Compositional statistical mechanics, entropy and variational inference

Grégoire Sergeant-Perthuis

▶ **To cite this version:**

Grégoire Sergeant-Perthuis. Compositional statistical mechanics, entropy and variational inference. Twelfth Symposium on Com- positional Structures (SYCO 12), Apr 2024, Birmingham (UK), United Kingdom. hal-04518736

HAL Id: hal-04518736

https://hal.sorbonne-universite.fr/hal-04518736v1

Submitted on 24 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Compositional statistical mechanics, entropy and variational inference

Grégoire Sergeant-Perthuis*[1]

[1]Laboratoire de Biologie Computationnelle et Quantitative, Sorbonne Université, Paris , France , ORCiD: 0000-0002-2079-3410

March 24, 2024

Abstract

In this document, accepted for the Twelfth Symposium on Compositional Structures (SYCO 12) [1], we aim to gather various results related to a compositional/categorical approach to *rigorous* Statistical Mechanics [2–9]. Rigorous Statistical Mechanics is centered on the mathematical study of statistical systems. Central concepts in this field have a natural expression in terms of diagrams in a category that couples measurable maps and Markov kernels [8]. We showed that statistical systems are particular representations of partially ordered sets (posets), that we call $\mathscr{A}$-specifications, and expressed their phases, i.e., Gibbs measures, as invariants of these representations. It opens the way to the use of homological algebra to compute phases of statistical systems. Two central results of rigorous Statistical Mechanics are, firstly, the characterization of extreme Gibbs measure as it relates to the zero–one law for extreme Gibbs measures, and, secondly, their variational principle which states that for translation invariant Hamiltonians, Gibbs measures are the minima of the Gibbs free energy. We showed in [9] how the characterization of extreme Gibbs measures extends to $\mathscr{A}$-specifications; we proposed in [10] an Entropy functional for $\mathscr{A}$-specifications and gave a message-passing algorithm, that generalized the belief propagation algorithm of graphical models (see O. Peltre's SYCO 8 talk, [11] and [12]), to minimize variational free energy.

1 Introduction

Context: Statistical physics is a framework that focuses on the probabilistic description of complex systems: a collection of interacting 'particles'

*gregoireserper@gmail.com

or components of a whole in most generality [13]. Its main feature is to introduce an energy function H for the system, that associates to any of its configurations a real value; the probability p of a configuration is given in terms of the Boltzmann distribution ($p = e^{-\beta H} / \int dx e^{-\beta H(x)}$), which associates high-energy configurations to unlikely events. We will call 'statistical system' the complex system provided with the Boltzmann distribution. Statistical physics serves as a rich framework for probabilistic modeling. It has several names depending on the community [14]; for example, it is called 'energy-based modeling' in machine learning, the probabilistic model is called a Gibbs Random Field in graphical modeling and is a particular case of a statistical system. It is widely used in engineering, for example: computational structural biology (in computational statistical physics: Hamiltonian Monte Carlo) [15, 16], robotics (reinforcement learning: Markov Chains, Markov Decision Processes) [17], and more generally for modeling interaction and dependencies of random variables using graphical models [18].

Applied Category Theory: New foundations, based on topology and geometry, were proposed for probability theory, information theory, and deep learning [19–23]. More generally, they fall in a field of research that has recently emerged, Applied Category Theory, which focuses on applying these principles to engineering [20–22, 24–26]. Recent results give a characterization of the zero–one law for independent random variables and for Markov chains in a categorical formulation [27, 28]. The zero–one law for extreme Gibbs measures is known to extend the ones of independent random variables and Markov chains [29], so it would be expected that the categorical formulation of extreme Gibbs measures we propose may also relate to the categorical formulation developed in the case of independent random variables and Markov chains.

Motivation: A landmark rigorous formulation of Statistical Mechanics can be found in Georgii's *Gibbs Measures and Phase Transition* [29]. Such a framework, which revolves around the concepts of 'specification' and Hamiltonian, is needed to define rigorously pure phases or, more generally, Gibbs measures of statistical systems. For a given Hamiltonian, there can be multiple phases only for infinitely many interacting particles. The main constructions of such a framework rely on the necessity to have a *universe*, denoted Ω, that encompasses all possible configurations of the system, i.e., all the possible joint configurations of the particles $\Omega := \prod_{i \in \mathbb{N}} E_i$, with E_i the state space of the particle i. There are important limitations to making references to a 'global' universe [19]. Let us state two of them. The first one is that it can be difficult to compute with Hamiltonians (e.g., computing free energies [15], expectations of observables [30], infinite volume Gibbs measures [29]) as it requires summing over a set which 'size' increases exponentially with the number of components (variables) involved in the system. The second one is that they don't account for heterogeneity, incompleteness, or incompatibility in the description of the system to model: by definition,

each configuration of the system corresponds to observing simultaneously and without loss of information the state of all its components. This second point is related to the first one as to model possible configurations of a system it is required to weight all possible simultaneous configurations of its particles; this forbids granularity of the description of the system: invariably, the dimension of the space of possible models increases exponentially with the number of (elementary) particles. Mathematically, Hamiltonians are sums of local potentials, however, the relation between potentials and their phases is highly nontrivial and noncompositional. In other words, the Hamiltonian point of view does not allow for building complex statistical systems from simple ones in a way that allows controlling and computing the phases of the associated statistical systems from the phases of the simpler one. The relation between the complexity of the interactions that appear in the statistical system and its number of (pure) phases (more than 2) is known only for very specific systems (e.g. Ising model); results in this direction are considered difficult problems [31, 32]. In our approach, we want to forget about the 'Ω' and we want to study statistical systems 'locally' focusing on their local interactions and how they compose. We propose a constructive approach: 'composing' statistical systems together by building in a controlled and computable manner nontrivial statistical systems from simpler ones. This is what reformulating specifications as presheaves allows us to do, in particular, by making use of operations such as coproducts and products but also geometric morphisms. The approach we propose also has the advantage to model statistical systems for which it is not possible to have complete knowledge of the states of the particles at the same time and that could also have conflicting local descriptions inconsistent with a global description.

Contribution: In this paper, we gather various results related to a compositional/categorical approach to 'rigorous' Statistical Mechanics [2–9].

Our motivation in [7] was to build a bridge between geometry and rigorous statistical physics by showing that the standard formal definition of a statistical system, a 'specification' (Definition 1.23 [29]), can be identified with a particular representation of a partially ordered set (poset), i.e., a functor of a poset in a certain category. For a given poset $\mathscr{A}$, we call such representation an $\mathscr{A}$-specification. Representations of posets have a precise geometric interpretation [33–37], and there is a rich literature coming from algebra, geometry, and topology to study them. We showed that phases of statistical systems are geometric invariants of these representations and computed them for 'projective' poset representations. We remarked that, in our setting, systems of 'finite size' can have multiple phases, which is not possible in the current formalism for phase transition, but we do not claim that those finite size systems capture correctly the possible phase transitions of infinite size systems. However, we believe that the algebraic treatment of statistical systems we propose to be fairly similar for finite size systems and

infinite size systems.

In [4], we gave a characterization of independent variables in terms of projective objects in the category of presheaves over a poset and an easy-to-verify condition that characterizes such objects. Injective representations were characterized in [3], and their relationship with the marginal extension problem is studied in [5]; a unifying perspective for Hilbert spaces can be found in [6]. In [10], we proposed a general framework for optimization of presheaves which in particular allows defining an Entropy functional for $\mathscr{A}$-specifications. In [9], we showed how the characterization of extreme Gibbs measures, one of the steps for proving a zero-one law for the extreme Gibbs measures, transfers to $\mathscr{A}$-specifications.

2 Structure of the Paper and Contribution

We start by recalling some important notions of rigorous statistical mechanics [29] (specifications, Gibbs measures). We then show that the standard concept that encodes the statistical system, which is the notion of 'specifications', can be extended into a poset representation; we also show that the Gibbs measures of a statistical system are geometric invariants of this representation. From there, we propose a novel categorical formulation for statistical mechanics; we define 'generalized specifications' and 'generalized Gibbs measures' for those specifications: the $\mathscr{A}$-specifications. We characterize the generalized Gibbs measures for projective $\mathscr{A}$-specifications and their Gibbs measures. We then recall what extreme Gibbs measures are and their characterization on the tail σ-algebra. We extend such characterization to extreme Gibbs measures of $\mathscr{A}$-specifications. Finally, we recall what the Bethe free energy is in the context of graphical models and what the Belief Propagation algorithm is; when the graphical model is a Hidden Markov model (HMM), the Belief Propagation relates to Kalman filtering. We explain how we can define such variational free energy for $\mathscr{A}$-specifications. We then give a message-passing algorithm for finding critical points of the variational free energy of $\mathscr{A}$-specifications.

3 Background: Rigorous Statistical Mechanics

We will follow the presentation of Georgii's reference book *Gibbs Measures and Phase Transitions* [29].

Definition 1 (Markov Kernel)**.** A Markov kernel k from the measurable space $(E, \mathscr{E})$ to the measurable space $(E_1, \mathscr{E}_1)$ is a function $k : \mathscr{E}_1 \times E \to [0, 1]$ such that

1. $\forall \omega \in E$, $k(.|\omega)$ is a measure on E_1

2. $\forall A \in \mathscr{E}_1$, $k(A|.)$ is a measurable map from E to $\mathbb{R}$

3. $\forall \omega \in E$, $k(E_1|\omega) = 1$, i.e., $k(.|\omega)$ is a probability measure.

We will denote a Markov kernel k from $(E, \mathscr{E})$ to $(E_1, \mathscr{E}_1)$ as $k : (E, \mathscr{E}) \to (E_1, \mathscr{E}_1)$. We denote $[(E, \mathscr{E}), (E_1, \mathscr{E}_1)]_K$, the set of kernels from $(E, \mathscr{E})$ to $(E_1, \mathscr{E}_1)$; if there is no ambiguity on the σ-algebras the spaces are provided with, we will simply denote it as $[E, E_1]_K$. We denote $\mathbb{P}(E)$, the space of probability distributions over E; it is a measurable space for the smallest σ-algebra that makes the evaluation maps, on measurable sets of E, measurable.

Markov kernels can be composed as follows. Let $k : E \to E_1$, $k_1 : E_1 \to E_2$ be two Markov kernels, then the composition $k_1 \circ k : E \to E_2$ is the following Markov kernel: for any $A \in \mathscr{E}_2$ and $\omega \in E$,

$$k_1 \circ k(A|\omega) = \int k_1(A|\omega_1) k(d\omega_1|\omega) \tag{3.1}$$

A measurable map $f : (E, \mathscr{E}) \to (E_1, \mathscr{E}_1)$ between two measurable spaces can be extended into the following Markov kernel: for any $A \in \mathscr{E}_1$ and $\omega \in E$,

$$k_f(A|\omega) = 1[f(\omega) \in A] \tag{3.2}$$

To avoid having too many notations, we will denote k_f as f and the context will specify if f refers to the measurable map or its extension; for example, a composition $k \circ f$ between a Markov kernel k and f necessarily means that, here, f refers to k_f. We will also denote $k_1 \circ k$ as $k_1 k$.

A probability measure $p \in \mathbb{P}(E)$ can also be identified with the following Markov kernel k_p from $*$, the measurable space with one element, to $(E, \mathscr{E})$; for any $A \in \mathscr{E}$, $k_p(A|*) = p(A)$. Similarly, we identify p and k_p.

Remark 1. Measurable spaces and measurable maps form a category. Giry [38] and Lawvere [39] are the first to have remarked that measurable spaces and Markov kernels also form a category; the latter category is the Kleisli category of the first.

Definition 2 (Proper Kernel, Section 1.1. [29])**.** Let $\mathscr{E}_1 \subseteq \mathscr{E}$ be two σ-algebras of a set E, a kernel $k \in [(E, \mathscr{E}_1), (E, \mathscr{E})]_K$ is proper if and only if, for any $A \in \mathscr{E}$, $B \in \mathscr{E}_1$,

$$k(A \cap B|.) = k(A|.)1_B \tag{3.3}$$

Let us set the notations. I is the set of components of a complex system. $(E_i, \mathscr{E}_i, i \in I)$ is a collection of measurable spaces, with each E_i being the space of configuration (state space) of the component $i \in I$. $(E, \mathscr{E})$ denotes the state space of the system; it is the product $E := \prod_{i \in I} E_i$ with the

product σ-algebra. For a sub-collection of components $a \subseteq I$, we denote $E_a := \prod_{i\in a} E_i$ the associated state space, and $\mathscr{E}_a$ the associated product σ-algebra. $i^a : E \to E_a$ is the projection that sends a configuration $\omega := (\omega_i, i \in I)$ to the configuration of the sub-collection a, $(\omega_i, i \in a)$. Finally, let us denote $\mathscr{P}_f(I)$ the set of finite subsets of I.

Definition 3 (Specification, Adaptation of Def. 1.23 [29])**.** A specification γ with state space $(E, \mathscr{E})$ is a collection $(\gamma_a, a \in \mathscr{P}_f(I))$ of proper Markov kernels such that for any $a \in \mathscr{P}_f(I)$, $\gamma_a \in [(E_{\overline{a}}, \mathscr{E}_{\overline{a}}), (E, \mathscr{E})]_K$ and which satisfies that for any $b \subseteq a$, i.e $\overline{a} \subseteq \overline{b}$ and any $A \in \mathscr{E}$,

$$\gamma_b \circ i^{\overline{b}} \circ \gamma_a(A|.) = \gamma_a(A|.) \tag{3.4}$$

Remark 2. In Definition (1.23) [29], E is the product over the same measurable space X over I, and I is a countably infinite set.

In the standard definition of a specification (Definition 3), the Markov kernels encode border conditions for experiments that only involve finite numbers of components ($a \in \mathscr{P}_f(I)$). It is what formally encodes the statistical system. The next definition defines the 'phases' of the system.

Definition 4 (Gibbs measures, Def. 1.23 [29])**.** Let γ be a specification with state space E; the set of probability measures,

$$\mathscr{G}(\gamma) := \{p \in \mathbb{P}(E) : \quad \mathbb{E}_p[A|\mathscr{E}_{\overline{a}}](\omega) = \gamma_a(A|\omega_a) \ p \text{ a.s.}\} \tag{3.5}$$

is the set of Gibbs measures of γ.

One of the central problems of (rigorous) statistical mechanics is to understand the relationship between a specification γ (statistical system) and its set of Gibbs measures $\mathscr{G}(\gamma)$ (its phases).

4 Statistical Systems as poset representations

Theorem 1. *Let γ be a specification with state space E. For any $a, b \in \mathscr{P}_f(I)$ such that $b \subseteq a$, there is a unique Markov kernel $F_b^a : E_{\overline{a}} \to E_{\overline{b}}$ such that the following diagram commutes,*

$$\begin{array}{ccc} E_{\overline{a}} & \xrightarrow{\gamma_a} & E \\ {\scriptstyle F_b^a}\downarrow & \nearrow_{\gamma_b} & \\ E_{\overline{b}} & & \end{array} \tag{4.1}$$

i.e. such that $\gamma_b \circ F_b^a = \gamma_a$. Furthermore for any collection $a, b, c \in \mathscr{P}_f(I)$ with $a \subseteq b \subseteq c$,

$$F_c^b \circ F_b^a = F_c^a \tag{4.2}$$

Proof. Let $b \subseteq a$, let F_b^a satisfy the commutative diagram 4.1, then,

$$i^{\overline{b}}\gamma_b F_b^a = i^{\overline{b}}\gamma_a \tag{4.3}$$

therefore,

$$F_b^a = i^{\overline{b}}\gamma_a \tag{4.4}$$

For any $a, b \in \mathscr{A}$ such that $b \subseteq a$ let $F_b^a = i^{\overline{b}}\gamma_a$ then for $c \subseteq b \subseteq a$,

$$F_c^b F_b^a = i^{\overline{c}}\gamma_b i^{\overline{b}}\gamma_a \tag{4.5}$$

Equation 3.4 can be rewritten as, for $b \subseteq a$,

$$\gamma_b i^{\overline{b}}\gamma_a = \gamma_a \tag{4.6}$$

therefore,

$$F_c^b F_b^a = F_c^a \tag{4.7}$$

□

Definition 5 (Partially ordered set)**.** A partially ordered set (poset), $(\mathscr{A}, \leq)$, is a set $\mathscr{A}$ provided with a binary relation $\leq\colon \mathscr{A} \times \mathscr{A} \to \{0, 1\}$, such that

1. reflexive: $\forall a \in \mathscr{A}, a \leq a$
2. transitive: if $c \leq b$ and $b \leq a$ then $c \leq a$
3. antisymmetric: if $b \leq a$ and $a \leq b$ then $a = b$

$(\mathscr{P}_f(I), \subseteq)$ is a poset for the inclusion relation; $(\mathscr{P}_f(I), \supseteq)$ is also a poset for the reversed inclusion relation: $b \supseteq a \iff a \subseteq b$. The convention is to denote $(\mathscr{P}_f(I), \supseteq)$ as $\mathscr{P}_f(I)^{op}$ because what relates $(\mathscr{P}_f(I), \supseteq)$ to $\mathscr{P}_f(I)$ is the fact that the order is 'opposed'; the same convention holds for any poset; $\mathscr{A}^{op}$ is the set $\mathscr{A}$ with opposed order.

We call a representation of the poset $\mathscr{A}$: a collection of 'spaces' $(G(a), a \in \mathscr{A})$ and a collection of 'maps' $(G_a^b : G(b) \to G(a); b, a \in \mathscr{A}, b \leq a)$ which satisfies for any $c \leq b \leq a$, $G_c^b \circ G_b^a = G_c^a$. We keep the notion of 'poset representation' a bit vague for now (we do not say what 'spaces' or 'maps' are); we keep this notion vague at this stage but in the next section will make this notion formal by introducing the concept of category and of functor.

We call a representation of the poset $\mathscr{A}$: a collection of 'spaces' $(G(a), a \in \mathscr{A})$ and a collection of 'maps' $(G_a^b : G(b) \to G(a); b, a \in \mathscr{A}, b \leq a)$ which

satisfies for any $c \leq b \leq a$, $G_c^b \circ G_b^a = G_c^a$. We keep the notion of 'poset representation' a bit vague for now (we do not say what 'spaces' or 'maps' are); we keep this notion vague at this stage but in the next section will make this notion formal by introducing the concept of category and of functor.

Theorem 1 implies that a specification γ can be promoted to a representation F of $(\mathscr{P}_f(I), \supseteq)$.

Theorem 2. *Let γ be a specification with state space E, let $p \in \mathscr{G}(\gamma)$. For any $a \in \mathscr{A}$ let $p_a := i^{\overline{a}} \circ p$; it is the marginal distribution on $E_{\overline{a}}$ of p. Then, for any $a, b \in \mathscr{P}_f(I)$ such that $b \subseteq a$,*

$$F_b^a \circ p_a = p_b \tag{4.8}$$

Proof. For any $a, b \in \mathscr{P}_f(I)$ such that $b \subseteq a$,

$$\gamma_b \circ (F_b^a p_a) = p \tag{4.9}$$

therefore, $F_b^a p_a = i^{\overline{b}} p$ and

$$F_b^a p_a = p_b \tag{4.10}$$

□

5 Categorical formulation of specification and Gibbs measures

We denote categories in bold, e.g. **C**. We will denote **Mes** as the category that has as objects measurable spaces and as morphisms measurable maps (Section 1 [38]); we will denote **Kern** as the category that has as objects measurable spaces and as morphisms Markov kernels (the Kleisli category of the monad **Mes**).

A poset, $(\mathscr{A}, \leq)$, can be seen as a category, **A**, with at most one morphism between two objects: the objects of **A** are the elements of $\mathscr{A}$ and for any two elements $b, a \in \mathscr{A}$ there is one morphism $b \to a$ when $b \leq a$. From now on, we will drop the bold notation for the category **A** and denote it simply as $\mathscr{A}$. A functor G from a poset $\mathscr{A}$ to a category **C** is precisely a collection of maps G_a^b for $b, a \in \mathscr{A}$ such that $b \leq a$, which satisfy $G_a^b \circ G_b^c = G_a^c$ for any three elements $c \leq b \leq a$. A functor from $\mathscr{A}$ to some target category is what we will call a representation of the poset $\mathscr{A}$; in general, the target category is the category of vector spaces or modules [33]. For this article, the target category will be **Mes** and **Kern**.

Consider a functor $G : \mathscr{A} \to \mathbf{Set}$ from a poset $\mathscr{A}$ to the category of sets. A collection $(\omega_a, a \in \mathscr{A})$ is called a section of G if for any $b \leq a$, $G_a^b(\omega_b) = \omega_a$; the set of sections of a poset representation is called the limit of G (III.4 [40]) and denoted $\lim G$. It is an 'invariant' of G that can

be computed using homological algebra when the target category of G is enriched with some algebraic structure (Chapter 13 [41]).

In Theorem 1 we showed that we can associate to a specification γ, a functor from $(\mathscr{P}_f(I), \supseteq)$ to **Kern**. The convention is to call a functor with source $\mathscr{A}^{op}$, a *presheaf*. In Theorem 2 we showed that Gibbs measures of γ are 'sections' of F. We will denote this set of 'sections' as $[*, F]_{K,\mathscr{A}}$; more precisely for a functor $F : \mathscr{A} \to \mathbf{Kern}$,

$$[*, F]_{K,\mathscr{A}} := \{(p_a \in \mathbb{P}(F(a)), a \in \mathscr{A}) | \quad \forall b \leq a, F_a^b p_b = p_a\} \tag{5.1}$$

Introducing the presheaf F and $[*, F]_{K,\mathscr{A}}$ is our way to emphasize that the compatible measures $p \in [*, F]_{K,\mathscr{A}}$ don't have to be measures over the whole space E. There is no need to define statistical systems 'globally', one can also define them 'locally'.

Let us now introduce the more general, categorical setting we propose for statistical systems.

Definition 6 (Generalized Specification, $\mathscr{A}$-Specifications)**.** Let $\mathscr{A}$ be a poset, a generalized specification over $\mathscr{A}$, or simply $\mathscr{A}$-specification, is a couple (G, F) of a presheaf and a functor where $G : \mathscr{A}^{op} \to \mathbf{Mes}$ and $F : \mathscr{A} \to \mathbf{Kern}$ are such that for any $a, b \in \mathscr{A}$ with $b \leq a$,

$$G_b^a F_a^b = \mathrm{id} \tag{5.2}$$

In the previous definition, G, in the particular case of $E = \prod_{i \in I} E_i$, encodes the collection of projections $i_b^a : E_a \to E_b$ for $b \subseteq a$; it is in some way the 'skeleton' of the spaces of observables of the statistical system. It is a key ingredient for the generalization of (rigorous) statistical mechanics to a categorical framework.

Definition 7 (Gibbs measures for $\mathscr{A}$-specifications)**.** Let $\gamma = (G, F)$ be an $\mathscr{A}$-specification, we call the Gibbs measures of γ the sections of F,

$$\mathscr{G}_g(\gamma) := [*, F]_{K,\mathscr{A}} \tag{5.3}$$

6 Gibbs measure of projective $\mathscr{A}$-specifications

Let E be a measurable space; we denote $L^\infty(E)$ the set of bounded, real-valued, measurable functions over E. One associates to a Markov kernel $F : E_b \to E_a$ a linear map $\pi : L^\infty(E_a) \to L^\infty(E_b)$ defined as follows: for any $f \in L^\infty(E_a)$,

$$\forall \omega_b \in E_b, \quad \pi(f)(\omega_b) = \int f(\omega_a) F(d\omega_a | \omega_b) \tag{6.1}$$

This association is 'functorial', we may denote the underlying functor $L^\infty : \mathbf{Kern}^{op} \to \mathbf{Vect}$ which is presheaf from the category of Markov kernels to the category of vector spaces. It is the presheaf that associates spaces to their space of observables. Let us denote $L^\infty \circ G : \mathscr{A} \to \mathbf{Vect}$ as i and $L^\infty \circ F : \mathscr{A}^{op} \to \mathbf{Vect}$ as π. In these notation one has that for any $a, b \in \mathscr{A}$ such that $b \leq a$ then $\pi^a_b \circ i^b_a = \mathrm{id}$.

For the definition and characterization of projective presheaf over a poset see [34–36].

Definition 8 (Projective $\mathscr{A}$-specifications)**.** An $\mathscr{A}$-specification (G, F) is called projective when $L^\infty \circ F$ is a projective presheaf (in **Vect**). In other words, there is a collection of presheaves $(S_a, a \in \mathscr{A})$ such that, $L^\infty \circ F \cong \bigoplus_{a \in \mathscr{A}} S_a$ where for any $b \geq a$, $S_a(b)$ is a constant vector space denoted S_a and $S_{ac}^{\ b} = \mathrm{id}$ for any $a \leq c \leq b$ and $S_a(b) = 0$ if $b \not\geq a$. The collection of presheaves $(S_a, a \in \mathscr{A})$ is called the decomposition of (G, F).

For any poset $\mathscr{A}$, symmetrizing the order defines the following equivalence relation,

$$\forall a, b \in \mathscr{A},\ a \sim b \iff a \leq b \text{ or } b \leq a \tag{6.2}$$

The equivalence classes of this equivalence relation are the connected components of $\mathscr{A}$ that we will denote as $\mathscr{C}(\mathscr{A})$. To each element of $a \in \mathscr{A}$ one can associate its connected component $\mathscr{C}(a)$. If each connected component has a minimum element, in other words, if for any $C \in \mathscr{C}(\mathscr{A})$, and any $b \in C$, there is $c \in C$ such that, $c \leq b$, then we shall denote, $\mathscr{C}_*(\mathscr{A})$ as the collection of these minimum elements; if not $\mathscr{C}_*(\mathscr{A}) = \emptyset$.

To conclude this article let us characterize Gibbs measures of projective $\mathscr{A}$-specifications.

Theorem 3. *Let $\gamma = (G, F)$ be a projective $\mathscr{A}$-specification. If at least one of the connected components of $\mathscr{A}$ does not have a minimum element, i.e. when,*

$$\mathscr{C}_*(\mathscr{A}) = \emptyset \tag{6.3}$$

then,

$$\mathscr{G}_g(\gamma) = \emptyset \tag{6.4}$$

if not,

$$\mathscr{G}_g(\gamma) = \prod_{a \in \mathscr{C}_*(\mathscr{A})} \mathbb{P}(G(a)) \tag{6.5}$$

Proof. Let us denote $L^\infty G$ as i and, $L^\infty F$ as π; (i,π) is decomposable, let $(S_a, a \in \mathscr{A})$ be its decomposition. For any $a, b \in \mathscr{A}$ such that $b \leq a$ and $\mu \in \mathscr{G}(\gamma)$, let us denote $L^\infty \mu$ as ν. For any $v \in L^\infty F(a)$,

$$\nu_b \pi_b^a (\sum_{c \leq a} S_c(a)(v)) = \nu_a (\sum_{c \leq a} S_c(a)(v)) \tag{6.6}$$

where $S_c(a)(v)$ denotes the projection of v on $S_c(a)$; therefore,

$$\nu_b (\sum_{c \leq b} S_{cb}^a(v)) = \nu_a i_a^b (\sum_{c \leq b} S_{cb}^a(v)) = \nu_a \left(\sum_{c \leq a} S_c(a)(v) \right) \tag{6.7}$$

and so,

$$\nu_a (\sum_{\substack{c \leq a \\ c \not\leq b}} S_c(a)) = 0 \tag{6.8}$$

Therefore for any $a \notin \mathscr{C}_*(\mathscr{A})$, $\nu_a|_{S_a(a)} = 0$. Furthermore,

$$\operatorname{colim} i \cong \bigoplus_{a \in \mathscr{A}} S_a(a). \tag{6.9}$$

ν is uniquely determined by $(\nu_a|_{S_a(a)}, a \in \mathscr{A})$; if there is a connected component $C \in \mathscr{C}(\mathscr{A})$ that does not have a minimal element, for any $a \in C$,

$$\nu_a|_{S_a} = 0 \tag{6.10}$$

Therefore for any $a \in C$, $\nu_a = 0$; this is contradictory with the fact that $\mu_a \in \mathbb{P}(\gamma(a))$ and so,

$$\mathscr{G}_g(\gamma) = \emptyset \tag{6.11}$$

When $\mathscr{C}_*(\mathscr{A})$ is non empty for any functor, H, from $\mathscr{A}$ to **Set**,

$$\lim H \cong \prod_{a \in \mathscr{C}_*(\mathscr{A})} H(a) \tag{6.12}$$

therefore,

$$\mathscr{G}_g(\gamma) = \prod_{a \in \mathscr{C}_*(\mathscr{A})} \mathbb{P}(G(a)) \tag{6.13}$$

□

7 Characterization of extreme Gibbs measures of $\mathscr{A}$-specifications

We now turn to the characterization of extreme Gibbs measures (Theorem 7.7 [29]), which is one of the steps for proving a zero-one law for the extreme Gibbs measures. We will show how it transfers to $\mathscr{A}$-specifications. In the classical theory of rigorous statistical mechanics, the tail σ-algebra generates the observables for which a 'generalized' law of large numbers (zero–one law) holds. Following [9], we give a candidate for such a σ-algebra in the categorical setting and show the associated extreme Gibbs measures decomposition.

Let us recall the definition of the tail σ-algebra in the classical formulation. Let us first consider the case of time series, i.e. $\Omega = \prod_{i\in\mathbb{N}} E_i$ with E_i are measurable spaces; let us denote $\mathscr{E}_{\geq k}$ as the σ-algebra generated by the cylinders $\prod_{n\geq k} E_k$. The tail σ-algebra is defined as $\bigcap_{k\in\mathbb{N}} \mathscr{E}_{\geq k}$. For specifications, I is any set and $\mathscr{E}_{\overline{a}}$ is indexed by a subset $a \subseteq I$ that is finite. We defined $\mathscr{E}_{\overline{a}}$ to be a σ-algebra of $E_{\overline{a}}$, however it can also be identified with the smallest σ-algebra on E that make $i^{\overline{a}} : E, \mathscr{E} \to E_{\overline{a}}, \mathscr{E}_{\overline{a}}$ measurable. Through this identification one defined $\mathscr{E}_\infty := \cap_{a\in\mathscr{P}_f(I)} \mathscr{E}_{\overline{a}}$.

For a functor from $\mathscr{A}$ to **Mes**, let us denote $\sigma(G(a))$ as the σ-algebra of the measurable space $G(a)$, where $a \in \mathscr{A}$, and $\sigma(G)$ as the underlying functor defined as $\sigma(G)^b_a A_b := G^{a\,-1}_b A_b$, with $b \leq a$. We propose that one candidate that plays the role of the tail σ-algebra for a given specification $\gamma = (G, F)$ is $\lim \sigma(G)$ defined as,

$$\lim \sigma(G) := \{(A_a \in \sigma(G(a)), a \in \mathscr{A}) | \forall a, b \in \mathscr{A}, \quad A_a = G^{a\,-1}_b A_b\} \tag{7.1}$$

Let us denote $1_A : E \to \{0, 1\}$ the indicator function over the set A that sends $\omega \in A$ to 1 and $\omega \notin A$ to 0. Let us remark that $1_{A_b} \circ G^a_b = 1[G^a_b(\omega_a) \in A_b] = 1_{G^{a\,-1}_b A_b}$. Remark that $A \in \lim \sigma(G)$ is equivalent to $1_A \in \lim i$; in other words, $\lim \sigma(G)$ is the restriction of $\lim i$ to indicator functions of the form $1_{A_a}, a \in \mathscr{A}$.

Finally, we also need to recall that for any $f \in L^\infty(E)$ and $\mu \in \mathbb{P}(E)$, one can define a measure $f \cdot \mu$ as $f \cdot \mu(d\omega) = f(\omega)d\omega$.

The key proposition of this section is Proposition 1; the proof of that proposition is given for $G(a), a \in \mathscr{A}$ finite measurable sets.

Assumption: Therefore, we assume in what follows that the measurable sets $G(a)$ are finite.

However, there is no finiteness constraint on $\mathscr{A}$. A weaker version holds when $G(a)$ is not finite. We will say that $F > 0$ when for any $a, b \in \mathscr{A}$, such

that $b \leq a$, $F(\omega_a|\omega_b) > 0$ for any ω_b such that $G_b^a(\omega_a) = \omega_b$; $G \circ F = \mathrm{id}$ requires that $F(\omega_a|\omega_b) = 0$ when $G_b^a(\omega_a) \neq \omega_b$.

The following lemma is an extension of the classical result that states that conditioning over a σ-subalgebra $\mathscr{F}_1 \subseteq \mathscr{F}$ defines a morphism of modules when finer ($\mathscr{F}$ measurable) observables are seen as modules over the coarser ($\mathscr{F}_1$ measurable observables).

Lemma 1. *Let E_1, E_2 be two measurable spaces, let $g : E_2 \to E_1$ be a measurable map and $f : E_1 \to E_2$ be a Markov kernel so that, $f \circ g = \mathrm{id}$. Let us denote respectively i and π the induced linear maps on $L^\infty(E_1), L^\infty(E_2)$. Let $h \in L^\infty(E_2)$ and $k \in L^\infty(E_1)$, then,*

$$\pi(h).k = \pi(h.i(k)) \tag{7.2}$$

Proof. Let us first prove the result in the particular case when $h = 1_B$ with $B \in \sigma(E_2)$ and $k = 1_A$ with $A \in \sigma(E_1)$. Let us denote $\overline{A}$ the complement of A, then $1_A + 1_{\overline{A}} = 1$ and $1_A.1_{\overline{A}} = 0$. Furthermore $i(1_A) = 1_{g^{-1}A}$

$$\pi(1_B) = \pi(1_B.1_{g^{-1}A}) + \pi(1_B.1_{\overline{g^{-1}A}}) \tag{7.3}$$

and

$$\pi(1_B.1_{g^{-1}A}) \leq \pi(1_{g^{-1}A}) = \pi \circ i(1_A) = 1_A \tag{7.4}$$
$$\pi(1_B.1_{\overline{g^{-1}A}}) \leq \pi(1_{\overline{g^{-1}A}}) = 1_{\overline{A}} \tag{7.5}$$

Therefore,

$$\pi(1_B)1_A = \pi(1_B.1_{g^{-1}A})1_A + \pi(1_B.1_{g^{-1}A})1_A \tag{7.6}$$

But, $\pi(1_B.1_{g^{-1}A})1_A \leq 1_{\overline{A}}.1_A = 0$ so, $\pi(1_B)1_A = \pi(1_B.1_{g^{-1}A})1_A$. Furthermore, $\pi(1_B.1_{g^{-1}A}) = \pi(1_B.1_{g^{-1}A})1_A + \pi(1_B.1_{g^{-1}A})1_{\overline{A}}$ therefore,

$$\pi(1_B.1_{g^{-1}A})1_{\overline{A}} \leq \pi(i(1_A)).1_{\overline{A}} = 0 \tag{7.7}$$

We just showed that,

$$\pi(1_B.1_{g^{-1}A}) = \pi(1_B.1_{g^{-1}A})1_{\overline{A}} \tag{7.8}$$

So $\pi(1_B).1_A = \pi(1_B.i(1_A))$. The result then extends by linearity directly to $h = \sum_{k \leq n} 1_{B_k}$ and $k = \sum_{k \leq n_1} 1_{A_n}$, which ends the proof. □

Let us remark that if $A \in \lim \sigma(G)$ then $\overline{A} := (\overline{A}_a, a \in \mathscr{A})$ is also in $\lim \sigma(G)$.

Proposition 1. *Let $\gamma = (G, F)$ be a specification, let $G(a)$ be finite sets for any $a \in \mathscr{A}$, let $F > 0$. Let $\mu \in \mathscr{G}(\gamma)$, for any $f \in \prod_{a\in A} L^\infty(G(a))$, such that $\forall a \in \mathscr{A}, \mu_a(f_a) = 1$,*

$$f.\mu \in \mathscr{G}_g(\gamma) \iff \exists \tilde{f} \in \lim i,\ s.t.\ f.\mu = \tilde{f}.\mu \tag{7.9}$$

Proof. Let us assume that $f.\mu \in \mathscr{G}_g(\gamma)$, then for any $a, b \in \mathscr{A}$ such that $b \leq a$, and for any $g_a \in L^\infty(G(a))$, by hypothesis, $(f.\mu)_b \pi_b^a(g_a) = (f.\mu)_a(g_a)$; it can be rewritten as,

$$\mu_b\left(\pi_b^a(g_a).f_b\right) = \mu_a\left(f_a.g_a\right) \tag{7.10}$$

By Lemma 1, $\pi_b^a(g_a).f_b = \pi_b^a(g_a.i_a^b f_b)$; therefore,

$$\mu_b \pi_b^a(g_a.i_a^b f_b) = \mu_a\left(f_a.g_a\right) \tag{7.11}$$

Therefore $f_a = i_a^b f_b$ μ_a-almost surely.

We will now show that there is $\tilde{f} \in \lim i$ such that $\tilde{f}.\mu = f.\mu$. It is in this part of the proof that we assume that $G(a)$ are finite sets and that $F > 0$. Let's call $S_a := \text{supp}\mu_a$, the support of μ_a, i.e., the set $\text{supp}\mu_a := \{\omega_a \in G(a) | \mu_a(\omega_a) > 0\}$. Let us denote $N_a := \overline{S_a}$ its complement and $M_a = 1_{N_a}$. We will now show that $(N_a, a \in \mathscr{A}) \in \lim \sigma(G)$.

For any $b, a \in \mathscr{A}$ such that $b \leq a$, $\mu_a i_a^b = \mu_b$; therefore, as $\mu_b(M_b) = 0$, one has that $\mu_a i_a^b(M_b) = 0$. Recall that $i_a^b(M_b)$ is the indicator function of the set $G_b^{a-1} N_b$; the previous remark implies that $i_a^b(M_b) \leq M_a$. Furthermore, $\mu_b \pi_b^a(M_a) = \mu_a(M_a) = 0$; therefore, $\pi_b^a(M_a) \leq M_b$.

Hence, as $\pi_b^a(i_a^b M_b) = M_b$ and $i_a^b M_b \leq M_a$, then by applying π_b^a on both sides, $M_b \leq \pi_b^a(M_a)$. And so $\pi_b^a(M_a) = M_b$.

Recall that we showed that $\pi_b^a(M_a) = M_b$ and $i_a^b(M_b) \leq M_a$. In particular, $M_a - i_a^b(M_b) \geq 0$; furthermore, $\pi_b^a(M_a - i_a^b(M_b)) = 0$ so $M_a = i_a^b(M_b)$. To be more explicit: $\forall \omega_b \in G(b)$,

$$\pi_b^a(M_a - i_a^b(M_b))(\omega_b) = \sum_{\omega_a \in G(a)} F(\omega_a|\omega_b)[M_a - i_a^b(M_b)](\omega_a) \tag{7.12}$$

As, by hypothesis, for any ω_b such that $G_b^a(\omega_a) = \omega_b$ one has that $F(\omega_a|\omega_b) > 0$, then $M_a = i_a^b(M_b)$. This implies that $M \in \lim i$ and $N \in \lim \sigma(G)$. This also implies that $S \in \lim \sigma(G)$.

Let $\tilde{f} = f 1_S$. Then $f 1_S \in \lim F$ and for any $a \in \mathscr{A}$, $f_a = \tilde{f}_a$ μ_a-a.s., which ends the proof.

□

One remarks that $\lim i$ is a subset of $\lim \pi$: for any $f \in \lim i$, by definition for any $a, b \in \mathscr{A}$ such that $b \leq a$, $i_a^b f_b = f_a$ and so $\pi_b^a i_a^b f_b = \pi_b^a f_a$ so $f_b = \pi_b^a f_a$.

Let us also remark that for any $b \leq a \in \mathscr{A}$, $\mu_a(A_a) = \mu_b(A_b)$ for $\mu \in \mathscr{G}(\gamma)$, $A \in \lim G$.

Theorem 4 (Extreme measure characterisation (Generalisation of Theorem 7.7 [29]))**.** *Let $\gamma = (G, F)$ be a specification, let $G(a)$ be finite sets for any $a \in \mathscr{A}$, let $F > 0$. $\mathscr{G}_g(\gamma)$ is a convex set. Each $\mu \in \mathscr{G}_g(\gamma)$ is uniquely determined by it's restriction to $\lim \sigma(G)$. Furthermore μ is extreme in $\mathscr{G}(\gamma)$ if and only if for any $A \in \lim \sigma(G)$, $\forall a \in \mathscr{A}$, $\mu_a(A_a) = 0$ or 1.*

Proof. Let us denote π^* the functor from $\mathscr{A}$ to **Vect** for which for any $b \leq a$, $\pi^{*b}_a : (L^\infty F(b))^* \to (L^\infty F(a))^*$ is the dual of π^a_b that send linear forms to linear forms. Then $\mathscr{G}_g(\gamma)$ is a subspace of the vector space $\lim F^*$ and furthermore for any $a \in \mathscr{A}$ and $p \in [0,1]$, $p\mu_a + (1-p)\nu_a \in \mathbb{P}(G(a))$ whenever $\mu_a, \nu_a \in \mathbb{P}(G(a))$. Therefore $\mathscr{G}_g(\gamma)$ is a convex set.

Proposition 1, allows us to apply a similar proof, when done with caution, to the one found of Theorem 7.7 in [29]. Let us recall the proof. Let $\mu, \nu \in \mathbb{G}(\gamma)$ such that $\mu_{|\lim i} = \nu_{|\lim i}$. Let $\overline{\mu} = \frac{\mu+\nu}{2}$, then $\overline{\mu} \in \mathscr{G}(\gamma)$. But μ and ν are absolutely continuous with respect to $\overline{\mu}$ therefore for any $a \in \mathscr{A}$ there is $f_a, g_a \in L^\infty(G(a))$ such that $\mu_a = f_a\overline{\mu}_a$ and $\nu_a = g_a\overline{\mu}_a$. By Proposition 1, $f, g \in \lim i$. By hypothesis, for any $h \in \lim i$ $\overline{\mu}_a(h_a) = \mu_a(h_a) = \nu_a(h_a)$. Importantly i^b_a is a ring morphism of $L^\infty(G(b))$, i.e. $i^b_a(k_b.h_b) = i^b_a(k_b).i^b_a(h_b)$. Therefore for any $h, k \in \lim i$, $k.h \in \lim i$; as $f - g \in \lim i$, then it is also the case that $(f-g)^2 \in \lim i$; but for any $a \in \mathscr{A}$,

$$\overline{\mu}_a[(f_a - g_a)^2] = 0 \tag{7.13}$$

so $f_a = g_a \quad \overline{\mu}_a -$ a.s. Therefore $f\overline{\mu} = g\overline{\mu}$ and $\mu = \nu$.

Showing that $\mu \in \mathscr{G}_g(\gamma)$ is extreme is equivalent to μ being trivial on $\lim i$ is a direct generalization of Corollary 7.4 [29] thanks to Proposition 1. Let $\mu \in \mathscr{G}(\gamma)$ be not trivial on $\lim \sigma(G)$ then there is $A = (A_a, a \in \mathscr{A}) \in \lim \sigma(G)$ such that,

$$\forall a \in \mathscr{A}, \quad 0 < \mu_a(A_a) < 1 \tag{7.14}$$

Therefore for any $a \in \mathscr{A}$,

$$\mu_a = \mu_a(A_a)\frac{1_{A_a}.\mu_a}{\mu_a(A_a)} + \mu_a(\overline{A}_a)\frac{1_{\overline{A}_a}.\mu_a}{\mu_a(\overline{A}_a)} \tag{7.15}$$

Furthermore, for any $b \leq a$,

$$i^b_a\frac{1_{A_b}}{\mu_a(A_b)} = \frac{1_{A_a}}{\mu_a(A_b)} = \frac{1_{A_a}}{\mu_a(A_a)} \tag{7.16}$$

Therefore $\frac{1_{A_a}}{\mu_a(A_a)}, a \in \mathscr{A}$ is in $\lim i$ and so by Lemma 1, $\left(\frac{1_{A_a}.\mu_a}{\mu_a(A_a)}, a \in \mathscr{A}\right) \in \mathscr{G}_g(\gamma)$. Similarly $\left(\frac{1_{\overline{A}_a}.\mu_a}{\mu_a(\overline{A}_a)}, a \in \mathscr{A}\right) \in \mathscr{G}_g(\gamma)$. In particular there is $0 < p < 1$

so that $\mu = p\nu + (1-p)\nu_1$ with $\nu, \nu_1 \in \mathscr{G}_g(\gamma)$. Therefore μ is not an extreme measure.

Assume now that $\mu \in \mathscr{G}_g(\gamma)$ is such that for any $A \in \lim \sigma(A)$, and any $a \in \mathscr{A}$, $\mu_a(A_a) = 0$ or 1. Suppose that there is $0 < p < 1$ such that $\mu = p\nu + (1-p)\nu_1$ with $\nu, \nu_1 \in \mathscr{G}_g(\gamma)$. Then for any $a \in \mathscr{A}$, ν_a, ν_{1a} is absolutely continuous with respect to μ_a. Therefore, there are $(f_a \geq 0, a \in \mathscr{A})$, $(g_a \geq 0, a \in \mathscr{A})$ both in $\prod_{a\in\mathscr{A}} L^\infty(G(a))$ such that $\nu = f\mu$ and $\nu_1 = g\mu$. As $\nu, \nu_1 \in \mathscr{G}_g(\gamma)$, then by Lemma 1, $f, g \in \lim i$. Therefore, for all $a \in \mathscr{A}$, $\mu_a(f_a) = 0$ or for all $a \in \mathscr{A}$, $\mu_a(g_a) = 0$. So, $\mu = \nu$ or $\mu = \nu_1$ and μ is extreme in $\mathscr{G}_g(\gamma)$.

□

Let us remark that if $\mathscr{A}$ has only one connected component, then for $A \in \lim \sigma(G)$, satisfying $\forall a \in \mathscr{A}, \mu_a(A_a) = 0$ or 1 is equivalent to $\exists a \in \mathscr{A}, \mu_a(A_a) = 0$ or 1. Indeed, if a, b are in the same connected component, i.e., $a \leq b$ or $b \leq a$, then $\mu_a(A_a) = \mu_b(A_b)$.

8 Background: Variational inference for graphical model and more

Consider a joint distribution $P_{X,Y} \in \mathbb{P}(E \times E_1)$ over two random variables $X \in E, Y \in E_1$. A classical problem is, given an observation ω_1 on Y, to compute the posterior $P_{X|Y}(\omega, \omega_1) = \frac{P_{X,Y}(\omega,\omega_1)}{P_Y(\omega_1)}$ with $P_Y(\omega_1) = \sum_{\omega\in E} P_{X|Y}(\omega, \omega_1)$, the marginal distribution of Y. However, doing so requires summing over all possible configurations of X, which can be computationally too costly. This is the case, for example, when $X = X_0, ..., X_T$ with $X_i \in S$ and $E = \prod_{i\leq T} B$. Instead, one resorts to variational inference to compute $P_{X|Y}$ approximately [30]. We will now explain what variational inference is, but first let us introduce entropy and Gibbs free energy. When E is a finite set, the entropy of a probability distribution Q on E is defined as:

$$S(Q) = -\sum_{\omega\in E} Q(x) \ln Q(x) \tag{8.1}$$

Let H be a measurable function $H : E \to \mathbb{R}$. For $Q \in \mathbb{P}(E)$, one calls $\mathbb{E}_Q[H] - \frac{1}{\beta}S(Q)$ the Gibbs free energy; in general $\beta = 1$. An important property is that,

$$-\ln \sum_{\omega\in E} e^{-\beta H(\omega)} = \inf_{Q\in\mathbb{P}(E)} \mathbb{E}_Q[\beta H] - S(Q) \tag{8.2}$$

The optimal solution to Equation 8.2 is given by the Boltzmann distri-

bution

$$Q^*(\omega) = \frac{e^{-H(\omega)}}{\sum_{\omega \in E} e^{-\beta H(\omega)}} \tag{8.3}$$

Let $H(\omega) = -\ln P_{X,Y}(\omega, \omega_1)$ and $\beta = 1$, then $Q^*(\omega) = P_{X|Y}(\omega|\omega_1)$. Therefore, solving the optimization problem of Equation 8.2 is equivalent to computing the posterior $P_{X|Y}(\omega|\omega_1)$. Solving Equation 8.2 over a subset of distributions $Q \in \Theta \subseteq \mathbb{P}(E)$ is called variational inference. If furthermore the Gibbs free energy is replaced by an approximation, we call it approximate variational inference.

One remarks that $\inf_{Q \in \mathbb{P}(E)} \mathbb{E}_Q[\beta H] - S(Q)$ is equivalent to $\sup_{Q \in \mathbb{P}(E)} S(Q) - \mathbb{E}_Q[\beta H]$. And this last optimization problem relates, through Lagrange multipliers, to maximizing entropy under energy constraints $U \in \mathbb{R}$,

$$\sup_{\substack{Q \in \mathbb{P}(E) \\ \mathbb{E}_Q[H] = U}} -\sum_{x \in E} Q(x) \ln Q(x) \tag{8.4}$$

In the physics literature, one refers to Equation 8.4 as MaxEnt [42], which stands for the principle of maximum entropy and such principle has many application see [42, 43]. Variational inference is called the variational principle.

Graphical models translate relations on graphs into conditional independence relations between variables [2, 44, 45]. The Hammersley-Clifford Theorem states that for strictly positive distributions, these conditional independence relations are equivalent to factorizing the joint distributions on the cliques of the graph. When a joint distribution factors according to an acyclic graph, computing the posterior $P_{X|Y}$ over some nodes of the graph can be done efficiently through dynamic programming with an algorithm called Belief Propagation. In the particular case where the graphical model is a Gaussian Hidden Markov Model, this dynamic programming algorithm is the smoother Kalman filter [46].

The Belief Propagation algorithm solves an approximate variational inference problem for a variational free energy called the Bethe free energy [10, 11, 18, 47].

We now specify the previous statement and explain what we mean by a probability distribution that factors according to a collection of subsets of variables. Then, we recall the remarkable property that for probability distributions that factor according to an acyclic graph, entropy can be decomposed into sums of 'local' entropies. This will allow us to introduce what the Bethe free energy is and how it relates to entropy.

It is important to consider the case of graphical models to understand the more general setting of *factor graphs* or *factorization models* Yedidia, Freeman, Weiss consider in their seminal article, *Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms*, extending the correspondence between a Generalized Belief Propagation algorithm and

the associated variational free energy that we will call the *Generalized Bethe free energy*. We present their work and O. Peltre's complemented version and extension of their results (see SYCO 8 talk, [11,48]) in the next section. Doing so will then motivate the entropy we introduce for $\mathscr{A}$-specifications and give intuition on the message passing algorithm we introduce for such specifications.

Let $E = \prod_{i\in I} E_i$ where I is finite. $\mathbb{P}$ is said to factor according to a subset $\mathscr{A}$ of $\mathscr{P}(I)$ when for any $a \in \mathscr{A}$ there is $f_a : E_a \to \mathbb{R}$ such that,

$$P(\omega) = \prod_{a\in\mathscr{A}} f_a(\omega_a) \tag{8.5}$$

We will denote the space of probability distributions that factor according to $\mathscr{A}$ as $\mathrm{Fac}_{\mathscr{A}}$. We will call such space the space of ($\mathscr{A}$-)factorization models.

A graph $G = (I, A)$ is a collection of vertices I and edges A; one can see it as a poset $\mathscr{A}(G)$ with the relation of inclusion. In other words, for any $v \in I$ and $e \in A$, $v \leq e$ if and only if $v \in e$.

Let us now recall a proposition that states that a strictly positive probability distribution on a finite set E, i.e., $Q(\omega) > 0$ for $\omega \in E$, factors according to its marginal distribution on the edges and vertices [12, 49]. We denote $\mathbb{P}_{>0}$ as the space of strictly positive probability distributions.

Proposition 2 (Factorization on acyclic graphs)**.** *Let I be a finite set and let $E = \prod_{i\in I} E_i$ be a product of finite sets, and let $G = (I, A)$ be a finite acyclic graph. $Q \in \mathbb{P}_{>0}(E)$ factors accordingly to $\mathscr{A}(G)$, i.e., $Q \in G_{\mathscr{A}(G)}$ if and only if for any $\omega \in E$,*

$$Q(\omega) = \frac{\prod_{e\in A} i^e{}_* Q(\omega_e)}{\prod_{v\in I} i^v{}_* Q^{d(v)-1}(\omega_v)}, \tag{8.6}$$

where $i^e : E \to E_e$ is the projection onto the state space $E_{v_1} \times E_{v_2}$ associated with the edge $e = (v_1, v_2)$, and similarly for $i^v : E \to E_v$.

For a given sub-poset $\mathscr{A} \subseteq \mathscr{P}(I)$, let us denote

$$[*, i]_{\mathscr{A}} := \{(Q_a \in \mathscr{P}(E_a), a \in \mathscr{A}) \,|\, \forall b \leq a,\, i^b_a Q_b = Q_a\} \tag{8.7}$$

as the collection of 'local' probability measures that have compatible marginals with respect to the projections $i^a_b : E_a \to E_b$ when $b \subseteq a$ in $\mathscr{P}(I)$. The next proposition expresses the free energy of probability distributions that factor according to an acyclic graph as a weighted sum of free energies of its marginals on edges and vertices. The weights are given by the inclusion-exclusion formula of the graph when seen as a poset. First, let us recall the inclusion-exclusion formula for a poset.

Definition 9 (Zeta operator of a poset)**.** Let $\mathscr{A}$ be a finite poset. We call the 'zeta-operator' of a poset $\mathscr{A}$, denoted ζ, the operator from $\bigoplus_{a\in\mathscr{A}}\mathbb{R}\to\bigoplus_{a\in\mathscr{A}}\mathbb{R}$ defined as, for any $\lambda\in\bigoplus_{a\in\mathscr{A}}\mathbb{R}$ and any $a\in\mathscr{A}$,

$$\zeta(\lambda)(a)=\sum_{b\leq a}\lambda_b \tag{8.8}$$

Proposition 3 (Reformulation of Proposition 2 [50], Rota' 64)**.** *Let $\mathscr{A}$ be a finite poset. The zeta-operator of $\mathscr{A}$ is invertible. We will call its inverse the Möbius inversion of $\mathscr{A}$, denoted μ. Furthermore, there is a collection $(\mu(a,b);b,a\in\mathscr{A}\ s.t.\ b\leq a)$ such that, for any $\lambda\in\bigoplus_{a\in\mathscr{A}}\mathbb{R}$ and $a\in\mathscr{A}$,*

$$\mu(\lambda)(a)=\sum_{b\leq a}\mu(a,b)\lambda_b \tag{8.9}$$

We call the coefficient $(\mu(a,b),b,a$ s.t. $b\leq a)$ the Möbius coefficients of $\mathscr{A}$. In particular Proposition 3 implies that, for any $b,a\in\mathscr{A}$ such that $b\leq a$,

$$\sum_{c:b\leq c\leq a}\mu(a,c)=1[b=a] \tag{8.10}$$

$$\sum_{c:b\leq c\leq a}\mu(c,b)=1[b=a] \tag{8.11}$$

For a collection of values $\lambda_a\in\mathbb{R}$, $a\in\mathscr{A}$, we call the following expression $\sum_{a\in\mathscr{A}}\sum_{b\leq a}\mu(a,b)\lambda_b$ the inclusion-exclusion formula over a poset $\mathscr{A}$; this formula corresponds to the value one would attribute to a maximal element, denoted 1, added to $\mathscr{A}$:

$$\lambda_1:=\sum_{a\in\mathscr{A}}\sum_{b\leq a}\mu(a,b)\lambda_b \tag{8.12}$$

which can be rewritten as $\lambda_1=\sum_{a\in\mathscr{A}}[\sum_{b\geq a}\mu(b,a)]\lambda_a$. We will denote $c(a)=\sum_{b\leq a}\mu(a,b)$ the weighted coefficients

To find the classical inclusion-exclusion formula, consider I a finite set; the poset $\mathscr{A}$ is $(\mathscr{P}(I),\supseteq)$ with the reversed order. The quantities λ_a, for $a\in\mathscr{A}$, are $|A_i|$ when $a=i\in I$, and represent the cardinality of the sets A_i, as well as the cardinality of all possible intersections $|\cap_{i\in a}A_i|$ when $a\subseteq I$. In this setting, the maximal element 1 has a value $\lambda_1=|\cup_{i\in I}A_i|$.

Proposition 4. *Let I be a finite set, and let $E=\prod_{i\in I}E_i$ be a product of finite sets. Consider a finite acyclic graph $G=(I,A)$. Let $(H_a:E_a\to\mathbb{R},a\in\mathscr{A})$ be a collection of Hamiltonians (measurable maps) that respectively factor through the projection $i^a:E\to E_a$. The following map,*

$$\begin{array}{rccc} \phi & : & [*,i]_{\mathscr{A}} & \to & \mathrm{Fac}_{\mathscr{A}(G)} \\ & & (Q_a, a \in \mathscr{A}(G)) & \mapsto & \frac{\prod_{e\in A} Q_e}{\prod_{v\in I} Q_v^{d(v)-1}} \end{array} \tag{8.13}$$

is a bijection; here $d(v)$ is the degree of node v. Furthermore,

$$\mathbb{E}_{\phi(Q)}\left[\sum_{e\in A} H_e + \sum_{v\in I} H_i\right] - S(\phi(Q)) = F\left(Q_a, a \in \mathscr{A}(G)\right) \tag{8.14}$$

where

$$F\left(Q_a, a \in \mathscr{A}(G)\right) = \sum_{a\in\mathscr{A}(G)} c(a)\left[S(i_*^a Q) - \mathbb{E}_{i_*^a Q}[H_a]\right] \tag{8.15}$$

Proof. Let for any $p \in \mathbb{P}_{>0}(E) \cap \mathrm{Fac}_{\mathscr{A}(G)}$,

$$\begin{array}{rccc} \psi & : & \mathbb{P}(E) \cap \mathrm{Fac}_{\mathscr{A}(G)} & \to & \lim F \\ & & P & \mapsto & (\pi_{a,*}P, a \in \mathscr{A}(G)) \end{array} \tag{8.16}$$

Then $\phi\psi = \mathrm{id}$ and $\psi\phi = \mathrm{id}$, furthermore for any $P \in \mathbb{P}(E) \cap \mathrm{Fac}_{\mathscr{A}(G)}$,

$$\mathbb{E}_Q[\sum_{a\in\mathscr{A}} H_a] = \mathbb{E}_Q[\sum_{a\in\mathscr{A}}\sum_{b\le a} \mu(a,b)H_a] = \sum_{a\in\mathscr{A}}\sum_{b\le a}\mu(a,b)\mathbb{E}_{i_*^a Q}[H_a] \tag{8.17}$$

and by Proposition 2,

$$S(Q) = \sum_{e\in A} S(i_*^e Q) - \sum_{v\in I}(d(v)-1)S(i_*^v Q) = \sum_{a\in\mathscr{A}(G)} c(a) S(i_*^a Q) \tag{8.18}$$

which ends the proof.

□

$F\left(Q_a, a \in \mathscr{A}(G)\right) = \sum_{a\in\mathscr{A}(G)} c(a)\left[S(i_*^a Q) - \mathbb{E}_{i_*^a Q}[H_a]\right]$ is called the Bethe free energy and depends only on the marginal distributions $i_*^a Q$. Proposition 4 implies that variational inference on acyclic graphical models can be done over marginal distributions and still give the exact posterior. This remark is very important as, for a graph (G, E), the state space of $(E_e, e \in A)$ is of size $|A| \times N^2$ with N the maximal size of the $E_v, v \in I$, whereas the cardinality of E is $O(N^{|I|})$. The collection of marginal distributions serves therefore as a compressed representation of the joint distribution. When G is acyclic, the Belief Propagation algorithm finds the minima of the Bethe free energy exploiting the compressed reformulation of $\mathrm{Fac}_{\mathscr{A}(G)}$; the complexity of the algorithm is $O(|A|)$.

The Bethe free energy can be defined on any graph, even those that may contain cycles; in such cases, the equality between the Bethe free energy and the Gibbs free energy does not hold in general. The key idea is to replace the Gibbs free energy with the Bethe free energy in variational inference and solve the associated optimization problem, hoping that the true posterior will be correctly approximated.

We introduced the Bethe free energy in a similar fashion to [11]. This also allows us to see how one can define the Bethe free energy for any factorization model. In this context, the Bethe free energy is called the Generalized Bethe free Energy or region-based approximation of free energy in [12]. In the next section, we define the (Generalized) Belief Propagation algorithm for the Generalized Bethe free Energy.

9 Background: Generalized Bethe free energy

The Generalized Bethe free energy is an approximation of the Gibbs free energy that generalizes the Bethe free energy [12]. Consider the configuration space $E = \prod_{i \in I} E_i$ of finite sets over a finite set I. Let $\mathscr{A} \subseteq \mathscr{P}(I)$, and let $(H_a \in E_a \to \mathbb{R}, a \in \mathscr{A})$ be a collection of (measurable) Hamiltonians. For $Q = (Q_a \in \mathbb{P}(E_a), a \in \mathscr{A})$, the General Bethe free energy is defined as:

$$F_{\text{Bethe}}(Q) = \sum_{a \in \mathscr{A}} \sum_{b \geq a} \mu(b, a) \left(\mathbb{E}_{Q_a}[H_a] - S_a(Q_a) \right) \tag{9.1}$$

Here we considered $Q = (Q_a \in \mathbb{P}(E_a), a \in \mathscr{A})$ to be any collection of distributions over the local variables $X_a \in E_a$; however, let's keep in mind that the collection we are interested in are $Q \in [*, i]_{\mathscr{A}}$ that are compatible with respect to marginalization.

In the Bethe Free energy, the entropy functional is replaced by a reconstruction of the 'local' entropies that only makes use of the entropy of the marginals. This is the term $S_{GB}(p) = \sum_{a \in \mathscr{A}} c(a) S(p_a)$ in F_{Bethe}. It is economical in the sense that it is a non-redundant way of computing entropy, similar to counting the cardinality of the union of sets using the Inclusion-Exclusion principle (see [51] for a detailed presentation of this idea).

The General Belief Propagation is an algorithm that enables us to find the critical points of the Generalized Bethe Free Energy. A classical result states that fixed points of this algorithm correspond to critical points of that free energy. Let us now recall the expression of this algorithm.

For $\mathscr{A} \subseteq \mathscr{P}(I)$, E a finite product of finite sets, and $(H_a, a \in \mathscr{A})$ a collection of Hamiltonians. For two elements of $\mathscr{A}$, a and b such that $b \subseteq a$, two types of messages are considered: top-down messages $m_{a \to b} \in \mathbb{R}^{E_b}$ and bottom-up messages $n_{b \to a} \in \mathbb{R}^{E_a}$. The update rule is as follows: consider

messages at times t, $(n^t_{b\to a}, m^t_{a\to b} | b, a \in \mathscr{A} \text{ s. t. } b \leq a)$, they are related by the following relation,

$$\forall a, b \in \mathscr{A}, \text{s.t. } b \leq a, \quad n^t_{b\to a} = \prod_{\substack{c: b\leq c \\ c \not\geq a}} m^t_{c\to b} \tag{9.2}$$

One can define beliefs as ,

$$\forall a \in \mathscr{A}, \quad b^t_a \propto e^{-H_a} \prod_{\substack{b\in\mathscr{A}: \\ b\leq a}} n^t_{b\to a} \tag{9.3}$$

The beliefs are sometimes interpreted as probability distributions up to a multiplicative constant; here to make the presentation clearer, we require that b_a is a probability distribution. The update rule is given by,

$$\forall a, b \in \mathscr{A}, \text{s.t.} b \leq a \quad m^{t+1}_{a\to b}(x_b) = m^t_{a\to b}(x_b) \frac{\sum_{y_a : \pi^a_b(y_a) = x_b} b^t_a(y_a)}{b^t_b(x_b)} \tag{9.4}$$

The multiplication of function $n_{b\to a}$ that have different domains is made possible because there is an the embedding of $\mathbb{R}^{E_b}$ into $\mathbb{R}^{E_a}$ implicitly implied in the last equation.

The algorithm can be rewritten in a more condensed manner, updating only the top-down messages, for all $a, b \in \mathscr{A}$, such that $b \leq a$,

$$m^{t+1}_{a\to b}(x_b) = m^t_{a\to b}(x_b) \frac{\sum_{y_a : i^a_b(y_a) = x_b} e^{-H_a(y_a)} \prod_{\substack{c\in\mathscr{A}: \\ c\subseteq a}} \prod_{\substack{d: c\subseteq d \\ d\not\subseteq a}} m^t_{d\to c}(x_c)}{e^{-H_b(x_b)} \prod_{\substack{c\in\mathscr{A}: \\ c\subseteq b}} \prod_{\substack{d: c\subseteq d \\ d\not\subseteq b}} m^t_{d\to c}(x_c)} \tag{9.5}$$

We will denote this update rule as GBP, $m(t+1) = GBP(m(t))$

Theorem 5 (Yedidia, Freeman, Weiss, Peltre)**.** *Let $(m_{a\to b}, a, b \in \mathscr{A} \text{ s.t. } b \subseteq a)$ be a fix point of the Generalized Belief Propagation, i.e.*

$$m = GBP(m) \tag{9.6}$$

Let $(b_a, a \in \mathscr{A})$ be the associated beliefs and let, for $a \in \mathscr{A}$, $p_a = b_a / \sum_{x\in E_a} b_a(x)$ be the associated normalized beliefs. Then $(p_a, a \in \mathscr{A})$ is a critical point of F_{Bethe} under the constraint that $p \in \lim F$.

Proof. Theorem 5.15 [11], Theorem 5 [12].

□

10 Entropy of $\mathscr{A}$-specifications and variational principle

Let $\mathscr{A}$ be a finite poset and $\gamma = (G, F)$ be a specifications with $G(a)$ being a finite set for any $a \in \mathscr{A}$. We propose that the entropy of $Q \in \mathscr{G}_g(\gamma)$ to be $S_{GB}(Q) = \sum_{a\in\mathscr{A}} c(a)S(Q_a)$ and that the free energy is $F_{\text{Bethe}}(Q) = \sum_{a\in\mathscr{A}} c(a)\left(\mathbb{E}_{Q_a}[H_a] - S_a(Q_a)\right)$.

It might seem at first glance that it is sufficient to apply GBP to find the critical points of the free energy for specifications, but it is actually much trickier than that. In GBP the free energy is constrained over $Q \in [*, i]_{\mathscr{A}}$ but for $Q \in \mathscr{G}_g(\gamma)$, $Q \in [*, F]_{K,\mathscr{A}}$. The first significant difference is that for GBP the presheaf is prescribed to the one associated with marginalization whereas here F can be anything. In Section 3.2 [10] (version 1) the case where F is any presheaf is treated in detail and corresponds to a fair amount of generalization with respect to GBP as presented in [11, 12]. The second difference is that F is a functor and not a presheaf; it is an essential difference tackled in the most general setting by Theorem 2.3 [10] (version 1). The intuition behind such a difference is as follows. The dual of the F acts on Lagrange multipliers; the Lagrange multipliers are analogous to the messages '$m_{a\to b}$' of GBP. When F is a presheaf its dual is a functor and one can send the Lagrange multipliers in a cell $F(a)$ but applying F_b^{a*} to $m_{a\to b}$ for any $b \leq a$ and so one can do the product of such messages (or the sum if one takes the logarithm). One cannot send the Lagrange multipliers for the b's smaller than a into the cell $F(a)^*$ when the F is a functor. However, when there is G such that F is a section of G, G can send the multipliers into $F(a)$ and one can build a message-passing algorithm that finds the critical points of the Bethe free energy for specifications. We now explicitly state Theorem 2.3 [10] (version 1) in the particular case of specifications and detail the message-passing algorithms for finding optimal Gibbs measures.

Problem to solve: The optimization problem we want to solve is the following,

$$\inf_{Q\in\mathscr{G}_g(\gamma)} F_{\text{Bethe}}(Q) \tag{10.1}$$

We will need to generalize the ζ and Möbius inversion of a poset to the ones for functors and presheaves.

Definition 10 (Mobius inversion associated to a functor)**.** Let $G : \mathscr{A} \to$ **Mod** be a functor from a finite poset to the category of R-modules with R a ring; let $\zeta_G : \bigoplus_{a\in\mathscr{A}} G(a) \to \bigoplus_{a\in\mathscr{A}} G(a)$ be such that for any $a \in \mathscr{A}$ and $v \in \bigoplus_{a\in\mathscr{A}} G(a)$,

$$\mu_G(v)(a) = \sum_{b\leq a} \mu(a, b)G_a^b(v_b) \tag{10.2}$$

Proposition 5. *Let $G : \mathscr{A} \to \textbf{\textit{Mod}}$ be a functor from a finite poset to the category of modules, μ_G is invertible and its inverse, denoted ζ_G, is defined as follows, for any $a \in \mathscr{A}$ and $v \in \bigoplus_{a\in\mathscr{A}} G(a)$,*

$$\zeta_G(v)(a) = \sum_{b\leq a} G_a^b(v_b) \tag{10.3}$$

Proof. Let $v \in \bigoplus_{a\in\mathscr{A}} G(a)$ and $a \in \mathscr{A}$,

$$\zeta_G\mu_G(v)(a) = \sum_{b\leq a}\sum_{c\leq b} \mu(b,c) G_a^b G_b^c(v_c) \tag{10.4}$$

therefore,

$$\zeta_G\mu_G(v)(a) = \sum_{c\leq a} \left(\sum_{b:\ c\leq b\leq a} \mu(b,c) \right) G_a^c(v_c) = G_a^a(v_a) \tag{10.5}$$

Furthermore,

$$\mu_G\zeta_G(v)(a) = \sum_{b\leq a} \mu(a,b) \sum_{c\leq b} G_a^c(v_c) = v_a \tag{10.6}$$

□

Remark 3. Let us remark for any poset $(\mathscr{A}, \leq)$ one can reverse the relations, in other words, for any $a, b \in \mathscr{A}$,

$$a \leq_{op} b \iff b \leq a \tag{10.7}$$

We sall also denote $\leq_{op}$ as $\geq$ and the corresponding poset as $\mathscr{A}^{op}$ or $(\mathscr{A}, \geq)$. One has that, for any $a, b \in \mathscr{A}$ such that $a \geq b$,

$$\zeta_{\mathscr{A}^{op}}(b,a) = \zeta_{\mathscr{A}}(a,b) \tag{10.8}$$
$$\mu_{\mathscr{A}^{op}}(b,a) = \mu_{\mathscr{A}}(a,b) \tag{10.9}$$

In particular for any $G : \mathscr{A} \to \textbf{Vect}$ functor from a finite poset to the category of modules,

$$\mu_{G^*} = (\mu_G)^* \tag{10.10}$$

as for any $(l_a \in G(a)^*, a \in \mathscr{A})$,

$$\sum_{a\in\mathscr{A}}\sum_{b\leq a} \mu(a,b) l_a G_a^b = \sum_{b\in\mathscr{A}}\sum_{a\geq b} \mu(a,b) G^{*a}_{\ b}(l_a) \tag{10.11}$$

In what follows, we go back to the convention for specification, which is that G is a presheaf (to the category of measurable spaces). Let us define the function $FE(Q) : \prod_{a\in\mathscr{A}} \mathbb{P}(E_a) \to \prod_{a\in\mathscr{A}} \mathbb{R}$ as $FE(Q) = (\mathbb{E}_{Q_a}[H_a] - S_a(Q_a),\, a \in \mathscr{A})$, which sends a collection of probability measures over $\mathscr{A}$ to their Gibbs free energies. For any $Q \in \prod_{a\in\mathscr{A}} \mathbb{P}(E_a)$, let us denote $\mathrm{d}_Q FE$ as the differential of FE at the point Q.

Theorem 6. *Let $\mathscr{A}$ be a finite poset, let $\gamma = (G, F)$ be a $\mathscr{A}$-specification such that $G(a)$ is a finite set for any $a \in \mathscr{A}$. Let $H_a : G(a) \to \mathbb{R}$ be a collection of (measurable) Hamiltonians. The critical points of the Generalized Bethe free energy are the $Q \in [*, F]_{K,\mathscr{A}}$ such that,*

$$\mu_{G^*} d_Q FE|_{[*,F]_{K,\mathscr{A}}} = 0 \tag{10.12}$$

Let us now present the message-passing algorithm we consider. For two elements of $\mathscr{A}$, a and b such that $b \leq a$, two types of messages are considered: top-down messages $m_{a\to b} \in \mathbb{R}^{G(a)}$ and bottom-up messages $n_{b\to a} \in \mathbb{R}^{G(a)}$. Consider messages at times t, $(n^t_{b\to a}, m^t_{a\to b}; b, a \in \mathscr{A} \text{ s.t. } b \leq a)$, they are related by the following relation,

$$\forall b \leq a, \forall \omega_1 \in G(a) \quad n^t_{b\to a}(\omega_1) = \prod_{\substack{c: b\leq c \\ c\not\leq a}} \sum_{\omega\in G(b)} m^t_{c\to b}(\omega) G^a_b(\omega|\omega_1) \tag{10.13}$$

One then defines beliefs as ,

$$\forall a \in \mathscr{A}, \quad b^t_a \propto e^{-H_a} \prod_{\substack{b\in\mathscr{A}: \\ b\leq a}} n^t_{b\to a} \tag{10.14}$$

where $b_a \in \mathbb{P}(G(a)$. The update rule is,

$$\forall a, b \in \mathscr{A}, \text{s.t.} b \leq a \quad m^{t+1}_{a\to b} = m^t_{a\to b} \frac{F^b_a b^t_b}{b^t_a} \tag{10.15}$$

Theorem 2.2 [10] (version 2) applied to specifications implies that fixed points of the previous message-passing algorithm are in correspondence with critical points of the Bethe free energy over the space of Gibbs measures.

11 Acknowledgement

We would like to thank the reviewers for their comments.

References

[1] G. Sergeant-Perthuis, "Compositional statistical mechanics, entropy and variational inference," in *Twelfth Symposium on Compositional Structures*, 2024.

[2] G. Sergeant-Perthuis, "Bayesian/graphoid intersection property for factorisation spaces," 2021. https://arxiv.org/abs/1903.06026.

[3] G. Sergeant-Perthuis, "Intersection property and interaction decomposition," Apr. 2019. https://arxiv.org/abs/1904.09017.

[4] G. Sergeant-Perthuis, "Interaction decomposition for presheaves," Aug. 2020. https://arxiv.org/abs/2008.09029.

[5] D. Bennequin, O. Peltre, G. Sergeant-Perthuis, and J. P. Vigneaux, "Extra-fine sheaves and interaction decompositions," Sept. 2020. https://arxiv.org/abs/2009.12646.

[6] G. Sergeant-Perthuis, "Interaction decomposition for Hilbert spaces," 2021. https://arxiv.org/abs/2107.06444.

[7] G. Sergeant-Perthuis, *Intersection property, interaction decomposition, regionalized optimization and applications.* PhD thesis, Université de Paris, 2021. Link.

[8] G. Sergeant-Perthuis, "A categorical approach to statistical mechanics," in *Geometric Science of Information* (F. Nielsen and F. Barbaresco, eds.), (Cham), pp. 258–267, Springer Nature Switzerland, 2023.

[9] G. Sergeant-Perthuis, "Characterization of extreme Gibbs measures for a Categorical Approach to Statistical Mechanics," Feb. 2024. https://hal.sorbonne-universite.fr/hal-04456412.

[10] G. Sergeant-Perthuis, "Regionalized optimization," 2022. https://arxiv.org/abs/2201.11876.

[11] O. Peltre, "Message passing algorithms and homology," 2020. Ph.D. thesis, Link.

[12] J. Yedidia, W. Freeman, and Y. Weiss, "Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms," *IEEE Transactions on Information Theory*, vol. 51, pp. 2282–2312, July 2005.

[13] D. Ruelle, *Statistical Mechanics.* Imperial College Press, 1999.

[14] Y. Lecun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, *A tutorial on energy-based learning.* MIT Press, 2006.

[15] T. Lelièvre, M. Rousset, and G. Stoltz, *Free Energy Computations*. Imperial College Press, 2010.

[16] C. Chipot and A. Pohorille, eds., *Free energy calculations: theory and applications in chemistry and biology*. No. 86 in Springer series in chemical physics, Berlin ; New York: Springer, 2007. OCLC: ocm79447449.

[17] M. Wiering and M. van Otterlo, eds., *Reinforcement Learning: State-of-the-Art*, vol. 12 of *Adaptation, Learning, and Optimization*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[18] M. Mezard and A. Montanari, *Information, Physics, and Computation*. USA: Oxford University Press, Inc., 2009.

[19] M. Gromov, "In a search for a structure, part 1: On entropy.," 2013.

[20] T. Fritz, "A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics," *Advances in Mathematics*, 2020.

[21] J. P. Vigneaux, "Information structures and their cohomology," *Theory and Applications of Categories*, vol. 35, no. 38, pp. 1476–1529, 2020.

[22] T. Fritz and D. I. Spivak, "Internal probability valuations." Workshop Categorical Probability and Statistics, 2020.

[23] J.-C. Belfiore and D. Bennequin, "Topos and stacks of deep neural networks," *ArXiv*, 2021.

[24] T. Fritz and P. Perrone, "A probability monad as the colimit of spaces of finite samples," *Theory and Applications of Categories*, 2019.

[25] M. Marcolli, "Gamma spaces and information," *Journal of Geometry and Physics*, 2019.

[26] P. Baudot and D. Bennequin, "The homological nature of entropy," *Entropy*, vol. 17, no. 5, pp. 3253–3318, 2015.

[27] T. Fritz, T. Gonda, and P. Perrone, "De finetti's theorem in categorical probability," *arXiv preprint arXiv:2105.02639*, 2021.

[28] S. Moss and P. Perrone, "A category-theoretic proof of the ergodic decomposition theorem," *Ergodic Theory and Dynamical Systems*, vol. 43, no. 12, p. 4166–4192, 2023.

[29] H.-O. Georgii, *Gibbs Measures and Phase Transitions*. Berlin, New York: De Gruyter, 2011.

[30] P. Alquier, "Approximate Bayesian Inference," *Entropy*, vol. 22, p. 1272, Nov. 2020.

[31] L. Onsager, "Crystal statistics. i. a two-dimensional model with an order-disorder transition," *Phys. Rev.*, vol. 65, pp. 117–149, Feb 1944.

[32] S. El-Showk, M. F. Paulos, D. Poland, S. Rychkov, D. Simmons-Duffin, and A. Vichi, "Solving the 3d ising model with the conformal bootstrap," *Phys. Rev. D*, vol. 86, p. 025022, Jul 2012.

[33] D. Simson, "Linear representations of partially ordered sets and vector space categories," 1993.

[34] C.-S. Hu, "A brief note for sheaf structures on posets," 2020.

[35] K. Yanagawa, "Sheaves on finite posets and modules over normal semigroup rings," *Journal of Pure and Applied Algebra*, vol. 161, no. 3, pp. 341–366, 2001.

[36] A. Brown and O. Draganov, "Computing minimal injective resolutions of sheaves on finite posets," 2021.

[37] E. Spiegel and C. J. O'Donnell, *Incidence Algebras*. CRC Press, 1997.

[38] M. Giry, "A categorical approach to probability theory," in *Categorical aspects of topology and analysis*, vol. 915 of *Lecture notes in Mathematics*, pp. 68–85, Springer, 1982.

[39] E. W. Lawvere, "The category of probabilistic mappings," 1962. Link.

[40] S. Mac Lane, *Categories for the working mathematician*, vol. 5. Springer Science & Business Media, 2013.

[41] J. Gallier and J. Quaintance, *Homology, Cohomology, and Sheaf Cohomology for Algebraic Topology, Algebraic Geometry, and Differential Geometry*. World Scientific, 2022.

[42] H. K. Kesavan, *Jaynes' maximum entropy principle*, pp. 1779–1782. Boston, MA: Springer US, 2009.

[43] A. De Martino and D. De Martino, "An introduction to the maximum entropy approach and its application to inference problems in biology," *Heliyon*, vol. 4, no. 4, p. e00596, 2018.

[44] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[45] S. L. Lauritzen, *Graphical Models*. Oxford Science Publications, 1996.

[46] E. Sudderth and J. Pacheco, "Probabilistic graphical models." Brown University CSCI 2950-P, Spring 2013, Lecture 12: Gaussian Belief Propagation, State Space Models and Kalman Filters, Guest Kalman Filter Lecture by Jason Pacheco, 2013.

[47] T. Morita, "Cluster variation method of cooperative phenomena and its generalization i," *Journal of the Physical Society of Japan*, vol. 12, no. 7, pp. 753–755, 1957.

[48] O. Peltre, "A homological approach to belief propagation and bethe approximations," in *Geometric Science of Information* (F. Nielsen and F. Barbaresco, eds.), (Cham), pp. 218–227, Springer International Publishing, 2019.

[49] T. P. Speed, "A note on nearest-neighbour gibbs and markov probabilities," *Sankhyā: The Indian Journal of Statistics, Series A*, 1979.

[50] G.-C. Rota, "On the foundations of combinatorial theory I. Theory of Möbius functions," *Probability theory and related fields*, vol. 2, no. 4, pp. 340–368, 1964.

[51] H. K. T., "On the amount of information," *Theory of Probability and its Applications*, 1962.