Inference on diagrams in the category of Markov kernels

Gregoire Sergeant-Perthuis and Nils Ruet

LCQB Sorbonne Université

ACT 7
Oxford, UK 18 June, 2024

Paper arXiv:2201.11876

Geometric Deep Learning

Introduction to geometric deep learning:

- Deep learning $\leftarrow$ curse of dimensionality
- Accounting for symmetry
  - $\rightarrow$ Translation $\rightsquigarrow$ CNN
- Geometry $\rightsquigarrow$ discretize
  - $\rightarrow$ Graph NN [BBCV21]
  - $\rightarrow$ Nodes share same features
  - $\rightarrow$ Limitations: heterogeneous data
- Heterogeneity
  - $\rightarrow$ Cellular sheaves [Cur13]
  - $\rightarrow$ cell complex, faces $\rightsquigarrow$ feature space, inclusions $\rightsquigarrow$ linear maps
  - $\rightarrow$ Functor from a poset to **Vect**
  - $\rightarrow$ Sheaf Neural Networks [BGC+22]

First remark: limitation $\rightarrow$ cell complexes.

Bayesian inference

- We will not talk about geometric deep learning today.
- Bayesian inference: graphical models, Markov random fields, factor graphs
- Limitation: no heterogeneity, no locality in the description of variables

$\rightarrow$ Extend Bayesian inference to account for probabilistic modeling with heterogeneity and local descriptions.

$\rightsquigarrow$ Independent work from PhD [SP21]

Structure of the Presentation

1. Graphical models
2. Factor graphs
3. Inference and (General) Belief Propagation
4. Graphical modesl, Factor graph as contravariant functor
5. New!: Heterogeneous structures and probabilistic modeling
6. New!: Inference on diagrams in the category of Markov kernels

Graphical models

Definition (Undirected Graphical model)

A graphical model is the data of

- an undirected graph $G = (I, A)$,
- a collection of variables $X = (X_i \in E_i, i \in I)$, one per node and one variable corresponds exactly to one node

Definition (Markov properties)

Let $G = (I, A)$ be a finite graph. Let $X = (X_i, i \in I)$ be a collection of random variables taking respectively values in the finite sets E_i. A strictly positive probability $P_X \in \mathbb{P}(X)$ on the finite set $\Omega = \prod_{i \in I} E_i$ obeys,

1. (P) the pairwise Markov property relative to G, if for any pair (i, j) of non-adjacent vertices

$$X_i \perp\!\!\!\perp X_j | X_{I \setminus \{i,j\}}.$$

2. (L) the local Markov property relative to G, if for any vectex $i \in V$,

$$X_i \perp\!\!\!\perp X_{I \setminus (i \cup \partial i)} | X_{\partial i}$$

And we call the respective sets $P(G)$, $L(G)$.

Definition (Factorisation space)

Let I be a finite set, let $\mathscr{A} \subseteq \mathscr{P}(I)$, where $\mathscr{P}(I)$ is the set of subsets of I. Let $(E_i, i \in I)$ be a collection of sets, let $E_a = \prod_{i \in a} E_i$ for any $a \in \mathscr{P}(I)$; for $x \in \Omega$, we will denote x_a its projection onto E_a. The factorisation space over $\mathscr{A}$ is defined as follows,

$$\mathrm{Fac}_{\mathscr{A}} = \{P \in \mathbb{P}(\Omega) : \quad \exists (f_a \in \mathbb{R}_{>0}^{E_a}, a \in \mathscr{A}), \text{s. t.} \forall x \in \Omega \; P = \prod_{a \in \mathscr{A}} f_a(x_a)\} \tag{0.1}$$

- How to relate the Markov properties to factorizations of the underlying distribution?

Definition (Cliques of a graph)

Let $G = (I, A)$ be a graph; a clique of G is a subset of G such that every two distinct vertices are adjacent. We will note $\mathscr{C}$ the set of its cliques.

Theorem (Hammersley-Clifford)

Let $G = (I, A)$ be a finite graph. For all P_X strictly positive probability law on a finite set $\prod_{i \in I} E_i$,

$$P_X \in P(G) \iff P_X \in L(G) \iff P_X \in \mathrm{Fac}_{\mathscr{C}} . \tag{0.2}$$

- Taking the 'log': product $\to$ sum
  - $\to$ $\prod_a f_a \to \sum_a H_a$
  - $\to$ Relation to statistical mechanics

Inference $\to$ Belief propagation (in few slides)

- Directed graphical models: Bayesian networks
- Inference on Bayesian networks:
  - $\rightarrow$ Define an undirected graphical model
  - $\rightarrow$ Inference on undirected graphical model

Factor Graphs

- We want more general interaction than pairwise interaction
- Factor graphs:
  - Bipartite graphs, nodes $V = V_0 \sqcup V_1$
  - V_1 collection of $a \subseteq I$ with $a \to f_a$
  - V_0 the set of indices $i \in a$ for some $a \in V_1$
  - Edges $i \to a$ when $i \in a$

An example:

- Graphical model: $X - Y - Z$
- Factor graph: $X \to f_{X,Y} \leftarrow Y \to f_{Y,Z} \leftarrow Z$
- Factor graphs generalize graphical models
- Spaces of factorization generalize both $\rightsquigarrow$ statistical mechanics.

Exemple de modèle graphique et Belief Propagation

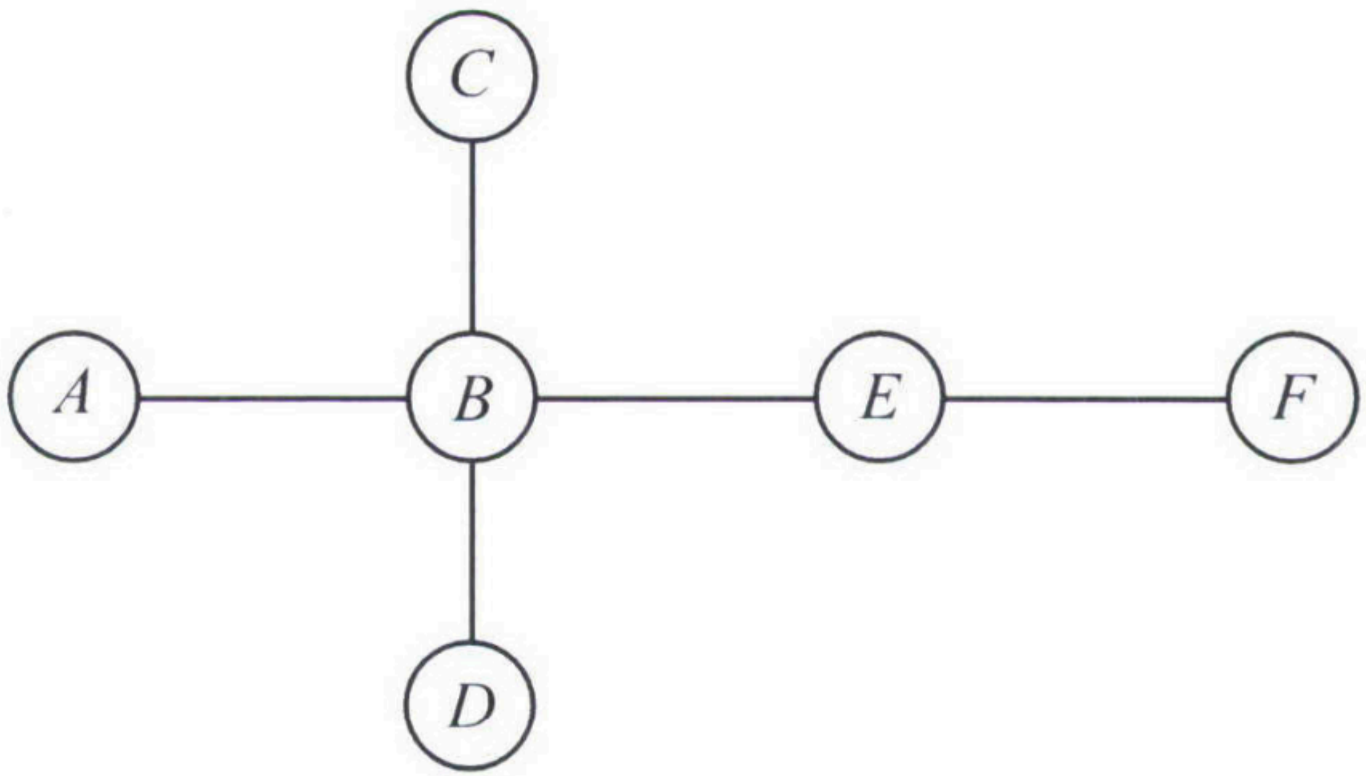

Figure 15. An acyclic undirected graphical model.

Reference: Statistical Inference in Graphical Models, K. Gimpel, D. Rudoy

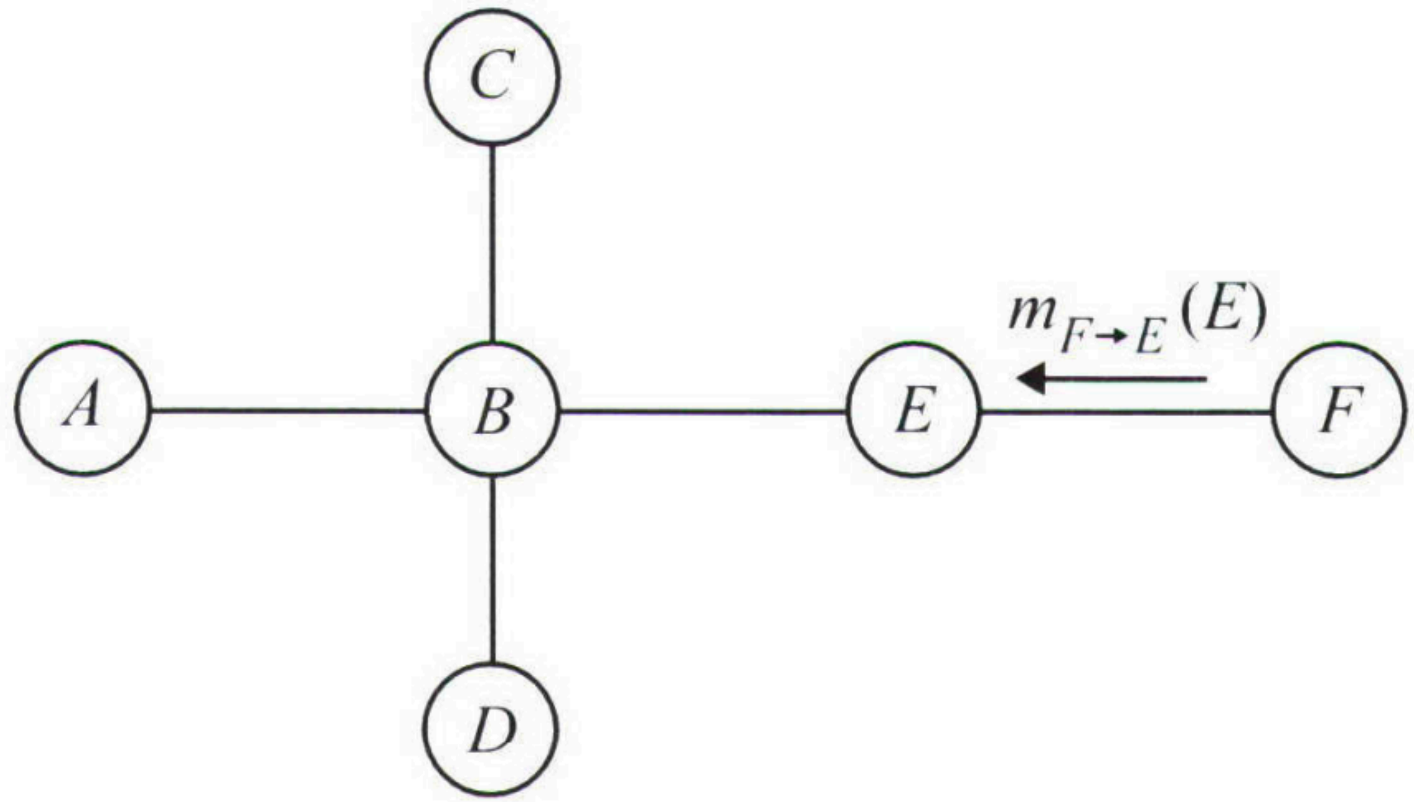

Figure 16. A message passed from node F to node E.

$$
\begin{aligned}
p(A) &= \sum_B p(AB) \sum_C p(C \mid B) \sum_D p(D \mid B) m_{E \to B}(B) \\
&= \sum_B p(AB) m_{E \to B}(B) \sum_C p(C \mid B) \sum_D p(D \mid B) \\
&= \sum_B p(AB) m_{E \to B}(B) \sum_C p(C \mid B) m_{D \to B}(B) \\
&= \sum_B p(AB) m_{E \to B}(B) m_{D \to B}(B) \sum_C p(C \mid B) \\
&= \sum_B p(AB) m_{E \to B}(B) m_{D \to B}(B) m_{C \to B}(B) \\
&= m_{B \to A}(A).
\end{aligned}
$$

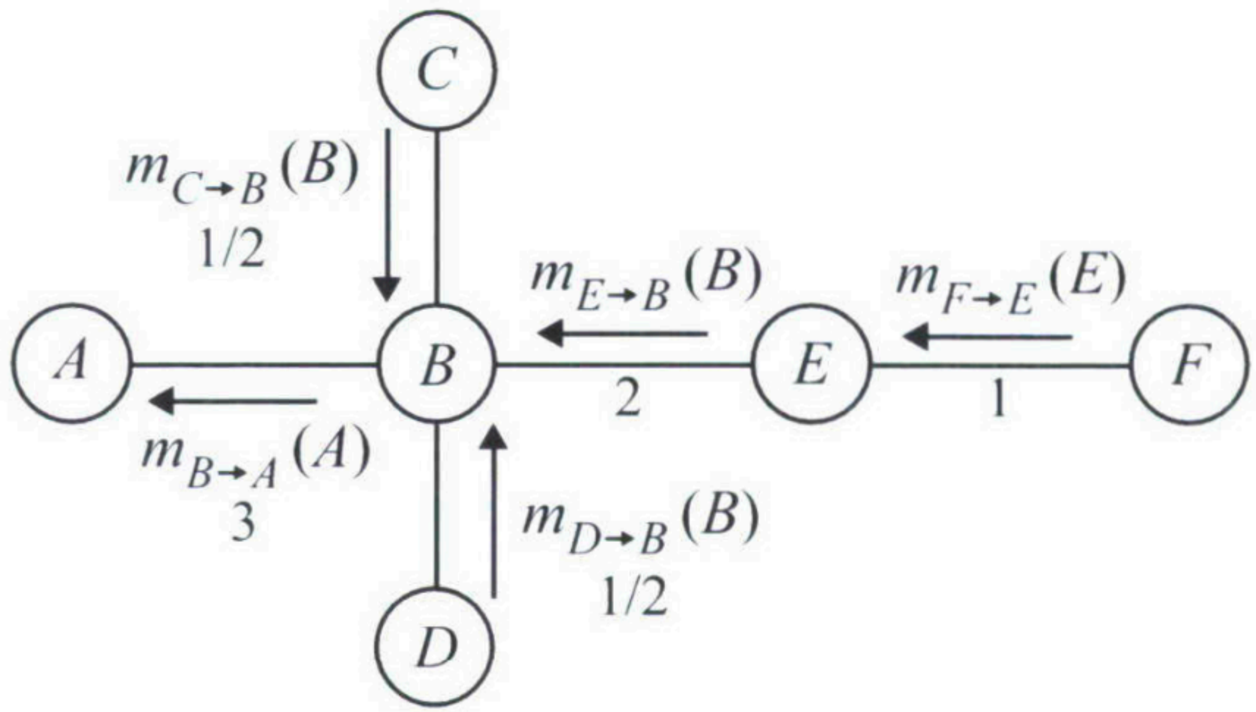

Figure 17. Message passes from an example run of the variable elimination algorithm to compute $p(A)$. A node x can only send a message to a neighbor y after x has received a message from each of its neighbors (besides y). The numbers associated with each message indicate the order in which the message must be computed. If two messages have the same number, they can be computed at the same time in parallel.

Inference on graphical models: Belief Propagation

Simpler case: on graphical models $\rightarrow$ use Belief Propagation

- Belief propagation computes the marginal distributions on edges and nodes

- Consider a collection of random variables $(X_i, i \in I)$ and an undirected graphical model $G : (I, A)$ that is **acyclic.**

$$P_{X_i, i \in I}(x_i, i \in I) = \prod_{e \in A} f_{X_e}(x_e)$$

- for each edge: two messages, $m_{X_i \rightarrow X_j} \in \mathbb{P}(E_j)$

Start with $m_{X_i \to X_j} = 1$ for all directed edges $(i \to j)$,

$$m^{t+1}_{X_i \to X_j}(x_j) = \sum_{x_i \in E_i} f_{\{i,j\}}(x_i, x_j) \prod_{Z \in \partial X_i \setminus X_j} m^t_{Z \to X_i}(x_i) \tag{0.3}$$

The stopping criteria for the algorithm is when $m_{X_i \to X_j}(x_j)^{t+1}$, which is a function over E_j, is proportional to $m_{X_i \to X_j}(x_j)^t$.
Once the algorithm has finished, the marginal distributions are computed as

$$P_{X_i}(x_i) \propto \prod_{X_j \in \partial X_i} m_{X_j \to X_i}(x_i) \tag{0.4}$$

- Importantly, the algorithm is exact: inference is exact.
  - $\rightarrow$ For Gaussian HMM $\rightsquigarrow$ (smoothed) Kalman filtering

Proposition (Factorization on acyclic graphs)

Let I be a finite set and let $\Omega = \prod_{i\in I} E_i$ be a product of finite sets and $X_i, i \in I$ a collection of random variables taking values respectively in E_i. Let $G = (I, A)$ be a finite acyclic graph. $P_X \in \mathbb{P}_{>0}(E)$ factors accordingly to $\mathscr{A}(G)$, i.e., $P_X \in \mathrm{Fac}_{\mathscr{A}(G)}$ if and only if for any $\omega \in \Omega$,

$$P_X(\omega) = \frac{\prod_{e\in A} P_{X_e}(\omega_e)}{\prod_{i\in I} P_{X_i}^{d(i)-1}(\omega_i)}, \tag{0.5}$$

where $d(i)$ is the degree of node $i \in I$.

- Bayesian inference is maximizing entropy.
- Entropy:

$$S(Q) = -\sum_{\omega \in E} Q(x) \ln Q(x) \tag{0.6}$$

- Bayesian inference:

$$\inf_{Q} DKL(Q \| P)$$

- The same as minizing Gibbs free energy

$$\inf_{Q \in \Theta} \mathbb{E}_Q[\beta H] - S(Q)$$

- But entropy:

$$S(P_X) = \sum_{e \in A} S(P_{X_e}) - \sum_{i \in I} (d(i) - 1) S(P_{X_i})$$

- Inclusion exclusion formula $c(e) = 1$, $c(i) = -(d(i) - 1)$
- Remarkably, Bayesian inference is the same as minimizing [YFW05, YFW03],

$$F_{\text{Bethe}}(Q) = \sum_{a \in V} c(a) S(Q_a) - \mathbb{E}_{Q_a}[H_a]$$

where $Q := (Q_a \in \mathbb{P}(X_a), a \in V)$ with compatibility by marginization:

$\rightarrow$ if a is an edges and i an edge in a

$\rightarrow$ $\pi_i^e : E_e \rightarrow E_i$

$\rightarrow$ we ask $\pi_{i\,*}^{e}(Q_e) = Q_i$

- Belief Propagation (BP) is a discrete-time gradient descent (on Lagrange multipliers) that solves

$$\min_{Q} F_{\text{Bethe}}(Q)$$

under 'marginal' compatibility.
- Fixed points of BP correspond to critical points of F_{Bethe}.

Graphical presheaves: what underlies Bayesian inference

$\rightarrow$ I did not invent it [Pel20]... but I call it...

Definition (Graphical presheaves)

Let I be a finite set and $\mathscr{A} \subseteq \mathscr{P}(I)$ be a sub-poset of the powerset of I. Let $E_i, i \in I$ are finite sets. For $a \in \mathscr{A}$ $E_a := \prod_{i \in a} E_i$, let $F(a) := E_a$, and for $b \subseteq a$, let $F_b^a : E_a \rightarrow E_b$ be the projection map from $\prod_{i \in a} E_i$ to $\prod_{i \in b} E_i$. F is called a graphical presheaf from $\mathscr{A}$ to $\mathbf{Mes}^f$.

- Only projections
- Only products of variables, and subcollection of variables

Recall the Structure of the Presentation

1. Graphical models
2. Factor graphs
3. Inference and (General) Belief Propagation
4. Graphical modesl, Factor graph ... as contravariant functor
5. New!: Heterogeneous structures and probabilistic model
6. New!: Inference on diagrams in the category of Markov kernels

- Consider any map, not just projections:
  - $\rightarrow$ Measurable maps for $b \rightarrow a$ and even Markov kernels
- Account for possible heterogeneity, incompleteness, and incompatibility in the description of variables:
  - $\rightarrow$ Agents with different world models that communicate their beliefs
  - $\rightarrow$ Broader class of effective models for potential computational chemistry

- **Kern**^{f}: objects are finite measurable spaces, morphisms are Markov kernels (stochastic matrices).
- F is a contravariant functor from $\mathscr{A}$ to **Kern**^{f}; $F^a_b : F(a) \to F(b)$ is denoted element-wise as $F^a_b(\omega_b \mid \omega_a)$, with $\omega_b \in F(b)$, $\omega_a \in F(a)$.
  - $\rightarrow$ F encodes all the ways our data can interact.
  - $\rightarrow$ $\mathscr{A}$ is any poset, not just a collection of subsets.
  - $\rightarrow$ Maps are not just projections.
- $Q = (Q_a \in \mathbb{P}(F(a)), a \in \mathscr{A})$
- $F_{Bethe}(Q) = \sum_{a\in\mathscr{A}} c(a)\,(\mathbb{E}_{Q_a}[H_a] - S(Q_a))$; $c(a) = \sum_{b\geq a} \mu(b,a)$ is the generalization of the inclusion-exclusion formula associated with $\mathscr{A}$.

For a finite poset $\mathscr{A}$,

- the 'zeta-operator' of $\mathscr{A}$, denoted ζ, from $\bigoplus_{a\in\mathscr{A}}\mathbb{R}$ to $\bigoplus_{a\in\mathscr{A}}\mathbb{R}$ is defined as, for any $\lambda \in \bigoplus_{a\in\mathscr{A}}\mathbb{R}$ and any $a \in \mathscr{A}$, $\zeta(\lambda)(a) = \sum\limits_{b\leq a}\lambda_b$
- its inverse denoted μ; $(\mu(a,b), b \leq a)$ Möbius function of $\mathscr{A}$.

We want to do Bayesian inference on these diagram.

- Constraint: the Q_a must be compatible under the actions of the F^a_b, i.e. $F^a_b \circ Q_a = Q_b$
- Problem: find an algorithm to 'solve' the optimization problem.
  - $\rightarrow$ New message passing algorithm!

F induces several actions: on probabilities, on probabilites seen as vectors, on their dual...

- $\tilde{F}_b^a : \mathbb{P}(F(a)) \to \mathbb{P}(F(b))$ is the linear map that sends probability distributions $p \in \mathbb{P}(F(a))$ to $F_b^a \circ p$
- $\tilde{F}^*$ is the functor obtained by dualizing the morphisms $\tilde{F}_b^a$, i.e. $\tilde{F}_a^{*,b} : \tilde{F}(b)^* \to \tilde{F}(a)^*$ sends linear maps $l_b : \tilde{F}(b) \to \mathbb{R}$ to $l_b \circ \tilde{F}_b^a : \tilde{F}(a) \to \mathbb{R}$.

μ can be extended to account for $\tilde{F}$, $\tilde{F}^*$

- for a functor G from $\mathscr{A}$ to $\mathbb{R}$-vector spaces, we define μ_G as, for any $a \in \mathscr{A}$ and $v \in \bigoplus_{a \in \mathscr{A}} G(a)$, $\mu_G(v)(a) = \sum_{b \leq a} \mu(a,b) G_a^b(v_b)$.

Frame Title

Recall $\min F_{\text{Bethe}} = \sum_a F(Q_a)$ under

- Constraint: the Q_a must be compatible under the actions of the F^a_b, i.e., $F^a_b \circ Q_a = Q_b$
  - i.e., $Q \in \lim \tilde{F}$
  - In fact, no... need to add the condition that the distribution sums to one.
  - But it's okay!

- $FE : \prod_{a\in\mathscr{A}} \mathbb{P}(E_a) \to \prod_{a\in\mathscr{A}} \mathbb{R}$ 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.
- $\mathrm{d}_Q FE \to$ differential of FE at the point Q.

Theorem

Let $\mathscr{A}$ be a finite poset, let F be a presheaf from $\mathscr{A}$ to $\textbf{Kern}^f$*. Let $H_a : F(a) \to \mathbb{R}$ be a collection of (measurable) functions. The critical points of $\mathscr{F}$ are the $Q \in \lim \tilde{F}$ such that,*

$$\mu_{\tilde{F}*} d_Q FE|_{\lim \tilde{F}} = 0 \tag{0.7}$$

Algorithm 1: Message passage algorithm for presheaves from $\mathscr{A}$ to $\mathbf{Kern}^f$

Data: Initialization: $(m^0_{a\to b} \in \mathbb{R}^{F(b)}, b, a \in \mathscr{A}$ s.t. $b \leq a)$, a poset $\mathscr{A}$, a presheaf $F : \mathscr{A} \to \mathbf{Kern}^f$;

for $t \leq T$ **do**

    for $a \in \mathscr{A}, b \in \mathscr{A}$ *such that* $b \leq a$ **do**

        $\forall \omega_a \in F(a), \quad n_{b\to a}(\omega_a) \leftarrow \prod_{\substack{c: b\leq c \\ c \nleq a}} \sum_{\omega'_b \in F(b)} m_{c\to b}(\omega'_b) \cdot F^a_b(\omega'_b|\omega_a)$

    end

    for $a \in \mathscr{A}, b \in \mathscr{A}$ *such that* $b \leq a$ **do**

        $b_a = e^{-H_a} \prod_{\substack{b \in \mathscr{A}: \\ b \leq a}} n_{b\to a}$

        $p_a = \frac{b_a}{\sum_{\omega_a} b_a(\omega_a)}$

        $m_{a\to b} \leftarrow m_{a\to b} \cdot \frac{\tilde{F}^a_b(p_a)}{p_b}$

    end

end

- Fix point of this message passing algorithm are critical point of F_{Bethe}

References I

- Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković, Geometric deep learning: Grids, groups, graphs, geodesics, and gauges, 2021.
- Cristian Bodnar, Francesco Di Giovanni, Benjamin Paul Chamberlain, Pietro Lio, and Michael M. Bronstein, Neural sheaf diffusion: A topological perspective on heterophily and oversmoothing in GNNs, Advances in Neural Information Processing Systems (Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, eds.), 2022.
- Justin Curry, Sheaves, cosheaves and applications, Ph.D. thesis, The University of Pennsylvania, 2013, arXiv:1303.3255.
- Olivier Peltre, Message passing algorithms and homology, 2020, Ph.D. thesis, Link to manuscript.
- Grégoire Sergeant-Perthuis, Intersection property, interaction decomposition, regionalized optimization and applications, 2021, PhD thesis, 10.13140/RG.2.2.19278.38729, Link to manuscript.

References II

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss, Understanding belief propagation and its generalizations, p. 239–269, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003.

J.S. Yedidia, W.T. Freeman, and Y. Weiss, Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms, IEEE Transactions on Information Theory **51** (2005), no. 7, 2282–2312 (en).

Proof of Characterization of Critical Points

Understanding expression of critical points:

Zeta function ζ and Möbius functions μ for functors:

- for $u \in \bigoplus_{a \in \mathscr{A}} G(a)$, and $a \in \mathscr{A}$,

$$\zeta_G(u)(a) = \sum_{b \leq a} G_a^b(u_b)$$

-

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

μ_G is the inverse of ζ_G

Proof of Characterization of Critical Points

Understanding expression of critical points:

For F a functor from $\mathscr{A}^{op}$ to vector spaces, critical points u of 'global' regionalized loss are such that:

$$[\mu_{F^*} d_u l]|_{\lim F} = 0$$

Proof of Characterization of Critical Points

Understanding expression of critical points:

$$0 \to \lim F \to \bigoplus_{a \in \mathscr{A}} F(a) \xrightarrow{\delta_F} \bigoplus_{\substack{a,b \in \mathscr{A} \\ a \geq b}} F(b)$$

where for any $v \in \bigoplus_{\substack{a,b \in \mathscr{A} \\ a \geq b}} F(b)$ and $a, b \in \mathscr{A}$ such that $b \leq a$,

$$\delta_F(v)(a, b) = F^a_b(v_a) - v_b$$

This is simply stating that $\ker \delta = \lim F$.

Proof of Characterization of Critical Points

Understanding expression of critical points:

$$0 \leftarrow (\lim F)^* \leftarrow \bigoplus_{a\in\mathscr{A}} F(a)^* \overset{d_F}{\leftarrow} \bigoplus_{\substack{a,b\in\mathscr{A}\\ a\geq b}} F(b)^*$$

Pose $d = \delta^*$. For any $l_{a\to b} \in \bigoplus_{\substack{a,b\in\mathscr{A}\\ a\geq b}} F(b)^*$ and $a \in \mathscr{A}$,

$$dm(a) = \sum_{a\geq b} F_b^{a*}(m_{a\to b}) - \sum_{b\geq a} m_{b\to a}$$

Proof of Characterization of Critical Points

Rewriting condition on fix points:

$$\mu_F^* d_u l \in \operatorname{im} \mathrm{d}$$

is the same as the fact that there is $(m_{a\to b} \in F(b)^* | a, b \in \mathscr{A}, b \leq a)$ such that,

$$d_u l = \zeta_{F^*} \mathrm{d} m$$

Proof of Characterization of fix points of algorithm

Understanding this choice of message passing algorithm:

g Lagrange multipliers m to $u \in \bigoplus_{a\in\mathscr{A}} F(a)$. $\delta_F(u) = 0$ defines the constraints on u.
$\delta_F g \zeta_{F^*} \mathrm{d}_F$ sends a Lagrange multiplier $m \in \bigoplus_{\substack{a,b\in\mathscr{A}\\ a\geq b}} F(b)^*$ to a constraint $c \in \bigoplus_{\substack{a,b\in\mathscr{A}\\ a\geq b}} F(b)$ defined as, for $a, b \in \mathscr{A}$ such that $b \leq a$,

$$c(a,b) = \delta_F g \zeta_{F^*} \mathrm{d}_F m(a,b) = F^a_b g_a(\zeta_{F^*} \mathrm{d}_F m(a)) - g_b(\zeta_{F^*} \mathrm{d}_F m(b))) \tag{0.1}$$

We are interested in $c = 0$, i.e.

$$\delta_F g \zeta_{F^*} \mathrm{d}_F m = 0$$

Proof of Characterization of fix points of algorithm

Understanding this choice of message passing algorithm:

Choice of algorithm on the Lagrange multipliers so that $\delta_F g \zeta_{F*} \mathrm{d}_F m = 0$,

$$m(t+1) - m(t) = \delta_F g \zeta_{F*} \mathrm{d}_F m(t)$$

Any other choice would also be a good candidate!