}}{=} \log Z[\mathbf{h}, \mathbf{J}] - \sum_i h_i \hat{m}_i - \sum_{ij} J_{ij} \hat{m}_{ij} \tag{4.1.3}$$

is the log likelihood. This leads to invert the linear response equations:

$$\frac{\partial \log Z}{\partial h_i}[\mathbf{h}, \mathbf{J}] = \hat{m}_i \tag{4.1.4}$$

$$\frac{\partial \log Z}{\partial J_{ij}}[\mathbf{h}, \mathbf{J}] = \hat{m}_{ij}, \tag{4.1.5}$$

$\hat{m}_{ij} = \hat{m}_i \hat{m}_j + \hat{\chi}_{ij}$ being the empirical expectation of $s_i s_j$. As noted e.g. in [34], the solution is minimizing the cross entropy, a Kullback-Leibler distance between the empirical distribution $\hat{P}$ based on historical data and the Ising model:

$$D_{KL}[\hat{\mathcal{P}} \| \mathcal{P}] = \log Z[\mathbf{h}, \mathbf{J}] - \sum_i h_i \hat{m}_i - \sum_{i<j} J_{ij} \hat{m}_{ij} - S(\hat{\mathcal{P}}). \tag{4.1.6}$$

The set of Equations (4.1.4,4.1.5) cannot be solved exactly in general because the computational cost of Z is exponential. Approximations resorting to various mean-field methods can be used to evaluate $Z[\mathbf{h}, \mathbf{J}]$.

Plefka's expansion To simplify the problem, it is customary to make use of the Gibbs free-energy, i.e. the Legendre transform of the free-energy, to impose the individual expectations $\mathbf{m} = \{\hat{m}_i\}$ for each variable:

$$G[\mathbf{m}, \mathbf{J}] = \mathbf{h}^T(\mathbf{m})\mathbf{m} + F[\mathbf{h}(\mathbf{m}), \mathbf{J}],$$

(with $F[\mathbf{h}, \mathbf{J}] \stackrel{\text{def}}{=} -\log Z[\mathbf{h}, \mathbf{J}]$, $\mathbf{h}^T\mathbf{m}$ is the ordinary scalar product) where $\mathbf{h}(\mathbf{m})$ depends implicitly on $\mathbf{m}$ through the set of constraints

$$\frac{\partial F}{\partial h_i} = -m_i. \tag{4.1.7}$$

Note that by duality we have

$$\frac{\partial G}{\partial m_i} = h_i(\mathbf{m}), \tag{4.1.8}$$

and

$$\Big[\frac{\partial^2 G}{\partial m_i \partial m_j}\Big] = \Big[\frac{d\mathbf{h}}{d\mathbf{m}}\Big]_{ij} = \Big[\frac{d\mathbf{m}}{d\mathbf{h}}\Big]^{-1}_{ij} = -\Big[\frac{\partial^2 F}{\partial h_i \partial h_j}\Big]^{-1} = \big[\chi^{-1}\big]_{ij}, \tag{4.1.9}$$

i.e. the inverse susceptibility matrix. Finding a set of J_{ij} satisfying this last relation along with (4.1.8) yields a solution to the inverse Ising problem since the m's and χ's are given. A way to connect the couplings directly with the covariance matrix is also given by the relation

$$\frac{\partial G}{\partial J_{ij}} = -m_{ij}. \tag{4.1.10}$$

The Plefka expansion is used to expand the Gibbs free-energy in power of the coupling J_{ij} assumed to be small. Multiplying all coupling J_{ij} by some parameter $\alpha \in \mathbb{R}$ yields the following cluster expansion:

$$G[\mathbf{m}, \alpha\mathbf{J}] = \mathbf{h}^T(\mathbf{m}, \alpha)\mathbf{m} + F[\mathbf{h}(\mathbf{m}, \alpha), \alpha\mathbf{J}] \tag{4.1.11}$$

$$= G_0[\mathbf{m}] + \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} G_n[\mathbf{m}, \mathbf{J}] \tag{4.1.12}$$

where each term G_n corresponds to cluster contributions of size n in the number of links J_{ij} involved, and $\mathbf{h}(\mathbf{m}, \alpha)$ depends implicitly on α in order to always fulfill (4.1.7). This is the Plefka expansion, and each term of the expansion (4.1.12) can be obtained by successive derivation of (4.1.11). We have

$$G_0[\mathbf{m}] = \sum_i \frac{1+m_i}{2} \log \frac{1+m_i}{2} + \frac{1-m_i}{2} \log \frac{1-m_i}{2}.$$

Letting

$$H_J \overset{\text{def}}{=} \sum_{i<j} J_{ij} s_i s_j,$$

considered as a small perturbation and using (4.1.7), the two first derivatives of (4.1.11) w.r.t α read

$$\frac{dG[\mathbf{m}, \alpha\mathbf{J}]}{d\alpha} = -\mathbb{E}_\alpha\big(H_J\big), \tag{4.1.13}$$

$$\frac{d^2 G[\mathbf{m}, \alpha\mathbf{J}]}{d\alpha^2} = -\operatorname{Var}_\alpha\big(H_J\big) - \sum_i \frac{dh_i(\mathbf{m}, \alpha)}{d\alpha} \operatorname{Cov}_\alpha\big(H_J, s_i\big), \tag{4.1.14}$$

where subscript α indicates that expectations, variance and covariance are taken at given α. To get successive derivatives of $\mathbf{h}(\mathbf{m}, \alpha)$ one can use (4.1.8). Another possibility is to express the fact that $\mathbf{m}$ is fixed,

$$\begin{aligned}\frac{dm_i}{d\alpha} = 0 &= -\frac{d}{d\alpha}\frac{\partial F[\mathbf{h}(\alpha), \alpha\mathbf{J}]}{\partial h_i} \\ &= \sum_j h_j'(\alpha)\,\mathrm{Cov}_\alpha(s_i, s_j) + \mathrm{Cov}_\alpha(H_J, s_i),\end{aligned}$$

giving

$$h_i'(\alpha) = -\sum_j [\chi_\alpha^{-1}]_{ij}\,\mathrm{Cov}_\alpha(H_J, s_j), \tag{4.1.15}$$

where χ_α denotes the susceptibility delivered by the model when $\alpha \neq 0$. To get the first two terms in the Plefka expansion, we need to compute these quantities at $\alpha = 0$:

$$\begin{aligned}\mathrm{Var}\big(H_J\big) &= \sum_{i<k,j} J_{ij}J_{jk}m_i m_k(1-m_j^2) + \sum_{i<j} J_{ij}^2(1-m_i^2m_j^2), \\ \mathrm{Cov}\big(H_J, s_i\big) &= \sum_j J_{ij}m_j(1-m_i^2), \\ h_i'(0) &= -\sum_j J_{ij}m_j, \\ [\chi_0^{-1}]_{ij} &= (1-m_i^2)^{-1}\delta_{ij}\end{aligned}$$

(by convention $J_{ii} = 0$ in these sums). The first and second orders then finally read:

$$G_1[\mathbf{m}, \mathbf{J}] = -\sum_{i<j} J_{ij}m_i m_j, \qquad G_2[\mathbf{m}, \mathbf{J}] = -\sum_{i<j} J_{ij}^2(1-m_i^2)(1-m_j^2),$$

and correspond respectively to the mean-field and to the TAP approximation. Higher order terms have been computed in [89].

At this point finding an approximate solution to the inverse Ising problem can be done, either by inverting Equation (4.1.9) or (4.1.10). To get a solution at a given order n in the couplings, solving (4.1.10) requires G at order $n+1$, while it is needed at order n in (4.1.9).

Taking the expression of G up to second order gives

$$\frac{\partial G}{\partial J_{ij}} = -m_i m_j - J_{ij}(1-m_i^2)(1-m_j^2),$$

and (4.1.10) leads directly to the basic mean-field solution:

$$J_{ij}^{MF} = \frac{\hat{\chi}_{ij}}{(1-\hat{m}_i^2)(1-\hat{m}_j^2)}. \tag{4.1.16}$$

At this level of approximation for G, using (4.1.8) we also have

$$h_i = \frac{1}{2}\log\frac{1+m_i}{1-m_i} - \sum_j J_{ij}m_j + \sum_j J_{ij}^2 m_i(1-m_j^2)$$

which corresponds precisely to the TAP equations. Using now (4.1.9) gives

$$\frac{\partial h_i}{\partial m_j} = [\chi^{-1}]_{ij} = \delta_{ij}\big(\frac{1}{1-m_i^2} + \sum_k J_{ik}^2(1-m_k^2)\big) - J_{ij} - 2J_{ij}^2 m_i m_j. \quad (4.1.17)$$

Ignoring the diagonal terms, the TAP solution is conveniently expressed in terms of the inverse empirical susceptibility,

$$J_{ij}^{TAP} = -\frac{2[\hat{\chi}^{-1}]_{ij}}{1+\sqrt{1-8\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}}, \quad (4.1.18)$$

where the branch corresponding to a vanishing coupling in the limit of small correlation i.e. small $\hat{\chi}_{ij}$ and $[\hat{\chi}^{-1}]_{ij}$ for $i \neq j$, has been chosen.

Bethe approximate solution When the graph formed by the pairs (i,j), for which the correlations $\hat{\chi}_{ij}$ are given by some observations is a tree, the following form of the joint probability corresponding to the Bethe approximation:

$$\mathcal{P}(\mathbf{s}) = \prod_{i<j}\frac{\hat{p}_{ij}(s_i,s_j)}{\hat{p}(s_i)\hat{p}(s_j)}\prod_i \hat{p}_i(s_i), \quad (4.1.19)$$

yields actually an exact solution to the inverse problem (4.1.2), where the $\hat{p}$ are the single and pair variables empirical marginals given by the observations. Using the following parametrization of the $\hat{p}$'s

$$\hat{p}_i^x \stackrel{\text{def}}{=} \hat{p}\big(\frac{1+s_i}{2} = x\big) = \frac{1}{2}(1+\hat{m}_i(2x-1)), \quad (4.1.20)$$

$$\hat{p}_{ij}^{xy} \stackrel{\text{def}}{=} \hat{p}\big(\frac{1+s_i}{2} = x, \frac{1+s_j}{2} = y\big)$$

$$= \frac{1}{4}\big(1+\hat{m}_i(2x-1)+\hat{m}_j(2y-1)+\hat{m}_{ij}(2x-1)(2y-1)\big) \quad (4.1.21)$$

relating the empirical frequency statistics to the empirical "magnetizations" $m \equiv \hat{m}$, we obtain the mapping onto an Ising model (4.1.1) with

$$h_i = \frac{1-d_i}{2}\log\frac{\hat{p}_i^1}{\hat{p}_i^0} + \frac{1}{4}\sum_{j\in\partial i}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{10}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{00}}\Big), \qquad J_{ij} = \frac{1}{4}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{00}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}}\Big),\ \forall\,(i,j)\in\mathcal{E}, \quad (4.1.22)$$

where d_i is the number of neighbors of i, using the notation $j \in \partial i$ for "j neighbor of i". The partition function is then explicitly given by

$$Z_{Bethe}[\hat{p}] = \exp\Big[-\frac{1}{4}\sum_{(i,j)\in\mathcal{E}}\log\big(\hat{p}_{ij}^{00}\ \hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}\ \hat{p}_{ij}^{11}\big) - \sum_i \frac{1-d_i}{2}\log\big(\hat{p}_i^0\ \hat{p}_i^1\big)\Big]. \quad (4.1.23)$$

The corresponding Gibbs free-energy can thus be written explicitly using (4.1.22,4.1.23). With fixed magnetizations m_i's, and given a set of couplings $\{J_{ij}\}$, the parameters $m_{ij} = m_{ij}(m_i, m_j, J_{ij})$ are implicit function obtained by inverting the relations (4.1.22). For the linear response, we get [236] from (4.1.22):

$$\frac{\partial h_i}{\partial m_j} = \Big[\frac{1-d_i}{1-m_i^2} + \frac{1}{16}\sum_{k\in\partial i}\Big(\big(\frac{1}{\hat{p}_{ik}^{11}} + \frac{1}{\hat{p}_{ik}^{01}}\big)\big(1+\frac{\partial m_{ik}}{\partial m_i}\big) + \big(\frac{1}{\hat{p}_{ik}^{00}} + \frac{1}{\hat{p}_{ik}^{10}}\big)\big(1-\frac{\partial m_{ik}}{\partial m_i}\big)\Big)\Big]\delta_{ij}$$
$$+ \frac{1}{16}\Big(\big(\frac{1}{\hat{p}_{ij}^{11}} + \frac{1}{\hat{p}_{ij}^{10}}\big)\big(1+\frac{\partial m_{ij}}{\partial m_i}\big) + \big(\frac{1}{\hat{p}_{ij}^{00}} + \frac{1}{\hat{p}_{ij}^{01}}\big)\big(1-\frac{\partial m_{ij}}{\partial m_i}\big)\Big)\Big]\delta_{j\in\partial i}.$$

Using (4.1.22), we can also express

$$\frac{\partial m_{ij}}{\partial m_i} = -\frac{\frac{1}{\hat{p}_{ij}^{11}} + \frac{1}{\hat{p}_{ij}^{01}} - \frac{1}{\hat{p}_{ij}^{10}} - \frac{1}{\hat{p}_{ij}^{00}}}{\frac{1}{\hat{p}_{ij}^{11}} + \frac{1}{\hat{p}_{ij}^{01}} + \frac{1}{\hat{p}_{ij}^{10}} + \frac{1}{\hat{p}_{ij}^{00}}},$$

giving eventually [176]

$$[\hat{\chi}^{-1}]_{ij} = \Big[\frac{1-d_i}{1-m_i^2} + \sum_{k\in\partial i}\frac{1-m_k^2}{(1-m_i^2)(1-m_k^2)-\hat{\chi}_{ik}^2}\Big]\delta_{ij} - \frac{\hat{\chi}_{ij}}{(1-m_i^2)(1-m_j^2)-\hat{\chi}_{ij}^2}\,\delta_{j\in\partial i}. \tag{4.1.24}$$

The existence of an exact solution can therefore be checked directly as a self-consistency property of the input data $\hat{\chi}_{ij}$, for a given pair (i,j) either:

- $[\hat{\chi}^{-1}]_{ij} \neq 0$, then this self-consistency relation (4.1.24) has to hold and J_{ij} is given by (4.1.22) using $\hat{m}_{ij} = \hat{m}_i\hat{m}_j + \hat{\chi}_{ij}$.
- $[\hat{\chi}^{-1}]_{ij} = 0$ then $J_{ij} = 0$ but $\hat{\chi}_{ij}$, which can be non vanishing, is obtained by inverting $[\hat{\chi}^{-1}]$ defined by (4.1.24).

Finally, complete consistency of the solution is checked on the diagonal elements in (4.1.24). If full consistency is not verified, this equation can nevertheless be used to find approximate solutions. Remark that, if we restrict the set of Equations (4.1.24), e.g. by some thresholding procedure, in such a way that the corresponding graph is a spanning tree, then, by construction, $\chi_{ij} \equiv \hat{\chi}_{ij}$ will be solution on this restricted set of edges, simply because the BP equations are exact on a tree. The various methods proposed for example in [162, 241] actually correspond to different heuristics for finding approximate solutions to this set of constraints. As noted in [176], a direct way to proceed is to eliminate χ_{ij} in the equations obtained from (4.1.22) and (4.1.24):

$$\chi_{ij}^2 + 2\chi_{ij}(m_i m_j - \coth(2J_{ij})) + (1-m_i^2)(1-m_j^2) = 0, \tag{4.1.25}$$

$$\chi_{ij}^2 - \frac{\chi_{ij}}{[\chi^{-1}]_{ij}} - (1-m_i^2)(1-m_j^2) = 0. \tag{4.1.26}$$

This leads then to $BA + LR$ estimate of the couplings

$$J_{ij}^{BA+LR} = -\frac{1}{2}\,\mathrm{atanh}\Big(\frac{2[\hat{\chi}^{-1}]_{ij}}{\sqrt{1+4(1-\hat{m}_i^2)(1-\hat{m}_j^2)[\hat{\chi}^{-1}]_{ij}^2} - 2\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}\Big). \tag{4.1.27}$$

Note that J_{ij}^{BA+LR} and J_{ij}^{TAP} coincide at second order in $[\hat{\chi}^{-1}]_{ij}$.

4.2 Inverse covariance matrix estimation

Another case of interest for inverse problems is the one dealing with multivariate Gaussian distributions, i.e. Gaussian Markov random fields (GMRF). The GMRF distribution is naturally characterized by a mean vector $\boldsymbol{\mu} \in \mathbb{R}$ and a positive definite precision (or concentration) matrix $\mathbf{A}$, which is simply the inverse of the covariance matrix $\mathbf{C}$. Zero entries in the precision matrix $\mathbf{A}$ indicate conditionally independent pairs of variables. This gives a graphical representation of dependencies: two random variables are conditionally independent if, and only if, there is no direct edge between them. Observations are summarized in an empirical covariance matrix $\hat{\mathbf{C}} \in \mathbb{R}^{N\times N}$ of a random vector $\mathbf{X} = (X_i)_{i\in\{1,\dots,N\}}$, and we look for a GMRF model with sparse precision matrix $\mathbf{A}$. The model estimation problem can be expressed as the maximization of the log-likelihood:

$$\mathbf{A} = \underset{\mathbf{M}\in\mathcal{S}_{++}}{\mathrm{argmax}}\ \mathcal{L}(\mathbf{M}), \quad \mathcal{L}(\mathbf{M}) \overset{\text{def}}{=} \log\det(\mathbf{M}) - \mathrm{Tr}(\mathbf{M}\hat{\mathbf{C}}),$$

where $\mathcal{S}_{++}$ formally represents the set of positive definite matrices.

Without any constraint on $\mathbf{M}$, the maximum likelihood estimate is trivially $\mathbf{A} = \hat{\mathbf{C}}^{-1}$. However, enforcing sparsity with simple thresholding of small magnitude entries may easily ruin the positive definiteness of the estimated precision matrix. In the context of structure learning, where meaningful interactions have to be determined, for instance among genes in genetic networks, the maximization is classically performed on the set of positive definite matrices, after adding to the log-likelihood a continuous penalty function P that imitates the L_0 norm. The Lasso penalty, a convex relaxation of the problem, uses the L_1 norm, measuring the amplitudes of off-diagonal entries in $\mathbf{A}$ [71, 108]. Various optimization schemes have been proposed to solve it efficiently [13, 71]. However, the L_1 norm penalty suffers from a modeling bias, due to excessive penalization of truly large magnitudes entries of $\mathbf{A}$. To overcome this issue, concave functions, that perform constant penalization to the large magnitudes, have been proposed [61, 143].

Chapter 5

Road traffic modelling applications

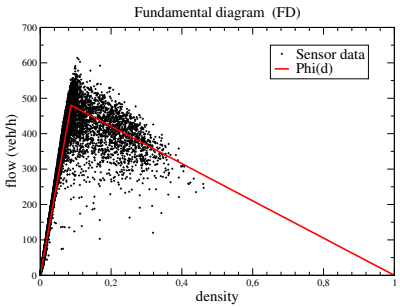


Figure 5.1.1: Fundamental diagram used in the LWR model (in red) fitting observations from sensors.

5.1 Various traffic models

Hydrodynamical models Let $q(x,t) \stackrel{\text{def}}{=} \rho(x,t)V(x,t)$ represent the flow of vehicles, where $\rho(x,t)$ and $V(x,t)$ are respectively the density and speed measured at position x and time t. A fundamental hypothesis made commonly in traffic modelling is that $q(x,t)$ is a function of the density $\rho(x)$ solely [149, 193]. This is the so-called fundamental diagram (FD) relationship, which leads to the following Lighthill-Whitham-Richards (LWR) first order model of traffic

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial q[\rho(x,t)]}{\partial x} = 0, \tag{5.1.1}$$

when expressing the conservation on vehicles along the segment. Steady state solutions of such equations can exist in the form of traveling waves $\rho(x,t) = \rho(x-ct)$ with

$$c = \frac{dq}{d\rho},$$

defined piecewise along characteristic curves $x(t)$ corresponding to $x'(t) = c$, where q is constant. This allows the apparition of shock waves, which obviously is a feature of traffic jam. However this model has many shortcomings as many traffic features are absent like platoon diffusion or the instability associated to heavy traffic and associated oscillatory phenomena. In order to cure this many continuous model based on kinetic equations have been proposed, most of them falling in the category of second-order models of the following form where in addition to equation 5.1.1 the speed is obeying a second equation:

$$\frac{\partial V(x,t)}{\partial t} + V(x,t)\frac{\partial V(x,t)}{\partial x} = -\nu \frac{1}{\rho(x,t)}\frac{\partial \rho(x,t)}{\partial x} + \frac{1}{\tau}\big(V_e[\rho(x,t)] - V(x,t)\big)$$

where ν is a viscosity coefficient which can depend on the density, τ a relaxation time to reach the "equilibrium" speed $V_e[\rho]$ for a given density ρ. Some controversy [96] has taken place in the past about un-physical effects of such models [39] and their possible resolution [11].

Microscopic models These equations can be seen in some cases as the continuous limits of so-called *car following* (see e.g. [26]) or *follow the leader* models [10] where the motion of single vehicles obey some kinematic rule intended to take into account distance and relative speed to the next vehicle.

A new family of models appeared in the mid 90's implementing cellular automaton (CA) rules [173] which were making it possible to simulate efficiently the traffic at the level of a conurbation. Let us give the rules of the original Nagel-Schreckenberg model which are extremely simple and similar to exclusion processes presented in Chapter 1. Space and time are discretized, on each site there is at most one vehicle present carrying a discrete speed label v, taking integer values between 0 and v_{max}. At each time step first the speed of each vehicle is updated as follows (parallel dynamics):

- Acceleration: for $v < v_{max}$ if the headway Δ is larger than $v+1$ then the speed is increased by one unit.
- Braking: if $\Delta \leq v$ then the speed is updated to $v = \Delta - 1$.
- Randomization: if positive, v is decreased by one unit with probability p.

After that each vehicle advances v sites. While very simple, this model exhibits the property that some spontaneous symmetry breaking among identical vehicles may occur (phantom jams), as can be seen experimentally on a ring geometry for example [216].

In the validation of traffic models [205], among many observables like for instance headway distributions [9], properties of the FD plays a central role. The three phases traffic theory of Kerner [130], states that the traffic phase diagram on highways should consist of three different kind of flows: the free flow, the synchronized flow and the wide moving jam. In the free flow regime, at low density, the flow is simply proportional to the density of cars; in the congested one, at large density, massive clusters of cars are present, and the flow decreases more or less linearly with this density; in the intermediate regime, the relation between flow and density is largely of stochastic nature, due to the presence of a large amount of small clusters of cars propagating at various random speeds. It is not clear however whether in this picture these phases, and especially the synchronized flow phase, are genuine dynamical or thermodynamical phases, meaningful in some large size limit in the stationary regime, or are intricate transient features of a slowly relaxing system. In fact the nature of the synchronized flow has been subject to controversy, of whether it should be considered as a phase on its own or only as fluctuations within the congested phase [204].

A mechanism present at the microscopic level suspected to be responsible for the properties of the FD is the fact that vehicles in the traffic may accelerate or brake in a dissymmetrical way. This should cause in particular spontaneous congestion to occurs. This mechanism is referred to as the *slow-to-start* mechanism, which is implicitly present in the Nagel-Schreckenberg model since the speed can decrease by an arbitrary amount while it can increase only by one unit. In its refined versions like the velocity dependant randomization (VDR)

model [14] the mechanism is explicitly introduced into the CA rules. VDR exhibits in particular a first order phase transition between the fluid and the congestion phase and some hysteresis phenomena [22] associated to metastable states.

5.2 Traffic inference

A very different approach to traffic is based on statistical models rather physical ones. Flow models to be made operational need a high level of details which translate into an enormous amount of parameters to be calibrated. The calibration of those parameters is so challenging and time consuming that the effectiveness of flow models can decrease sharply with system size. In this respect, data driven models can be an efficient alternative. In addition the availability of traffic data has drastically increased. Data-driven models can be divided into two main categories.

- *parametric models*: vector auto-regressive models (VAR), ARIMA, STARIMA, probabilistic models, Bayesian models, MRF-based models;
- *non-parametric models*: k-NN, random forest, Gaussian process, support vector regression, neural networks.

Parametric models, based on ordinary statistical considerations, are more traditional. They are sometimes preferred to their non-parametric counterparts owing to their interpretability. Machine learning is potentially offering a very large variety of non-parametric models with a wide range of complexity and potential efficiency. General references on these various approaches can be found in [229]. A lot of methods are targeted toward independent segment modeling. Methods trying to leverage spatial dependencies are less numerous but many have appeared recently [57]. If we focus more specifically on forecasting models which attempt to address the problem at the network scale, the requirements we can think of for such models to be ideally deployed in online applications are the following

- *accuracy*: predictions should be significantly better than a simple persistent predictor combined with historical day-time dependent average for instance, in use when data are incomplete.
- *missing data*: we cannot expect to have at any time a complete information of the network state, which means that both the learning and the running of the model have to be able to be done in a setting with missing data.
- *scaling*: the model should scale up to high systems size, i.e. networks of the size of a conurbation, where number of road segments to be tackled can be around one or two hundreds thousands. Actually, if we think in terms of detectors, this requirement might be lower. At the moment, the number of effective detectors covering a given urban area is smaller by one or two orders of magnitude than the number of road segments.

Traditional methods based on autoregressive models [151] have been adapted recently to this context, e.g. for treating floating car data (FCD) at small scale (120 points location in central Rome) [81]. While yielding a good level of interpretability, this type of methods do not seem suitable to scale up to large network sizes. In order to capture local spatial features of traffic patterns, several studies (see e.g. [217]) proposed hybrid machine learning methods involving neural network and L1 regularization of the weight matrix connecting the input to the hidden layer. They remained however limited to scale of the order of a few hundred of detectors. More recently, deep learning approaches have been proposed: in [152], a stacked auto-encoder is trained layer-wise on highway data at a coarse grain level, by considering the aggregation of traffic flow along each freeway direction. In order to address forecasting at a more detailed level, graph convolutional neural networks – a generalization of convolutional neural networks to graph structured data – have been proposed [244, 147] with various specifications and combinations with other RNN architectures like LSTM, in order to encode the temporal dynamics of spatial features extracted by the GCNN. Most of them show convincing performance improvement over traditional methods, though often demonstrated on small scale problems involving again a few hundreds of variables, presumably due to the heavy computationally training procedure [190].

Another limit of such methods, aside from computational resources that are needed for training and the seemingly limited network scale of application, is the assumption that the data are complete. Missing values have to be imputed beforehand in a way or another in order to train the model and to use it [55].

Part II

Particle systems and traffic modelling

Chapter 6

From fluctuating planar paths to exlusion processes

This chapter is based on the following papers:

G. Fayolle and C. Furtlehner, Dynamical Windings of Random Walks and Exclusion Models. Part I: Thermodynamic Limit in Z^2. *J.Stat.Phys. 114*, 1-2 (2004), 229-260.

G. Fayolle and C. Furtlehner, Stochastic Dynamics of Discrete Curves and Multi-Type Exclusion Processes.*J.Stat.Phys. 127*, 5 (2007), 1049-1094.

G. Fayolle and C. Furtlehner, Stochastic deformations of sample paths of random walks and exclusion models. *In Proc. of 3rd Colloquium of Mathematics and Computer Science. Mathematics and Computer Science III: Algorithms, Trees, Combinatorics and Probabilities, Birkhäuser, Basel* (2004) 415-428.

6.1 Steady states and continuous limits of a fluctuating planar path

Simple processes can sometimes present a rich phenomenology. Let us illustrate this with the following stochastic process [62] which will give us some insight on various tools and techniques of subsequent use in this document. We consider the stochastic evolution of a planar oriented path $\mathbf{C}_N$ of length N, supported by a square lattice and subject to local transformations. The path is encoded as a N-sequence of four letters A, B, C and D. We introduce the notation $\{(A_i, B_i, C_i, D_i), i = 1, \dots N\}$ to represent the path, where A_i, B_i, C_i and D_i are all Boolean representing the presence of a link of the corresponding type at position i, with the exclusion constraint $A_i + B_i + C_i + D_i = 1$. Once the initial configuration (supposedly random) is given, the system evolves according to the four local pattern transformations depicted in Figure 6.1.1. Only a single point of the path can be moved at a time, with the constraint that no link is broken (i.e. the path remains always connected). Geometrically, to each point of the path is

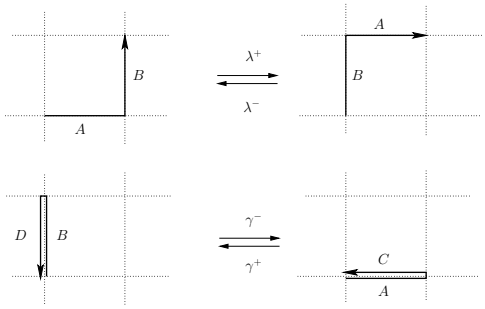


Figure 6.1.1: Pattern transition rates.

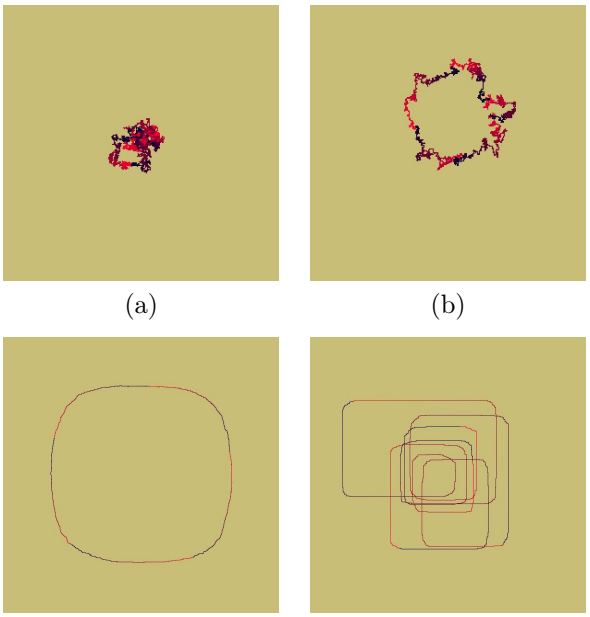
associated a pattern $M \in \{M1, M2, M3, M4\}$ with $M1 \in \{AB, BC, CD, DA\}$ (left bend), $M3 \in \{BA, CB, DC, AD\}$ (right bend), $M2 \in \{AC, BD, CA, DB\}$ (vertical or horizontal fold) and $M4 \in \{AA, BB, CC, DD\}$ (straight $\rightarrow\rightarrow$), and the following local transitions can occur:

$$\begin{cases} M1 \to M3, \text{ with rate } \lambda^+, \\ M3 \to M1, \text{ with rate } \lambda^-, \\ \text{rotation of } M2 \text{ of angle } \pm\frac{\pi}{2}, \text{ with rate } \gamma^{\pm}. \end{cases}$$

With exponentially distributed jump times, these events generate a global Markovian continuous time evolution of the system. Interesting things happen when we break the chiral symmetry by imposing a *detuning* between λ^+ and λ^-, measured by the scaling parameter

$$\eta = N\frac{\lambda^+ - \lambda^-}{\lambda^+ + \lambda^-} = \mathcal{O}(1). \tag{6.1.1}$$

For closed walks, four different situations can be observed (see Figure 6.1.2). The study of this process proceed first by observing that it can be decomposed



(a) (b) (c) (d)

Figure 6.1.2: Pictures of a path of $N = 5000$ steps, for different values of η. Each colored segment represents 1000 steps. (a) $\eta = 0$ (scale=1). (b) $\eta = 5$ (scale=1). (c) $\eta = 12.5$ (scale $= 1/6$). (d) $\eta = 250$ (scale $= 1/2$).

into two coupled exclusion processes (τ^a, τ^b) (see Chapter 1) with

$$\tau_i^a = B_i + C_i \qquad \text{and} \qquad \tau_i^b = C_i + D_i.$$

Indeed each transition corresponds to a move of either a particle of type (a) or a of type (b) (never both at the same time). Transition rates $\lambda_{a,b}^{\pm}(i)$ for a particle of either type to move forward or backward depend locally on the presence of particle of the other type. In the particular case $\gamma^{\pm} = \lambda^{\pm}$, we get simple expressions:

$$\begin{cases} \lambda_a^{\pm}(i) = \lambda \pm (2s_i^b - 1)\mu, \\ \lambda_b^{\pm}(i) = \lambda \mp (2s_i^a - 1)\mu, \end{cases} \tag{6.1.2}$$

with

$$\lambda = \frac{1}{2}(\lambda^+ + \lambda^-) \qquad \text{and} \qquad \frac{1}{2}(\lambda^+ - \lambda^-).$$

The condition for the process to be reversible reads simply

$$N_a = N_b = N/2,$$

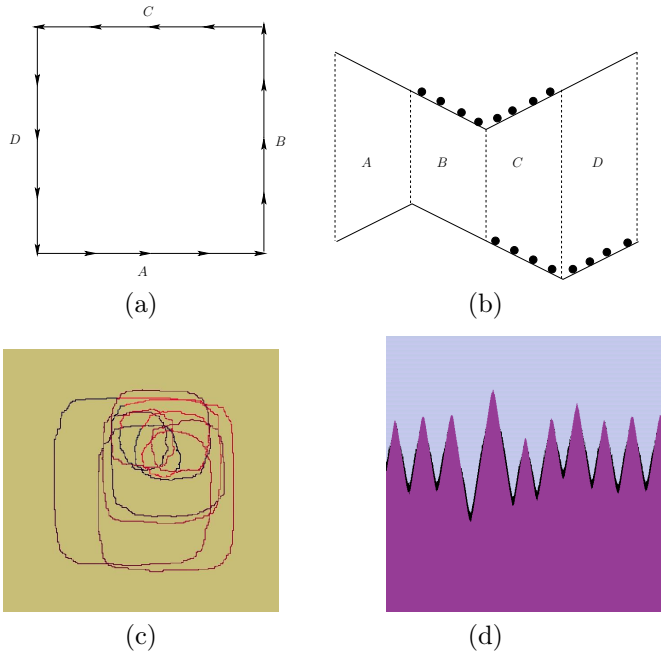


Figure 6.1.3: (a) Stable configuration for closed paths. (b) Corresponding representation in terms of exclusion processes. (c) Metastable state with $N = 5000$. (d) Corresponding KPZ landscape with the density of trapped particles appearing in black.

which is interpreted geometrically as having closed paths. In that case the invariant measure reads

$$P(\tau^a, \tau^b) = \frac{1}{Z} \exp\Big[\beta \sum_{i<j} (\tau_i^a \bar{\tau}_j^b - \tau_i^b \bar{\tau}_j^a)\Big] \delta\Big(\sum_i \tau_i^a - N_a\Big) \delta\Big(\sum_i \tau_i^b - N_b\Big), \quad (6.1.3)$$

using the Boolean notation $\bar{\tau} = 1 - \tau$, with $\beta = \log \frac{\lambda - \mu}{\lambda + \mu}$ and where the constraint on the conservation of the number of particles of both types is explicitly written. Interestingly we get here a bipartite Boltzmann measure between type (a) and type (b) particles, similar in form with the restricted Boltzmann machines (see Chapter 2), except for the constraints on the number of particles. Consider

$$q_i^a = P(\tau_i^a = 1 | \tau^b) \qquad \text{and} \qquad q_i^b = P(\tau_i^b = 1 | \tau^a)$$

the conditional marginal probabilities for each species, conditionally to the other one. In thermodynamic limit, we let $N \to \infty$ with fixed η and have the existence of the weak limits (with $x = i/N$)

$$\rho^a(x) = \lim_{N \to \infty} q_{Nx}^a \qquad \text{and} \qquad \rho^b(x) = \lim_{N \to \infty} q_{Nx}^b$$

thereby defining continuous particle densities. These happen to obey deterministic Lotka-Volterra type of equations:

$$\begin{cases} \dfrac{\partial \rho^a(x)}{\partial x} = 4\eta\rho^a(x)(1-\rho^a(x))\Big(\rho^b(x)-\dfrac{1}{2}\Big), \\ \dfrac{\partial \rho^b(x)}{\partial x} = -4\eta\rho^b(x)(1-\rho^b(x))\Big(\rho^a(x)-\dfrac{1}{2}\Big), \end{cases} \tag{6.1.4}$$

with constraints $\int_0^1 \rho^a(x)dx = \int_0^1 \rho^b(x)dx = \frac{1}{2}$ and $\rho^{a,b}(x+1) = \rho^{a,b}(x)$. They represent the condition to minimize the large deviation free energy functional

$$\begin{aligned} \mathcal{F}(\rho^a,\rho^b) = -\int_0^1 dx \Big[& \rho^a(x)\log\big(\rho^a(x)\big) + \big(1-\rho^a(x)\big)\log\big(1-\rho^a(x)\big) \\ & + \rho^b(x)\log\big(\rho^b(x)\big) + \big(1-\rho^b(x)\big)\log\big(1-\rho^b(x)\big)\Big] \\ +2\eta\int_0^1 dx\int_0^x dy \Big[& \rho^b(x)\big(1-\rho^a(y)\big) - \rho^a(x)\big(1-\rho^b(y)\big)\Big], \end{aligned}$$

resulting from (6.1.3) in the thermodynamic limit. These equations are solved by means of the Jacobi elliptic function

$$\rho_a(x) = \frac{1}{2} + \frac{1}{2\sqrt{1-C}}\,\mathrm{sn}(\eta x, \sqrt{1-C}),$$

where C is a constant of motion

$$\rho_a(x)\big(1-\rho_a(x)\big)\rho_b(x)\big(1-\rho_b(x)\big) = \frac{C}{16}.$$

Finding C is obtained by imposing the fundamental period $X(C)$ of these functions to be 1. We have

$$X(C) = \frac{1}{\eta}F(\frac{\pi}{2}, \sqrt{1-C}) = \frac{4}{\eta}\int_0^1 \frac{d\nu}{\sqrt{[1-\nu^2][1-(1-C)\nu^2]}},$$

where F is the elliptic integral of the first kind. $X(C)$ is a decreasing function of C on $]0,1]$, reaching its minimum for $C=1$, so that

$$X(C) \geq X(1) = \frac{2\pi}{\eta}.$$

Thus appears a critical value for η, namely

$$\eta_c = 2\pi.$$

This is actually the transition point between the "Brownian" regime and the stretched one where one loop may appear, the degenerate solution $\rho^a(x) = \rho^b(x) = 1/2$ becoming unstable at this point, as seen by linear stability analysis.

The transition identified in [33] for the ABC model introduced in Chapter 1 is in fact analogous. Here the sequences can be mapped onto paths on a triangular lattice. With the following specification of the rates

$$\log \frac{\lambda_{ab}}{\lambda_{ba}} = \frac{\alpha}{N}, \quad \log \frac{\lambda_{bc}}{\lambda_{cb}} = \frac{\beta}{N}, \quad \log \frac{\lambda_{ca}}{\lambda_{ac}} = \frac{\gamma}{N},$$

and the mean densities verifying

$$\widetilde{\rho}_a = \frac{\alpha}{\alpha+\beta+\gamma}, \quad \widetilde{\rho}_b = \frac{\beta}{\alpha+\beta+\gamma}, \quad \widetilde{\rho}_c = \frac{\gamma}{\alpha+\beta+\gamma},$$

there exist then a critical value of the parameter $\eta = (\alpha+\beta+\gamma)/3$

$$\eta_c \overset{\text{def}}{=} \frac{2\pi}{3\sqrt{\widetilde{\rho}_a\widetilde{\rho}_b\widetilde{\rho}_c}},$$

corresponding to the transition between Brownian and stretched paths [63].

6.2 Non-reversibility and cycle currents

All the preceding considerations are valid when the process is reversible. Reversibility is broken when there exists at least one cycle in the state graph for which the Kolmogorov criterion fails. Using the network theory of Schnakenberg [202] summarized in Chapter 1 we have a way to express the probability currents at steady-state from the expression (1.1.2) upon choosing a cycle basis $\mathcal{C}$ of the state graph $\mathcal{G}$. Let $\mathcal{C}_{\eta\eta'}$ the set of cycles in $\mathcal{G}$ containing the oriented edge (η,η') and having a positive orientation w.r.t. this edge. Let $\mathcal{T}_C$ a set of subgraph of $\mathcal{G}$, s.t. when $C \in \mathcal{C}$ is glued into a single node η_C, $\mathcal{T}_C$ represents the set of oriented spanning trees rooted in η_C, as defined in Chapter 1. With these notations, the steady-state current between η and η' can be cast as [64]

$$J_{\eta\eta'} = 2 \sum_{C\in\mathcal{C}_{\eta\eta'}} \frac{\sum_{t\in\mathcal{T}_C} w(t)}{\sum_{\eta''\in\mathcal{V}}\sum_{t\in\mathcal{T}_{\eta''}} w(t)} D(C) \sinh\Big(\frac{\mathcal{A}_C}{2}\Big)$$

with

$$D(C) = \prod_{(\eta,\eta')\in C} \big(W_{\eta\eta'}W_{\eta'\eta}\big)^{1/2}$$

and $\mathcal{A}_C$ the affinity associated to cycle C as defined in (1.1.5).

We explicitly see here that only the cycles for which the Kolmogorov criteria is violated contribute with a positive entropy production 1.1.3. In distinguishing among cycles the ones which are not reversible from the other ones comes the notion of non-trivial cycles being topologically equivalent[1], i.e. which can be made identical by combining them with trivial cycles. Then it is interesting to consider a cycle basis such that each topological class is represented by one

[1] Slightly different also in spirit from the topological currents introduced in [29].

single element of the basis. Specified to the ABC exclusion processes these considerations lead to consider a cycle basis where non-reversible cycles are the ones in which one particle performs a round trip w.r.t. the others. Then the decomposition (1.1.4) takes the following form for e.g. an $AB \to BA$ transition:

$$J_{\eta\eta'} = \lambda_{ab}\pi_\eta - \lambda_{ba}\pi_{\eta'} = \Phi(\eta_{\setminus a}) - \Phi(\eta_{\setminus b}) \tag{6.2.1}$$

where π_η denote the steady-state measure and where $\Phi(\eta_{\setminus a})$ and $\Phi(\eta_{\setminus b})$ are quantities which depend on the configuration obtained from η (or equivalently η') by removing respectively the particle A or the particle B which are moving. $\eta_{\setminus a}$ [resp. $\eta_{\setminus b}$] indeed specify the cycle obtained by letting A [resp. B] performing a positive [resp. negative] round trip. Note that for the simple asymmetric exclusion process with open boundary (and more evidently on the ring) where the exact stationary measure can be expressed in terms of a trace of a matrix product, the form (6.2.1) holds with $\Phi(\eta_{\setminus a}) = \pi_{\eta_{\setminus a}}$ and $\Phi(\eta_{\setminus a}) = -\pi_{\eta_{\setminus b}}$, i.e. the invariant of measure of the reduced system, where one particle or one empty site has been removed. Based on this observation, if N is the number of particles we postulate that a similar relation

$$\Phi_a^{(N)}(\eta_{\setminus a}) = C_a^{(N)}\pi_{\eta\setminus a}^{(N-1)}, \tag{6.2.2}$$

holds for multi-type exclusion processes at least asymptotically when $N \to \infty$.

6.3 More on continuous limits

In this section we examine how the microscopic coefficients $C_a^{(N)}$, whenever (6.2.2) holds, can be transposed at macroscopic level and how they are related to important coefficients showing up in the Lotka-Volterra equations of the fluid limit. We consider hereafter the n-type exclusion process.

Let $\phi_a, a = 1 \dots n$ a set of arbitrary functions in $\mathbf{C}^2[0,1]$. For $i \in \{1, \dots N\}$, X_i^a is a binary random variable and, at time t, the presence of a particle of type a at site i is equivalent to $X_i^a(t) = 1$, with the exclusion constraint $\sum_{a=1}^n X_i^a(t) = 1$. We are interested in the time dependent moment generating function of the process

$$f_t^{(N)}(\phi) \stackrel{\text{def}}{=} \mathbb{E}\left[\exp\Big(\frac{1}{N}\sum_{a=1,i=1}^{n,N} \phi_a\big(\frac{i}{N}\big)X_i^a(t)\Big)\right],$$

where ϕ denotes the set $\{\phi_a, a = 1 \dots n\}$. The hydrodynamic scaling assumes an asymptotic expansion of the form

$$\lambda_{ab}(N) = D\big(N^2 + \frac{\alpha_{ab}}{2}N\big) + \mathcal{O}(1), \qquad \forall a, b \; a \neq b,$$

where $\alpha_{ab} = -\alpha_{ba}$ are real constants. With D independent of the specific pair (a, b) the system is *equidiffusive*. When $N \to \infty$, f_t satisfies the functional

integral equation [65]

$$\frac{\partial f_t}{\partial t} = D\int_0^1 dx \sum_{a=1}^n \phi_k(x)\ \frac{\partial}{\partial x}\Big[\ \frac{\partial}{\partial x}\frac{\partial f_t}{\partial \phi_a(x)} - \sum_{a\neq b}\alpha_{ab}\Big(\frac{\partial^2 f_t}{\partial\phi_a(x)\partial\phi_b(x)}\Big)\Big].$$

Assume at time 0 the given initial profile $\rho_a(x,0)$ to be twice differentiable w.r.t. x. Then

$$\log[f_t(\phi)] = \int_0^1 dx \sum_{a=1}^n \rho_a(x,t)\phi_a(x),$$

where $\rho_a(x,t)$ satisfy the hydrodynamic system of coupled Burger's equations

$$\frac{\partial \rho_a}{\partial t} = D\Big[\frac{\partial^2\rho_a}{\partial x^2} + \frac{\partial}{\partial x}\Big(\sum_{a\neq b}\alpha_{ab}\rho_a\rho_b\Big)\Big], \qquad a = 1,\dots,n. \tag{6.3.1}$$

If we assume now (6.2.2) to hold, then the limit functional $f_\infty[\phi] = \lim_{N\to\infty} f_\infty^{(N)}[\phi]$, where

$$f_\infty^{(N)}[\phi] = \sum_{\{\eta\}}\pi_\eta \exp\Big(\frac{1}{N}\sum_{a=1,i=1}^{n,N} X_i^a\phi_a(\frac{i}{N})\Big),$$

satisfies the equation

$$\frac{\partial}{\partial x}\frac{\partial f_\infty}{\partial\phi_a(x)} + \sum_{a\neq b}\alpha_{ab}\frac{\partial^2 f_\infty}{\partial\phi_a(x)\partial\phi_b(x)} = c_a f_\infty - v\frac{\partial f_\infty}{\partial\phi_a(x)}, \tag{6.3.2}$$

with

$$\lim_{N\to\infty}\frac{N^2 C_a^{(N)}}{\lambda_{ab}^{(N)}} = \lim_{N\to\infty}\frac{C_a^{(N)}}{D} = c_a, \qquad \text{with} \qquad v \stackrel{\text{def}}{=} \sum_{a=1}^n c_a.$$

Now we can make the link between (6.2.2) and the fluid limit description of stationary states. A solution is sought of the form

$$f_\infty(\phi) = \exp\Big(\int_0^1 dx\sum_{a=1}^n \rho_a^\infty(x)\phi_a(x)\Big),$$

which, instantiated into (6.3.2), gives the Lotka-Volterra type of equations

$$\frac{\partial\rho_a^\infty}{\partial x} - \rho_a^\infty\sum_{b\neq a}\alpha^{ab}\rho_b^\infty = c_a - v\rho_a^\infty, \qquad a = 1\dots n.$$

This is a particular stationary solution of the system formed by the coupled Burger's equations (6.3.1) where the functions ρ_a are sought in the class

$$\rho_a(x,t) \stackrel{\text{def}}{=} \rho_a^\infty(x - vt),$$

the variable $(x - vt)$ being taken modulo 1. Hence, there is a rotating frame at velocity v, in which ρ_a^∞ is periodic. Moreover, in this frame, the stationary currents do not vanish and have constant values

$$J_a(x) = c_k$$

Therefore, while the macroscopic constants $\{c_a, a = 1, \ldots, n\}$ are in principle determined from the periodic boundary conditions constraints and from the fixed average values of each particle species, they can also be related to microscopic quantities.

Chapter 7

Modelling the fundamental diagram of traffic with solvable models

This chapter is based on the following paper:

C. Furtlehner, J.M. Lasgouttes and M. Samsonov, One-dimensional particle processes with acceleration/braking asymmetry. *J.Stat.Phys 147*, 6 (2012), 1113–1144.

From coupled exclusion processes to multi-type exclusion processes relevant to traffic modelling there is a small leap that we take now in this chapter. Cellular automaton models have been successful for large scale simulations of traffic but not much amenable to precise theoretical analysis because of non-local moves and parallel dynamics. Here we transpose basic mechanisms that make them successful into the definition of multi-type exclusion processes in order to obtain (almost) solvable models of traffic.

7.1 Multi-type exclusion processes for traffic modelling

We consider a multi-type exclusion process, generalizing the simple exclusion process on the line introduced in Chapter 1, combining the braking and accelerating feature of the Nagel-Schreckenberg models [173], with the locality of the simple ASEP model, in which only two consecutive sites do interact at a given time. For this we allow each car to change stochastically its hopping rate, depending on the state of the next site. For a 2-speed model, let A (resp. B) denote a site occupied by a fast (resp. slow) vehicle, let E denote an empty site and $O = A$ *or* B an occupied site; the model is defined by the following set of reactions, involving pairs of neighbouring sites:

$$AE \xrightarrow{\mu_a} EA \qquad \text{simple move of fast vehicle} \tag{7.1.1}$$

$$BE \xrightarrow{\mu_b} EB \qquad \text{simple move of slow vehicle} \tag{7.1.2}$$

$$BE \xrightarrow{\gamma} AE \qquad \text{slow vehicle spontaneously accelerates} \tag{7.1.3}$$

$$AO \xrightarrow{\delta} BO \qquad A \text{ brakes behind } O = A \text{ or } B \tag{7.1.4}$$

μ_a, μ_b, γ and δ denote the transition rates of the associated Markov process. The dynamics is purely random sequential, as opposed to the parallel dynamics of the Nagel-Schreckenberg model. It encodes the tendency of a vehicle to accelerate when there is space ahead (7.1.3), and to slow down otherwise (7.1.4). The main mechanism behind congestion, namely the asymmetry between braking and acceleration is potentially present in the model when γ is different from δ. Our model is in fact similar to the model of Appert-Santen [8], in which there is a single speed, but particles have 2 states (at rest and moving), with possible transitions between these 2 states. We consider the model on the ring geometry with a total size denoted $S = N + L$, N being the (fixed) number of vehicles and L the number of empty sites. This model contains and generalizes several sub-models which are known to be integrable with particular rates. The hopping part (7.1.1,7.1.2) of the models is just the TASEP when $\mu_a = \mu_b$, which is known to be integrable with help of the Bethe Ansatz (see e.g. [92] and reference therein). A generalization including multiparticle dynamics with overtaking is provided by the Karimipour model [124, 27], which turns out to be integrable as well. As explained in Chapter 1, the matrix ansatz allows one to describe the stationary

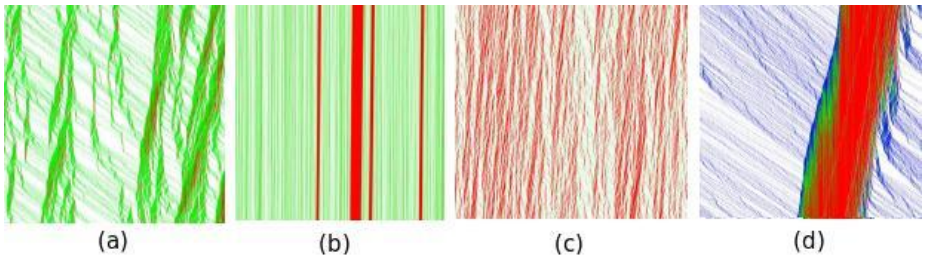


Figure 7.1.1: (a)-(c) Space-time diagrams: time is going downward and particles travel to the right. Slow ones ($\mu_b = 10$) are colored in red and fast ones $\mu_a = 100$ in green. The system size is $S = 3000$ for (a) and (c) and $S = 100000$ for (b), all with same density $\rho = 0.2$. Other parameters are $\gamma = 10$, $\delta = 1$ for (a) and (b) and $\delta = 10$ for (c). (d) is a jam obtained for a model with an additional label colored in blue corresponding to a faster speed level μ_c.

regime of the model. The acceleration/deceleration dynamics is equivalent to the coagulation/decoagulation models, which are known to be solvable by the empty interval method and by free fermions for particular sets of rates [198], but the whole process is presumably not integrable. As seen on Figure 7.1.1 when no asymmetry between braking and accelerating is present ($\gamma = \delta$) no spontaneous large jam structure is observed. As the density $\rho \stackrel{\text{def}}{=} N/S$ of cars increases, one observes a smooth transition between a TASEP of fast particles for small ρ to a TASEP of slow particles around $\rho \simeq 1$. Instead, when the ratio δ/γ is reduced, there is a proliferation of small jams. Below some threshold of this ratio, we observe apparition of large jams above some threshold value of the density.

It is tempting to interpret this as a condensation mechanism at equilibrium in the canonical ensemble [60], by combining the Nagel-Paczuski [172] interpretation of competing queues with some results [126, 75] which, in the context of tandem queues on a ring, allows this condensation mechanism to take place if the appearance of slow vehicles is a sufficiently rare event. In fact in [120] a general criterion for having phase separation is conjectured for conserved systems, based on the asymptotic behavior of the current passing through clusters of large size. We will come back to this point in the next section when considering the queuing interpretation of jam formation in this model. A way to observe the effect of asymmetry is to allow particles to enter or quit the system with some very low rate when compared to the others. The global density of cars then performs a random walk, and by looking at trajectories in the FD plane, we see hysteresis effect for $\delta < \gamma$. There are two different quantities of interest here:

$$\Phi_1 \stackrel{\text{def}}{=} \frac{1}{S}\sum_{i=0}^{S-1}(\mu_a A_i + \mu_b B_i)E_{i+1}, \qquad \text{and} \qquad \Phi_2 \stackrel{\text{def}}{=} \frac{1}{S}\sum_{i=0}^{S-1}\mu_a A_i + \mu_b B_i,$$

(with Boolean variables $A_i + B_i + E_i = 1$). Φ_1 represents the flow of particles,

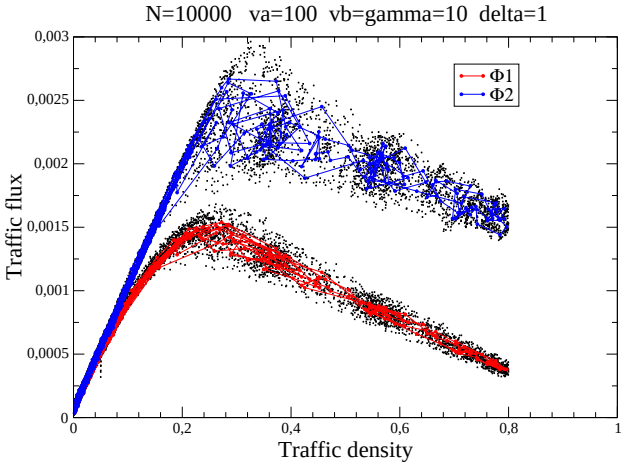


Figure 7.1.2: Fundamental diagram with $\gamma = 10\ \delta$ with small spontaneous rate of emission and escape of particles. Trajectories of Φ_1 and Φ_2 are displayed. Φ_1 is the actual particle flow and Φ_2 is interpreted as the "traffic flow".

while Φ_2 counts particles with their speed, regardless of whether the next site is occupied or not. Φ_2 is more representative of the traffic flow than Φ_1 because in reality, cars are obeying a parallel dynamics and clusters of particles should be though as moving platoons of cars. This is reflected in the FD of Figure 7.1.2 where Φ_2 gives a much more realistic FD than Φ_1. This suggests that, when comparing the FD diagram of sequential exclusion processes with parallel dynamical cellular automata, Φ_2 should be considered rather than Φ_1.

7.2 Queuing processes with dynamically coupled service rates

In the context of exclusion processes, jams are represented as clusters of particles. Clustering phenomena can be analyzed in some cases by mapping the process to a Jackson queueing network as explained in Chapter 1. In principle two dual mappings are possible:

(i) the queues are associated to empty sites and the clients are the particles in contact behind this site,

(ii) the queues are associated with particles and the clients are the empty sites in front of this particle.

Concerning the model of the preceding section, the mapping of type (i) is exact up to a slight extension of the ordinary definition of a queueing process. In this

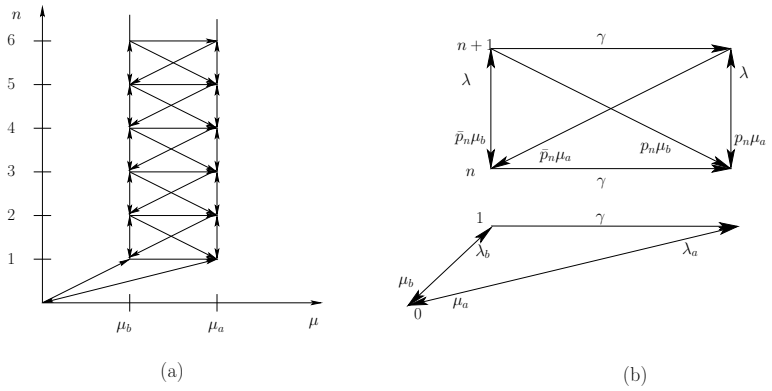


Figure 7.2.1: (a) State flow diagram of a queue with stochastic service rate corresponding to model (7.1.1,. . . 7.1.4). (b) Details of the transition rates.

new process by contrast to standard ones, the service rate is a stochastic Markov process as well: it can take two values $\mu_a > \mu_b$, and slow queues become fast at rate γ conditionally to having at least one client, while empty fast queues become slow at rate δ. However, the process corresponding to each queue taken in isolation will not be reversible in general, which means that the steady state of the joint process is not in product form. In this setting, jamming will be associated to long range correlation between (empty) queues, which make this mapping useless.

The mapping of type (i) is more interesting with respect to jam distribution because in that case a large jam can be represented by a single filled queue, so that the product form may constitute a good approximation to the joint measure of the queuing network. Unfortunately the mapping in that case can be only an approximate one. Heterogeneity of speed labels within clusters get lost when encoded as queues. For this mapping, we use a simple approximation which consists in to estimate first the typical profile of a cluster containing n vehicles. According to this profile we can then estimate conditionally to n the (stationary) probability p_n of the front vehicle to be of type A or B. Given an arrival rate $\lambda = \lambda_a + \lambda_b$ decomposed into incoming rates of type respectively A and B particles and δ the rate of decay from type A to B we have

$$p_n = \frac{\lambda_a}{\lambda} r^n \qquad \text{with} \qquad r = \frac{\lambda}{\lambda + \delta}.$$

Since the front end interface of the cluster has no causal effect on the rest of the queue, except for the front vehicle which may accelerate with rate γ, we can consider the dynamics of the sequence independently of the motion of the front interface. With the rather crude additional assumption of independence of the local speed labels in the bulk we are able to write a master equation of the joint process $(n(t), \mu(t))$ corresponding to the queue taken in isolation.

This equation governs the evolution of $P_t(n,\tau) = P(n(t) = n, \mu = \mu_a\tau + \mu_b\bar{\tau})$, the joint probability that the queue has n clients and its front car is of type A ($\tau = 1$) or B ($\bar{\tau} \stackrel{\text{def}}{=} 1 - \tau = 1$). Given $p_n(\tau) \stackrel{\text{def}}{=} p_n\tau + \bar{p}_n\bar{\tau}$, using $\bar{p}_n$ to denote $1 - p_n$, the master equation reads

$$\begin{aligned}
\frac{dP_t(n,\tau)}{dt} &= \lambda\big(P_t(n-1,\tau) - P_t(n,\tau)\big) + \big(\mu_a P_t(n+1,1) + \mu_b P_t(n+1,0)\big)p_n(\tau) \\
&\quad - (\mu_a\tau + \mu_b\bar{\tau})P_t(n,\tau) + \gamma(\tau - \bar{\tau})P_t(n,0), \qquad n \geq 2 \\
\frac{dP_t(1,\tau)}{dt} &= (\lambda_a\tau + \lambda_b\bar{\tau})P_t(0) - \lambda P_t(1,\tau) + \big(\mu_a P_t(2,1) + \mu_b P_t(2,0)\big)p_1(\tau) \\
&\quad - (\mu_a\tau + \mu_b\bar{\tau})P_t(1,\tau) + \gamma(\tau - \bar{\tau})P_t(1,0), \\
\frac{dP_t(0)}{dt} &= -\lambda P_t(0) + \mu_a P_t(1,1) + \mu_b P_t(1,0).
\end{aligned}$$

It is a special case of a queuing process with a 2-level dynamically coupled stochastic service rate, as defined in [77], which state-graph is represented on Figure 7.2.1. Its stationary regime can be analyzed by introducing the following generating functions

$$g_{a,b}(z) \stackrel{\text{def}}{=} \sum_{n=1}^{\infty} \pi_n^{a,b} z^n \qquad \text{and} \qquad g(z) \stackrel{\text{def}}{=} \pi_0 + g_a(z) + g_b(z).$$

where $\pi_n^{a,b} \stackrel{\text{def}}{=} P(n(t) = n, \mu = \mu_{a,b})$ and $\pi_0 = P(n(t) = 0)$, represents the stationary state. The solution reads:

$$g_a(z) = \frac{\mu_b}{\mu_a - \mu_b}\pi_0 + \frac{\lambda z - \mu_b}{\mu_a - \mu_b}g(z), \qquad g_b(z) = \frac{\mu_a}{\mu_b - \mu_a}\pi_0 + \frac{\lambda z - \mu_a}{\mu_b - \mu_a}g(z). \tag{7.2.1}$$

with

$$g(z) = \sum_{n=0}^{\infty}(-u)^n \frac{w - vr^n z}{\prod_{k=0}^{n}(zr^k - z^+)(zr^k - z^-)}. \tag{7.2.2}$$

and

$$\begin{aligned}
z^{\pm} &= \frac{1}{2\lambda}\big(\mu_a + \mu_b + \lambda + \gamma \pm \sqrt{(\lambda - \gamma + \mu_a - \mu_b)^2 + 4\lambda\gamma}\big), \\
\lambda^2 u &= \lambda_a(\mu_a - \mu_b), \\
\lambda^2 v &= (\lambda_a\mu_a + \lambda_b\mu_b)\pi_0, \\
\lambda^2 w &= (\mu_a\mu_b + \lambda_a\mu_a + \lambda_b\mu_b + \gamma\mu_a)\pi_0.
\end{aligned}$$

Upon using Cauchy integrals, the $\pi_n^{a,b}$ are then given as sums of geometric laws. From the radius of convergence z^- of g, the limit of ergodicity is obtained for $z^- \geq 1$, i.e. for

$$\lambda \leq \mu_b + \gamma\frac{\mu_a - \mu_b}{\mu_a + \gamma}.$$

7.3 Large deviation functional of the Fundamental Diagram

In practice, points plotted in experimental FD studies as shown on Figure 5.1.1 are obtained by averaging data from static loop detectors over a few minutes. This is difficult to compute from our queue-based model, for which a space average is much easier to obtain. The equivalence between time and space averaging is not an obvious assumption [22], but since jams are moving, space and time correlations are combined in some way [172] and we consider this assumption to be quite safe. In this section, we want to extend the traditional study of the FD to the analysis of the fluctuations, i.e. the departure from the deterministic function relating the flow to the density. Experimentally, the congestion region of the FD is seen to be dominated by fluctuations, while the free flow part is rather deterministic. In the following we consider a probabilistic version of the FD, where the deviation from the deterministic FD is analyzed in the large deviation framework using the mapping of the preceding section. We consider the conditional probability $P(\phi|d)$, where d represents the spatial density of cars and ϕ the normalized flow:

$$\begin{cases} d = \dfrac{N}{N+L}, \\ \phi = \dfrac{\Phi}{N+L}, \end{cases} \quad \text{with} \quad \begin{cases} L & \text{number of queues} \\ N = \displaystyle\sum_{i=1}^{L} n_i & \text{number of vehicles} \\ \Phi = \displaystyle\sum_{i=1}^{L} \mu_i \mathbb{1}_{\{n_i>0\}} & \text{integrated flow} \end{cases}$$

The numbers N of vehicles and L of queues are fixed, meaning that we are working with the canonical ensemble. The closed tandem formed out of the effective queuing processes of the preceding section do not satisfy the conditions given in [77]. At this stage we simply take it as a crude approximation, hence assuming the following form of the joint probability measure:

$$P(\{n_i,\mu_i\}) = \frac{\delta\big(N - \sum_{i=1}^{L} n_i\big)}{Z_L(N)} \prod_{i=1}^{L} \pi^{\lambda}(n_i,\mu_i),$$

with canonical partition function

$$Z_L(N) \stackrel{\text{def}}{=} \sum_{\{n_i,\mu_i\}} \delta\big(N - \textstyle\sum_{i=1}^{L} n_i\big) \displaystyle\prod_{i=1}^{L} \pi^{\lambda}(n_i,\mu_i),$$

where δ denotes now the usual Dirac function. When ϕ is interpreted as a continuous variable, the properly normalized density-flow conditional probability distribution takes the form

$$P(\phi|d) = \frac{L}{1-d} \frac{Z_L\big[L\frac{d}{1-d}, L\frac{\phi}{1-d}\big]}{Z_L\big[L\frac{d}{1-d}\big]},$$

with

$$Z_L(N,\Phi) \stackrel{\text{def}}{=} \sum_{\{n_i,\mu_i\}} \delta\big(N - \textstyle\sum_{i=1}^L n_i\big)\delta\big(\Phi - \sum_{i=1}^L \mu_i \mathbb{1}_{\{n_i>0\}}\big) \prod_{i=1}^L \pi^\lambda(n_i,\mu_i). \tag{7.3.1}$$

Note (by simple inspection, see e.g. [129]) that $P(\phi|d)$ is independent of λ. $Z_L(N)$ and $Z_L(N,\Phi)$ represent respectively the probability of having N vehicles and the joint probability for having at the same time N vehicles and a flow Φ, under the unconstrained product form. Under this product form, on general ground, we expect d and ϕ to satisfy a large deviation principle (see e.g. [223]), i.e. that there exist two rate functions $I(d)$ and $J(d,\phi)$ such that, for large L,

$$Z_L(N) \asymp e^{-LI(d)},$$

$$Z_L(N,\Phi) \asymp e^{-LJ(d,\phi)},$$

where "$\asymp$" stands for logarithmic equivalence. This leads to a large deviation version of the fundamental diagram

$$P(\phi|d) \asymp e^{-LK(\phi|d)}, \quad \text{with} \quad K(\phi|d) \stackrel{\text{def}}{=} J(d,\phi) - I(d). \tag{7.3.2}$$

These rate function can be expressed with help of the moment and cumulant generating function g and h associated to π^λ,

$$g(s,t) \stackrel{\text{def}}{=} \sum_{n=0,\mu}^{\infty} \pi^\lambda(n,\mu) e^{sn+t\mu} \qquad \text{and} \qquad h(s,t) \stackrel{\text{def}}{=} \log[g(s,t)],$$

already encountered in (7.2.2) for our specific problem, where it is assumed by convention that the rate μ is zero in absence of client.

We denote by λ_d (both for I and J) and λ_ϕ the Lagrange multipliers associated respectively to the density and flux constraints. The rate functions are given by

$$I(d) = \frac{d}{1-d}\lambda_d(d) - h\big(\lambda_d(d),0\big),$$

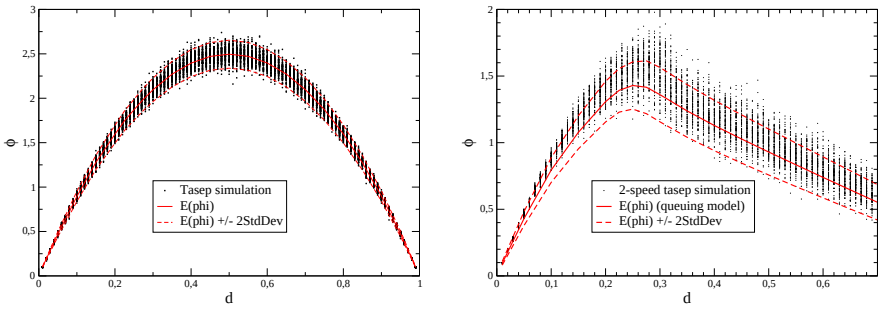
$$J(d,\phi) = \frac{d}{1-d}\lambda_d(d,\phi) + \frac{\phi}{1-d}\lambda_\phi(d,\phi) - h\big(\lambda_d(d,\phi),\lambda_\phi(d,\phi)\big),$$

with the Lagrange multipliers implicitly given by

$$\frac{\partial h}{\partial s}\big(\lambda_d(d),0\big) = \frac{d}{1-d}, \tag{7.3.3}$$

for I and

$$\begin{cases} \dfrac{\partial h}{\partial s}\big(\lambda_d(d,\phi),\lambda_\phi(d,\phi)\big) = \dfrac{d}{1-d}, \\ \dfrac{\partial h}{\partial t}\big(\lambda_d(d,\phi),\lambda_\phi(d,\phi)\big) = \dfrac{\phi}{1-d}, \end{cases} \tag{7.3.4}$$



(a) (b)

Figure 7.3.1: Comparing FD based on exclusion processes and approximate queuing process. Size is $L = 10^3$, $\delta = 0$ (TASEP) on panel (a), $\mu_a = 10 \times \mu_b = 10 \times \gamma = 100 \times \delta$ on panel (b).

for J. The ordinary FD $\phi(d)$ is the minimizer of $K(\phi|d)$ and actually corresponds to

$$K(\phi(d)|d) = 0.$$

The small i.e. Gaussian fluctuations are then obtained by expanding K at second order in $\phi - \phi(d)$. Denoting by $H^\star(s,t)$ the 2 by 2 dual Hessian corresponding to second derivatives of $h(s,t)$, representing the covariant matrix between the charges of the queues and the flux, we find the following expression for the variance of the FD:

$$\mathrm{Var}(\phi|d) = \frac{(1-d)^2}{L}\big(H^{\star -1}{}_{tt}\big)^{-1}.$$

As a check, computing the FD rate function for TASEP reads ($\bar{d} \stackrel{\text{def}}{=} 1-d$)

$$K(\phi|d) = \frac{d}{\bar{d}}\log\frac{\mu d - \phi}{\mu d^2} + \frac{\phi}{\mu\bar{d}}\log\frac{\phi^2}{(\mu d-\phi)(\mu\bar{d}-\phi)} - \log\frac{\mu\bar{d}^{\,2}}{\mu\bar{d}-\phi},$$

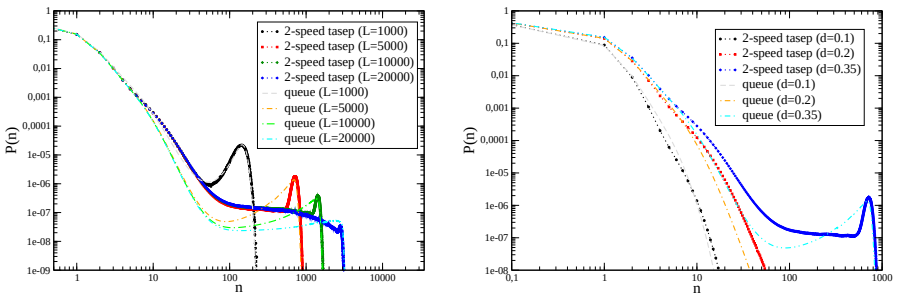
yielding

$$\phi(d) = \mathbb{E}(\phi|d) = \mu d\bar{d}, \qquad \mathrm{Var}(\phi|d) = \frac{\mu^2}{N+L}\, d^2\bar{d}^{\,2}.$$

In the case of the 2-speed process, from (7.2.1,7.2.2), we get for the cumulant generating function

$$h(s,t) = \log\Big(\pi_0\big(1+\frac{\mu_b e^{\mu_a t} - \mu_a e^{\mu_b t}}{\mu_a - \mu_b}\big) + \frac{(\mu_a - \lambda e^s)e^{\mu_b t} + (\lambda e^s - \mu_b)e^{\mu_a t}}{\mu_a - \mu_b}\, g(e^s)\Big),$$

with $g(z)$ given by (7.2.2), from which the Legendre transform as well as the Hessian $H^\star(s^\star, 0)$ for the small fluctuations can be obtained, where $s^\star$ is the



(a) (b)

Figure 7.3.2: Comparing cluster vs queue size distributions. (a) $\mu_a = 10\mu_b = 100$, $\gamma = 10$ and $\delta = 1$, varying size L with fixed density $d = 0.35$. (b) Varying densities with fixed $L = 5000$.

point satisfying

$$e^{s^\star} g'(s^\star) = \frac{d}{1-d}.$$

Solving these equations numerically gives us the plots of Figure 7.3.1. In addition to the FD, it is also interesting to determine the single queue distribution in canonical ensemble [60]. With help of the partition function in the large deviation framework we have

$$\begin{aligned} p_{\mathrm{CE}}(n,\mu) &= \pi^\lambda(n,\mu)\frac{Z_{L-1}(N-n)}{Z_L(N)} \\ &\simeq \pi^\lambda(n,\mu)\exp\Big[L\big(h(\lambda_d(d-x),0) - h(\lambda_d(d),0) \\ &\quad - \frac{d-x}{1-d-x}\lambda_d(d-x) + \frac{d}{1-d}\lambda_d(d)\big)\Big], \end{aligned}$$

with $x \stackrel{\text{def}}{=} n/(N+L)$ and the density constraint 7.3.3 satisfied by $\lambda_d(d)$. A comparison of this queuing formulation with the original exclusion process is given on Figure 7.3.2. The correspondence between the cluster size distribution observed on the bi-speed TASEP (7.1.1,...,7.1.4) , with the single queue distribution obtained from the generalized queuing process is rather accurate. In particular, in both cases, a bump is observed in the distributions at the same location, for small size systems. It indicates that condensation is observed as a finite size phenomena. In the thermodynamic limit, macroscopic jams are absent. In this respect, it is different from the type of condensation analyzed in [60], which is obtained under some conditions on the service rate, as a large deviation principle but with different scaling than L.

Part III

Traffic inference with Belief Propagation

In this part we turn to the traffic prediction application, which we have been tackling with help of MRF combined with belief propagation. This has been also a source of motivation to address the inverse Ising problem with mean-field methods.

Some years ago we started to investigate [79] the possibility of building an MRF which could encode both spatial and temporal dependencies and come with a linear computational time when running the probabilistic inference task. We considered two types of models involving either latent binary variables [158] or Gaussian copula models, both coming with an associated learning algorithm to generate compliant models, respectively with Generalized Belief Propagation for binary variables [73] and Gaussian Belief Propagation for real-valued ones [157]. In this approach, a graphical model representing relations of conditional independence between segments at different locations and different time steps is determined through a graph of pairwise interactions.

Chapter 8

Ising based approach

This chapter is based on the following papers:

V. Martin, J.M. Lasgouttes and C. Furtlehner, Latent binary MRF for online reconstruction of large scale systems, *Annals of Mathematics and Artificial Intelligence* (2015),1–32.

C. Furtlehner, J.M. Lasgouttes and A. Auger, Learning multiple belief propagation fixed points for real time inference.*Physica A: Statistical Mechanics and its Applications 389*, 1 (2010), 149–163.

8.1 latent congestion variables and traffic index

When looking at standard traffic information systems, the representation of the congestion network suggests two main traffic states: non-congested (green) or congested (red) as shown on Figure 8.1.1. At the level of a single traffic segment, the two states free flow/congested clearly represent distinct regimes of traffic flow, well identified on the fundamental diagram, with different statistical properties. Viewing it as a latent state, we incorporate it explicitly in our modelling by attaching to each segment i at any given time t a binary latent variable $s_{i,t} \in \{-1, 1\}$ representing the congested/non-congested state. To relate this state to observations we could, given a distribution $\hat{f}$ of travel time for instance, take the mean or the median travel time as a separator of the two states. Actually we proceed in a way that encompasses this possibility, but is not limited to it. The idea is to define the latent binary state τ $(= \frac{1+s}{2})$ associated to some travel time x in a more abstract way through the mapping:

$$\Lambda(x) \stackrel{\text{def}}{=} P(\tau = 1|x). \tag{8.1.1}$$

This means that an observation x is translated into a conditional probability for the considered segment to be congested. This number $\Lambda(x) \in [0, 1]$, represents our operational definition for the *traffic index*.

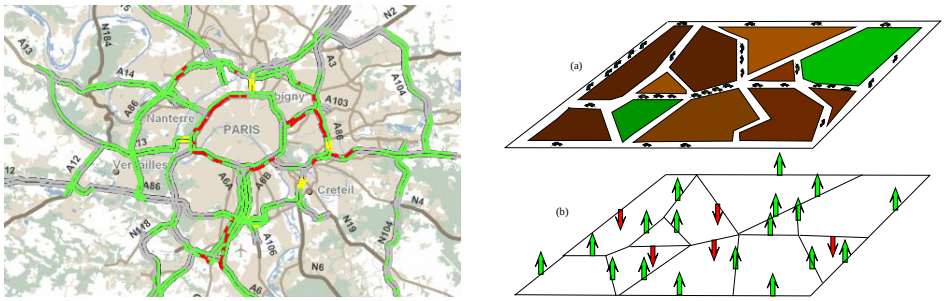


Figure 8.1.1: Underlying Ising modelling of traffic configurations.

Using Bayes rules and the Boolean notation $\bar{\tau} \stackrel{\text{def}}{=} 1 - \tau$, we obtain

$$P(x|\tau) = \Big(\frac{\Lambda(x)}{p_\Lambda}\tau + \frac{1 - \Lambda(x)}{1 - p_\Lambda}\bar{\tau}\Big)\hat{f}(x). \tag{8.1.2}$$

where $p_\Lambda \stackrel{\text{def}}{=} P(\tau = 1)$. The normalization constraint imposes

$$p_\Lambda = \int \Lambda(x)\hat{f}(x)dx. \tag{8.1.3}$$

A certain amount of information can be stored in this mapping. A special case mentioned before corresponds to having for Λ a step function, i.e.

$$\Lambda(x) = \mathbb{1}_{\{x>x^*\}}, \tag{8.1.4}$$

with an adjustable parameter corresponding to the threshold x^*. Another parameter free possibility is to use the empirical cumulative distribution:

$$\Lambda(x) = \hat{F}(x) \stackrel{\text{def}}{=} P(\hat{x} < x). \tag{8.1.5}$$

The main advantage of introducing this map Λ, is the possibility which is offered to convert back a probability of congestion $u = P(\tau = 1)$ into an estimation of the variable x of interest, like e.g. speed or travel time. If Λ is invertible, given u we then simply have:

$$\hat{x} = \Lambda^{-1}(u). \tag{8.1.6}$$

If not, another legitimate way to proceed that we seek for Λ is that the mutual information $I(x, \tau)$ between x and τ be maximal. This reads

$$I(x, \tau) = \int du \; h\big[\Lambda\big(\hat{F}^{-1}(u)\big)\big] - h(p_\Lambda),$$

after introducing the binary information function $h(x) \stackrel{\text{def}}{=} x \log x + (1-x)\log(1-x)$. The step function (8.1.4) with $x^* = \hat{F}^{-1}(1/2)$ corresponding to the median observation is the limit function which maximizes $I(x, \tau)$. If instead we use the inverse map Λ^{-1}, the mutual information between x and τ is not relevant.

Without any specific hypothesis on the distribution of beliefs that BP should generate, another possible requirement is to impose a minimum information i.e. a maximum entropy contained in the variable $u = \Lambda(x)$, which probability density is given by

$$\begin{aligned} dF(u) &\stackrel{\text{def}}{=} \int \delta\big(u - \Lambda(x)\big) d\hat{F}(x) \\ &= \frac{d\hat{F}}{d\Lambda}\big(\Lambda^{-1}(u)\big). \end{aligned}$$

Using this and the change of variable $x = \Lambda^{-1}(u)$ yields the entropy

$$S[u] = -\int d\hat{F}(x) \frac{d\hat{F}}{d\Lambda}(x) = -D_{KL}(\hat{F} \| \Lambda),$$

expressed as the opposite of the relative entropy between F and Λ. Without any further constraint, this leads to the fact that $\Lambda = F$ is the optimal mapping. In both cases, additional constraints comes from the fact that we want a predictor $\hat{x}$ minimizing a loss function $\|\hat{x} - x\|_r$ which depends on the choice of the Euclidean norm $\mathbb{L}_r$ (see [158] for details).

8.2 Ising inference model

The mapping between real-valued observations and the binary latent states is only one element of the model. The general schema of our Ising based inference model is sketched on Figure 8.2.1. It can be decomposed into 4 distinct pieces:

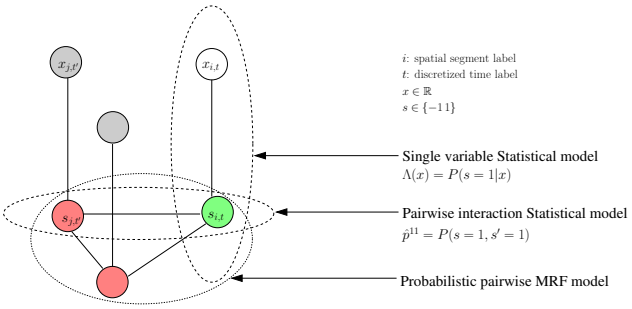


Figure 8.2.1: Sketch of the Ising based inference schema.

- A single variable statistical model translating real-valued observations $x \in \mathbb{R}$ into binary latent states $\tau \in \{0, 1\}$.
- A pairwise statistical model of the dependency between latent states.
- A MRF model to encode the network of dependencies.
- The Belief propagation algorithm to decode a partially observed network.

It is based on a statistical description of traffic data which is obtained by spatial and temporal discretization, in terms of road segments i and discrete day time slots t corresponding to time intervals of typically a few minutes, yielding a set of labels $\mathcal{V} = \{\nu = (i, t)\}$. Then there are two layers of variables, the visible ones (x_ν) and the latent variables (τ_ν). The model itself is based on historical data in form of empirical marginal distributions $\hat{p}(x_\nu)$, $\hat{p}(x_\nu, x_{\nu'})$, giving reference states and statistical interactions between degrees of freedom. Finally, reconstruction and prediction are produced in the form of conditional marginal probability distribution $p(x_\nu|\mathcal{V}^*)$ of unobserved variables in $\mathcal{V}\backslash\mathcal{V}^*$, conditionally to the actual state of the observed variables in the set $\mathcal{V}^*$.

The binary latent states are used to model the interactions in a simplified way enabling for large scale applications. Trying to model exactly the pairwise dependencies at the observation level is potentially too expensive from the statistical as well as the computational viewpoint. So the pairwise model sketched on Figure 8.2.1 corresponds to

$$P(x_\nu, x_{\nu'}) = \sum_{\tau,\tau'} \hat{p}_{\nu\nu'}(\tau, \tau') P(x_\nu|\tau) P(x_{\nu'}|\tau'),$$

with $P(x|\tau)$ given in (8.1.2) and $\hat{p}_{\nu\nu'}$ to be determined from empirical frequency statistics. A probability law of two binary variables requires three independent parameters; two of them are already being given by individuals marginals probabilities $\hat{p}^1_\nu \overset{\text{def}}{=} P(\tau_\nu = 1)$ according to (8.1.3). For each pair of variables, one parameter remains therefore to be fixed. By convenience we consider the coefficient $p^{11}_{\nu\nu'} \overset{\text{def}}{=} P(\tau_\nu = 1, \tau_{\nu'} = 1)$ and obtain thanks to a moment matching

constraint

$$\hat{p}^{11}_{\nu\nu'} = \hat{p}^1_\nu \hat{p}^1_{\nu'} + \frac{\widehat{\mathrm{cov}}\big[\Lambda_\nu(x_\nu), \Lambda_{\nu'}(x_{\nu'})\big]}{\big(2\hat{p}^1_\nu - 1\big)\big(2\hat{p}^1_{\nu'} - 1\big)},$$

involving the empirical covariance between latent states $\widehat{\mathrm{cov}}\big[\Lambda_\nu(x_\nu), \Lambda_{\nu'}(x_{\nu'})\big]$ obtained from observation data.

The next step is to define the Ising model itself, on which to run BP with good inference properties. Recall that we try to answer two related questions:

- Given the set of coefficients $\hat{p}(\tau_\nu)$ and $\hat{p}(\tau_\nu, \tau_{\nu'})$, considered now as model input, what is the joint law $P(\{\tau_\nu, \nu \in \mathcal{V}\})$?
- Given actual observations $\{x^*_\nu, \nu \in \mathcal{V}^*\}$, how to infer $\{x_\nu, \nu \in \mathcal{V} \backslash \mathcal{V}^*\}$?

For this we have to solve in principle an inverse Ising problem. A simple heuristic solution that we have been exploring first in [79] is based on the the Bethe approximation described in Section 3.2. It consists to use this approximation (3.2.4) for the encoding and the belief-propagation for the decoding, such that the calibration of the model is coherent with the inference algorithm. This leads to write boldly:

$$\mathcal{P}(\tau) \propto \prod_{\nu \in \mathcal{V}} \hat{p}_\nu(\tau_\nu) \left(\prod_{(\nu,\nu') \in \mathcal{F}} \frac{\hat{p}_{\nu\nu'}(\tau_\nu, \tau_{\nu'})}{\hat{p}_\nu(\tau_\nu)\hat{p}_{\nu'}(\tau_{\nu'})} \right)^\beta, \tag{8.2.1}$$

where $\mathcal{F}$ is a well selected subset of all pairs of nodes, various possible heuristic being possible for the selection of the most important pairs. β is an adjustable parameter, representing an inverse temperature in the Ising model, tuned to compensate for saturation effects ($\beta < 1$), when the coupling between variables are too large. This is due to some over-counting of the dependencies between variables occurring in a multiply connected graph. By construction $m_{\nu \to \nu'}(x_{\nu'}) \equiv 1$ is a particular BP fixed point when $\beta = 1$.

Next, for the decoding part, information is inserted in real time in the model in the form of probabilities owing to the map (8.1.1) relating the observation x to the latent state τ. The optimal way of inserting this quantity into the BP equations is obtained variationally by imposing the additional constraint $p_\nu(\tau_\nu) = p^*(\tau_\nu)$, which results in modified messages sent from $\nu \in \mathcal{V}^*$, now reading [74]

$$n_{\nu \to \nu'}(x_\nu) = \frac{p^*_\nu(\tau_\nu)}{m_{\nu' \to \nu}(\tau_\nu)}.$$

This leads to a new version of BP which convergence properties have been analyzed in [158].

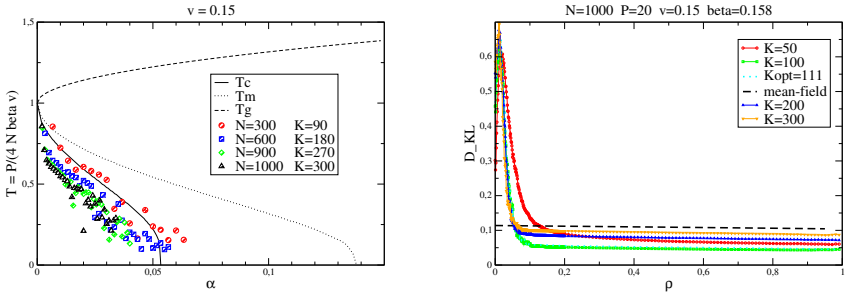
8.3 Belief propagation fixed points as macroscopic traffic states

Experiments with a preliminary version of this model [79] indicate that many BP fixed point can coexist in absence of information, each one being interpreted as

specific congestion pattern on the network i.e. macroscopic traffic states. This situation can be analyzed [78] as follows by assuming the traffic distribution as a simple probabilistic mixture of P patterns:

$$P_{\text{hidden}}(\tau) \stackrel{\text{def}}{=} \frac{1}{P}\sum_{q=1}^{P}\prod_{\nu\in\mathcal{V}} p_\nu^q(\tau_\nu). \tag{8.3.1}$$

The single sites probabilities $p_\nu^q \stackrel{\text{def}}{=} p_\nu^q(1)$, corresponding to each pattern q, are generated randomly as i.i.d. variables $p_\nu^q = \frac{1}{2}(1+\tanh h_\nu^q)$ with h_ν^q uniformly distributed in some fixed interval $[-h_{max}, +h_{max}]$. The mean of p_ν^q is therefore $1/2$ and its variance denoted by $v \stackrel{\text{def}}{=} \frac{1}{4}\mathbb{E}_h\big(\tanh^2(h)\big) \in [0, 1/4]$. We perform reconstruction experiments, where given a randomly sampled configuration from (8.3.1), the variables $\tau_{\mathcal{V}^*}$ are gradually revealed in a random order and conditional predictions for the remaining unknown variables are computed. We then compare the beliefs obtained with the true conditional marginal probabilities $P(\tau_\nu = \tau|\tau_{\mathcal{V}^*})$ computed with (8.3.1). A sample test shown on Figure 8.3.1.b



(a) (b)

Figure 8.3.1: (a) Phase diagram of the Hopfield model and optimal points found experimentally.(b) D_{KL} error as a function of observed variables ρ for the single parameter model with $N = 1000$ and $P = 20$. K is the mean connectivity.

indicates for example that, on a system with 10^3 variables, it is possible with our model to infer with good precision a mixture of 20 components by observing 5% of the variables. To interpret these results, letting $s_\nu = 2\tau_\nu - 1$, we first identify the parameters of the corresponding Ising model at temperature T with Hamiltonian given by

$$H[\mathbf{s}] \stackrel{\text{def}}{=} -\frac{1}{2}\sum_{\nu\nu'} J_{\nu\nu'}s_\nu s_{\nu'} - \sum_\nu h_\nu s_\nu.$$

Identifying the parameters and considering the limit $P \gg 1$, $N \gg P$ and fixed average connectivity K, we get asymptotically a mapping to the Hopfield model introduced in Chapter 2. The relevant parameters in this limit are $\alpha = P/N$

and the variance $v \in [0, 1/4]$ of the variable bias in the components of the mixture. In this limit, the Hamiltonian is indeed similar to the one governing the dynamics of the Hopfield neural network model:

$$H[\mathbf{s}] = -\frac{1}{2N} \sum_{\nu,\nu',q} u_\nu \xi^q_\nu u_{\nu'} \xi^q_{\nu'} s_\nu s_{\nu'} - \sum_{\nu,q} h^q_\nu \xi^q_\nu s_\nu,$$

$$\text{with} \quad \xi^q_\nu \stackrel{\text{def}}{=} \frac{p^q_\nu(1) - \frac{1}{2}}{\sqrt{v}} \quad \text{and} \quad h^q_\nu = \frac{P}{2\beta K\sqrt{v}} - \frac{2P\sqrt{v}}{K} \sum_{\nu' \in \nu} \mathrm{Cov}(\xi^q_\nu, \xi^q_{\nu'}),$$

and the temperature given by the mapping reads $T = \frac{P}{4\beta v K}$. The coefficients u_ν are the components of the Perron vector (normalized to $\sqrt{N}$), associated to the largest eigenvalue K of the incidence matrix. The various models that

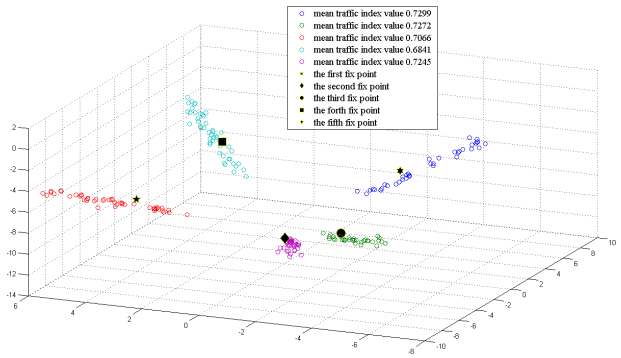


Figure 8.3.2: Segmentation and BP fixed point identification for synthetic travel time data corresponding to a mixture with five components with internal correlations.

we obtain are consistently located in the retrieval phase as shown on the phase diagram on Figure 8.3.1.a.

An example of BP fixed points associated to the mixture's components is given on Figure 8.3.2. Note that in this figure, the 3-d projection space corresponds to the first principal components of the travel time vectors. The set of beliefs corresponding to each fixed point is converted into travel time through the inverse mapping given in 8.1.6 and projected on this 3-d space.

Chapter 9

Dealing with cycles via dual representations

This chapter is based on the following papers:

C. Furtlehner, Approximate inverse Ising models close to a Bethe reference point. *J.Stat.Mech.* (2013), P09020.

C. Furtlehner and A. Decelle, Cycle-based cluster variational method for direct and inverse inference. *J.Stat.Phys. 164*, 3 (2016), 531–574.

The approach based on the Ising model combined with belief propagation of the preceding chapter would be to some extent exact with tree-like factor graphs. Both the inverse Ising problem and the inference task with BP can be performed exactly in that case. Unfortunately the dependencies between variables requires more complex Ising models with denser graphs to be efficient. In this chapter we turn to a more theoretical study in order to deal with loop corrections both for the inference algorithm and the inverse Ising problem.

9.1 Dual representation via cycle basis

In absence of external fields a traditional way to deal with the low temperature regime is given by a duality transformation [199] which for the Ising model coincide with the so-called high temperature expansion by rewriting

$$e^{J_{ij}s_is_j} = \cosh(J_{ij})(1+\tanh(J_{ij})s_is_j). \tag{9.1.1}$$

This leads to re-express the partition function as:

$$Z(\mathbf{J}) = Z_0 \times \sum_{\{\tau_{ij}\in\{0,1\}\}} \prod_{ij} \left(\bar{\tau}_{ij} + \tau_{ij}\tanh(J_{ij})\right) \prod_i \mathbb{1}_{\{\sum_{j\in\partial i}\tau_{ij} = 0 \mod 2\}},$$

with

$$Z_0 = \prod_{(ij)} \cosh(J_{ij}).$$

The summation over bond variables $\tau_{ij} \in \{0,1\}$ ($\tau_{ij} \stackrel{\text{def}}{=} 1-\tau_{ij}$), corresponds to choosing one of the 2 terms in the factor (9.1.1). The summation over spin variables then selects bonds configurations having an even number of bonds $\tau_{ij}=1$ attached to each vertex i. From this condition it results that the paths formed by these bonds must be closed. The contribution of a given path is simply the product of all bond factor $\tanh(J_{ij})$ along the path. As such the partition function is expressed as

$$Z(\mathbf{J}) = Z_0 \times Z_{loops}$$

with

$$Z_{loops} \stackrel{\text{def}}{=} \sum_{\ell} Q_\ell,$$

where the last sum runs over all possible closed loops $\mathcal{G}_\ell$, i.e. subgraphs for which each vertex has an even degree, including the empty graph and

$$Q_\ell \stackrel{\text{def}}{=} \prod_{(ij)\in\mathcal{E}_\ell} \tanh(J_{ij}),$$

where $\mathcal{E}_\ell$ denotes the set of edges involved in loop $\mathcal{G}_\ell$. This is a special case of the loop expansion around a belief propagation fixed point proposed by Chertkov and Chernyak in [31, 30]. Loops which contribute have a simple combinatorial

structure once a cycle basis is given (see Chapter 1). If we associate dual variables $\sigma_c \in \{-1, 1\}$ to each element c of the cycle basis we end up with the dual expression for the partition function:

$$Z = \sum_{\sigma} \prod_{c \in \mathcal{C}_f} \sigma_c \exp\Big(\sum_{e \in \mathcal{E}} J_e \prod_{c \in e} \sigma_c\Big).$$

with the dual coupling

$$J_e = \log\big(|\tanh(J_e)|\big),$$

generalizing to arbitrary graphs an observation made long time ago in [68]. Given this factorized form it is natural to formulate a belief propagation on the dual graph whose nodes are given by the elements of the cycle basis and the factors by the edges of the primal graph contained in more than one cycles [72]. The extension of such considerations to arbitrary pairwise models with local fields led us to consider a generalized belief propagation (see Chapter 3) based on such cycle basis.

9.2 Cycle based Kikuchi approximation

The idea is to define a GBP with regions attached to the elements of a cycle basis. This is motivated by the observation that the Bethe approximation violates the "global unit sum rule" (3.3.1) for counting numbers, except on singly connected graphs, precisely by an amount corresponding to the cyclomatic number of the graph. Completing the regions set with elements of a cycle basis restores the unit sum rule property [235].

As explained in Section 3.3 all mean-field type approximations underlying BP or GBP, consist in assuming a factorized form of the joint measure in term of some of its marginal distributions. So given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a cycle basis $\mathcal{C}$, the CVM with maximal clusters representing cycle basis elements leads to consider the following factorization of the joint measure:

$$P_{\text{GBP}}(\mathbf{x}) = \prod_{c \in \mathcal{C}} p_c(\mathbf{s}_c) \prod_{\ell \in \mathcal{E}} p_\ell(\mathbf{s}_\ell)^{\kappa_\ell} \prod_{v \in \mathcal{V}} p_v(s_v)^{\kappa_v}, \tag{9.2.1}$$

where p_c, p_ℓ and p_v are marginal probabilities respectively associated with cycles, links and single variables, with corresponding arguments respectively noted $\mathbf{s}_c$, $\mathbf{s}_\ell$ and s_v. The probability p_c associated with a cycle is itself expressed as a pairwise MRF, with each factor corresponding to one edge of the cycle:

$$p_c(\mathbf{s}_c) = \prod_{\ell \in c} \varphi_\ell(\mathbf{s}_\ell). \tag{9.2.2}$$

In (9.2.1) counting numbers respectively of cycles, edges and vertices are set to $\kappa_c = 1$, $\kappa_\ell = 1 - d^\star_\ell$ and $\kappa_v = 1 - \sum_{c \ni v} \kappa_c - \sum_{\ell \ni v} \kappa_\ell$. $d^\star_\ell$ is the number of cycles in $\mathcal{C}$ containing edge ℓ.

Given a reference graphical model defined on $\mathcal{G}$,

$$P(\mathbf{x}) = \prod_{\ell \in \mathcal{E}} \psi_\ell^0(\mathbf{x}_\ell) \prod_{v \in \mathcal{V}} \phi_v(x_v), \tag{9.2.3}$$

and an associate cycle basis, we define a dual bipartite graph $\mathcal{G}^\star = (\mathcal{V}_c^\star, \mathcal{V}_t^\star, \mathcal{E}^\star)$, where $\mathcal{V}_c^\star$ are the dual nodes or cycle-nodes, elements of $\mathcal{V}_t^\star$ represent connected intersection between cycles, i.e. either single nodes, links or sub-trees corresponding to bridges connecting distant cycles. Elements of $\mathcal{E}^\star$ connect intersecting elements of $\mathcal{V}_c^\star$ and $\mathcal{V}_t^\star$ (see Figure 9.2.1). The variational problem that

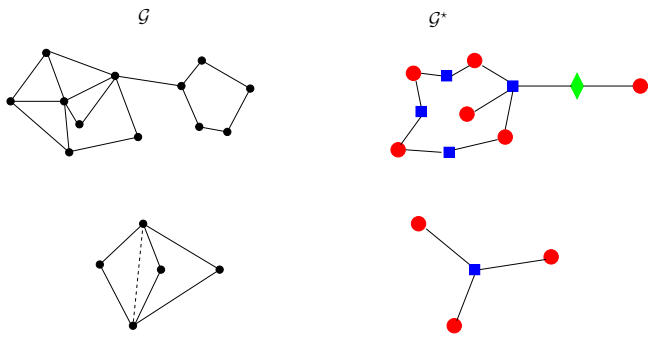


Figure 9.2.1: Dual graph construction. Dashed link correspond to one virtual link introduced to eliminate dual loops.

GBP aims at solving, is to find the closest distribution of the form (9.2.1) to the reference distribution (9.2.3), in the sense of the Kullback-Leibler divergence $D_{\mathrm{KL}}(P_{\mathrm{GBP}} \| P)$. It is a simple matter to convince oneself that when $\mathcal{G}^\star$ is acyclic, the factorization (9.2.1) is exact. This minimization problem concern the set of marginals p_v, p_ℓ and p_c obeying compatibility conditions among each others, which can be solved in dual form. After introducing three sets of Lagrange multipliers, $\lambda_{c\ell}(\mathbf{x}_\ell)$, $\lambda_{\ell v}(x_v)$ and $\lambda_{cv}(x_v)$ to enforce respectively cycle-edge, edge-variable and cycle-variable marginals compatibility, the minimum is then reparameterized as:

$$\begin{cases} p_c(\mathbf{x}_c) \propto \Psi_c(\mathbf{x}_c) \exp\big[\sum_{\ell \in c} \lambda_{c\ell}(\mathbf{x}_\ell) + \sum_{v \in c} \lambda_{cv}(x_v)\big] \\ p_\ell(\mathbf{x}_\ell) \propto \psi_\ell(\mathbf{x}_\ell) \exp\big[\frac{1}{\kappa_\ell}\big(\sum_{v \in \ell} \lambda_{\ell v}(x_v) - \sum_{c \ni \ell} \lambda_{c\ell}(\mathbf{x}_\ell)\big)\big] \\ p_v(x_v) \propto \phi_v(x_v) \exp\big[-\frac{1}{\kappa_v}\big(\sum_{c \ni v} \lambda_{cv}(x_v) + \sum_{\ell \ni v} \lambda_{\ell v}(x_v)\big)\big] \end{cases}$$

9.3 Counting numbers and dual loops

Before turning to the generalized belief propagation algorithm allowing one to find these Lagrange multipliers, let us make a remark concerning the topology of

the dual graph. The counting number κ_v contains some information about the local structure of the dual graph. In order to unravel it we define the local dual graph $\mathcal{G}_v^\star \subset \mathcal{G}^\star$ attached to v as $\mathcal{G}_v^\star = (\mathcal{V}_{v;c}^\star, \mathcal{V}_{v;t}^\star, \mathcal{E}_v^\star)$, where $\mathcal{V}_{v;c}^\star$ are dual vertices corresponding to cycles containing v; $\mathcal{V}_{v;t}^\star$ are dual vertices corresponding to all edges containing v with non-zero counting number; $\mathcal{E}_v^\star$ is the set of dual edges

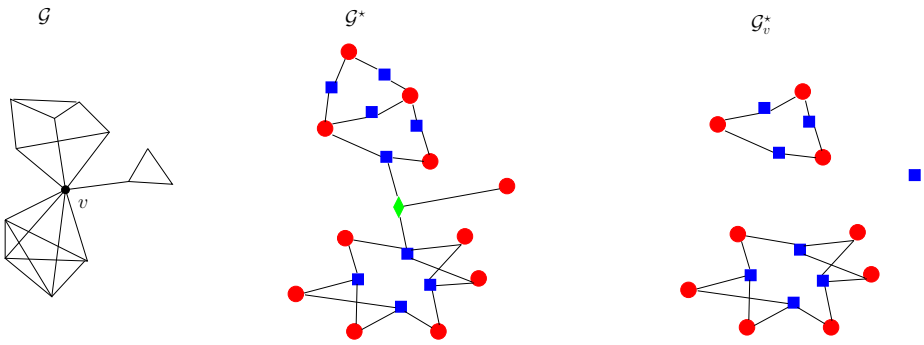


Figure 9.3.1: Local dual graph construction. In this case the choice of cycle basis leads to $\kappa_v = 2$ with $d_v^\star = 3$ and $\mathcal{C}_v^\star = 4$.

connecting ℓ-nodes in $\mathcal{V}_{v;t}^\star$ to their corresponding c-nodes in $\mathcal{V}_{v;c}^\star$ they belong to in the primal graph. This construction is illustrated on Figure 9.3.1

Proposition 9.3.1. *Let $d_v^\star$ be the number of components of $\mathcal{G}_v^\star$ and $\mathcal{C}_v^\star$ its cyclomatic number. We have*

$$\kappa_v = 1 - d_v^\star + \mathcal{C}_v^\star. \tag{9.3.1}$$

Intuitively $\mathcal{C}_v^\star$ represents the number of dual cycles "centered" on v. This decomposition will prove useful for building our cycle based region graph.

For instance we have $\mathcal{C}_v^\star = 1$ for nodes in the bulk of a planar graph we have, $\mathcal{C}_v^\star = d(d-1)/2$ on a d-dimensional square lattice. Using regular cycle basis, we have on a $N/2 + N/2$ bipartite graph typically $\mathcal{C}_v^\star = 3N/2 - 1$.

9.4 Generalized cycle based belief propagation (GCBP)

At this point, following the region-based algorithm [242] prescriptions, a message passing algorithm can be set-up which rules are associated with the Hasse diagram of the regions hierarchy. As noticed in [183], dependencies between Lagrange multipliers are present in the parent-to-child algorithm. This results in a more complex factor graph with more feed-back loops than necessary which in turn may cause convergence failures of GBP. In [183] a minimal graphical representation construction is proposed to settle such problems, in order to eliminate all redundant Lagrange multipliers. In our setting this leads to an essentially

unstable algorithm for graphs containing at least one single dual loop. This problem of redundant Lagrange multipliers has actually also been discussed in the context of the 2-D Edward Anderson (EA) model in [54]. In this context the authors propose a solution based on a specific gauge choice for the message definition in order to regularize GBP. Our approach to this problem is different. It is solely based on topological properties of the graph of interactions, yielding a generic method independent of the graph or the type of interactions.

We introduce here a specification of the region graph which on the one hand eliminates all unnecessary feed-back loops present in the parent-to-child algorithm, but on the other hand prevents instabilities associated with dual loops. Additional "clone variables" need to be introduced for variables at the center of dual loops, i.e. for which $C_v^\star \neq 0$, as defined in Section 9.3, to prevent some instability. The region graph which we refer to as the mixed factor graph (MFG) has the following specifications:

- (i) Each term in (9.2.1) having a non-zero counting number is associated with a node in the MFG. There are three families of nodes, c-nodes, ℓ-nodes and v-nodes, respectively associated with cycles, links and vertices of the original graph. c-nodes are always factors while v-nodes are always variables. Instead, ℓ-nodes associated with links are composite nodes, i.e. can be of both types.
- (ii) Edges of the MFG represent Lagrange multipliers and relate variables to factors. A v-node can be linked to ℓ-nodes, considered then as factors nodes. ℓ-nodes considered as variable nodes can be linked to c-nodes.
- (iii) all links of a given cycle c with non-vanishing counting numbers are linked as variables to this c-node.
- (iv) to a variable v we associate in general two types of v-nodes depending on $d_v^\star$ and $\mathcal{C}_v^\star$ defined in Section 9.3:
  - (a) if $d_v^\star > 1$ one v-node is associated with v, which connects exactly to one single arbitrary ℓ-node of each components of $\mathcal{G}_v^\star$, its degree being therefore $d_v^\star$ and a counting number of $1 - d_v^\star$ is attributed to it. If necessary an ℓ-node with zero counting number can be inserted into the MFG in order to ensure that this v-node is properly connected to all components it needs to be.
  - (b) if $\mathcal{C}_v^\star > 0$, to each ℓ containing v we associate one $v^\star$-node that is singly connected to ℓ as long as this ℓ-node is in a component of $\mathcal{G}_v^\star$ containing at least one dual loops. Each clone is attributed a counting number $\kappa_{v^\star} = \mathcal{C}_v^\star/q$ if q is the number of clones.

This set of rules is illustrated on Figure 9.4.1. Rule (iii) ensures that all marginal probabilities of cycles are compatibles at link intersections. Rule (iv)(a) is applied to cut-vertices, i.e. vertices which separate $\mathcal{G}$ in multiple components when removed as shown on the example of Figure 9.4.1. Rule (iv)(b) is there to take into account dual loop corrections. The prescription (iv)(b) is there to ensure a better convergence of GCBP by making use of replicas of v-nodes, while preserving the minimal use of Lagrange multipliers.

After a change of variables, as a direct generalization of the one used in BP,

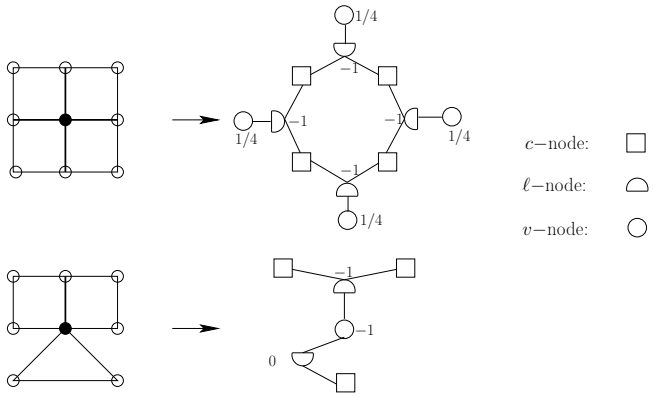


Figure 9.4.1: Pairwise MRF (left). Variables and links with non-zero counting number are in bold. Corresponding mixed factor graph (right) with counting numbers.

we get the following expression for the beliefs

$$p_v(x_v) = \phi_v(x_v)\exp\big[-\frac{1}{1-d_v^\star}\sum_{\ell\ni v}\lambda_{\ell v}(x_v)\big] = \phi_v(x_v)\prod_{\ell\ni v} m_{\ell\to v}(x_v),$$

$$p_{v^\star}(x_v) = \phi_v(x_v)\exp\big[-\frac{1}{\kappa_{v^\star}}\lambda_{\ell_{v^\star}v^\star}(x_v)\big] = \phi_v(x_v) m_{\ell_{v^\star}\to v^\star}(x_v),$$

$$p_\ell(\mathbf{x}_\ell) = \psi_\ell(\mathbf{x}_\ell)\exp\big[\frac{1}{\kappa_\ell}\big(\sum_{v\in\ell}\lambda_{\ell v}(x_v) - \sum_{c\ni\ell}\lambda_{c\ell}(\mathbf{x}_\ell)\big)\big] = \psi_\ell(\mathbf{x}_\ell)\prod_{c\ni\ell} m_{c\to\ell}(\mathbf{x}_\ell)\prod_{v\in\ell} n_{v\to\ell}(x_v),$$

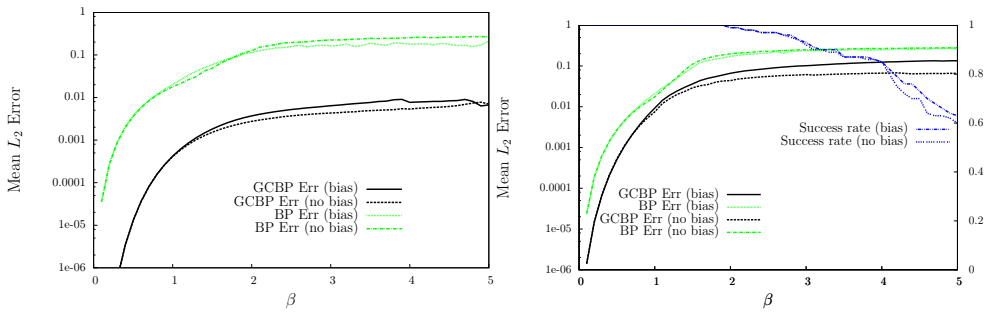
$$p_c(\mathbf{x}_c) = \Psi_c(\mathbf{x}_c)\exp\big[\sum_{\ell\in c}\lambda_{c\ell}(\mathbf{x}_\ell)\big] = \Psi_c(\mathbf{x}_c)\prod_{\ell\in c}\big[n_{\ell\to c}(\mathbf{x}_\ell)\prod_{v\in\ell} n_{v\to\ell}(x_v)\big], \tag{9.4.1}$$

where $\ell_{v^\star}$ denotes the ℓ-node connected to $v^\star$. From this we get the following message passing rules:

$$m_{c\to\ell}(\mathbf{x}_\ell) \longleftarrow \sum_{\mathbf{x}_c\setminus\mathbf{x}_\ell}\frac{\Psi_c(\mathbf{x}_c)}{\psi_\ell(\mathbf{x}_\ell)}\prod_{\ell'\in c\setminus\ell}\big[n_{\ell'\to c}(\mathbf{x}_{\ell'})\prod_{v\in\ell'} n_{v\to\ell'}(x_v)\big], \tag{9.4.2}$$

$$m_{\ell\to v}(x_v) \longleftarrow \sum_{\mathbf{x}_\ell\setminus x_v}\frac{\psi_\ell(\mathbf{x}_\ell)}{\phi_v(x_v)}\times\prod_{c\ni\ell} m_{c\to\ell}(\mathbf{x}_\ell)\prod_{v'\in\ell\setminus v} n_{v'\to\ell}(x_{v'}), \tag{9.4.3}$$

$$m_{\ell\to v^\star}(x_v) \longleftarrow \Big(\sum_{\mathbf{x}_\ell\setminus x_v}\frac{\psi_\ell(\mathbf{x}_\ell)}{\phi_v(x_v)}\times\prod_{c\ni\ell} m_{c\to\ell}(\mathbf{x}_\ell)\prod_{v'\in\ell\setminus v^\star} n_{v'\to\ell}(x_{v'})\Big)^{1/(1+\kappa_{v^\star})}. \tag{9.4.4}$$



(a) (b)

Figure 9.4.2: Mean error for the direct inference of 2-D random Ising model comparing GCBP with BP as a function of β, on a 5×5 square grid (left) and on random $20 + 20$ bipartite graphs of mean connectivity 4 (right) with or without local fields averaged over 100 instances. Couplings J_{ij} and local fields h_i are i.i.d sampled uniformly respectively in the range $[-\beta, \beta]$ and $[-0.2\beta, 0.2\beta]$

With this formulation GCBP can be seen mainly as an ordinary belief propagation defined on the MFG, where (9.4.2,9.4.3) are direct generalization on a MFG of ordinary BP update rules (3.2.2,3.2.3), with an additional peculiarity given by dual loop corrections carried by clone variables in (9.4.4). For non-regular graph a cycle basis has to be determined algorithmically. The best choice is a basis which leads to the smaller possible number of dual loops. As shown in [73] this is provided by the minimum cycle basis (MCB), i.e. the one having the lowest mean cycle size. Albeit exact polynomial algorithms do exist [127], we resort to an heuristic with quadratic complexity in $|\mathcal{V}|$ providing us with good enough approximate MCB.

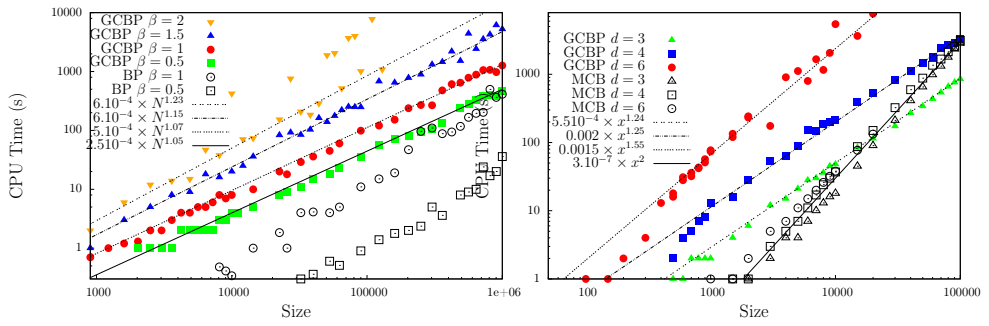


Figure 9.4.3: (left) Convergence behaviour of GCBP and BP regarding computational time on 2-D EA models of large sizes. Cases corresponding to $\beta = 0.5, 1$ have local random fields in $[-0.1\beta, 0.1\beta]$ while other cases are without external fields. (Right) Computational times of GCBP and the MCB search algorithm on random bipartite graphs at $\beta = 1$ for different mean connectivity d.

9.5 Loop corrections and associate inverse Ising algorithm

In this framework, the loop corrections are represented by the c-node to ℓ-node messages. To compute them we need to solve explicitly the Ising model on a loop:

$$P_c(\mathbf{s}) = \frac{1}{Z_c} \exp\Big(\sum_{i=1}^{n} h_i^c s_i + \sum_{i=1}^{n-1} J_i^c s_i s_{i+1}\Big), \tag{9.5.1}$$

where $h_i^c \in \mathbb{R}$ is the local field exerted on variable i and $J_i^c \in \mathbb{R}$ denotes the coupling between s_i and s_{i+1}. Addressed in [187], this problem can actually be solve more directly by starting from the BP factorization of the joint measure:

$$P(\mathbf{s}) = \frac{1}{Z_{\mathrm{BP}}} \prod_{i=1}^{n} \frac{b_i^c(s_i, s_{i+1})}{b_i^c(s_i) b_{i+1}^c(s_{i+1})} \prod_{i=1}^{n} b_i^c(s_i), \tag{9.5.2}$$

where the $b_i^c(\cdot)$ and $b_i^c(\cdot,\cdot)$ are the single and pairwise approximate marginals delivered by BP. These can be parameterized as follows

$$b_i^c(s_i) = \frac{1}{2}(1 + \breve{m}_i s_i), \tag{9.5.3}$$

$$b_i^c(s_i, s_{i+1}) = \frac{1}{4}(1 + \breve{m}_i s_i + \breve{m}_j s_j + (\breve{m}_i \breve{m}_j + \breve{\chi}_i) s_i s_j), \tag{9.5.4}$$

where $m_i \overset{\text{def}}{=} \mathbb{E}(s_i)$ represents the "magnetization" of spin s_i and $\chi_i \overset{\text{def}}{=} \mathbb{E}(s_i s_{i+1}) - \mathbb{E}(s_i)\mathbb{E}(s_{i+1})$ the susceptibility coefficient, between s_i and s_{i+1}. We use the sign $\breve{}$ to denote a BP estimate, which is to be distinguished from the exact value. We then get the following relations

$$Z_{\mathrm{BP}} = 1 + Q, \tag{9.5.5}$$

$$m_i = \frac{1-Q}{1+Q}\, \breve{m}_i, \tag{9.5.6}$$

$$\chi_i = \frac{\breve{\chi}_i}{1+Q} + \frac{Q}{1+Q}\Big(\frac{(1-\breve{m}_i^2)(1-\breve{m}_{i+1}^2)}{\breve{\chi}_i} + 4\frac{\breve{m}_i \breve{m}_{i+1}}{1+Q}\Big), \tag{9.5.7}$$

between BP values and exact ones with

$$Q \overset{\text{def}}{=} \prod_{i=1}^{n} \frac{\breve{\chi}_i}{\sqrt{(1-\breve{m}_i^2)(1-\breve{m}_{i+1}^2)}}. \tag{9.5.8}$$

The corresponding loop corrected marginals p_i and p_{ii+1} are expressed from the loop corrected quantities (m_i, m_{i+1}, χ_i) through the same relations (9.5.3) and (9.5.4) and allow one to obtain all messages 9.4.2 sent by the c-node at once from the BP beliefs, so the cost per-message in this special case is now $O(1)$ instead of $O(n)$ if there are n messages to be sent.

In addition to this slight but non-crucial reduction in computational cost, one should note the scalar characterization in terms of $Q \in]-1,1]$ of the cycle which shows up. It is the product of "BP correlations" along the loop and characterizes its strength.

- $Q \simeq 0$ corresponds to weak loop correction, BP is nearly exact.
- $Q \to 1$ corresponds to a strongly correlated loop.
- $Q \to -1$ corresponds to a strongly correlated frustrated loop.

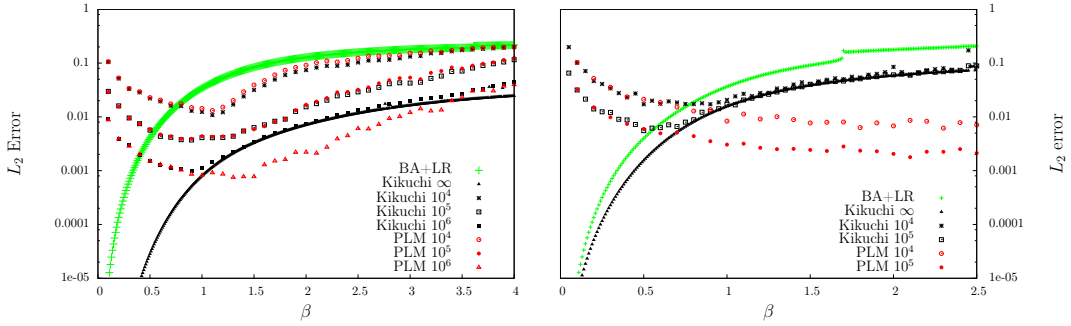


Figure 9.5.1: KIC compared with BA+LR (4.1.27) at infinite and with Pseudo Likelihood Methods at finite sampling on a 5×5 square grid (left) and on a bipartite model with connectivity 3(right) with varying β.

Equations (9.5.6,9.5.7) can be inverted which opens up the possibility for a simple inverse mean-field Ising algorithm (see Chapter 4) based on the explicit expression of the Kikuchi type approximation (9.2.1). Assuming we know the graph structure and have a cycle basis, it remains to determine the marginal probabilities p_c, p_ℓ and p_v associated with each region. We expect the p_ℓ's and p_v's to be given from the data, but the p_c's have to be constructed. This means that the global inverse problem is decomposed into $|\mathcal{C}|$ small inverse problems. In the Ising case, if we denote by h_i^c and J_ℓ^c the local field and coupling associated as in (9.5.1) with the marginal representing cycle c, $\hat{h}_i^\ell$, $\hat{J}_\ell$ associated with p_ℓ and finally $\hat{h}_i$ to p_i, then from (9.2.1) the corresponding Kikuchi cycle based (KIC) approximate inverse Ising solution reads

$$h_i^{(\mathrm{KIC})} = \kappa_i \hat{h}_i + \sum_{c \ni i} h_i^c + \sum_{\ell \ni i} (1 - d_\ell^\star) \hat{h}_i^\ell,$$

$$J_\ell^{(\mathrm{KIC})} = (1 - d_\ell^\star) \hat{J}_\ell + \sum_{c \ni \ell} J_\ell^c.$$

Obtaining h_i^c and J_ℓ^c for any cycle $c \in |\mathcal{C}|$ is then a matter of inverting equations (9.5.6,9.5.7). This can be done efficiently in practice by combining a fixed point method with a line search optimization of the log likelihood associated to that loop.

Chapter 10

Gaussian copula model

This chapter is based on the following papers:

V. Martin, C. Furtlehner, Y. Han and J.M. Lasgouttes, GMRF estimation under topological and spectral constraints. *In Proceedings of ECML PKDD* (2014), pp. 370–385.

C. Furtlehner, J.M. Lasgouttes, A. Attanasi, L. Meschini, and M. Pezzulla, Spatio-temporal Probabilistic Short-term Forecasting on Urban Networks, Research Report RR-9236, INRIA, 2018 (submitted).

In addition to Ising based models, we have developed over the years a second class of models compatible with Gaussian belief propagation, suitable as well to fast inference for traffic forecasting. The model is basically a sparse Gaussian copula learned with help of a specific algorithm called $\star$-IPS [157]. It uses a specific encoding of the data able to cope with daytime and seasonal variations, and has proven very efficient on various real world dataset for traffic forecasting with a prediction horizon of up to many hours [76].

10.1 $\star$-IPS for sparse inverse BP-compatible covariance matrices

Multivariate Gaussian distributions constitute a second type of MRF on which BP -then called GaBP- can be defined to run without approximations (see Chapter 3). Since GaBP may often encounter convergence issues, especially with non-sparse structures, it can be of practical interest to construct off-line a Gaussian MRF which is compatible with GaBP. By combining various methods proposed in the context of sparse inverse covariance matrix estimation [13, 71, 201] mentioned in Chapter 4, a way to do that as been elaborated in [157] in the form of the $\star$-IPS algorithm. The starting point is the likelihood maximization

$$\mathcal{L}(A) = \log\det(A) - \mathrm{Tr}(A\hat{C})$$

of the precision matrix A, given some covariance empirical matrix $\hat{C}$. Without any constraint on A, the maximum likelihood solution is trivially $A = \hat{C}^{-1}$. In our context, where compatibility with GaBP has to be imposed, one feature like sparsity can be desirable, albeit without much guarantee. Indeed, specific topological properties like the presence of short loops, are likely to damage the GaBP compatibility, even on a sparse graph. Additional spectral properties, e.g. walk-summability [153], can guarantee the compatibility with GaBP-based inference. $\star$-IPS incorporates these explicitly, by combining an approach based on the iterative proportional scaling (IPS) procedure [40, 212], with block-updates techniques used in [13, 71]. The rationale of $\star$-IPS is to construct the graphical model $P(\mathbf{x})$ link by link, by ensuring at each step that the constraints are satisfied. If P is the current approximate model after some steps, it turns out that

$$P'(\mathbf{x}) = P(\mathbf{x}) \times \frac{\hat{p}_{ij}(x_i, x_j)}{p_{ij}(x_i, x_j)}, \tag{10.1.1}$$

is the optimal deformation of link (i, j), where $\hat{p}_{ij}$ is the empirical pairwise marginal, while p_{ij} is the pairwise marginal of P. The corresponding log-likelihood gain is given by $\Delta\mathcal{L} = D_{KL}(\hat{p}_{ij} \| p_{ij})$. Sorting all the candidate new links w.r.t. this quantity yields the optimal 1-link correction to be made. In terms of precision matrix modification, this corresponds to a 2×2 update which involves the current covariance matrix $C = A^{-1}$ of the approximate model. This covariance matrix has to be maintained after each update, which can be done efficiently thanks to the Sherman–Morrison–Woodbury formula for low

rank modifications of the precision matrix A. Direct inspection of the modified precision matrix, shows that positive definiteness of the matrix is preserved by such updates.

Each modification is accepted only if it satisfies the constraints. The best candidate link can thus be discarded if the constraints are violated by this addition. Two families of constraints are considered:

- Topological constraints avoid the presence of small loops, with possibly the distinction between frustrated/non-frustrated loops, i.e. loops along which the product of partial covariances $(-A_{ij})$ is negative.

- Spectral constraints like walk-summability [resp. weak walk-summability] involve definite positiveness of matrix $\mathrm{Diag}(A) - |A - \mathrm{Diag}(A)|$ [resp. $2\,\mathrm{Diag}(A) - A$], where $\mathrm{Diag}(A)$ is the matrix containing only the diagonal elements of A.

When a new link is added, existing links can become detuned by a slight amount. In order to optimize existing links, (10.1.1) can be used. Other local updates are also available like block updates, via a single row-column update of the precision matrix, as originally proposed in [13] and refined in [71]. In practice, $\star$-IPS alternates many link additions, corresponding to significant mean connectivity increase, with block coordinate descent procedures. Overall, sparse precision matrix of good likelihood are generated in $O(N^3)$ steps, with the advantage of having available all the optimization path, by mean of many graphical models of intermediate connectivity. Note finally that we have discarded the standard way for generating sparse precision matrix based on Lasso penalty [71, 108]. There are two reasons for that: firstly the L_1 norm penalty suffers from a modeling bias, due to excessive penalization of truly large magnitudes entries of A; secondly it is not flexible enough for the kind of constraints we are interested in, in order to produce graphical models compatible with GaBP of high likelihood [157].

10.2 Gaussian copula models of traffic indexes

We turn now to our traffic inference model. A key feature of our method is a mapping of raw data, that can correspond to flow or speed for instance, to a standard normal variable, a kind of properly normalized traffic index on which the prediction is performed. The joint probability measure of these traffic indexes is approximated through a Gaussian copula. We define this index in the following way. Let t be a discretized time, measured in time steps δ_t of fixed length. N_t represents the number of such time steps contained in a single day. For a given t, the daytime $\tau \in \{0, \dots N_t - 1\}$ is given by $\tau = t$ modulo N_t. At each time step, we assume the system to be represented by N_v variables X_i^t, $i \in \{1, \dots N_v\}$ corresponding to traffic detectors. Now for each variable X_i and each day time τ we build from the historical data a running average $\bar{X}_i^\tau$ and a variance V_i^τ. If the dataset is clustered into a certain number of weekly or

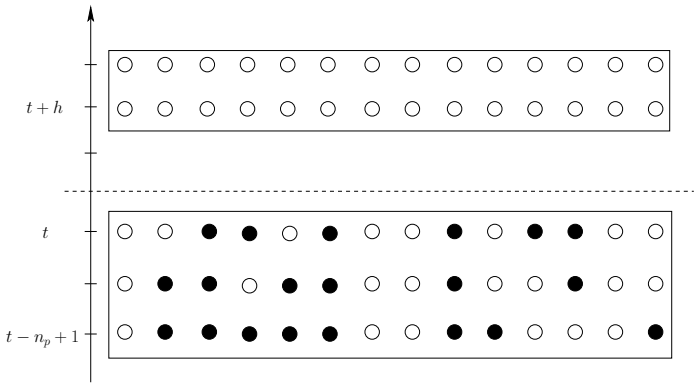


Figure 10.2.1: Time layers setting of one basic space time configuration. Each circle represents one single variable at a given time, filled circles correspond to observations.

seasonal patterns, then these quantities will be estimated for each cluster labeled by some extra index ℓ. Then, for each variable index i, time t (and associated day time τ) and cluster label ℓ, we map $X_i^{t,\ell}$ to the following variable

$$U_i^{t,\ell} \stackrel{\text{def}}{=} \frac{X_i^{t,\ell} - \bar{X}_i^{\tau,\ell}}{\sqrt{V_i^{\tau,\ell}}}, \tag{10.2.1}$$

which represents a centered and normalized variable for given τ and ℓ. From this transformed historical data, we build for each index i a single cumulative distribution function

$$F_i(x) = P(U_i < x),$$

which for a given realization $U_i^{t,\ell} = x$ represents the traffic index. The purpose of this index is to encapsulate all average time-dependent trends, week-day and seasonal dependencies, while the Gaussian copula will take care of the fluctuations around these trends. In order to build a sparse Gaussian copula of all the indexes we first transform each variable $U_i^{t,\ell}$ into a normal variable via the following mapping:

$$Y_i^{t,\ell} = F_{\mathcal{N}(0,1)}^{-1} \circ F_i(U_i^{t,\ell}), \tag{10.2.2}$$

where $F_{\mathcal{N}(0,1)}$ is the cumulative distribution of a standard normal variable.

The copula model corresponding to $n + h + 1$ time layers (n past layers, 1 present and h future layers) is then obtained by considering the vector

$$Z^t = (Y_i^{t+k,\ell}, i = 1 \dots N_v, k = -n, \dots, 0, h, \dots)$$

and constructing its associate sparse multivariate approximate model with $\star$-IPS. This model will be used to generate predictions $\hat{Y}_i$, which in turn can be converted into predictions $\hat{X}_i$ of the original variables by inverting (10.2.1,10.2.2).

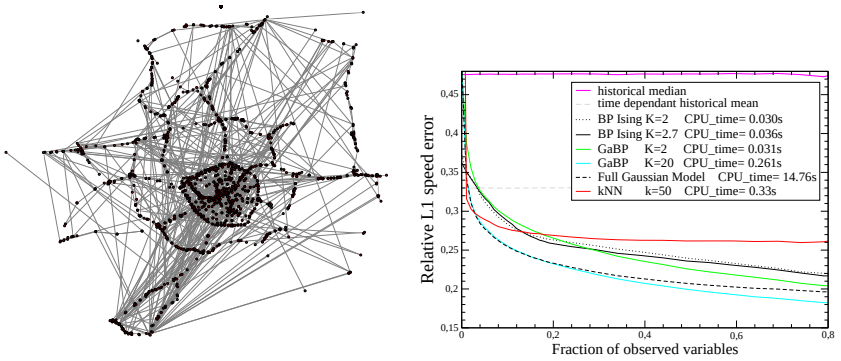


Figure 10.3.1: Experiment on the Sytadin dataset (Île de France) with 1,632 stations. (left) Graph obtained with $\star$-IPS of mean connectivity 2 used by BP-Ising. (right) L^1 relative speed error as a function of the proportion of observed variables. Comparison between BP-Ising, k-NN, GaBP and exact inverse Gaussian on the decimation experiment. Each curve is drawn averaged over 1,000 runs. The indicated running times in seconds represents the mean CPU time needed to perform one inference over the network, averaged over all the runs.

10.3 Experiments on real traffic dataset

First an imputation experiment as been performed on data obtained from the Sytadin platform (See Figure 10.3.1). In this experiment there is one single time layer, a random subset of variables representing a given fraction of all the variables is observed, and a score is computed based on the ability to predict the complementary set of variables being kept hidden.

Forecasting experiments done with GaBP are presented on Figures. 10.3.2 and 10.3.3 for a dataset of the Vienna agglomeration with 263 detectors and on Figures. 10.3.4 and 10.3.5 for a dataset of the Turin agglomeration with 685 detectors, out of which 566 correspond to flow and 119 to speed measurements. The results presented on Figure 10.3.2 are obtained with a model with 4 past layers and one single specialized future layer. Similar results can be obtained with a multi-step ahead model having 4 past and 4 future layers. Specialized or multi-step ahead models yield identical performances within error bars. For sake of comparison are given the performance of the predictor based on the daytime average ($mean(t)$), the one based on the last observed value (t_0) and the one predictor based on the k closest sample observations in the learning dataset (k-NN). A qualitative indication that the GaBP predictor is performing well is obtained by looking at the individual counting location's (CLOC) time series and associated predictions. As an example of the typical behavior of the model, a small sample of these time series is shown on Figure 10.3.3. We can see that GaBP follows very well without any delay the changes in traffic conditions

and realizes a kind of smoothing of the actual traffic flow signal.

The experiments on the Turin dataset yield similar results. The results for flow are clearly more impressive than those for speed predictions. This hides important disparities between various days. In fact the aggregated error for the speed is dominated by nighttime prediction errors, where the small amount of speed measurements leads to a very noisy signal.

On Figure 10.3.5 are shown some excerpts of single detectors prediction time series. As for the Vienna dataset, the model is able to anticipate correctly the changes in traffic flow even far from recurrent traffic conditions. Sudden drops in speed are not always anticipated, as shown on the last panel of this figure for instance.

Finally since GaBP comes with a variance estimate σ_i (3.2.7) of each prediction μ_i given in (3.2.6) we can exploit that feature to deliver levels of confidence on our predictions as shown on the left of Figure 10.3.6. These confidence intervals seems quite consistent and meaningful and may actually help to identify detector errors.

That taken aside, there are systematic errors that our model makes which we would like to identify and possibly cure. The main source of error comes from the Gaussian copula hypothesis. Recall first that, after the transformation (10.2.2) is performed, each $Y_i^{t,\ell}$ taken individually is by construction and up to numerical precision, a standard normal variable. However, the joint distribution has no specific reason in general to be multi-variate Gaussian. In order to estimate how far from a multivariate Gaussian is our model, we consider the main directions of fluctuations of Y by extracting the dominant eigenmodes of the covariance matrix. Then for each of these modes $e_{\cdot,\alpha}$, we compute the corresponding cumulative distribution $P(z_\alpha^{t,\ell}/\sigma_\alpha < x)$ of the components

$$z_\alpha^{t,\ell} = \sum_{i=1}^{N} e_{i,\alpha} Y_i^{t,\ell}$$

of the data normalized by the standard deviation

$$\sigma_\alpha \overset{\text{def}}{=} \sqrt{\mathbb{E}\big[(z_\alpha^{t,\ell})^2\big]}$$

along these modes. On Figure 10.3.6 is shown for the various datasets how this distributions compare to the expected cumulative distribution of a standard normal variable. As we can see, for all datasets the alignment is pretty good for most of the dominant modes over more or less 2 std. Beyond that, the model shows inadequacy in the distributions tails which are clearly non-Gaussian.

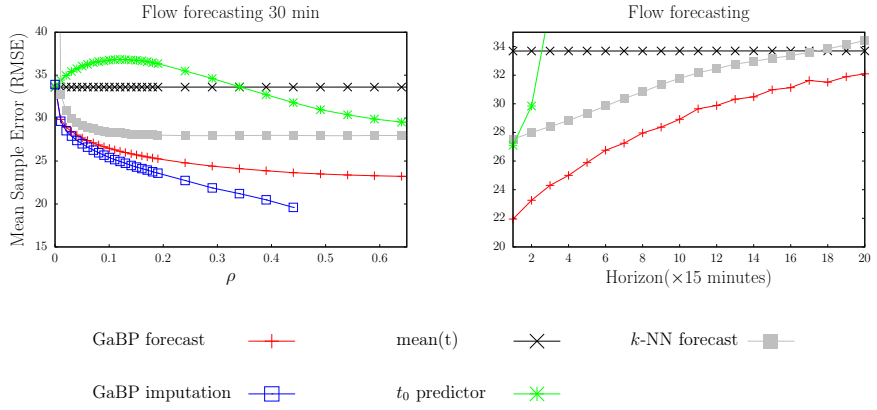


Figure 10.3.2: Vienna dataset: (left panel) average flow forecasting error for a given time lag of 30 minutes, as a function of the fraction ρ of observed variables in the past time layers, averaged over 5000 test samples. The reconstruction error for missing data is also shown (in blue). A point of comparison is given by the k-NN predictor, optimized for $k = 50$. (Right panel) Average flow forecasting error as a function of time lag at maximum possible observation rate $\rho \sim 0.65$ (average over the full test set).

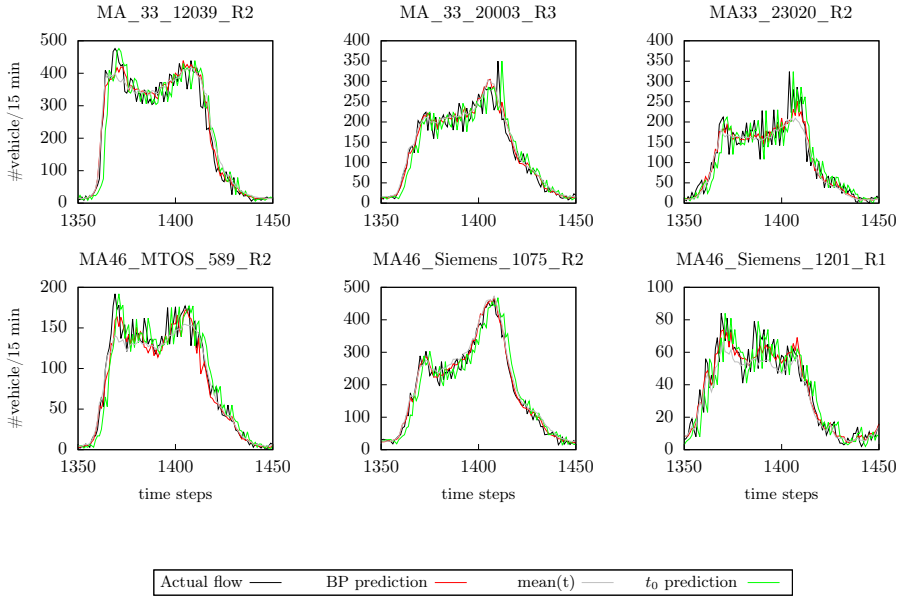


Figure 10.3.3: Vienna dataset: excerpt of flow time-series along with GaBP and t_0 prediction for 30 minutes horizon for 6 different CLOCs.

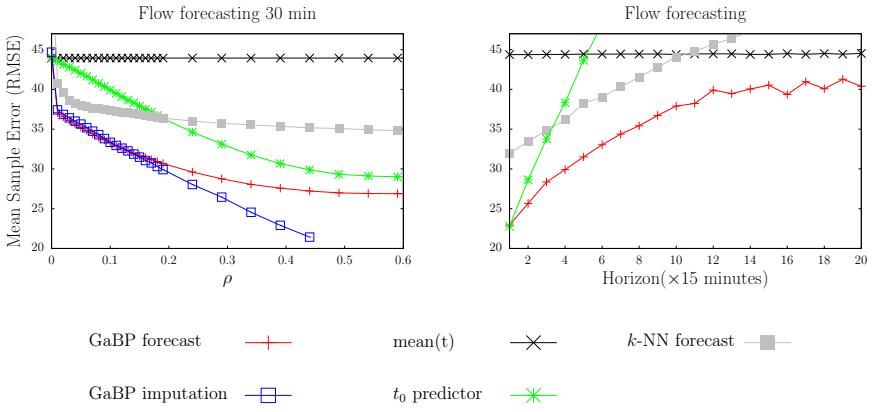


Figure 10.3.4: Turin dataset: (left panel) average flow forecasting error for a given time lag of 30 minutes, as a function of the fraction ρ of observed variables in the past time layers, averaged over 5000 test samples. The reconstruction error for missing data is also shown (in blue). A point of comparison is given by the k-NN predictor with $k = 50$. (Right panel) Average flow forecasting error as a function of time lag (right) at maximum possible observation rate $\rho \sim 0.6$, average over the full test set.

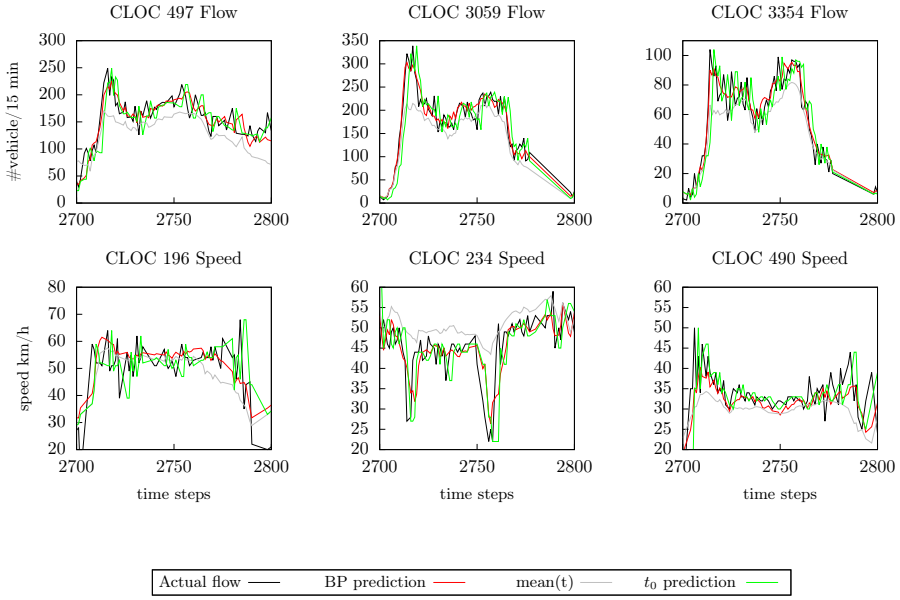


Figure 10.3.5: Turin dataset: excerpt of flow time-series along with GaBP and t_0 prediction for 30 minutes horizon for 6 different CLOCs of flows or speed detectors

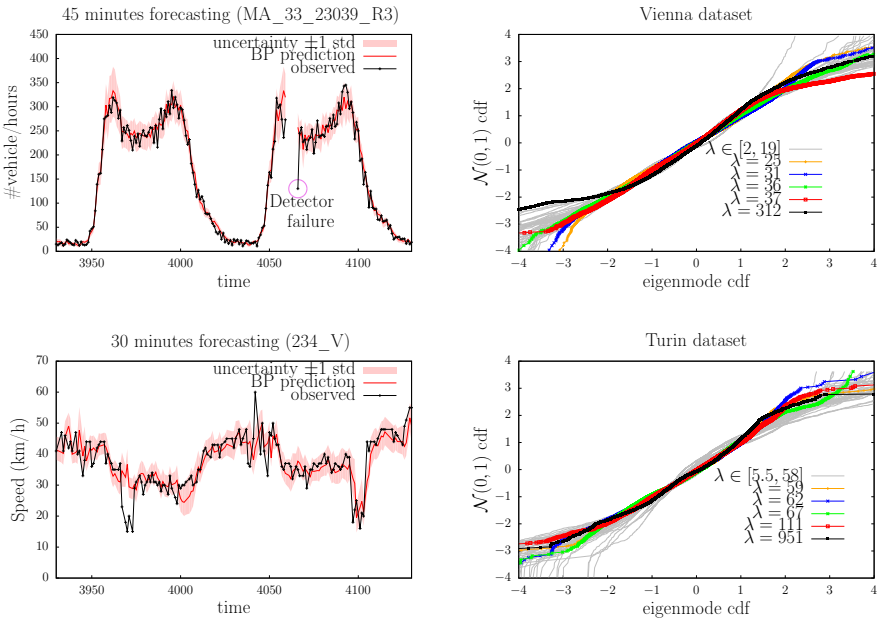


Figure 10.3.6: (left panel) Flow and speed forecast, including uncertainty estimate delivered by GaBP, for a given CLOC respectively in Vienna and Turin. The confidence interval corresponds to 1 std in copula space, corresponding i.e. to 84% of confidence that the true value is within this interval. (Right panel) Empirical cumulative distribution of the normalized projection of the (copula transformed) data along the 50 first principal modes against the cumulative distribution of a standard normal variable. Modes are ordered by decreasing eigenvalues λ_α (From top to bottom: Vienna and Turin)

Part IV

Statistical physics of simple ML algorithms

Chapter 11

Clustering with affinity propagation

This chapter is based on the following papers:

C. Furtlehner, M. Sebag and X. Zhang, Scaling Analysis of Affinity Propagation. Phys.Rev. E 81,066102 (2010) 14pp

X. Zhang,C. Furtlehner, J. Perez, C. Germain-Renaud and M. Sebag, Toward Autonomic Grids: Analyzing the Job Flow with Affinity Streaming. *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* KDD 2009

11.1 Clustering streams of data with AP

The clustering of large-scale dynamic datasets is a key issue for most application domains, at the crossroad of databases, data mining and machine learning [38]. High performance computers and large-size memory storage do not *per se* sustain scalable and accurate clustering. Typically, advances in large-scale clustering (see e.g., [118]) mainly proceed by distributing the dataset and processing the subsets in parallel; when dealing with dynamic datasets, such divide and conquer approaches face some limitations in terms of latency and/or communication costs.

Furthermore, the choice of a clustering method must reflect the applicative needs. The motivating application pertains to the strategic field of Autonomic Computing [194], aimed at providing large computational systems with self-modelling, self-configuring, self-healing and self-optimizing facilities. More specifically, the applicative goal here was to enable the administrator of a large-scale grid system, the EGEE Grid[1], to analyze the flow of jobs submitted to and processed by the grid. The input data thus is made of the Logging and Bookkeeping (L&B) files, automatically generated by the grid middleware. As noted by [83], modern data mining is more and more concerned with automatically generated datasets ("computers are fueling each other"); building understandable summaries thereof is even more critical. For this reason, it is highly desirable that a job cluster be summarized by an actual job (as opposed to an artefact, as done in K-means).

The AP algorithm mentioned in Chapter 2 and detailed in Chapter 3 does satisfy the above interpretability and stability constraint but its quadratic cost requires some adaptation to the streaming context. This leads to a new algorithm called Strap which we describe now. Formally, the stream model is encoded into a set of clusters $C_i = (e_i, n_i, \Sigma_i, t_i)$, where e_i is the cluster exemplar, n_i and Σ_i respectively stand for the cluster size and distortion, and t_i is the last time stamp when a data item joined the cluster. As the stream flows in, current data item e_t is checked against the model. If its distance to the nearest exemplar e_i is less than a threshold computed in the initialization step, e_t joins the C_i cluster. The C_i time stamp is set to the current time step t, while C_i size and distortion are updated by relaxation. The model update is parameterized from a (user supplied) time length Δ; the idea is that clusters which have not received any additional item during Δ consecutive time steps should disappear [247]. If data item e_t does not fit the model, it is considered to be an outlier and put in the reservoir. The reservoir gathers the last M outliers. A change point detection test is used to monitor the stability of the data distribution. The Page Hinkley (PH) statistical test [182, 99] is applied to the outlier rate. Upon triggering the PH test, the stream model is rebuilt using a weighted version (WAP) of AP from the current model (exemplars weighted by the current size of the associated cluster) and the outliers in the reservoir. In

[1] The EGEE grid was established in the EU project *Enabling Grid for E-SciencE*. It involves 41,000 CPUs, 5 Petabytes storage and concurrently supports 20,000 jobs on 24/24, 7/7 basis.

addition a divide and conquer strategy can be combined with AP to avoid an excessive computational cost, resulting in a hierarchical version of AP called HAP. The fact that at each time step the stream model is based on exemplars makes it natural to apply HAP on the overall set of exemplars gathered along time, thus extracting "super-exemplars". These super-exemplars capture the various trends of the data stream along time, enabling to characterize any period (day, week or month) after the representativity of each such super-exemplar. The

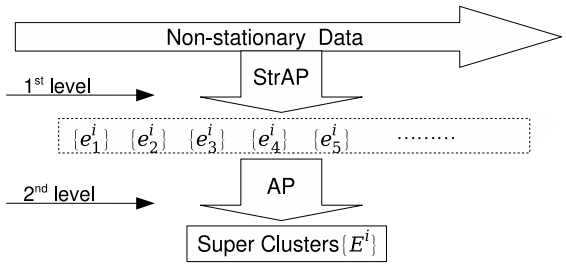


Figure 11.1.1: Online (1st level) and retrospective (2nd level) representation of a data stream with STRAP .

overall stream is visualized on Figure 11.1.2, each row corresponding to a given super-exemplar, and each column corresponding to a day (or a time period; a zooming functionality allows the administrator to adjust the granularity of the visualization). The color of the super-exemplar indicates the percentage (or number) of jobs associated to this super-exemplar in the time period, enabling the administrator to spot the load regularities.

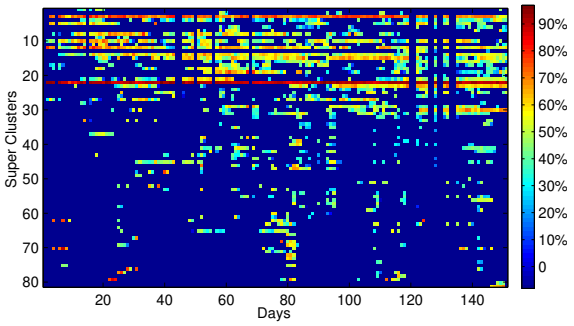


Figure 11.1.2: Visualization of the Stream Model along time (x axis: time; y axis: super-exemplars ordered by attribute 4)

11.2 Decreasing the complexity of affinity propagation

As already mentioned the AP computational complexity is expected to scale like $\mathcal{O}(N^2)$ which is not adapted to streams of data. This limitation can be overcome through a divide and conquer heuristics.

Dataset $\mathcal{E}$ is randomly split into b data subsets; AP is launched on every subset and outputs a set of exemplars; the exemplar weight is set to the number of initial samples it represents; finally, all weighted exemplars are gathered and clustered using WAP. This divide and conquer strategy can be pursued hierarchically in a self-similar way, as a branching process with b representing the branching coefficient of the procedure, defining the Hierarchical AP (HAP) algorithm.

Formally, let us define a tree of clustering operations, where the number h of successive random partitions of the data represents the height of the tree. At each level of the hierarchy, the penalty parameter s_* of AP is set in such a way that the expected number of exemplars extracted along each clustering step is upper bounded by a constant K.

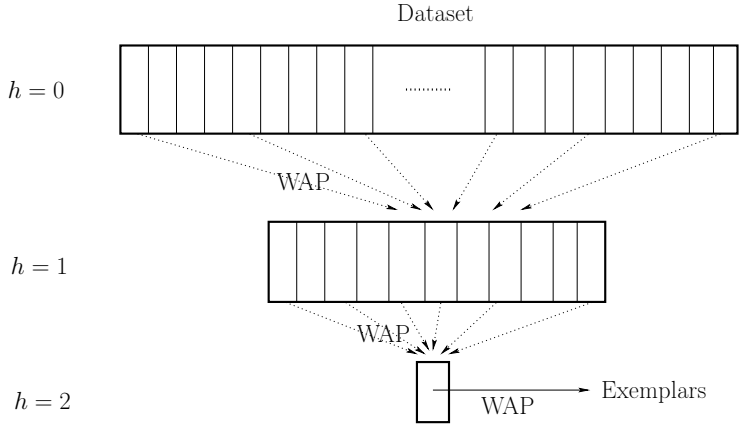


Figure 11.2.1: Sketch of the HAP procedure for 2 hierarchical levels. At each elementary clustering steps, items are weighted in proportion to what they represent as exemplars, i.e. WAP is in use instead of AP.

Proposition 11.2.1. *Let us define the branching factor b as*

$$b = \left(\frac{N}{K}\right)^{\frac{1}{h+1}},$$

Then the overall complexity $C(h)$ of HAP *is given by*

$$C(h) \propto K^{\frac{h}{h+1}} N^{\frac{h+2}{h+1}} \qquad N \gg K,$$

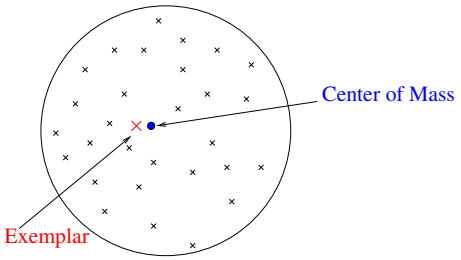


Figure 11.2.2: The point minimizing the energy cost for a single cluster.

up to logarithmic terms.

It is seen that $C(0) = N^2$, $C(1) \propto N^{3/2}$,..., and $C(h) \propto N$ for $h \gg 1$. Note that this procedure is naturally implemented in a streaming context; the partition is made automatically by buffering the data as they arrive in a buffer of size M. When it is full, AP is run on this set, and the exemplars are stored in another buffer of identical size M but corresponding to the next hierarchical level. The procedure can be continued indefinitely as long as the data flow is not too large, i.e. the run-time taken by AP to treat one single buffer at lowest hierarchical level should not exceed the time needed for the same buffer to be full again. Let us examine the price to pay for this complexity reduction. The distortion loss incurred by HAP w.r.t. AP is examined in the simple case where the data samples follow a centered distribution in $\mathbb{R}^d$. By construction, AP aims at finding the cluster exemplar $\mathbf{r_C}$ nearest to the center of mass of the sample points noted $\mathbf{r}_{cm}$: In the simple case where points are sampled along a centered distribution in $\mathbb{R}^d$, let $\tilde{\mathbf{r}}_\mathbf{C} = \mathbf{r_C} - \mathbf{r}_{cm}$ denote the relative position of exemplar $\mathbf{r_C}$ with respect to the center of mass $\mathbf{r}_{cm}$. By symmetry the probability distribution of $\mathbf{r}_{cm} + \tilde{\mathbf{r}}_\mathbf{C}$ is the convolution of a spherical with a cylindrical distribution. We denote by x the square distance to the origin, f its probability density and by F the cumulative distribution, while subscripts sd refer to sample data, ex to the exemplar, and cm to center of mass. Assuming

$$\sigma \stackrel{\text{def}}{=} \mathbb{E}[x_{sd}] = \int_0^\infty x f_{sd}(x) dx,$$

$$\alpha \stackrel{\text{def}}{=} -\lim_{x \to 0} \frac{\log(F_{sd}(x))}{x^{\frac{d}{2}}},$$

to exist and being finite, then the cumulative distribution of x_{cm} for a set of M samples satisfies

$$\lim_{M \to \infty} F_{cm}(\frac{x}{M}) = \frac{\Gamma\left(\frac{d}{2}, \frac{2x}{d\sigma}\right)}{\Gamma\left(\frac{d}{2}\right)},$$

by virtue of the central limit theorem, where $\Gamma(x, y)$ is the incomplete gamma function, In the meanwhile, $x_{\widetilde{ex}} = |\mathbf{r}_{ex} - \mathbf{r}_{cm}|^2$ has a universal extreme value

distribution (up to rescaling, see e.g. [41] for general methods):

$$\lim_{M\to\infty} F_{ex}^{\sim}\left(\frac{1}{M^{2/d}}x\right) = \exp\left(-\tilde{\alpha}x^{\frac{d}{2}}\right), \tag{11.2.1}$$

where $\tilde{\alpha} \neq \alpha$ stands for the fact that the extreme value parameter is possibly affected by the displacement of the center of mass. To see how the clustering error propagates along with the hierarchical process, one proceeds inductively. At hierarchical level h, M samples, spherically distributed with variance $\sigma^{(h)}$ are considered; the sample nearest to the center of mass is selected as exemplar. Accordingly, at hierarchical level $h+1$, the next sample data is distributed after the convolution of two spherical distributions, the exemplar and center of mass distributions at level h. The following scaling recurrence property holds:

$$\lim_{M\to\infty} F_{sd}^{(h+1)}\left(\frac{x}{M^{(h+1)\gamma}}\right) = \begin{cases} \frac{\Gamma\left(\frac{d}{2}, \frac{x}{\sigma^{(h+1)}}\right)}{\Gamma\left(\frac{d}{2}\right)} & \text{for} \quad d < 2\ , \text{ with } \gamma = 1 \\ \exp\left(-\alpha^{(h+1)}x^{\frac{d}{2}}\right) & \text{for} \quad d > 2\ , \text{with } \gamma = \dfrac{2}{d} \\ \exp(-\beta^{(h+1)}x) & \text{for} \quad d = 2\ , \text{with } \gamma = 1. \end{cases}$$

with

$$\sigma^{(h+1)} = \sigma^{(h)}, \quad \alpha^{(h+1)} = \alpha^{(h)}, \quad \beta^{(h+1)} = \frac{\beta^{(h)}}{2}.$$

It follows that the distortion loss incurred by HAP does not depend on the hierarchy depth h except in dimension $d = 2$. Fig. 11.2.3 shows the distribution of the clustering distortion depending on the hierarchy-depth h and the dimension d of the dataset. In dimension $d = 1$, the distribution is dominated by the variance of the center of mass, yielding the gamma law which is also stable with respect to the hierarchical procedure. In dimension $d = 2$ however, the Weibull and gamma laws do mix at the same scale; the overall effect is that the width of the distribution (of the distortion) increases like h^2, as shown in Fig. 11.2.3 (top right).

For finite number of data points per cluster we can also estimate corrections to this behaviour, which then also depends on the shape of the cluster. The parameter α defined in the preceding section is actually related to the density at the center of the cluster $p_{sd}(0)$ by

$$\alpha = p_{sd}(0)\frac{\Omega_d}{d}, \tag{11.2.2}$$

with $\Omega_d = 2\pi^{d/2}/\Gamma(d/2)$ the d-dimensional solid angle, as long as the distribution is locally spherical around this point. Still, the shape of the cluster has some influence on the final result and we characterize it by defining the following ad hoc shape factor

$$\omega \stackrel{\text{def}}{=} \frac{\sigma\alpha^{2/d}}{\Gamma\left(1+\frac{2}{d}\right)}, \tag{11.2.3}$$

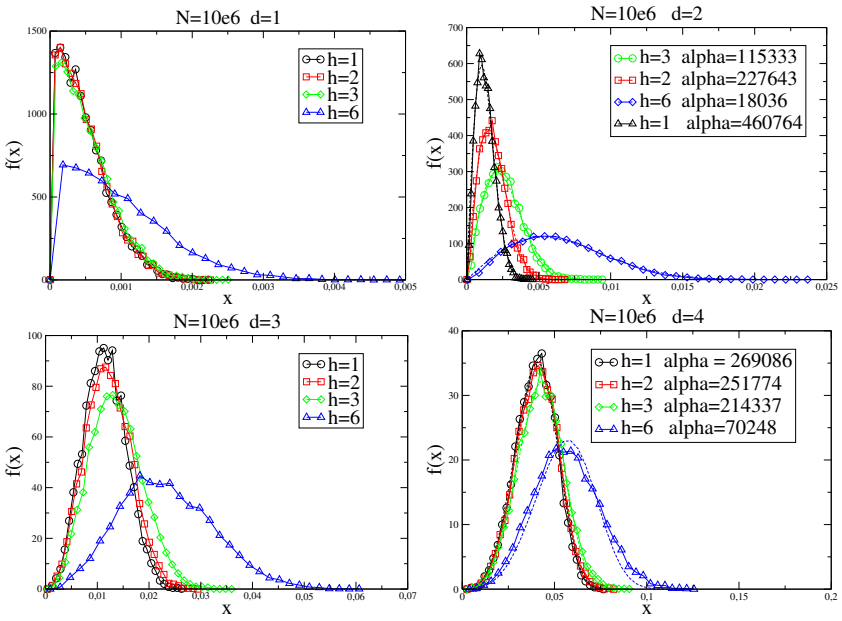


Figure 11.2.3: Radial distribution plot of exemplars obtained by clustering of Gaussian distributions of $N = 10^6$ samples in $\mathbb{R}^d$ in one single cluster exemplar, with hierarchical level h ranging in 1,2,3,6, for diverse values of d: $d = 1$ (upper left), $d = 2$ (upper right), $d = 3$ (bottom left) and $d = 4$ (bottom right). Fitting functions are of the form $f(x) = Cx^{d/2-1}\exp(-\alpha x^{d/2})$.

($= 1$ for the Weibull (11.2.1) distribution) relating the density at the center of the cluster to its variance. For $d > 2$, assuming $\alpha = \alpha^{(h)}$, $\sigma = \sigma^{(h)}$ and $\omega = \omega^{(h)}$ at level h we find:

$$\sigma^{(h+1)} = \frac{\sigma^{(0)}}{\omega^{(0)}}\Big(1 + \frac{1}{M^{1-2/d}}\Big) + o\big(M^{2/d-1}\big).$$

Compared with $\sigma^{(1)}$ obtained directly with AP , we get

$$\frac{\sigma^{(h)}}{\sigma^{(1)}} - 1 = M^{2/d-1} + o\big(M^{2/d-1}\big),$$

when d is larger than 2. This is consistent with the numerical check shown on Figure 11.2.4.

11.3 The number of clusters in AP: a renormalization group viewpoint

In Section 11.2 we left aside the question concerning the penalty coefficient s (see Chapter 2), how should it be modified from one hierarchical level to the

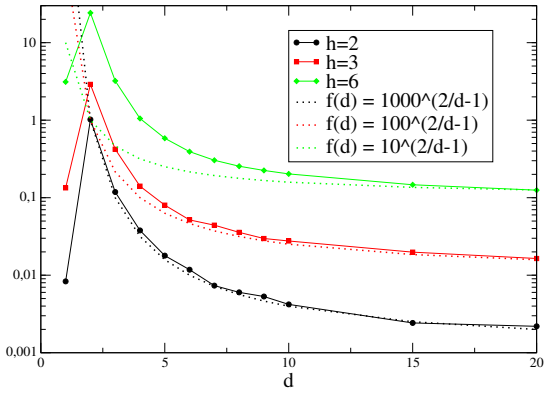


Figure 11.2.4: $\sigma^{(h)}/\sigma^{(1)} - 1$ for $h = 2, 3, 6$ as a function of the dimension, when finding exemplars of a single cluster of 10^6 points (repeated 10^4 times)

next one. We address this question in the present section by applying a simple and exact renormalization principle to AP, based on the results of the preceding section, to yield a way to determine the number of true underlying clusters in a dataset [80].

By convenience we setup a thermodynamic limit where data point and clusters are distributed in a large spatial volume V and go to infinity independently with a fixed density of underlying clusters. After dividing s by V, the clustering cost per datapoint (3.4.1) reads for large numbers of clusters n and data points N, $n << N$:

$$e(\rho) = \sigma(\rho) + s\rho, \tag{11.3.1}$$

with $\rho = n/V$ denoting a fixed density of clusters found by AP:

$$\sigma(\rho) \stackrel{\text{def}}{=} \sum_{c=1}^{\rho V} \nu_c \sigma_c, \tag{11.3.2}$$

denotes the distortion function, with $\nu_c = N_c/N$ the fraction of points in cluster c and σ_c the corresponding variance of the AP-cluster c.

Recall that penalty s implicitly fix the number of clusters. We assume that there exists a value $s^\star$ of s for which AP yields the true underlying structure of clusters assumed to be well separated. $s^\star$ therefore separates a coalescent phase for $s > s^\star$ where true clusters are merged into larger ones, from a fragmenting phase for $s < s^\star$, where true clusters are fragmented into smaller ones. In that case in thermodynamic limit we should be able in principle to identified the critical point $s^\star$ by a Kadanoff like decimation procedure. Indeed, consider a

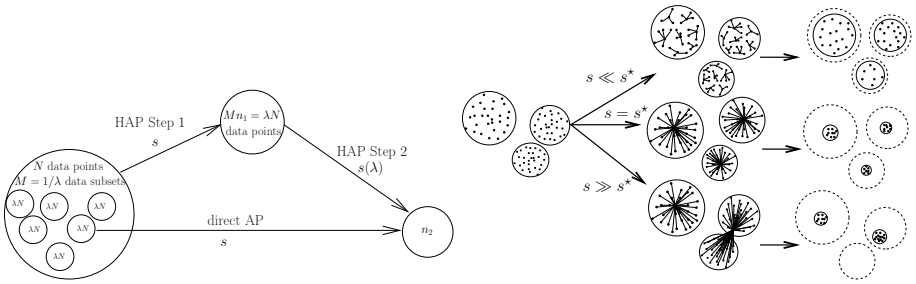


Figure 11.3.1: Divide and conquer strategy translated in a Kadanoff decimation procedure. Transformation of the clusters after the first HAP step depending on s. $s^{(\lambda)}$ is defined to insure clustering stability when $s \simeq s^\star$.

two level HAP of a dataset of size N. For the first clustering stage, the dataset is randomly partitioned into $M = 1/\lambda$ subsets of λN points each and where the reduced penalty s is fixed to some value such that each clustering procedure yields n exemplars on average. The obtained set of examplars constitutes a new dataset of n/λ items, which in turn is clustered with a penalty $s^{(\lambda)}$. $s^{(\lambda)}$ is adjusted in order to recover the same result as would be obtained by clustering the initial dataset directly in one single stage with penalty s. Performing the clustering using one or two hierarchical levels should yield the same result. This basic requirement indicates how s should be renormalized. It is obtained by reinterpreting the divide and conquer strategy as a decimation procedure by enforcing the self-consistency of HAP as illustrated in Figure 11.3.1. Let n_1 [resp. n_2] be the number of clusters obtained after the first [resp. second] clustering stage. Depending on s the proper rescaling may vary, but for $s \simeq s^\star$ this is supposed to behave in a universal way, because in that case, the clusters are preserved while their variance, as shown in the preceding section is simply multiplied by $\left(N\lambda/n_1\right)^{-2/d}/\omega = \lambda^{2/d}/\omega$ in dimension $d > 2$, ω being given in (11.2.3). Therefore we choose to rescale s as

$$s^{(\lambda)} = \frac{\lambda^{2/d}}{\omega}\, s. \tag{11.3.3}$$

When $\lambda^{2/d}/\omega \ll 1$, i.e. when there is a sufficient amount of data points per cluster, we expect the following property of HAP to hold:

$$\text{if} \quad \begin{cases} s < s^\star & \text{then} \quad n_2 \geq n_1 \geq n^\star, \\ s = s^\star & \text{then} \quad n_2 = n_1 = n^\star, \\ s > s^\star & \text{then} \quad n_2 = n_1 \leq n^\star. \end{cases} \tag{11.3.4}$$

Tests of this renormalized procedure are shown on Figure 11.3.2. On Figure 11.3.2.a 11.3.2.c and 11.3.2.d tests are done on artificial data with known

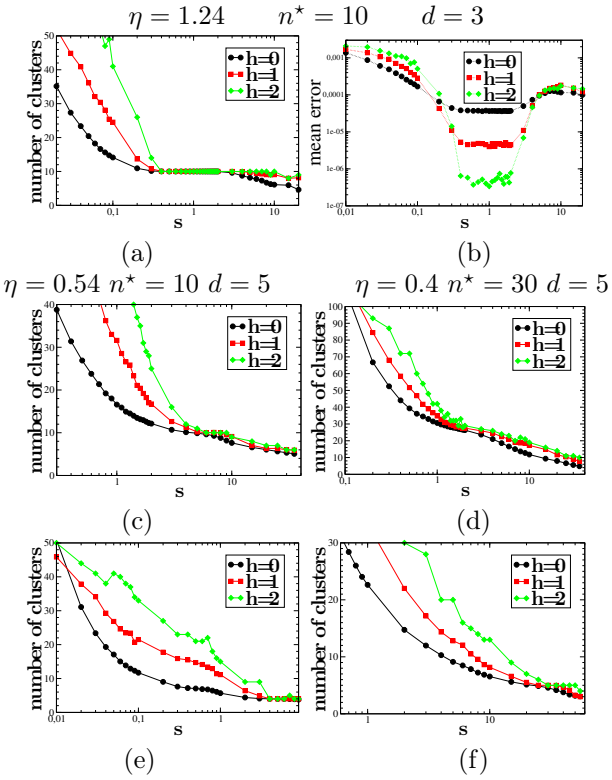


Figure 11.3.2: Number of clusters obtained at each hierarchical level as a function of s, with fixed size of individual partition $\lambda N = 300$, for various spatial dimension, separability indexes and number of underlying clusters (a), (c) and (d), for the EGEE dataset (e) and of a jpeg image (f) of 1.5 10^5 pixels size. Error distance of the exemplars from the true underlying centers (b) corresponding to clustering (a).

underlying number of clusters with different values of $\eta \stackrel{\text{def}}{=} \frac{d_{min}}{2R_{max}}$, where d_{min} is the minimal distance between clusters and R_{max} the maximal radius of the clusters. The self-similar point is clearly identified when plotting the number of clusters against the bare penalty, when the separation of clusters characterized by η is not to small. As expected from the scaling (11.3.3), the effect is less sensible when the dimension increases, but remains perfectly visible and exploitable at least up to $d = 30$. The absence of information loss of the hierarchical procedure can be seen on the mean-error plots on Figure 11.3.2.b , in the region of s around the critical value $s^\star$. Tests on real data, EGEE dataset and a natural colored image are shown respectively on Figure 11.3.2.e,f. Both show well defined structures according to the RG procedure.

Chapter 12

Pattern formation in Restricted Boltzmann machines

This chapter is based on the following papers:

A. Decelle, G. Fissore and C. Furtlehner, Thermodynamics of Restricted Boltzmann Machines and Related Learning Dynamics, *J.Stat.Phys. 172*,6 (2018) 1576–1608

A. Decelle, G. Fissore and C. Furtlehner, Spectral Dynamics of Learning Restricted Boltzmann Machines EPL (2017) 119,6: 60001.

In this chapter we study the learning process itself of a tractable machine learning algorithm, the RBM learning introduced in Chapter 2. We analyze this in the perspective of concept-formation [4], namely how information extracted from the data get encoded into the machine, by identifying linear instabilities responsible for the main patterns and their evolution in the non-linear regime of the learning process thanks to an RBM ensemble averaging.

12.1 Statistical ensembles of RBM's

As explained in chapter 2, RBM is an important ML tool which deserves to be studied in depth. To understand the learning process we have to analyze the learning equations (2.3.4,2.3.5,2.3.6). We develop an average case analysis of these equations. In order to do that we first have to define a statistical ensemble of RBM to average over. To be realistic this ensemble will be based on the following empirical observations [43].

Empirical observations Let us present results of learning an RBM on the MNIST dataset commonly used in pattern recognition. This famous dataset is composed of 60000 grey scale images of handwritten digits of binarized 28×28 pixels. Our first and main observation concerns the spectral density (SVD modes) of the weight matrix during the learning shown on Fig. 12.1.1 and Fig. 12.1.3. We see that after only a few updates the system has already learned many SVD modes from the data. Some modes escape from the Marchenko-

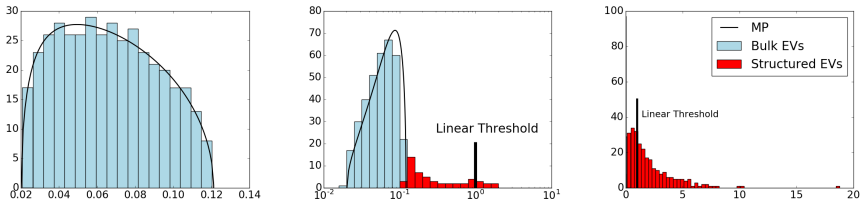


Figure 12.1.1: (left) Initial Marchenko-Pastur distribution of the weight matrix singular values. (middle) After a few epoch of training some singular pass the linear threshold. (right) Distribution of the singular values in the end of the training: we can see many outliers spread above threshold and a spike of below-threshold singular values near zero.

Pastur bulk while other condense down to zero. In particular, we can see that the dominant modes at the beginning of the learning correspond well to the SVD modes of the data (see Fig. 12.1.2) which will be justified later by a linear stability analysis. After many epochs, we observe on Fig. 12.1.2-f that non-linear effects have deformed the SVD modes of W. On Fig. 12.1.3 the quantitative evolution of the singular values is displayed. Dominant modes are amplified but other modes beyond some rank around 250 are dumped. We also see that the

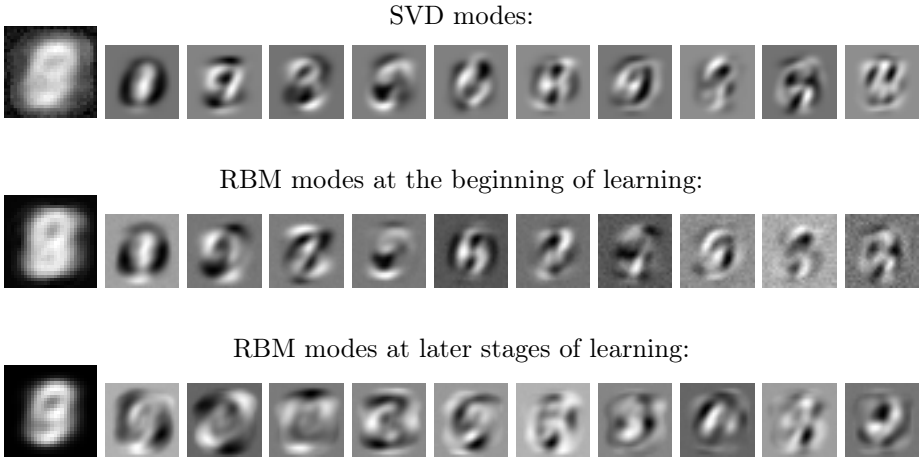


Figure 12.1.2: (top) First principal components extracted from the training set. (middle) First modes of an RBM trained for one epoch. (bottom) Same but after ten epochs of training.

top part of the spectrum of W appears flattened as compared to the empirical SVD spectrum. As we shall see, this presumably enables the expression of many states of similar free energy related to various digit configurations.

Statistical ensemble of RBM's : When analyzing the thermodynamical properties of RBMs, it is commonly assumed [168, 17, 110] that the weights W_{ij} are i.i.d. random variables. This generally leads to a Marchenko-Pastur (MP) distribution [156] of the singular values of W, which is unrealistic. In order to remedy this oversimplification we propose to consider instead the following form of the weight matrix

$$W_{ij} = \sum_{\alpha=1}^{K} w_\alpha u_i^\alpha v_j^\alpha + r_{ij}, \tag{12.1.1}$$

composed of a structured component and a random part. The first one is assumed to represent the information content of the RBM while the second represents uncorrelated noise. The $w_\alpha = O(1)$ are isolated singular values (describing a rank K matrix), the $\mathbf{u}^\alpha$ and $\mathbf{v}^\alpha$ are normalized vectors, representing the dominant singular vectors of the SVD decomposition and the $r_{ij} = \mathcal{N}(0, \sigma^2/L)$ are i.i.d. terms corresponding to noise, $L = \sqrt{N_h N_v}$ being the size of the system. The $\{u^\alpha\}$ and $\{v^\alpha\}$ are two sets of respectively N_v and N_h-dimensional orthonormal vectors, which means that their components are respectively $O(1/\sqrt{N_v})$ and $O(1/\sqrt{N_h})$, and $K \leq N_v, N_h$. We assume $N_h < N_v$ to be the rank of W, $w_\alpha > 0$ and $O(1)$ for all α. This form corresponds to the picture shown on Figure 12.1.1. Note that in the limit $N_v \to \infty$ and $N_h \to \infty$ with $\kappa \overset{\text{def}}{=} N_h/N_v$ fixed and $K/L \to 0$, WW^T has a spectrum density $\rho(\lambda)$ composed of a Marchenko-Pastur

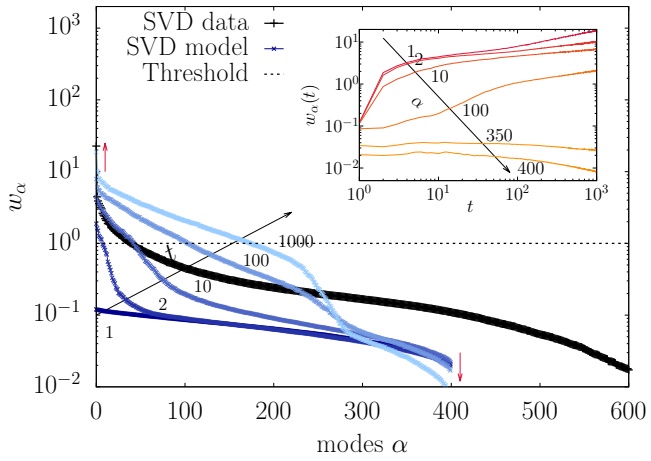


Figure 12.1.3: Singular values (SVD) ranked in decreasing order. In the inset, the time evolution of the modes 1, 2, 10, 100, 350, 400 during the learning as a function of the number of epochs. Dotted line indicates the linear threshold for mode condensation.

bulk of eigenvalues and of set of discrete modes:

$$\rho(\lambda) \approx \frac{L}{2\pi\sigma^2}\frac{\sqrt{(\lambda^+-\lambda)(\lambda-\lambda^-)}}{\kappa\lambda}\mathbb{1}_{\{\lambda\in[\lambda^-,\lambda^+]\}} + \sum_{\alpha=1}^{K}\delta(\lambda-w_\alpha^2),$$

with

$$\lambda^\pm \overset{\text{def}}{=} \sigma^2\left(\kappa^{\frac{1}{4}} \pm \kappa^{-\frac{1}{4}}\right)^2.$$

The interpretation for the noise term r_{ij} is given by the presence of an extensive number of modes at the bottom of the spectrum, along which the variables won't be able to condense but still contributing to the fluctuations. In the present form our model of RBM is similar to the Hopfield model and recent generalizations [160], the patterns being represented by the SVD modes outside of the bulk. The main difference, in addition to the bipartite structure of the graph, is the non-degeneracy of the singular values w_α. A simplification is made here by restricting the analysis to K finite, giving $W_{ij} = O(1/N)$ and coherently $\theta_j = \mathcal{O}(1)$. In addition we assume simple i.i.d distributions for the components of $\mathbf{u}^\alpha$ and $\mathbf{v}^\alpha$ like e.g. Gaussian, Bernoulli or Laplace. Altogether, this defines our statistical ensemble of RBM to which we restrict our analysis of the learning procedure.

With an extensive number of condensed modes we should instead consider an average over the orthogonal group which would lead to a different mean-field theory [184, 181].

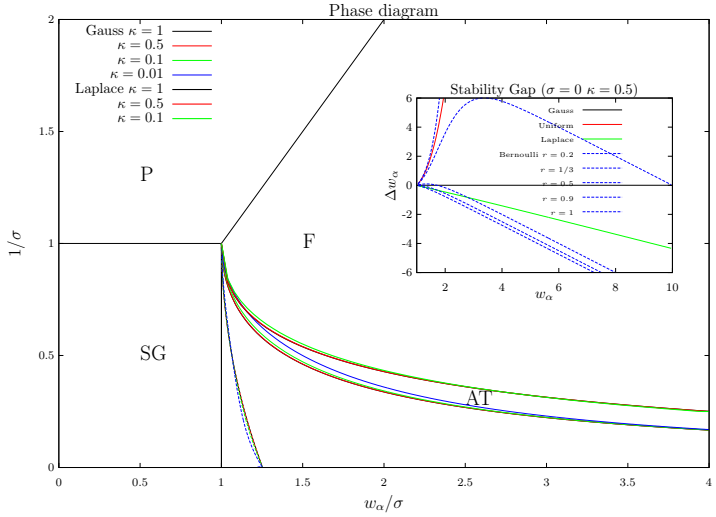


Figure 12.1.4: Phase diagram in absence of bias ($\theta = \eta = 0$). w_α is the highest singular value. Different colors show the sensitivity of the AT line to the distributions (here Gaussian or Laplace) of u and v, or to κ. Inset: high temperature ($\sigma = 0$) stability gap Δw_α expressed as a function of w_α for various distributions of u and v.

12.2 Mean-field theory and nature of the ferromagnetic phase

The mean-field properties of this model can be analyzed with help of replica to average the log partition function w.r.t. the noise r_{ij} and the components u_i^α, v_j^α of the SVD vectors. In the replica symmetric phase which is the one of interest in this context, this leads to the introduction of the following set of order parameters:

- *ferromagnetic* order parameters:

$$m_\alpha = E_{u,v,r}\big(\langle \sigma_\alpha \rangle\big) \qquad \bar{m}_\alpha = E_{u,v,r}\big(\langle s_\alpha \rangle\big)$$

- *spin-glass* order parameters:

$$q = E_{u,v,r}\big(\langle \sigma_j \rangle^2\big) \qquad \bar{q} = E_{u,v,r}\big(\langle s_i \rangle^2\big)$$

with spin projections $s_\alpha \stackrel{\text{def}}{=} \frac{1}{\sqrt{L}} \sum_i s_i u_i^\alpha$ and $\sigma_\alpha \stackrel{\text{def}}{=} \frac{1}{\sqrt{L}} \sum_j \sigma_j^a v_j^\alpha$ of the visible [resp. the hidden] configurations of the replica on the dominant modes. $\mathsf{E}_{u,r}$ and $\mathsf{E}_{v,r}$ denote an average w.r.t. u and v and the noise matrix r_{ij}. These variables represent the correlations of the hidden [resp. visible] states with the left [resp. right] singular vectors and the Edward-Anderson (EA) order parameters

measuring the correlation between replicas of hidden or visible states. Assuming a replica-symmetric (RS) phase, the free energy reads[1]:

$$f[m,\bar{m},q,\bar{q}] = \sum_{\alpha} w_\alpha m_\alpha \bar{m}_\alpha - \frac{\sigma^2}{2} q\bar{q} + \frac{\sigma^2}{2}(q+\bar{q}) \\ - \frac{1}{\sqrt{\kappa}} \mathsf{E}_{u,x}\Big[\log 2\cosh\big(h(x,u)\big)\Big] - \sqrt{\kappa}\mathsf{E}_{v,x}\Big[\log 2\cosh\big(\bar{h}(x,v)\big)\Big], \tag{12.2.1}$$

and the saddle-point equations are given by

$$m_\alpha = \kappa^{\frac{1}{4}} \mathsf{E}_{v,x}\Big[v^\alpha \tanh\big(\bar{h}(x,v)\big)\Big], \qquad q = \mathsf{E}_{v,x}\Big[\tanh^2\big(\bar{h}(x,v)\big)\Big] \tag{12.2.2}$$

$$\bar{m}_\alpha = \kappa^{-\frac{1}{4}} \mathsf{E}_{u,x}\Big[u^\alpha \tanh\big(h(x,u)\big)\Big], \qquad \bar{q} = \mathsf{E}_{u,x}\Big[\tanh^2\big(h(x,u)\big)\Big] \tag{12.2.3}$$

where

$$h(x,u) \stackrel{\text{def}}{=} \kappa^{\frac{1}{4}}\big(\sigma\sqrt{q}x + \sum_\gamma (w_\gamma m_\gamma - \eta_\gamma)u^\gamma\big),$$

$$\bar{h}(x,v) \stackrel{\text{def}}{=} \kappa^{-\frac{1}{4}}\big(\sigma\sqrt{\bar{q}}x + \sum_\gamma (w_\gamma \bar{m}_\gamma - \theta_\gamma)v^\gamma\big).$$

$\kappa = N_h/N_v$, $\mathsf{E}_{u,x}$ and $\mathsf{E}_{v,x}$ denote an average over the Gaussian variable $x = \mathcal{N}(0,1)$ and the rescaled components $u \sim \sqrt{N_v}u_i^\alpha$ and $v \sim \sqrt{N_h}v_j^\alpha$ of the SVD modes. We note that the equations are symmetric under the exchange $\kappa \to \kappa^{-1}$, simultaneously with $m \leftrightarrow \bar{m}$, $q \leftrightarrow \bar{q}$ and $\eta \leftrightarrow \theta$, given that u and v have the same distribution. In addition, for independently distributed u_i^α and v_j^α and vanishing fields ($\eta = \theta = 0$), solutions corresponding to non-degenerate magnetizations have symmetric counterparts: each pair of non-vanishing magnetizations can be negated independently as $(m_\alpha, \bar{m}_\alpha) \to (-m_\alpha, -\bar{m}_\alpha)$, generating new solutions. So to one solution presenting n condensed modes, there correspond 2^n distinct solutions.

The fixed point equations (12.2.2, 12.2.3) can be solved numerically to tell us how the variables condense on the SVD modes within each equilibrium state of the distribution and whether a spin-glass or a ferromagnetic phase is present. The important point here is that with K finite and a non-degenerate spectrum the mode with highest singular value dominates the ferromagnetic phase.

In absence of bias ($\eta = \theta = 0$) and once $1/\sigma$ is interpreted as temperature and w_α/σ as ferromagnetic couplings, we get a phase diagram similar to that of the Sherrington-Kirkpatrick (SK) model with three distinct phases (see Figure 12.1.4)

- a paramagnetic phase ($q = \bar{q} = m_\alpha = \bar{m}_\alpha = 0$) (P),

[1] Assuming the local fields to have vanishing transverse components $\eta^\perp = \theta^\perp = 0$ to the modes

- a ferromagnetic phase $(q, \bar{q}, m_\alpha, \bar{m}_\alpha \neq 0)$ (F),
- a spin glass phase $(q, \bar{q} \neq 0;\ m_\alpha = \bar{m}_\alpha = 0)$ (SG).

In general, the lines separating the different phases are not much sensitive to κ and to the specific choice of distribution for u and v.

Nature of the Ferromagnetic phase Some subtleties arise in the structure of the ferromagnetic phase when considering various ways of averaging over the components of the singular vectors [44]. In [2, 168] is emphasized the importance of compositional states characterized by the activation of a small number of hidden variables. In our representation we investigate a "dual compositional" property, namely how states may or may not result from a combination of modes. For this we first rewrite the mean-field equations in a convenient way. Let

$$p_\alpha(\mathbf{u}) \stackrel{\text{def}}{=} p^\star(u^\alpha) \prod_{\beta \neq \alpha} p(u^\beta),$$

where

$$p^\star(u) \stackrel{\text{def}}{=} -\int_{-\infty}^{u} xp(x)dx = \int_{|u|}^{\infty} xp(x)dx, \tag{12.2.4}$$

is also a normalized distribution, since p has unit variance and provided it is even. From this, define also

$$q_\alpha = \int dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} d\mathbf{v} p_\alpha(\mathbf{v}) \tanh^2\Big(\kappa^{-\frac{1}{4}}\big(\sigma\sqrt{\bar{q}}x + \sum_\gamma (w_\gamma \bar{m}_\gamma - \theta_\gamma)v^\gamma\big)\Big), \tag{12.2.5}$$

$$\bar{q}_\alpha = \int dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} d\mathbf{u} p_\alpha(\mathbf{u}) \tanh^2\Big(\kappa^{\frac{1}{4}}\big(\sigma\sqrt{q}x + \sum_\gamma (w_\gamma m_\gamma - \eta_\gamma)u^\gamma\big)\Big). \tag{12.2.6}$$

The mean-field equations (12.2.2,12.2.3) can be rewritten as follows

$$m_\alpha = (w_\alpha \bar{m}_\alpha - \theta_\alpha)(1 - q_\alpha), \tag{12.2.7}$$

$$\bar{m}_\alpha = (w_\alpha m_\alpha - \eta_\alpha)(1 - \bar{q}_\alpha), \tag{12.2.8}$$

with the immediate consequence (in absence of bias, $\theta = \eta = 0$) that $w_\alpha = 1/\sqrt{(1-q_\alpha)(1-\bar{q}_\alpha)}$ whenever the mode α condenses $(m_\alpha, \bar{m}_\alpha \neq 0)$. Let $w(q, \bar{q}) \stackrel{\text{def}}{=} 1/\sqrt{(1-q)(1-\bar{q})}$. Assuming that the mode α has condensed alone, with EA order parameters values $(q, \bar{q})$, the linear stability analysis of the fixed point displays the following quantity (stability gap)

$$\Delta w_\alpha \stackrel{\text{def}}{=} w(q, \bar{q}) - w(q_\alpha, \bar{q}_\alpha), \tag{12.2.9}$$

to indicate that a finite magnetization along any other modes β cannot develop when $w_\beta < w_\alpha + \Delta w_\alpha$. Let F and F_α be the cumulative distributions associated

respectively to p and p_α

$$F(u) \stackrel{\text{def}}{=} \int_{-\infty}^{u} p(x)dx$$

$$F_\alpha(u) \stackrel{\text{def}}{=} \int d\mathbf{u}\ \theta(u-u^\alpha)p_\alpha(\mathbf{u}) = -\int_{\infty}^{u} du^\alpha \int_{-\infty}^{u^\alpha} xp(x)dx.$$

We have the following property: if F_α (i) dominates [resp. (ii) is dominated by] F on $\mathbb{R}^+$ ($F_\alpha(u) > F(u), \forall u \in \mathbb{R}^+$) then the stability gap Δw_α is positive [resp. negative]. Letting κ_u the kurtosis of p we remark that property (i) [resp. (ii)] actually corresponds to having $\kappa_u < 3$ [resp. $\kappa_u > 3$]. Therefore distributions p with negative relative kurtosis $\gamma_u = \kappa_u - 3$ (w.r.t. the Gaussian case) favors the presence of metastable states. Instead, a positive relative kurtosis may lead to a situation where the fixed point associated to the highest mode α_{max} is not stable due to the presence of lower modes in the range $[w(q,\bar{q}), w_{\alpha_{max}}[$; as a result stable fixed points are necessarily associated to combinations of modes in that case.

Let us give some examples. The Gaussian distribution is a special case with $\gamma_u = 0$. In addition, for instance for p corresponding to Bernoulli, Uniform or Laplace, we have the following properties illustrated in the inset of Figure 12.1.4:

- Gaussian ($\gamma_u = 0$): $\Delta w = 0$, only the dominant mode is stable.
- Bernoulli ($\gamma_u = -2$): $\Delta w > 0$, metastable stable states can occur.
- Uniform ($\gamma_u = -6/5$): $\Delta w > 0$, metastable stable states can occur.
- Laplace ($\gamma_u = 3$): $\Delta w < 0$, compositional states can occur.

12.3 Dynamics of learning in thermodynamical limit

We re-express the learning dynamics in the reference frame defined by the singular vectors of W. Discarding stochastic fluctuations, we let the learning rate $\gamma \to 0$, and assume the continuous version of (2.3.4-2.3.6) to be well defined. In this limit we introduce the skew-symmetric rotations generators $\Omega^{v,h}_{\alpha\beta}(t)$ of the left and right singular vectors:

$$\Omega^v_{\alpha\beta}(t) = -\Omega^v_{\beta\alpha} \stackrel{\text{def}}{=} \frac{d\mathbf{u}^{\alpha,T}}{dt}\mathbf{u}^\beta,$$

$$\Omega^h_{\alpha\beta}(t) = -\Omega^h_{\beta\alpha} \stackrel{\text{def}}{=} \frac{d\mathbf{v}^{\alpha,T}}{dt}\mathbf{v}^\beta.$$

Projecting the continuous version of (2.3.4-2.3.6) yields:

$$\frac{1}{L}\frac{dw_\alpha}{dt} = \langle s_\alpha \sigma_\alpha \rangle_{\text{Data}} - \langle s_\alpha \sigma_\alpha \rangle_{\text{RBM}} \tag{12.3.1}$$

$$\frac{d\eta_\alpha}{dt} = \langle s_\alpha \rangle_{\text{RBM}} - \langle s_\alpha \rangle_{\text{Data}} + \sum_\beta \Omega^v_{\alpha\beta}\eta_\beta \tag{12.3.2}$$

$$\frac{d\theta_\alpha}{dt} = \langle \sigma_\alpha \rangle_{\text{RBM}} - \langle \sigma_\alpha \rangle_{\text{Data}} + \sum_\beta \Omega^h_{\alpha\beta}\theta_\beta \tag{12.3.3}$$

along with

$$\Omega^v_{\alpha\beta}(t) = -\frac{1}{w_\alpha + w_\beta}\left(\frac{dW}{dt}\right)^{\text{A}}_{\alpha\beta} + \frac{1}{w_\alpha - w_\beta}\left(\frac{dW}{dt}\right)^{\text{S}}_{\alpha\beta} \tag{12.3.4}$$

$$\Omega^h_{\alpha\beta}(t) = \frac{1}{w_\alpha + w_\beta}\left(\frac{dW}{dt}\right)^{\text{A}}_{\alpha\beta} + \frac{1}{w_\alpha - w_\beta}\left(\frac{dW}{dt}\right)^{\text{S}}_{\alpha\beta} \tag{12.3.5}$$

where

$$\left(\frac{dW}{dt}\right)^{\text{A,S}}_{\alpha\beta} \overset{\text{def}}{=} \frac{1}{2}\Big(\langle s_\alpha \sigma_\beta \rangle_{\text{Data}} \pm \langle s_\beta \sigma_\alpha \rangle_{\text{Data}} \mp \langle s_\beta \sigma_\alpha \rangle_{\text{RBM}} - \langle s_\alpha \sigma_\beta \rangle_{\text{RBM}}\Big).$$

These equations can be expressed exactly for linear RBM to recover known

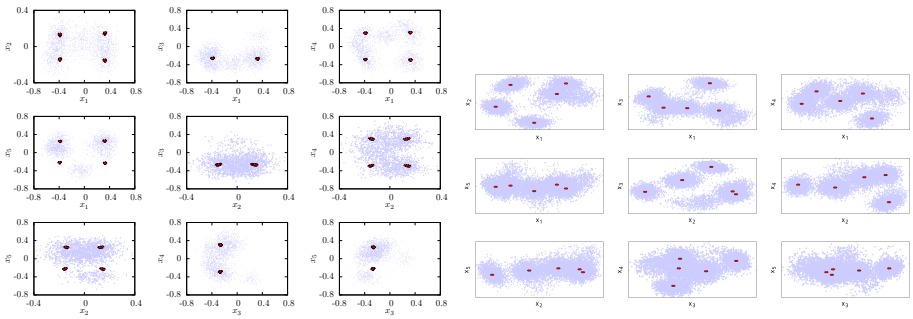


Figure 12.3.1: Comparison between RBM ensemble (left) and single RBM instance (right) showing scatter plots of the mean-field magnetizations (in red) and the samples (in blue) projected on left eigenvectors of W. The architecture is $(N_v, N_h) = (100, 50)$ and the synthetic dataset corresponds to 10^4 samples of size $N_v = 100$ obtained from a multimodal distribution with 11 clusters randomly defined on a submanifold of dimension $d = 5$.

results [222, 122], while for binary-binary RBM we make use of the thermody-

namic limit of the preceding section, to estimate the needed response terms:

$$\langle s_\alpha \rangle_{\text{RBM}} = \frac{1}{Z_{\text{Therm}}} \sum_\omega e^{-Lf(m^\omega, \bar{m}^\omega, q^\omega, \bar{q}^\omega)} \bar{m}_\alpha^\omega \stackrel{\text{def}}{=} \langle \bar{m}_\alpha \rangle_{\text{Therm}},$$

$$\langle s_\alpha \sigma_\beta \rangle_{\text{RBM}} = \frac{1}{Z_{\text{Therm}}} \sum_\omega e^{-Lf(m^\omega, \bar{m}^\omega, q^\omega, \bar{q}^\omega)} \bar{m}_\alpha^\omega m_\beta^\omega \stackrel{\text{def}}{=} \langle \bar{m}_\alpha m_\beta \rangle_{\text{Therm}},$$

$$\text{with} \qquad Z_{\text{Therm}} \stackrel{\text{def}}{=} \sum_\omega e^{-Lf(m^\omega, \bar{m}^\omega, q^\omega, \bar{q}^\omega)}.$$

The index ω runs over all the stable fixed point solutions of (12.2.2,12.2.3) weighted accordingly to the free energy given by (12.2.1). These are the dominant contributions as long as free energy differences are $O(1)$, and the internal fluctuations given by each fixed point are comparatively of order $O(1/L)$. In addition, the dynamics of the bulk can be characterized by empirically defining

$$\sigma^2 \stackrel{\text{def}}{=} \frac{1}{L} \sum_{ij} r_{ij}^2.$$

The goal is to find trajectory of the RBM ensemble in the form of a trajectory in the space $\{w_\alpha(t), \eta_\alpha(t), \theta_\alpha(t), \Omega_{\alpha\beta}^{v,h}(t)\}$, by averaging over u_i^α and v_i^α and r_{ij}. Components s_α of any given sample have to be kept fixed while averaging for this to be meaningful. Letting

$$q_\alpha[\mathbf{s}] \stackrel{\text{def}}{=} \int dx \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} d\mathbf{v} p_\alpha(\mathbf{v}) \tanh^2\Big(\kappa^{-\frac{1}{4}}\big(\sigma x + \sum_\gamma (w_\gamma s_\gamma - \theta_\gamma) v^\gamma\big)\Big),$$

the empirical counterpart measured on a single data point of the (mode dependent) EA order parameter (12.2.5), we end up with the following set of thermodynamical learning equations for the binary-binary RBM ensemble (12.1.1):

$$\frac{1}{L}\frac{dw_\alpha}{dt} = \langle s_\alpha (w_\alpha s_\alpha - \theta_\alpha)(1 - q_\alpha[\mathbf{s}])\rangle_{\text{Data}} - \langle \bar{m}_\alpha (w_\alpha \bar{m}_\alpha - \theta_\alpha)(1 - q_\alpha)\rangle_{\text{Therm}}, \tag{12.3.6}$$

$$\frac{d\eta_\alpha}{dt} = \langle \bar{m}_\alpha \rangle_{\text{Therm}} - \langle s_\alpha \rangle_{\text{Data}} + \sum_\beta \Omega_{\alpha\beta}^v \eta_\beta, \tag{12.3.7}$$

$$\frac{d\theta_\alpha}{dt} = \langle (w_\alpha \bar{m}_\alpha - \theta_\alpha)(1 - q_\alpha)\rangle_{\text{Therm}} - \langle (w_\alpha s_\alpha - \theta_\alpha)(1 - q_\alpha[\mathbf{s}])\rangle_{\text{Data}} + \sum_\beta \Omega_{\alpha\beta}^h \theta_\beta, \tag{12.3.8}$$

$$\frac{d\sigma^2}{dt} = \sigma^2\Big(\langle q \rangle_{\text{Therm}} - \langle q[\mathbf{s}]\rangle_{\text{Data}}\Big). \tag{12.3.9}$$

Note here that the w_α variables, with respect to the other variables, evolve on a faster time scale. Two things are noticeable in these equations

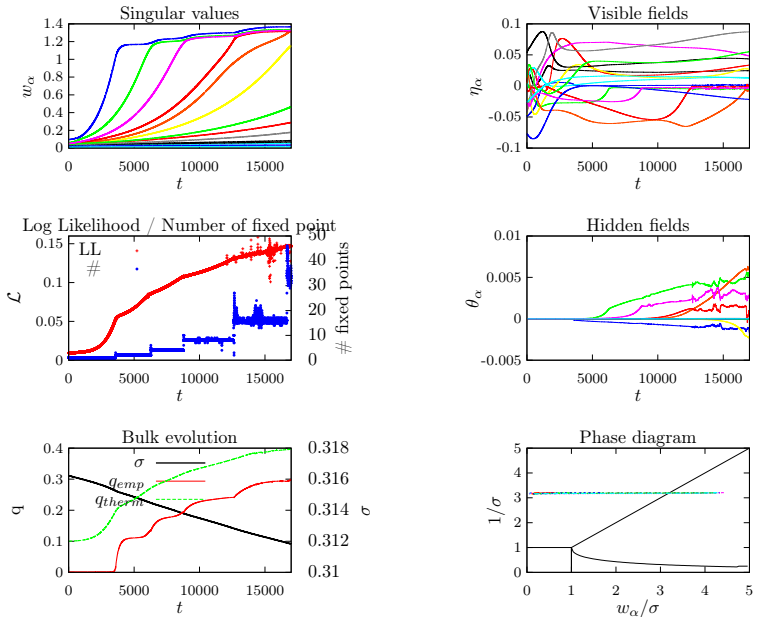


Figure 12.3.2: Predicted mean evolution of an RBM of size $(N_v, N_h) = (1000, 500)$ learned on a synthetic dataset of 10^4 samples of size $N_v = 1000$ obtained from a multimodal distribution with 20 clusters randomly defined on a submanifold of dimension $d = 15$. The dynamics follows the projected magnetizations in this reduced space with help of 15 modes. We observe a kind of pressure on top singular values from lower ones.

- non-linearities enters through the q_α coefficients. In particular the source term reduces to the empirical covariance matrix of the data in the linear regime which can be integrated exactly. The precise form of the non-linearity depend on the activation function (tanh() in the present case) which defines implicitly the similarity matrix between data points.

- the RBM is performing a clustering of the data, with centroids corresponding to solutions of the mean-field equations with magnetizations $\bar{m}_\alpha$ and EA parameters q_α corresponding respectively to their empirical counterparts $\langle s_\alpha \rangle$ and $\langle q_\alpha[\mathbf{s}] \rangle$ representing cluster magnetization and variance (See Figure 12.3.1).

For sake of illustration a synthetic dataset composed of 10^4 samples with an effective dimension $d = 15$ organized in 20 separate clusters was generated and the RBM ensemble learning trajectory integrated with (12.3.6,12.3.7,12.3.8,12.3.9) is compared with a single instance RBM learning process on the same data. (See Figures.12.3.2 and 12.3.3). The RBM ensemble dynamics of the singular values

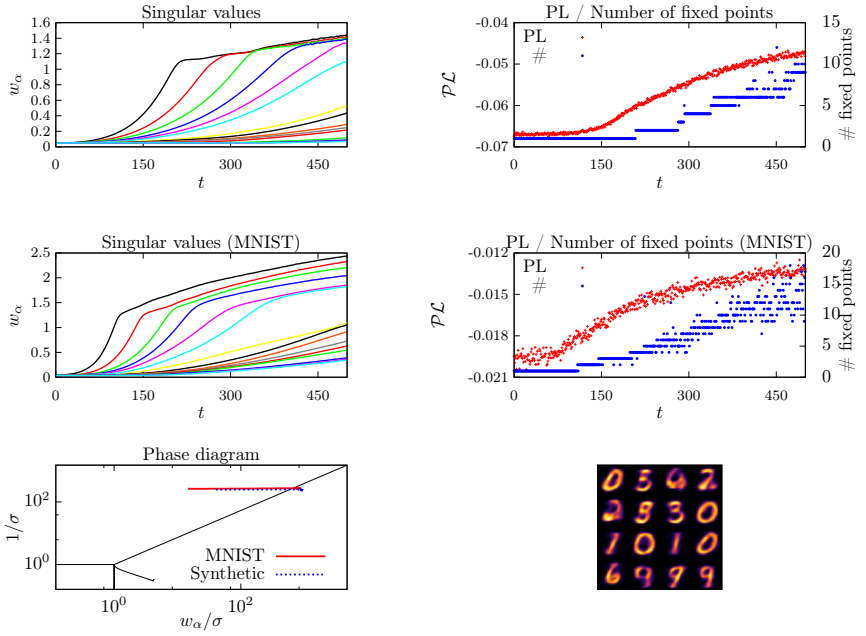


Figure 12.3.3: Experimental evolution of an RBM during training for a synthetic dataset (top plots, to compare to Fig. 12.3.2) and for MNIST (central plots). The bottom left plot shows the learning trajectories in the phase diagram, while the bottom right image shows some examples of fixed point solutions for MNIST.

reproduce faithfully the single RBM trajectory. The number of fixed point instead grows faster, due to a deficient selection mechanism, to be traced back to the neglect of $\theta^\perp$. The learning trajectory on the phase diagram goes from the paramagnetic to the ferromagnetic phase in straight line, keeping track of the initial temperature by lack of regularization.

Interestingly, similar trajectories of the singular values, showing as well their emergence by order of importance taken form the data as an expected consequence of the linear regime, can be obtained with a Gaussian-spherical RBM [45]. In that case an exact integration of the learning equation is performed, based on a particle process (detailed in Chapter 1 and studied in Chapter 6 and 7) representation of the partition function, making an explicit link between condensation mechanism and ferromagnetic order, thereby closing the statistical physics loop of this document.

Conclusion

To conclude this document let us try as promised in the preamble to find some guideline relating all different subjects or at least finding some recurrences. These works follow some tradition of research in statistical physics targeting phenomena observed outside physics laboratories as object studies, which started many decades ago with biological applications, error correcting codes, road traffic or econophysics for instance.

One of the guideline here is to develop practically working algorithms based on statistical physics tools. These are some usual expected tools from statistical physics: *mean-field* as a result of law of large numbers which is encountered both in traffic modelling when looking at hydrodynamic limits or in traffic prediction when using belief propagation or in machine learning when using TAP; *linear stability analysis* appears as well as a powerful tool to analyze complex behaviour like the learning process of a restricted Boltzmann machine; *scale invariance* as seen in the context of clustering problems can be turned into algorithmic considerations when imposing self-similarity; *dual transformation* has been also shown as a possible tool of interest when formulating generalized belief propagation on a dual graph associated to a cycle basis.

The recent trend in machine learning favours models of deep learning characterized by a tremendous increase in complexity, accompanied by a tremendous increase in perplexity of ordinary statistical physicist of my sort on the way to proceed to account for such models with our basic tools at disposal.

A different direction, the one I am trying to follow, consists instead to identify basic relevant mechanisms offering potentially the possibility to simplify machine learning models while degrading as little as possible the performances. The kind of trade-off that one should be looking for is typically obtained by constraining the models to stay in the validity domain of the mean-field approximation, which for an RBM for instance would mean to constraint the weights to remains of order $\mathcal{O}\left(1/\sqrt{N}\right)$ without losing the ability to learn the dataset. This was already the guideline followed in part III to perform traffic prediction with belief propagation, by looking for models compatible with BP. This sweet spot if it exists at all should be viewed as a “weak-coupling" machine learning theory where fluctuations are small enough beyond simple compositional mechanism and could be reached presumably by starting or after learning an adapted representation of the data.

This leads to another direction which consists in to include more symmetry, in the usual way which is done in physics, at the level of the definition of the model itself, directly in the weight matrix to be learned for the RBM for instance. For translation invariance this would lead to impose as in solid state physics from the Bloch theorem a Fourier envelope to the weight matrix spectrum, or for scale invariance to work with Parisi like matrices or more generally with various kind of wavelet basis to represent the weight matrices.

All this of course makes only sense after the dataset to model is specified. Looking at a particular model should therefore be motivated by some application, where choosing the model of adapted complexity makes the difference in performances. One application which has not been presented in this manuscript

concerns space weather forecasting. We already proposed some inference algorithm to predict solar wind speed with unknown delay from pictures of the sun [28], with a *linear stability analysis* helping to monitor and drive the algorithm. Still the peculiarity of solar images calls for more specific models than the one we used for that purpose, where the previous considerations could be operated. Another application concerns the use of probabilistic models as generative models for genetic data, with preliminary tests based on GAN and RBM in [243]. Again the high level of compressibility of genetic data suggest that more specific models should be investigated in this domain.

Bibliography

[1] Ackerman, M., and Ben-David, S. Measures of clustering quality: A working set of axioms for clustering. In *NIPS* (2008).

[2] Agliari, E., Barra, A., Galluzzi, A., Guerra, F., and Moauro, F. Multitasking associative networks. *Phys. Rev. Lett. 109* (2012), 268101.

[3] Alpert, S. Spectral partitioning: The more eigenvectors, the better. In *32nd Design Automation Conference* (1995), pp. 195–200.

[4] Amari, S.-I. Neural theory of association and concept-formation. *Biol. Cybern. 26*, 3 (1977), 175–185.

[5] Amit, D. J., Gutfreund, H., and Sompolinsky, H. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett. 55*, 14 (1985), 1530–1533.

[6] Amit, D. J., Gutfreund, H., and Sompolinsky, H. Statistical mechanics of neural networks near saturation. *Annals of Physics 173*, 1 (1987), 30–67.

[7] Andrieux, D., and Gaspard, P. Fluctuation theorem for currents and Schnakenberg network theory. *J. Stat. Phys. 127*, 1 (2006).

[8] Appert, C., and Santen, L. Boundary induced phase transitions in driven lattice gases with meta-stable states. *PRL 86* (2001), 2498.

[9] Appert-Rolland, C. Experimental study of short-range interactions in vehicular traffic. *Phys. Rev. E 80* (2009), 036102.

[10] Aw, A., Klar, A., Rascle, M., and Materne, T. Derivation of continuum traffic flow models from microscopic follow-the-leader models. *SIAM Journal on Applied Mathematics 63*, 1 (2002), 259–278.

[11] Aw, A., and Rascle, M. Resurrection of "second order" models of traffic flow. *SIAM Journal on Applied Mathematics 60*, 3 (2000), 916–938.

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

[13] Banerjee, O., El Ghaoui, L., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR 9* (2008), 485–516.

[14] Barlović, R., Santen, L., Schadschneider, A., and Schreckenberg, M. Metastable states in cellular automata for traffic flow. *Eur. Phys. J. B5* (1998), 793.

[15] Barra, A., Bernacchia, A., Santucci, E., and Contucci, P. On the equivalence of Hopfield networks and Boltzmann machines. *Neural Networks 34* (2012), 1–9.

[16] Barra, A., Genovese, G., Sollich, P., and Tantari, D. Phase transitions in restricted Boltzmann machines with generic priors. *Phys. Rev. E 96*, 4 (2017), 042156.

[17] Barra, A., Genovese, G., Sollich, P., and Tantari, D. Phase diagram of restricted Boltzmann machines and generalized Hopfield networks with arbitrary priors. *Phys. Rev. E 97* (2018), 022310.

[18] Baskett, F., Chandy, K., Muntz, R., and Palacios, F. Open, closed, and mixed networks of queues with different classes of customers. *J. ACM 22*, 2 (1975), 248–260.

[19] Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *PNAS 116*, 32 (2019), 15849–15854.

[20] Ben-David, S., and von Luxburg, U. Relating clustering stability to properties of cluster boundaries. In *COLT* (2008).

[21] Benassi, A., and Fouque, J. Hydrodynamical limit for the asymmetric simple exclusion process. *Ann. Prob. 15*, 2 (1987), 546–560.

[22] Blank, M. Hysteresis phenomenon in deterministic traffic flows. *J. Stat. Phys. 120* (2005), 627–658.

[23] Blythe, R. A., and Evans, M. R. Nonequilibrium steady states of matrix product form: A solver's guide. *J. Phys. A: Math. & Theor. 40* (2007), 333–441.

[24] Boser, B., Guyon, I., and Vapnik, V. A training algorithm for optimal margin classifiers. In *COLT* (1992), ACM, pp. 144–152.

[25] Bourlard, H., and Kamp, Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics 59*, 4 (1988), 291–294.

[26] Brackstone, M., and McDonald, M. Car-following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour 2*, 4 (1999), 181 – 196.

[27] Cantini, L. Algebraic Bethe Ansatz for the two species ASEP with different hopping rates. *J. Phys. A: Math. Theor.* **41** (2008), 095001.

[28] Chandorkar, M., Furtlehner, C., Poduval, B., Camporeale, E., and Sebag, M. Dynamic Time Lag Regression: Predicting What & When. In *proceedings of ICLR* (2020).

[29] Chernyak, V., Chertkov, M., Malinin, S., and Teodorescu, R. Non-equilibrium thermodynamics and topology of currents. *J.Stat.Phys. 137*, 109 (2009).

[30] Chertkov, M., and Chernyak, V. Loop series for discrete statistical models on graphs. *J.Stat.Mech.* (2006), P06009.

[31] Chertkov, M., and Chernyak, V. Y. Loop calculus in statistical physics and information science. *Phys. Rev. E 73* (2006), 065102.

[32] Cho, K., Ilin, A., and Raiko, T. Improved learning of Gaussian-Bernoulli restricted Boltzmann machines. In *International conference on artificial neural networks* (2011), Springer, pp. 10–17.

[33] Clincy, M., Derrida, B., and Evans, M. R. Phase transition in the ABC model. *Phys. Rev. E 67* (2003), 6115–6133.

[34] Cocco, S., and Monasson, R. Adaptive cluster expansion for the inverse Ising problem: convergence, algorithm and tests. *J. Stat. Phys. 147*, 2 (2012), 252–314.

[35] Cocco, S., Monasson, R., and Sessak, V. High-dimensional inference with the generalized Hopfield model: Principal component analysis and corrections. *Phys. Rev. E 83* (2011), 051123.

[36] Cohen, O., and Mukamel, D. Phase diagram of the ABC model with nonequal densities. *J. Phys. A: Math. & Theor. 44*, 41 (2011), 415004.

[37] Cooper, G. The computational complexity of probabilistic inference using Bayesian belief networks (research note). *A.I. 42*, 2-3 (1990), 393–405.

[38] Cormode, G., Muthukrishnan, S., and Zhuang, W. Conquering the divide: Continuous clustering of distributed data streams. In *ICDE* (2007), pp. 1036–1045.

[39] Daganzo, C. Requiem for second order fluid approximation of traffic flow. *Transportation Research B* (1995).

[40] Darroch, J., and Ratcliff, D. Generalized iterative scaling for log-linear models. *Ann. Math. Statistics 43* (1972), 1470–1480.

[41] de Haan, L., and Ferreira, A. *Extreme Value Theory*. Operations Research and Financial Engineering. Springer, 2006.

[42] De Masi, A., and Presutti, E. *Mathematical Methods for Hydrodynamic Limits*, vol. 1501 of *Lecture Notes in Mathematics*. Springer-Verlag, 1991.

[43] Decelle, A., Fissore, G., and Furtlehner, C. Spectral dynamics of learning in restricted Boltzmann machines. *EPL 119*, 6 (2017), 60001.

[44] Decelle, A., Fissore, G., and Furtlehner, C. Thermodynamics of restricted Boltzmann machines and related learning dynamics. *J.Stat.Phys. 172*, 18 (2018), 1576–1608.

[45] Decelle, A., and Furtlehner, C. Gaussian-spherical restricted Boltzmann machines. *J. Phys. A: Math. & Theor. 53*, 18 (2020), 184002.

[46] Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. Inference and phase transitions in the detection of modules in sparse networks. *Phy.Rev.Lett. 107*, 6 (2011), 065701.

[47] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood for incomplete data via the EM algorithm. *J. Royal Stat. Soc. B 39*, 1 (1977), 1–38.

[48] Derrida, B. An exactly soluble non-equilibrium system:the asymmetric simple exclusion process. *Phys. Rep. 301* (1998), 65–83.

[49] Derrida, B., and Appert, C. Universal large-deviation function of the KPZ equation in one dimension. *J. Stat. Phys. 94*, 1 (1999), 1–30.

[50] Derrida, B., Evans, M. R., Hakim, V., and Pasquier, V. Exact solution for 1d asymmetric exclusion model using a matrix formulation. *J. Phys. A: Math. Gen. 26* (1993), 1493–1517.

[51] Derrida, B., Lebowitz, J. L., and Speer, E. R. Exact large deviation functional of a stationary open driven diffusive system: The asymmetric exclusion process. *J. Stat. Phys. 110*, 3 (2003), 775–810.

[52] Dietrich, R., Opper, M., and Sompolinsky, H. Statistical mechanics of support vector networks. *Phys. Rev. Lett. 82* (1999), 2975–2978.

[53] Domany, E., and Meir, R. *Layered Neural Networks*. Springer Berlin Heidelberg, 1991, pp. 307–334.

[54] Domínguez, E., Lage-Castellanos, A., Mulet, R., Ricci-Tersenghi, F., and Rizzo, T. Characterizing and improving generalized belief propagation algorithms on the 2d Edwards-Anderson model. *J. Stat. Mech.: Theory and Experiment 2011*, 12 (2011), P12007.

[55] Duan, Y., Lv, Y., Liu, Y.-L., and Wang, F.-Y. An efficient realization of deep learning for traffic data imputation. *Transportation Research Part C: Emerging Technologies 72* (2016), 168 – 181.

[56] Dudoit, S., and Fridlyand, J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology 2*, 7 (2002), 0036.1–0036.21.

[57] Ermagun, A., and Levinson, D. Spatiotemporal traffic forecasting: review and proposed directions. *Transport Reviews 38*, 6 (2018), 786–814.

[58] Evans, M., Ferrari, P., and Mallick, K. Matrix representation of the stationary measure for the multispecies TASEP. *J. Stat. Phys. 135*, 2 (2009), 217–239.

[59] Evans, M., and Hanney, T. Nonequilibrium statistical mechanics of the zero-range process and related models. *J.Phys. A: Math. & Gen. 38*, 19 (2005), R195–R240.

[60] Evans, M. R., Majumdar, S. N., and Zia, R. K. P. Canonical analysis of condensation in factorized steady states. *J. Stat. Phys. 123*, 2 (2006), 357–390.

[61] Fan, J., and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. American Stat. Ass.* (2001).

[62] Fayolle, G., and Furtlehner, C. Dynamical windings of random walks and exclusion models. *J. Stat. Phys. 114*, 1-2 (2004), 229–260.

[63] Fayolle, G., and Furtlehner, C. Stochastic deformations of sample paths of random walks and exclusion models. In *Mathematics and computer science. III*, Trends Math. Birkhäuser, 2004, pp. 415–428.

[64] Fayolle, G., and Furtlehner, C. Stochastic dynamics of discrete curves and multi-type exclusion processes. *J. Stat. Phys. 127*, 5 (2007), 1049–1094.

[65] Fayolle, G., and Furtlehner, C. About hydrodynamic limit of some exclusion processes via functional integration. In *International Mathematical conference "50 years of IITP"*. 2011.

[66] Fayolle, G., Iasnogorodski, R., and Malyshev, V. *Random Walks in the Quarter-Plane*. Springer, 1999.

[67] Ferrari, P., and Martin, J. Stationary distributions of multi-type totally asymmetric exclusion processes. *Ann. Probab. 35*, 3 (2007), 807–832.

[68] Fradkin, E., Huberman, B., and Shenker, S. Gauge symmetries in random magnetic systems. *Phys. Rev. B 18* (1978), 4789–4814.

[69] Fraley, C., and Raftery, A. How many clusters?which clustering method? answer via model-based clustering. *The Computer Journal 41*, 8 (1998).

[70] Frey, B., and Dueck, D. Clustering by passing messages between data points. *Science 315* (2007), 972–976.

[71] Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*, 3 (2008), 432–441.

[72] Furtlehner, C. Approximate inverse Ising models close to a Bethe reference point. *J. Stat. Mech.* (2013), P09020.

[73] Furtlehner, C., and Decelle, A. Cycle-based cluster variational method for direct and inverse inference. *Journal of Statistical Physics 164*, 3 (2016), 531–574.

[74] Furtlehner, C., Han, Y., Lasgouttes, J.-M., Martin, V., Marchal, F., and Moutarde, F. Spatial and temporal analysis of traffic states on large scale networks. In *Proc. IEEE ITSC* (2010), vol. 13, pp. 1215 –1220.

[75] Furtlehner, C., and Lasgouttes, J. A queueing theory approach for a multi-speed exclusion process. In *Traffic and Granular Flow ' 07* (2007), pp. 129–138.

[76] Furtlehner, C., Lasgouttes, J., Attanasi, A., Meschini, L., and Pezzulla, M. Spatio-temporal Probabilistic Short-term Forecasting on Urban Networks. Research Report RR-9236, INRIA, 2018.

[77] Furtlehner, C., Lasgouttes, J., and Samsonov, M. One-dimensional particle processes with acceleration/braking asymmetry. *J. Stat. Phys 147*, 6 (2012), 1113–1144.

[78] Furtlehner, C., Lasgouttes, J.-M., and Auger, A. Learning multiple belief propagation fixed points for real time inference. *Physica A: Statistical Mechanics and its Applications 389*, 1 (2010), 149–163.

[79] Furtlehner, C., Lasgouttes, J.-M., and de La Fortelle, A. A belief propagation approach to traffic prediction using probe vehicles. In *Proc. IEEE ITSC* (2007), vol. 10, pp. 1022–1027.

[80] Furtlehner, C., Sebag, M., and Zhang, X. Scaling analysis of affinity propagation. *Phys. Rev. E 81* (2010), 066102.

[81] Fusco, G., Colombaroni, C., and Isaenko, N. Short-term speed predictions exploiting big data on large urban road networks. *Transportation Research Part C: Emerging Technologies 73* (2016), 183 – 201.

[82] Gabrié, M., Tramel, E., and Krzakala, F. Training restricted Boltzmann machine via the TAP free energy. In *Advances in Neural Information Processing Systems 28*. 2015, pp. 640–648.

[83] Gama, J., Rocha, R., and Medas, P. Accurate decision trees for mining highspeed data streams. In *SIGMOD* (2003), pp. 523–528.

[84] Gardner, E. Maximum storage capacity in neural networks. *EPL (Europhysics Letters) 4*, 4 (1987), 481.

[85] Gardner, E., and Derrida, B. Optimal storage properties of neural network models. *J. Phys. A: Math. & Gen. 21*, 1 (1988), 271.

[86] Gardner, E., and Derrida, B. Three unfinished works on the optimal storage capacity of networks. *J.Phys. A: Math. & Gen. 22*, 12 (1989), 1983–1994.

[87] Geiger, M., et al. Scaling description of generalization with number of parameters in deep learning. *J.Stat.Mech.: Theo. & Exp. 2020*, 2 (2020), 023401.

[88] Gelenbe, E., and Pujolle, G. *Introduction to Queueing Networks*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[89] Georges, A., and Yedidia, J. How to expand around mean-field theory using high-temperature expansions. *J. Phys. A: Math. & Gen. 24*, 9 (1991), 2173.

[90] Gerschenfeld, A., and Derrida, B. Current fluctuations at a phase transition. *EPL (Europhysics Letters) 96*, 2 (2011), 20001.

[91] Godrèche, C., and Luck, J. Nonequilibrium dynamics of urn models. *Journal of Physics: Condensed Matter 14*, 7 (2002), 1601–1615.

[92] Golinelli, G., and Mallick, K. The asymmetric simple exclusion process: an integrable model for non-equilibrium statistical mechanics. *J. Phys. A: Math. & Gen. 39*, 41 (2006), 12679.

[93] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS* (2014), pp. 2672–2680.

[94] Gwa, L.-H., and Spohn, H. Bethe solution for the dynamical-scaling exponent of the noisy Burgers equation. *Phys. Rev. A 46* (1992), 844–854.

[95] Haussler, D., Kearns, M., Opper, M., and Schapire, R. Estimating average-case learning curves using Bayesian, statistical physics and VC dimension methods. In *NIPS 4*. 1992, pp. 855–862.

[96] Helbing, D., and Johansson, A. F. On the controversy around Daganzo's requiem for and Aw-Rascle's resurrection of second-order traffic flow models. *The European Physical Journal B 69*, 4 (2009), 549–562.

[97] Heskes, T. On the uniqueness of loopy belief propagation fixed points. *Neural Computation 16* (2004), 2379–2413.

[98] Heskes, T., Albers, K., and Kappen, B. Approximate inference and constrained optimization. In *UAI* (2003).

[99] Hinkley, D. V. Inference about the change-point from cumulative sum tests. *Biometrika 58*, 3 (1971), 509–523.

[100] Hinton, G. *A Practical Guide to Training Restricted Boltzmann Machines*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 599–619.

[101] Hinton, G., and Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science 313*, 5786 (2006), 504–507.

[102] Hinton, G., and Zemel, R. Autoencoders, minimum description length and Helmholtz free energy. In *NIPS 6*. 1994, pp. 3–10.

[103] Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural computation 14* (2002), 1771–1800.

[104] Hinton, G. E., and Sejnowski, T. J. Learning and relearning in Boltzmann machines. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. MIT Press, 1986, pp. 282–317.

[105] Hjelm, R., Calhoun, V., Salakhutdinov, R., Allen, E., Adali, T., and Plis, S. Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks. *NeuroImage 96* (2014), 245–260.

[106] Höfling, H., and Tibshirani, R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihood. *JMLR 10* (2009), 883–906.

[107] Hopfield, J. J. Neural network and physical systems with emergent collective computational abilities. *PNAS 79* (1982), 2554–2558.

[108] Hsieh, C., Sustik, M. A., Dhillon, I. S., and Ravikumar, K. Sparse inverse covariance matrix estimation using quadratic approximation. In *NIPS* (2011).

[109] Hu, X., Huang, H., Peng, B., Han, J., Liu, N., Lv, J., Guo, L., Guo, C., and Liu, T. Latent source mining in fmri via restricted Boltzmann machine. *Human brain mapping 39*, 6 (2018), 2368–2380.

[110] Huang, H. Statistical mechanics of unsupervised feature learning in a restricted Boltzmann machine with binary synapses. *J. Stat. Mech.: Theor. & Exp. 2017*, 5 (2017), 053302.

[111] Huang, H., and Toyoizumi, T. Advanced mean-field theory of the restricted Boltzmann machine. *Phys. Rev. E 91*, 5 (2015), 050101.

[112] Iba, Y. The Nishimori line and Bayesian statistics. *J.Phys. A: Math. & Gen. 32*, 21 (1999), 3875–3888.

[113] Isaev, A., Pyatov, P. N., and Rittenberg, V. Diffusion algebras. *Journal of Physics A: Mathematical and General 34*, 29 (2001), 5815–5834.

[114] Jackson, J. R. Networks of waiting lines. *Operations Research 5*, 4 (1957), 518–521.

[115] Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *NIPS 31*. 2018, pp. 8571–8580.

[116] Jaynes, E. T. *Probability Theory: The Logic of Science (Vol 1)*. Cambridge University Press, 2003.

[117] Jiu-li, L., Van den Broeck, C., and Nicolis, G. Stability criteria and fluctuations around nonequilibrium states. *Zeitschrift für Physik B Condensed Matter 56*, 2 (1984), 165–170.

[118] Judd, D., McKinley, P., and Jain, A. Large-scale parallel data clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* (1998).

[119] Kabashima, Y., and Saad, D. Belief propagation vs. TAP for decoding corrupted messages. *Europhys. Lett. 44* (1998), 668.

[120] Kafri, Y., Levine, E., Mukamel, D., Schütz, G. M., and Török, J. Criterion for phase separation in one-dimensional driven systems. *Phys. Rev. Lett. 89* (2002), 035702.

[121] Kappen, H., and Rodríguez, F. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation 10*, 5 (1998), 1137–1156.

[122] Karakida, R., Okada, M., and Amari, S.-I. Dynamical analysis of contrastive divergence learning: Restricted Boltzmann machines with Gaussian visible units. *Neural Networks 79* (2016), 78 – 87.

[123] Kardar, M., Parisi, G., and Zhang, Y. Dynamic scaling of growing interfaces. *Phys. Rev. Lett. 56* (1986), 889–892.

[124] Karimipour, V. A multi-species ASEP and its relation to traffic flow. *Phys. Rev. E 59* (1999), 205.

[125] Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of CVPR* (2019), pp. 4401–4410.

[126] Kaupužs, J., Mahnke, R., and Harris, R. J. Zero-range model of traffic flow. *Phys. Rev. E 72* (2005), 056125.

[127] Kavitha, T., Liebchen, C., Mehlhorn, K., Michail, D., Rizzi, R., Ueckerdt, T., and Zweig, K. Cycle bases in graphs characterization, algorithms, complexity, and applications. *Computer Science Review 3*, 4 (2009), 199 – 243.

[128] Kelly, F. P. Networks of queues with customers of different types. *Journal of Applied Probability 12*, 3 (1975), 542–554.

[129] Kelly, F. P. *Reversibility and Stochastic Networks*. Cambridge University Press, New York, NY, USA, 2011.

[130] Kerner, B. *The Physics of Traffic*. Springer Verlag, 2005.

[131] Kikuchi, R. A theory of cooperative phenomena. *Phys. Rev. 81* (1951), 988–1003.

[132] Kingma, D., and Welling, M. Auto-encoding variational Bayes. In *ICLR* (2014).

[133] Kipnis, C., and Landim, C. *Scaling limits of Interacting Particles Systems*. Springer-Verlag, 1999.

[134] Kleinberg, J. An impossibility theorem for clustering. In *NIPS* (2002).

[135] Krauth, W., Mézard, M., and Nadal, J. Basins of attraction in a perceptron-like neural network. *Complex Systems 2* (1988), 387–408.

[136] Krauth, W., and Mézard, M. Storage capacity of memory networks with binary couplings. *J. Phys. France 50*, 20 (1989), 3057–3066.

[137] Krizhevsky, A. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009.

[138] Krzakala, F., Mézard, M., Sausset, F., Sun, Y., and Zdeborová, L. Statistical-physics-based reconstruction in compressed sensing. *Phys.Rev.X 2*, 2 (2012), 021005.

[139] Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L., and Zhang, P. Spectral redemption in clustering sparse networks. *PNAS 110*, 52 (2013), 20935–20940.

[140] Kschischang, F. R., Frey, B. J., and Loeliger, H. A. Factor graphs and the sum-product algorithm. *IEEE Trans. on Inf. Th. 47*, 2 (2001), 498–519.

[141] Lauritzen, S. *Graphical models*. Oxford University Press, USA, 1996.

[142] LeCun, Y., Bengio, Y., and Hinton, G. E. Deep learning. *Nature 521* (2015), 436–444.

[143] Lee, D., Baosheng, H., and Lin, X. Variable selection and estimation with the seamless-L_0 penalty. *Statistica Sinica 23*, 2 (2012), 929–962.

[144] Lee, S.-I., Ganapathi, V., and Koller, D. Efficient structure learning of Markov networks using L_1-regularization. In *NIPS* (2006).

[145] Leone, M., Sumedha, and Weigt, M. Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics 23* (2007), 2708.

[146] Leone, M., Sumedha, and Weigt, M. Unsupervised and semi-supervised clustering by message passing: Soft-constraint affinity propagation. *Eur. Phys. J. B* (2008), 125–135.

[147] Li, Y., Yu, R., Shahabi, C., and Liu, Y. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR* (2018).

[148] Liggett, T. M. *Interacting Particle Systems*. Springer, Berlin, 2005.

[149] Lighthill, M., and Whitham, G. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc. Royal Soc. London A. Math. & Phys. Sciences 229*, 1178 (1955), 317–345.

[150] Lin, H., Tegmark, M., and Rolnick, D. Why does deep and cheap learning work so well? *J.Stat.Phys. 168*, 6 (2017), 1223–1247.

[151] Lippi, M., Bertini, M., and Frasconi, P. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *Proc. IEEE ITSC 14* (2013), 871–882.

[152] Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F. Y. Traffic flow prediction with big data: A deep learning approach. *Proc. IEEE ITSC 16*, 2 (2015), 865–873.

[153] Malioutov, D., Johnson, J., and Willsky, A. Walk-sums and Belief Propagation in Gaussian graphical models. *JMLR 7* (2006), 2031–2064.

[154] Mallat, S. Understanding deep convolutional networks. *Phil. Trans. Royal Soc. A: Math., Phys. and Eng. Sciences 374*, 2065 (2016), 20150203.

[155] Mallick, K., and Sandow, S. Finite-dimensional representations of the quadratic algebra: Applications to the exclusion process. *J. Phys. A: Math. and Gen. 30*, 13 (1997), 4513–4526.

[156] Marčenko, V. A., and Pastur, L. A. Distribution of eigenvalues for some sets of random matrices. *Math. USSR-Sbornik 1*, 4 (1967), 457.

[157] Martin, V., Furtlehner, C., Han, Y., and Lasgouttes, J. GMRF estimation under topological and spectral constraints. In *In Proceedings of ECML PKDD* (2014), pp. 370–385.

[158] Martin, V., Lasgouttes, J., and Furtlehner, C. Latent binary MRF for online reconstruction of large scale systems. *Ann. of Math. and A.I.* (2015), 1–32.

[159] Mehta, P., Bukov, M., Wang, C., Day, A., Richardson, C., Fisher, C., and Schwab, D. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports 810* (2019), 1 – 124.

[160] Mézard, M. Mean-field message-passing equations in the Hopfield model and its generalizations. *Phys. Rev. E 95* (2017), 022117.

[161] Mézard, M., and Montanari, A. *Information, physics, computation: Probabilistic approaches*. Cambridge University Press, Cambridge, 2008.

[162] Mezard, M., and Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris 103*, 1-2 (2009), 107 – 113.

[163] Mézard, M., Parisi, G., and Virasoro, M. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

[164] Mézard, M., and Zecchina, R. The random K-satisfiability problem: from an analytic solution to an efficient algorithm. *Phys.Rev.E 66* (2002), 56126.

[165] Mézard, M., Nadal, J.P., and Toulouse, G. Solvable models of working memories. *J. Phys. France 47*, 9 (1986), 1457–1462.

[166] Mézard,M., and Nadal,J.P. Learning in feedforward layered networks: the tiling algorithm. *J.Phys.A: Math. & Gen. 22*, 12 (1989), 2191–2203.

[167] Minka, T. Expectation propagation for approximate Bayesian inference. In *Proceedings UAI* (2001), pp. 362–369.

[168] Monasson, R., and Tubiana, J. Emergence of compositional representations in restricted Boltzmann machines. *Phys. Rev. Let. 118* (2017), 138301.

[169] Mooij, J. M., and Kappen, H. J. On the properties of the Bethe approximation and loopy belief propagation on binary network. *J. Stat. Mech.* (2005), P11012.

[170] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D., Sander, C., Zecchina, R., Onuchic, J., Hwa, T., and Weigt, M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *PNAS 108*, 49 (2011), E1293–E1301.

[171] Morita, T. Cluster variation method and Möbius inversion formula. *J. Stat. Phys. 59*, 3-4 (1990), 819–825.

[172] Nagel, K., and Paczuski, M. Emergent traffic jams. *Phys. Rev. E 51*, 4 (1995), 2909–2918.

[173] Nagel, K., and Schreckenberg, M. A cellular automaton model for freeway traffic. *J. Phys. I,2* (1992), 2221–2229.

[174] Nair, V., and Hinton, G. Rectified linear units improve restricted Boltzmann machines. In *ICML '10* (2010), pp. 807–814.

[175] Ng, A., Jordan, M., and Yair, W. On spectral clustering: Analysis and an algorithm. In *NIPS 14*. 2002, pp. 849–856.

[176] Nguyen, H., and Berg, J. Bethe-Peierls approximation and the inverse Ising model. *J. Stat. Mech.*, 1112.3501 (2012), P03004.

[177] Nguyen, H., and Berg, J. Mean-field theory for the inverse Ising problem at low temperatures. *Phys. Rev. Lett. 109* (2012), 050602.

[178] Nishimori, H. *Statistical Physics of Spin Glasses and Information Processing: An Introduction*. Oxford University Press, 2001.

[179] Opper, M. Learning and generalization in a two-layer neural network: The role of the VC dimension. *Phys. Rev. Lett. 72* (1994), 2113–2116.

[180] Opper, M., and Haussler, D. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Phys. Rev. Lett. 66* (1991), 2677–2680.

[181] Opper, M., and Winther, O. Adaptive and self-averaging TAP mean field theory for probabilistic modeling. *Phys. Rev. E 64* (2001), 056131.

[182] Page, E. S. Continuous inspection schemes. *Biometrika 41*, 1-2 (1954), 100–115.

[183] Pakzad, P., and Anantharam, V. Estimation and marginalization using the Kikuchi approximation methods. *Neural Computation 17*, 8 (2005), 1836–73.

[184] Parisi, G., and Potters, M. Mean-field equations for spin models with orthogonal interaction matrices. *J. Phys. A: Math. & Gen. 28*, 18 (1995), 5267.

[185] Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. Morgan Kaufmann, 1988.

[186] Pelizzola, A. Cluster variation method in statistical physics and probabilistic graphical models. *J.Phys.A: Math. & Gen. 38*, 33 (2005), R309–R339.

[187] Percus, J. K., and Zhang, M. Q. One-dimensional inhomogeneous Ising model with periodic boundary conditions. *Phys. Rev. B 38* (1988), 11737–11740.

[188] Personnaz, L., Guyon, I., and Dreyfus, G. Information storage and retrieval in spin-glass like neural networks. *J. Phys. Lett. 46*, 8 (1985), 359–365.

[189] Plefka, T. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J. Phys. A: Mathematical and General 15*, 6 (1982), 1971.

[190] Polson, N. G., and Sokolov, V. O. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies 79* (2017), 1 – 17.

[191] Priezzhev, V. B. Exact nonstationary probabilities in the asymmetric exclusion process on a ring. *Phys. Rev. Lett. 91* (2003), 050601.

[192] Prigogine, I., and Kondepudi, D. *Thermodynamique: des moteurs thermiques aux structures dissipatives*. Sciences. Odile Jacob, Paris, 1999.

[193] Richards, P. Shock waves on the highway. *Operations Research 4*, 1 (1956), 42–51.

[194] Rish, I., Brodie, M., Ma, S., Odintsova, N., Beygelzimer, A., Grabarnik, G., and Hernandez, K. Adaptive diagnosis in distributed systems. *IEEE Transactions on Neural Networks 16* (2005), 1088–1109.

[195] Ruozzi, N. *Message Passing Algorithms for Optimization*. PhD thesis, Yale University, 2011.

[196] Salakhutdinov, R., and Hinton, G. Deep Boltzmann machines. In *Artificial Intelligence and Statistics* (2009), pp. 448–455.

[197] Salazar, D. Nonequilibrium thermodynamics of restricted Boltzmann machines. *Phys. Rev. E 96* (2017), 022131.

[198] Samsonov, M., Furtlehner, C., and Lasgouttes, J. Exactly solvable stochastic processes for traffic modelling. Tech. Rep. 7278, INRIA, 2010.

[199] Savit, R. Duality in field theory and statistical systems. *Rev. Mod. Phys. 52*, 2 (1980), 453–487.

[200] Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, 2014.

[201] Scheinberg, K., and Rish, I. Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach. In *ECML-PKDD* (2010).

[202] Schnakenberg, J. Network theory of behavior of master equation systems. *Rev. Mod. Phys. 48*, 4 (1976).

[203] Schneidman, E., Berry, M., Segev, R., and Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature 440* (2006), 1007–1012.

[204] Schönhof, M., and Helbing, D. Criticism of three-phase traffic theory. *Transportation Research 43* (2009), 784–797.

[205] Schreckenberg, M., Schadschneider, A., Nagel, K., and Ito, N. Discrete stochastic models for traffic flow. *Phys. Rev. E 51* (1995), 2339.

[206] Schütz, G. Exactly solvable models for many-body systems far from equilibrium. In *Phase Transitions and Critical Phenomena*, C. Domb and J. Lebowitz, Eds., vol. 19. (Academic Press, San Diego), 2001.

[207] Seung, H. S., Sompolinsky, H., and Tishby, N. Statistical mechanics of learning from examples. *Phys. Rev. A 45* (1992), 6056–6091.

[208] Shimony, S. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence 68*, 2 (1994), 399 – 410.

[209] Shwartz-Ziv, R., and Tishby, N. Opening the black box of deep neural networks via information. arXiv 1703.00810, 2017.

[210] Silver, D., et al. Mastering the game of Go with deep neural networks and tree search. *Nature 529* (2016), 484–489.

[211] Smolensky, P. *In Parallel Distributed Processing: Volume 1 by D. Rumelhart and J. McLelland.* 194-281. MIT Press, 1986, ch. 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory.

[212] Speed, T., and Kiiveri, H. Gaussian Markov distributions over finite graphs. *The Annals of Statistics 14*, 1 (1986), 138–150.

[213] Spitzer, F. Interaction of Markov processes. *Adv. Math.* **5** (1970), 246.

[214] Spohn, H. *Large Scale Dynamics of Interacting Particles.* Springer, 1991.

[215] Still, S., and Bialek, W. How many clusters?:an information-theoretic perspective. *Neural Computation 16* (2004), 2483–2506.

[216] Sugiyama, Y., et al. Traffic jams without bottlenecks: experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics 10* (2008), 1–7.

[217] Sun, S., Huang, R., and Gao, Y. Network-scale traffic modeling and forecasting with graphical lasso and neural networks. *Journal of Transportation Engineering 138*, 11 (2012), 1358–1367.

[218] Takahashi, C., and Yasuda, M. Mean-field inference in Gaussian restricted Boltzmann machine. *Journal of the Physical Society of Japan 85*, 3 (2016), 034001.

[219] Talagrand, M. Rigorous results for the Hopfield model with many patterns. *Probab. Th. Relat. Fields 110* (1998), 177–276.

[220] Thouless, D. J., Anderson, P. W., and Palmer, R. G. Solution of 'solvable model of a spin glass'. *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics 35*, 3 (1977), 593–601.

[221] Tieleman, T. Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML '08* (2008), pp. 1064–1071.

[222] Tipping, M. E., and Bishop, C. M. Mixtures of probabilistic principal component analyzers. *Neural Comput. 11*, 2 (1999), 443–482.

[223] Touchette, H. The large deviation approach to statistical mechanics. *Physics Reports 478* (2009), 1–69.

[224] Tramel, E., Gabrié, M., Manoel, A., Caltagirone, F., and Krzakala, F. A Deterministic and Generalized Framework for Unsupervised Learning with Restricted Boltzmann Machines. *Phys. Rev. X 8* (2018), 041006.

[225] Tubiana, J. *Restricted Boltzmann machines: from compositional representations to protein sequence analysis*. PhD thesis, 2018.

[226] Valiant, L. A theory of the learnable. In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing* (1984), STOC '84, pp. 436–445.

[227] Vanicat, M. *Approche intégrabiliste des modèles de physique statistique hors d'équilibre*. PhD thesis, Université Grenoble Alpes, 2016.

[228] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, Berlin, Heidelberg, 1995.

[229] Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. Short-term traffic forecasting: Where we are and where we're going. *Transportation Research Part C: Emerging Technologies 43, Part 1* (2014), 3 – 19. Special Issue on Short-term Traffic Flow Forecasting.

[230] Wainwright, M., and Jordan, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn. 1*, 1-2 (2008), 1–305.

[231] Wang, K., Zhang, J., Li, D., Zhang, X., and Guo, T. Adaptive affinity propagation clustering. *Acta Automatica Sinica 33*, 12 (2007), 1242–1246.

[232] Watanabe, Y., and Fukumizu, K. Graph zeta function in the Bethe free energy and loopy belief propagation. In *NIPS* (2009), vol. 22, pp. 2017–2025.

[233] Weiss, Y., and Freeman, W. T. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Comput. 13*, 10 (2001), 2173–2200.

[234] Welling, M. On the choice of regions for generalized belief propagation. In *UAI '04* (2004), pp. 585–592.

[235] Welling, M., Minka, T., and Teh, Y. Structured region graphs: Morphing EP into GBP. In *UAI* (2005), vol. 21.

[236] Welling, M., and Teh, Y. Approximate inference in Boltzmann machines. *A.I. 143*, 1 (2003), 19–50.

[237] Wiseman, S., Blatt, M., and Domany, E. Super-paramagnetic clustering of data. *Phys. Rev. E 57* (1998), 3767–3787.

[238] Yamashita, T., Tanaka, M., Yoshida, E., Yamauchi, Y., and Fujiyoshii, H. To be Bernoulli or to be Gaussian, for a restricted Boltzmann machine. In *ICPR 22* (2014), pp. 1520–1525.

[239] Yasuda, M., and Tanaka, K. The mathematical structure of the approximate linear response relation. *J. Phys. A: Math. and Theor. 40*, 33 (2007), 9993.

[240] Yasuda, M., and Tanaka, K. Approximate learning algorithm in Boltzmann machines. *Neural Comp. 21* (2009), 3130–3178.

[241] Yasuda, M., and Tanaka, K. Susceptibility propagation by using diagonal consistency. *Phys. Rev. E 87* (2013), 012134.

[242] Yedidia, J. S., Freeman, W. T., and Weiss, Y. Generalized belief propagation. *NIPS* (2001), 689–695.

[243] Yelmen, B., Decelle, A., Ongaro, L., Marnetto, D., Tallec, C., Montinaro, F., Furtlehner, C., Pagani, L., and Jay, F. Creating artificial human genomes using generative models. *bioRxiv* (2019), 769091.

[244] Yu, H., Wu, Z., Wang, S., Wang, Y., and Ma, X. Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks. *Sensors 17*, 7 (2017).

[245] Yuille, A. L. CCCP algorithms to minimize the Bethe and Kikuchi free energies: convergent alternatives to belief propagation. *Neural Computation 14* (2002), 1691–1722.

[246] Zdeborová, L., and Krzakala, F. Statistical physics of inference: thresholds and algorithms. *Advances in Physics 65*, 5 (2016), 453–552.

[247] Zhang, X., Furtlehner, C., and Sebag, M. Data streaming with affinity propagation. In *ECML/PKDD* (2008), pp. 628–643.