Functoriality of inference on diagrams in the category of Markov kernels.

Grégoire Sergeant-Perthuis

Functoriality of inference on diagrams in the category of Markov kernels.

Gregoire Sergeant-Perthuis

LCQB Sorbonne Université

Non Commutative Geometry & Higher Structures
Thessaloniki 21/9/24

Papers: arXiv:2201.11876, arXiv:2403.16104 (SYCO 12), hal-04527780 (ACT 7)

Introduction to geometric deep learning [BBCV21]:

- Deep learning $\leftarrow$ curse of dimensionality
- Accounting for symmetry
  - $\rightarrow$ Translation $\rightsquigarrow$ CNN
  - $\rightarrow$ Other groups [SPMBO22]
- 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** [SP22, SP24a, SPR24, SP24b]
  - $\rightarrow$ Sheaf Neural Networks [BGC$^+$22]

- Our focus:
  - $\rightarrow$ Bayesian inference, graphical models, Markov random fields, factor graphs

Context:

Let X be a random variable taking values in a finite measurable space E_X, and $\theta \in \Theta \subseteq \mathbb{R}^d$ that parametrizes a collection of probability measures $P(x|\theta)$, where $x \in E_X$.
Assume that one is given a prior $Q \in \mathbb{P}(\Theta)$, where $\mathbb{P}(\Theta)$ denotes the space of probability measures on Θ. For any observation $x_0 \in E_X$, one computes the posterior using Bayes' rule:

$$P(\theta|x) = \frac{P(x|\theta)Q(\theta)}{\int P(x|\theta)Q(\theta)\, d\theta}$$

- Problem: When $\theta = (\theta_i, i \in [0, N])$ is a collection of variables, where each $\theta_i \in E$.
- $\sum_\theta P(x|\theta)Q(\theta) = \sum_{\theta_0} \cdots \sum_{\theta_N} P(x|\theta_0, \ldots, \theta_N)Q(\theta_0, \ldots, \theta_N)$

  $\rightarrow$ Number of operations: $\mathscr{O}(|E|^N)$
- Notation: For a set of indices I and a subset $a \subseteq I$, $\theta_a := (\theta_i \in E_i, i \in a)$.
- In what follows, all the sets in which variables take values will be finite: E_i are finite sets.

- One relation to statistical mechanics:

Let $\theta = (Y_i \in E_i, i \in I)$ be the unobserved variables, and $X = (X_j \in F_j, j \in J)$ the observed variables. Both I and J are finite sets, and E_j for $j \in J$ and F_i for $i \in I$ are finite sets.

$$\ln P(\theta, X) = -\beta \sum_{a \subseteq I \sqcup J} H_a(X_{a\cap J}, Y_{a\cap I})$$

Given an observation $x = (x_i, i \in I)$, computing $\ln P(\theta|x)$ is equivalent to computing:

$$\ln \sum_{(y_i, i\in I)} e^{-\beta \sum_{a\subseteq I\sqcup J} H_a(x_{a\cap J}, y_{a\cap I})}$$

This is the same as:

$$\ln Z(x) := \ln \sum_{\theta \in \prod_i E_i} e^{-\beta \tilde{H}_x(\theta)}$$

- Similar frameworks but different names: Bayesian networks, graphical models, factor graphs, Markov random fields.

Definition (Factorisation Space)

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

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

Consider an undirected graph $G = (I, A)$ that is **acyclic.** Denote $\mathscr{A}(G)$ as the *partially ordered set* (poset) with elements $V = I \sqcup A$ and the following relations:

- $\forall i \in I,\ i \leq i$, and $\forall e \in A,\ e \leq e$
- $\forall i \in I,\ \forall e \in A,\ i \leq e \iff i \in e$

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 according to $\mathscr{A}(G)$, i.e., $P_X \in \mathrm{Fac}_{\mathscr{A}(G)}$, if and only if for any $\omega \in \Omega$,

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

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

- Bayesian inference is maximizing (relative) entropy.
- **Entropy:**

$$S(Q) = -\sum_{\omega \in E} Q(\omega) \ln Q(\omega) \tag{0.1}$$

- Recall that minimizing Gibbs free energy gives Helmholtz free energy:

$$\beta \frac{-\ln Z}{\beta} = \inf_{Q \in \mathbb{P}(E)} \left(\mathbb{E}_Q[\beta H] - S(Q)\right)$$

- Set $\beta = 1$.

- 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]:

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

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

→ If a is an edge and i a vertex in a

→ $\pi_i^e : E_e \to E_i$

→ We ask that ${\pi_i^e}_*(Q_e) = Q_i$

- Bayesian inference corresponds to computing $\ln Z$ for a Hamiltonian $H : \prod_{i \in I} E_i \to \mathbb{R}$, with $Z = \sum_x e^{-\beta H(x)}$.
- From now on, set $\beta = 1$; notation $E = \prod_{i \in I} E_i$.
- It is computationally costly to compute directly, but note that

$$-\ln Z = \inf_{Q \in \mathbb{P}(E)} \left(\mathbb{E}_Q[H] - S(Q)\right)$$

- The previous problem can be reformulated as minimizing:

$$F_{\mathrm{Bethe}}(Q_a, a \in \mathscr{A}(G)) = \sum_{a \in \mathscr{A}(G)} c(a) \left(\mathbb{E}_{Q_a}[H_a] - S(Q_a)\right)$$

with $Q_i(x_i) = \sum_{y \in X_{i'}} Q_e(x_i, y)$ when $e = \{i, i'\}$.

- Belief propagation is an algorithm of complexity $\mathscr{O}(|A||E_i|^2)$ to solve this optimization problem, when $E_i = E_j$ for all $i, j \in I$.

→ Extension to higher-order interactions: not just graphs.

→ I did not invent it [Pel20, YFW05]... but no name?

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 $E_i, i \in I$ be finite sets. For $a \in \mathscr{A}$, define $E_a := \prod_{i \in a} E_i$. Let $F(a) := E_a$, and for $b \subseteq a$, let $F^a_b : E_a \to E_b$ be the projection map from $\prod_{i \in a} E_i$ to $\prod_{i \in b} E_i$. We call F a graphical presheaf from $\mathscr{A}$ to $\mathbf{Mes}^f$.

- Only projections.
- Only products of variables, and subcollections of variables.

- Consider any map, not just projections:
  - $\rightarrow$ Any measurable maps for $b \rightarrow a$ and even Markov kernels, i.e., stochastic matrices when the source and target are finite sets.
- 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 potentials in computational chemistry.

Extension done in previous work [SP22, SPR24, SP24a]

- **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 $\omega_b \in F(b)$, $\omega_a \in F(a)$.
  - $\to$ F encodes all the ways our data can interact.
  - $\to$ $\mathscr{A}$ is any poset, not just a collection of subsets.
  - $\to$ 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 $a\in\mathscr{A}$, $\zeta(\lambda)(a)=\sum_{b\leq a}\lambda_b$
- its inverse is denoted as 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...

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

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

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

Recall we want to solve $\inf F_{\text{Bethe}} = \sum_a c(a) F(Q_a)$ under

- Constraint: the Q_a must be compatible under the actions of the F_b^a, i.e., $F_b^a \circ Q_a = Q_b$
  - 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!

image-act.png

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

Proposition

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

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

$T \lim \tilde{F}$ is the underlying vector space of the affine space $\lim \tilde{F}$

Pose $I_a(Q_a) = \mathbb{E}_{Q_a}[H_a] - S(Q_a)$

Theorem (GSP)

F a functor from $\mathscr{A}^{op}$ to vector spaces. An element $u \in \lim \tilde{F}$ is a critical point of the F_{Bethe} if and only if there is $(m_{a\to b} \in \bigoplus_{\substack{a,b:\\ b\leq a}} \tilde{F}(b)^*)$ such that for any $a \in \mathscr{A}$,

$$d_u I_a = \sum_{b\leq a} \tilde{F}_b^{a*} \left(\sum_{c\leq b} \tilde{F}_c^{b*} m_{b\to c} - \sum_{c\geq b} m_{c\to b} \right) \tag{CP}$$

- To understand in greater detail these propositions and the previous algorithm, we need to extend the setting of the optimization problem.
- Change the loss:
  - $\rightarrow$ Replace entropy with a ”local loss.”
  - $\rightarrow$ $S(Q_a) \rightsquigarrow l_a(v_a)$
- Change the functor:
  - $\rightarrow$ Replace F with a contravariant functor from a poset $\mathscr{A}$ to **Vect**.
- Result: we can extend the message passing algorithm to solve:

$$\min_{v} \sum_{a \in \mathscr{A}} c(a) l_a(v_a)$$

  with $v := (v_a, a \in \mathscr{A})$ under the constraint $v \in \lim F$.
- This approach is different and on some points more general than decentralized optimization on cellular sheaves [HG19].

For F a functor from $\mathscr{A}^{op}$ to vector spaces, critical points u of $\sum_{a\in\mathscr{A}} c(a) l_a(v_a)$ are $u \in \bigoplus_{a\in\mathscr{A}} F(a)$ such that:

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

where, $l(v) = (l_a(v_a), a \in \mathscr{A})$, $d_u l(a) = d_{u_a} l_a$ and,

$$[\mu_{F^*} d_u l](a) = \sum_{b\leq a} \mu(a,b) d_{u_b} l_b \circ F_b^a$$

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

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 $b \leq a$,

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

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

Understanding expression of critical points:

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

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

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

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

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,

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

Assume that the local losses $l_a, a \in \mathscr{A}$ are such that there is a collection of functions $g_a, a \in \mathscr{A}$ that inverses the relation induced by differentiating the local losses, i.e.

$$d_{u_a} l_a = y_a \iff u_a = g_a(y_a)$$

It is the case for the free energy $\mathbb{E}_{Q_a}[H_a] - S(Q_a)$.
Messages:

$$m(t) \in \bigoplus_{\substack{a,b:\\ b \leq a}} F(b)^*\colon m_{a \to b} \text{ for } b \leq a$$

Understanding this choice of message passing algorithm:

g sends 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 $b \leq a$,

$$c(a,b) = \delta_F g \zeta_{F^*} \mathrm{d}_F m(a,b) = F_b^a g_a(\zeta_{F^*} \mathrm{d}_F m(a)) - g_b(\zeta_{F^*} \mathrm{d}_F m(b))) \tag{0.2}$$

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

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

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!

The message passing algorithm is defined as:

$$\delta m := \delta_{\tilde{F}} g \zeta_{\tilde{F}*} d_{\tilde{F}} m$$

Define $BP_{F,H} := \delta_{\tilde{F}} g \zeta_{\tilde{F}*} d_{\tilde{F}}$.
When differentiating the free energy:

$$y_a = H_a + \ln q_a + 1$$

Therefore, $g_a(y_a) = e^{y_a - H_a - 1}$.

- Joint work with Toby St Clere Smithe, in progress.
- Consider a natural transformation $\phi : F \to F_1$ where ϕ_a is a deterministic map, not a Markov kernel.
- The map ϕ extends into maps between $\tilde{F} \to \tilde{F}_1$ and $\tilde{F}^* \to \tilde{F}_1^*$.
- ϕ induces maps between $\bigoplus_b \tilde{F}(b) \to \bigoplus_b \tilde{F}_1(b)$ and $\bigoplus_{b\leq a} \tilde{F}(b) \to \bigoplus_{b\leq a} \tilde{F}_1(b)$. It also induces a map $\phi^* : \bigoplus_b \tilde{F}_1^*(b) \to \bigoplus_b \tilde{F}^*(b)$ and $\phi^* : \bigoplus_{b\leq a} \tilde{F}_1^*(b) \to \bigoplus_{b\leq a} \tilde{F}^*(b)$.

Pose:

$$\tilde{H}_a = \ln \sum_{\omega' : \phi_a(\omega') = \omega} e^{-H_a(\omega')}$$

Then we showed that:

$$BP_{F_1, \tilde{H}} = \phi \circ BP_{F,H} \circ \phi^*$$

- $\rightarrow$ Few results on characterizing critical points of the Bethe free energy.
- $\rightarrow$ Use transformations on the underlying functor to reduce to simpler cases (Hamiltonians, posets).

What about base change? $\phi : \mathscr{A} \rightarrow \mathscr{A}_1$

- $\rightarrow$ When a right adjoint to the pullback exists, results on natural transformations can be reused.
- $\rightarrow$ When $\mathscr{A}$ is isomorphic to a full subposet of $\mathscr{A}_1$, similar result holds.

Thank you for your attention!

- 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.
- Jakob Hansen and Robert Ghrist, Distributed optimization with sheaf homological constraints, 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2019, pp. 565–571.
- Olivier Peltre, Message passing algorithms and homology, 2020, Ph.D. thesis, Link to manuscript.
- Grégoire Sergeant-Perthuis, Regionalized optimization, 2022.

References II

- Grégoire Sergeant-Perthuis, Compositional statistical mechanics, entropy and variational free energy (Extended a 7th International Conference on Applied Category Theory (ACT 7) (Oxford, UK, France), David Jaz Myers and Michael Johnson, June 2024.
- Grégoire Sergeant-Perthuis, Compositional statistical mechanics, entropy and variational inference, 2024, Presented at the Twelfth Symposium on Compositional Structures (SYCO 12).
- Grégoire Sergeant-Perthuis, Jakob Maier, Joan Bruna, and Edouard Oyallon, On non-linear operators for geometric deep learning, Advances in Neural Information Processing Systems (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, Curran Associates, Inc., 2022, pp. 10984–10995.
- Grégoire Sergeant-Perthuis and Nils Ruet, Inference on diagrams in the category of Markov kernels (Extended abstract), 7th International Conference on Applied Category Theory (ACT 7) (Oxford (UK), United Kingdom), David Jaz Myers and Michael Johnso, June 2024.

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).