Optimization over presheaves and message passing algorithms (with applications).

Gregoire Sergeant-Perthuis

INRIA & IMJ-PRG

January 17,2023

Reference

Presentation based on work in:

Regionalized optimization, arXiv:2201.11876 [SP22]

Outline of the presentation

I. Compositional structure: what and why?
II. Optimization on compositional structure: what and how?

I. Compositionality for modeling structured data

Definition

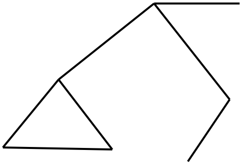
Definition of compositionality (in computer science):

> *The ability to determine properties of the whole [system] from properties of the parts together with the way in which the parts are put together.*

John Baez, Compositionality, The n-Category Café
What I will be presenting: functor over a partially ordered set.

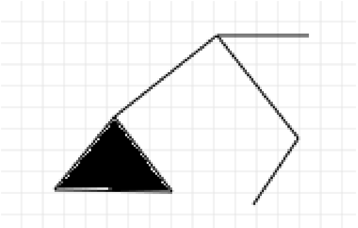
Example of compositionality: Combinatorial objects I

- Graphs (V, E):
  - Object are nodes, 'glued' by edges.
  - Relation between objects: there is a path between two nodes.



Example of compositionality: Combinatorial objects II

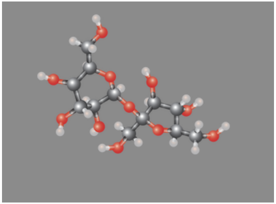
- Simplicial complexes of dimension $n \in \mathbb{N}$
  - $*$ Object are simplicies: $k \leq n$ simplices Δ_k
  - $*$ 'glued' on their borders: $k-1$ simplices $\Delta_{k-1} \in \partial\Delta_k$
  - $*$ Relation between objects: inclusion $\Delta_{k-1} \subseteq \Delta_k$.

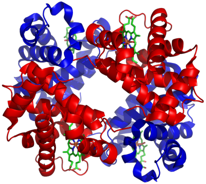


- E.g.: Graphs are simplicial complexes of dimension 1.

Combinatorial objects as discretized geometry: leverage geometry of data I

- How to process data with geometrical properties?
  - 3D shapes: human body
  - Networks: recommendation, knowledge, traffic...
  - Chemical structures: molecules
  - Structural biology: protein networks

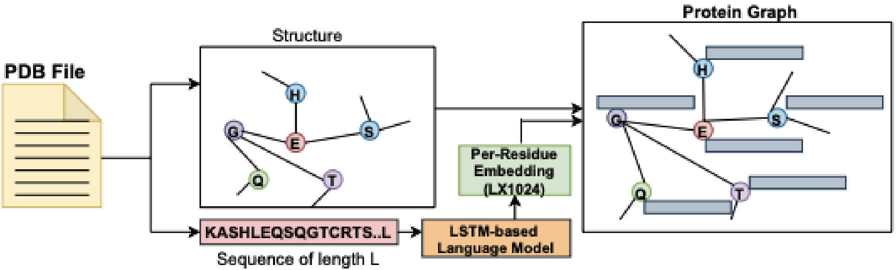




Combinatorial objects as discretized geometry: leverage geometry of data II

- As images?$\to$ Convolution Neural Networks
  - $*$ good for data structured over Euclidean space ('Euclidean data': sound, images, videos)
  - $*$ Downside: does not leverage the knowledge of specific geometry (if more complex than a grid)
- Solution:
  - $*$ discretize geometry (more general than grids)
  - $*$ geometry $\to$ combinatorial structure (e.g. graph)

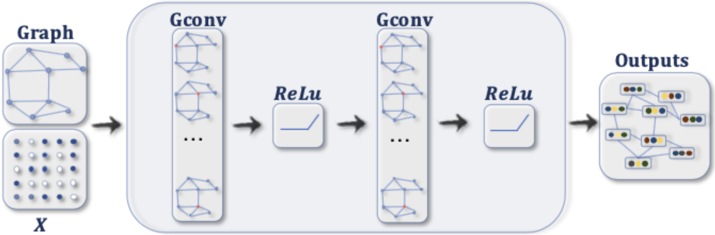
Geometric Data processing: Geometric deep learning!



Graph representation of a protein with node features. Reproduced from [JSS22]

- Amino-acids/residues contact network: Graph (V, E)
- Features are descriptors: $x_v \in \mathbb{R}^{20}$ for $v \in V$

- Convolution $\mathbb{R}^2 \to$ Graph convolution
  - Graph $G = (V, E)$, $v \in \partial u$ are neighbors to u
  - Features $x : V \to \mathbb{R}^{d_0}$, k-th layer output $h^{(k)} : V \to \mathbb{R}^{d_k}$
  - $h_u^{(0)} = x_u$ for $u \in V$
  - $h_u^{(k+1)} = \sum_{v \in \partial u} f_\theta(h_v^{(k)}, h_u^{(k)})$, learned parameter θ



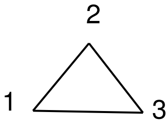
GNN image reproduced from [WPC+21]

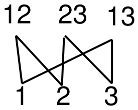
Extension: Sheaf Neural Networks

- More evolved data representation compared to GNN.
- Extension when data is heterogeneous and local consistent embedding through edges.
  - * Graph $G = (V, E)$
  - * $x_v \in F(v)$, $F(v)$ vector space for $v \in V$
  - * Edge embedding vector space $F(e)$ for $e \in E$
  - * Embedding linear functions: $F_e^v : F(v) \to F(e)$

Extension to partially ordered sets

- Limitations of 'Sheaf' representation over graphs:
  - $*$ Restricted version of sheaf: only over graphs
  - $*$ Probabilistic Methods in machine learning: **hierarchies**
  - $*$ Graphs are particular hierarchies (height 1).





- We proposed independently, in PhD thesis (Chapter 9 [SP21]), to represent data with local consistency properties:
  - $*$ $\to$ presheaves over a poset (abstraction of a hierarchy)
  - $*$ Posets more general than graph, presheaf over poset stronger modeling power

Probabilistic data processing: energy based modeling

- I collection of objects to model
- X_i random variable describes object $i \in I$
- Notation: $X_I = (X_i, i \in I)$, an event: $x_I = (x_i, i \in I)$
- Probability of an event x_I given by an energy function H_I:

$$\mathbb{P}_{X_I}(x_I) \cong e^{-H_I(x_I)}$$

- Some variables $a \subseteq I$ are observed, the rest $\overline{a}$ is not

$$\mathbb{P}_{X_a}(x_a) \cong \sum_{y_{\overline{a}}} e^{-H_I(x_a, y_{\overline{a}})}$$

Models of dependencies between variables: Graphical models

Example of structure:

- Dependencies between variables $\rightarrow$ Graphical Model
  - $*$ Graph $G = (V, E)$, V vertices, E edges
  - $*$ $V \leftarrow$ variables $(X_i, i = 1...n)$
  - $*$ $E \leftarrow$ modeled dependencies between variables (undirected)

$$V = (X_1, X_2, X_3)$$
$$E = \{(X_1, X_2), (X_2, X_3)\}$$

$$X_1 - -X_2 - -X_3$$

Hammersley–Clifford theorem
(e.g. see [SP19])

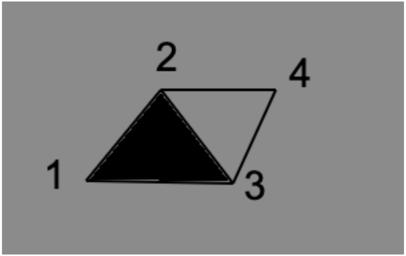
$$\mathbb{P}_{X_1,X_2,X_3} = f_{12}(X_1, X_2) f_{23}(X_2, X_3)$$

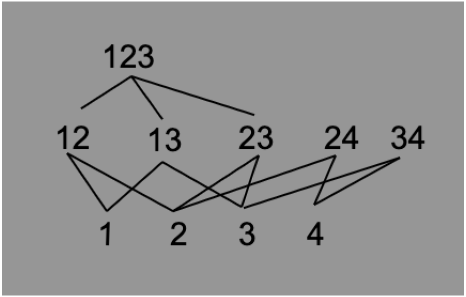
Example of Structure: dependencies between variables II

Inference on graphical models? $\to$ Bioinformatics

* Viterbi algorithm
* Em algorithm for HMM: Baum-Welch algorithm
* Forward-Backward algorithm $\rightsquigarrow$ Message Passing algorithms.
* Efficient variational inference $\to$ Belief Propagation

More general: hierarchies and factorization spaces





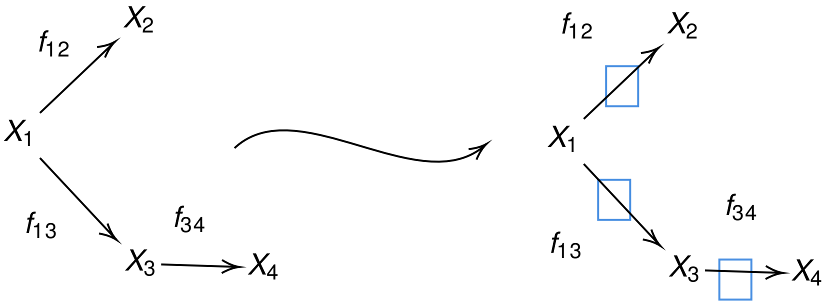
- $I = \{1, 2, 3, 4\}$
- $\mathbb{P}_{X_I}$ factors according to the simplicial complex (**hierarchy**):

$$\mathbb{P}_{X_I} = f_{123}(X_1, X_2, X_3) f_{24}(X_2, X_4) f_{34}(X_3, X_4)$$

Limitation of 'graphical models' & factorization spaces.

- Description of graphical model is global: the whole distribution factors accordingly to model.
  - Refers to global probability distribution: **local** $=$ **'parts'**, local features embedded in global probability space.
  - Somehow breaks the idea of 'compositionality': local probabilistic descriptors could correspond to **no** global model (not the case for graphical models)
- How to make description local?
  - No global description in GNN, SNN
  - Several answers

Local description of energy based models

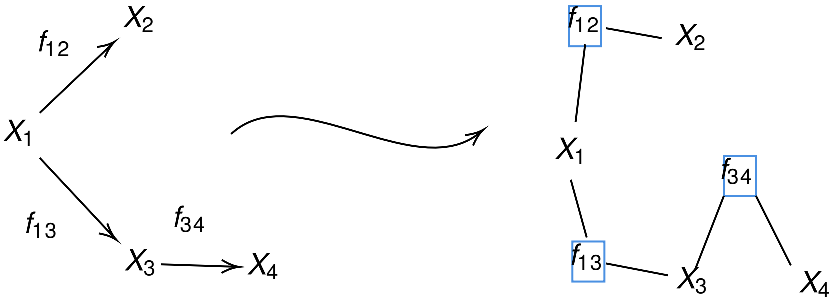


Transformation of graphical model to factor graph

Representation of graphical model I

Graphical Model

Factor graph : Bipartite

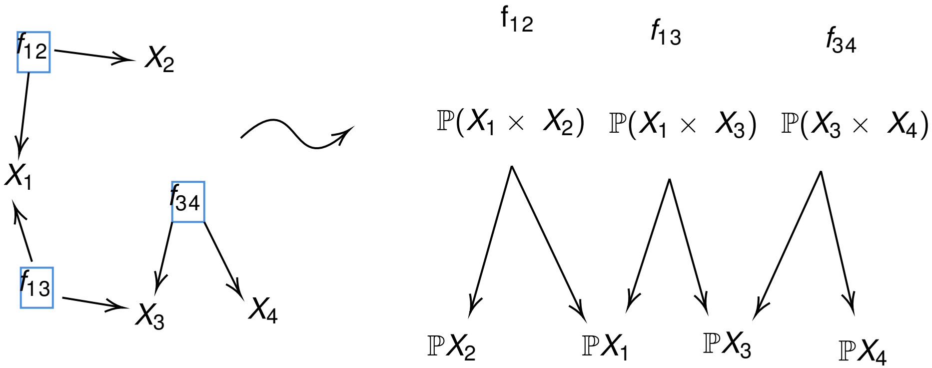


Transformation of Graphical Model to factor graph

Representation of graphical model II

Factor graph : choice of a direction

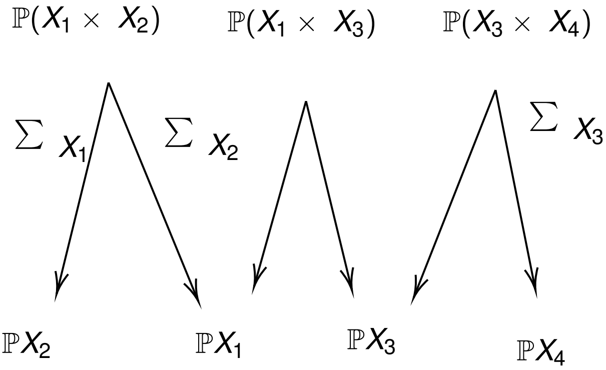
Enriched nodes



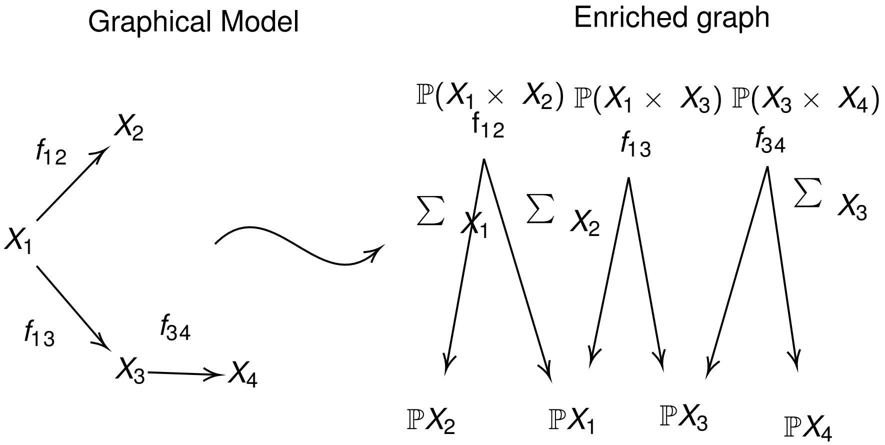
Transformation of factor graph to enriched graph

Representation of graphical model III

Enriched edges



Enriched graph



Transformation of Graphical Model to enriched graph

Local features of probabilitic model: Graph (Poset) of marginalizations

From enriched graph to a constrained space

- Each arrow is a constraint on ‘q’:

  $$\sum_{X_2} : \mathbb{P}(X_1 \times X_2) \to \mathbb{P}(X_1) \quad \longleftrightarrow \quad \sum_{y_2} q_{X_1,X_2}(x_1, y_2) = q_{X_1}(x_1)$$

- Replace $\mathbb{P}_{X_I}$ by local probabilities $(\mathbb{P}_{X_v}, \mathbb{P}_{X_e}, i \in V, e \in E)$
- This local version of graphical models relates to celebrated **Belief Propagation** [YFW05]

Other model: Graph (Poset) of Markov kernels (Conditioning) I

- A graph G is acyclic if there are no cycles: loops inside the graph
- Graphical models over acyclic graphs: decomposition into conditional distribution
  - $*$ Belief Network
  - $*$ Choose a directed version of the graph

$$X_1 \longrightarrow X_2 \longrightarrow X_3 \;\Big|\; \quad \mathbb{P}_{X_1,X_2,X_3} = \mathbb{P}_{23}(X_3|X_2)\mathbb{P}_{12}(X_2|X_1)\mathbb{P}_1(X_1)$$

- Markov Kernel $\pi : X \to Y$: generalization of $\mathbb{P}_{Y|X}$

$$\sum_{y \in Y} \pi(y|x) = 1$$

Other model: Poset of Markov kernels (Conditioning) II

- More generally, to an energy base model, $\mathscr{A} \subseteq \mathscr{P}(I)$,

$$H_I(x_I) = \sum_{a\in\mathscr{A}} H_a(x_a)$$

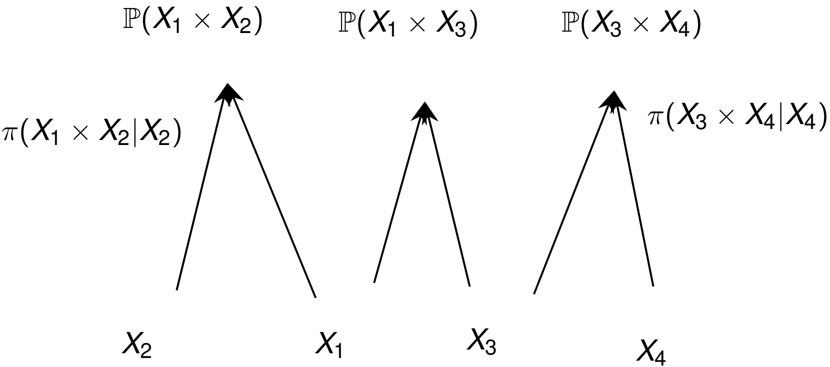
One can associate the (probability) kernels (Chapter 9 [SP21]): for $x_a \in X_a, x'_b \in X_b$

$$\pi^a_b(x_a|y_b) \cong \sum_{z:z_a=x_a} e^{-\sum\limits_{\substack{c\in\mathscr{A}\\ c\cap\bar{b}\neq\emptyset}} H_c(z_{c\cap\bar{b}},y_b)}$$

where $b \subseteq a$

Other model: Poset of Markov kernels (Conditioning) III

- As for graph (poset) of marginalizations, the poset of Markov kernels represents a local description of the energy based model (*noisy channel networks*).



Proposed Framework: Compositional Data

All previous structured data representation are of the following type:

Partially ordered set $\mathscr{A}$: a relation $\leq$ ($\subseteq \mathscr{A} \times \mathscr{A}$) such that,

1. $a \leq a$
2. (Transitivity) $b \leq a$ and $c \leq b$ then $c \leq a$
3. $b \leq a$ and $a \leq b$ then $a = b$

Functor G over a poset:

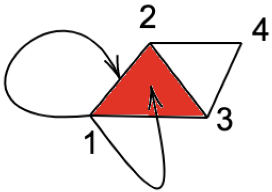
1. sends elements $a \in \mathscr{A}$ to a (vector) space $G(a)$
2. relations $b \leq a$ to (linear) morphisms between spaces

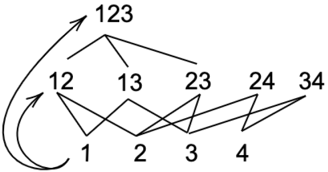
$$G_a^b : G(b) \to G(a)$$

3. Respects Transitivity:

$$G_a^b G_b^c = G_a^c$$

Illustration and remarks





- Presheaf: opposite relation on the poset ($G : \mathscr{A}^{op} \to \mathbf{Vect}$)
  - $b \leq_{op} a \iff b \geq a$
- Natural topology on $\mathscr{A}$
  - Alexandrov topology
  - 'Make' it a sheaf: sheafification

End of part I

Any Questions?

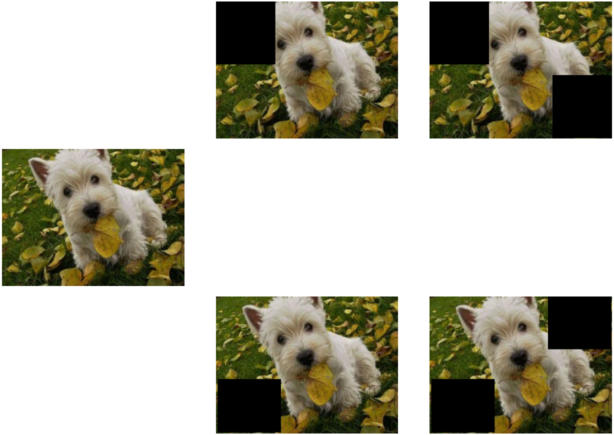
Part II: Optimization over compositional data

Problem

- Many optimization problem ’make sense’ at any place of the hierearchy (e.g. Regression, classification, MLE, MaxENT).
- How to define a loss on the whole structure (compositional data)?

Toy model (1)

Data with multiple point of view on it: for example cropped images of Dog.



Toy model (2)

Crop Cat images.

How to classify dogs and cats taking into account the extra data given by the different point of views?

Data: collection of images $(u_{i,j}, i \in \{0,1,2\}, j \in \{0,1\})$

$u_{1,0}$

$u_{1,1}$

$u_{0,0}$

$u_{2,0}$

$u_{2,1}$

Denote crop as C.

$u_{0,0}$

$u_{1,0}$

To go from $u_{0,0}$ to $u_{1,0}$,

$$u_{1,0} = C(left, top)[u_{0,0}]$$

Compatibility relations:

$$u_{0,0} = id[u_{0,0}]$$

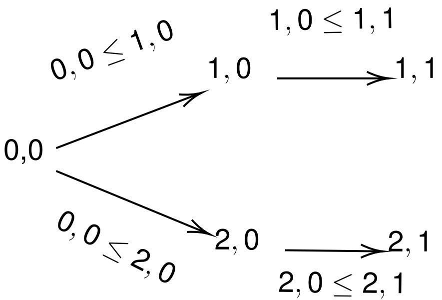
$$u_{1,0} = C(left, top)[u_{0,0}] \qquad u_{1,1} = C(r, b)[u_{1,0}]$$

$$u_{2,0} = C(left, bottom)[u_{0,0}] \qquad u_{2,1} = C(right, top)[u_{2,0}]$$

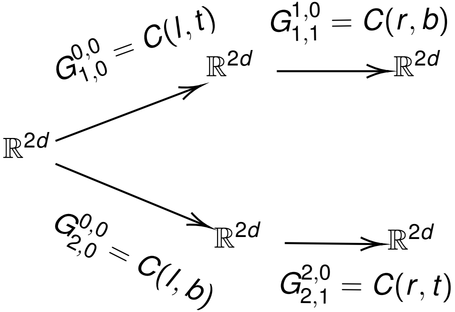
Compositional data (3)

Formally, compatibility relations are equivalent to saying that: $(u_{i,j}, i \in \{0,1,2\}, j \in \{0,1\})$ is a *section* of a *functor G* over a *partially ordered set* (poset) $\mathscr{A}$.

Poset $\mathscr{A}$:



Functor G:



where $\mathbb{R}^{2d}$ is the space in which the images live.

Compositional data (4)

Limit of a functor: ($\lim G$) set of collections $(u_a \in G(a), a \in \mathscr{A})$ that are compatible under the functor :

$$\forall b \leq a, \quad G_a^b(u_b) = u_a$$

- Implies compatibility of different points of view
- **Now:** Data is the limit of a functor over a poset.

Combinatorial Loss (1)

To classify cats or dogs over a dataset $D = [(x^i, y^i), i = 1..N]$ of size N:
Cross entropy

$$l(\theta) = \frac{1}{N} \sum_{i=1..N} \ln p_\theta(y^i | x^i)$$

where $y = 0$ for a cat and $y = 1$ for a dog.

Combinatorial Loss (2)

In our case there are multiple points of views on the images: the dataset is a collection of samples $[(x^i_{a(i)}, y^i), i = 1..N]$ over **different view points** $a \in \mathscr{A}$ where $a(i)$ is the view point on the image, recall that possible values are:

$$(0,0),(1,0),(1,1),(2,0),(2,1)$$

For example for the following sample

$$a(i) = (1,0)$$

Combinatorial Loss (3)

Dataset can be reorganized as collection of datasets $[(x_a^i, y^i), i = 1..N_a]$ for $a \in \mathscr{A}$.

The expression of the loss does not change with the point of view on the data,

$$l_{0,0} = \frac{1}{N_{0,0}} \sum_{i=1..N_{0,0}} \ln p_{\theta_{0,0}}(y^i | x^i_{0,0})$$

$$l_{1,0} = \frac{1}{N_{1,0}} \sum_{i=1..N_{1,0}} \ln p_{\theta_{1,0}}(y^i | x^i_{1,0})$$

For a given point of view *a*, the previous loss is simply the cross entropy for the dataset restricted to this point of view:

$$l_a(p_{\theta_a}) = \frac{1}{N_a} \sum_{i=1..N_a} \ln p_{\theta_a}(y^i | x_a^i)$$

- **Formally**, for each element of the poset $a \in \mathscr{A}$, we consider a collection of losses (functions) $l_a : G(a) \to \mathbb{R}$. We now call the points of view 'local'.

Combinatorial Loss (5)

Problem: How to optimize l_a for all points of view at the same time?
Answer?: Total loss is the sum of the losses? $l = \sum_{a \in \mathscr{A}} l_a$.

- Very redundant!
- $u \in \lim G$ is a 'global' reconstruction of 'local' points of view $u_a, a \in \mathscr{A}$ we want the loss to 'behave' the same way
- In our example, the non cropped image $u_{0,0}$ is enough to index the sections of G:

$$G \cong \mathbb{R}^{2d}$$

- However $l \neq l_{0,0}$. This loss **does not** behave well under 'global reconstruction'
- **NOT an answer**

Combinatorial Loss (6)

We follow the construction of Yedidia,Freeman, Weiss in the celebrated article *Constructing free-energy approximations and generalized belief propagation algorithms*[YFW05]. They use inclusion–exclusion principle to build an entropy on probability distribution compatible by marginalization.

- Good properties under 'global reconstruction' Proposition 2.2 [SP22]

Combinatorial Loss (7)

Inclusion–exclusion principle: simplest version for two set A, B then,

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Rota in his celebrated article *On the foundations of combinatorial theory I. Theory of Möbius functions* [Rot64], extended inclusion–exclusion to any poset by introducing Möbius inversion.

Combinatorial Loss (8)

Two important functions for poset $\mathscr{A}$:

- ζ funtion of the poset, for any $f \in \bigoplus_{a \in \mathscr{A}} \mathbb{Z}$,

$$\forall a \in \mathscr{A} \quad \zeta(f)(a) = \sum_{b \leq a} f(b)$$

- Its inverse (Proposition 2 [Rot64]), Möbius inversion μ,

$$\forall a \in \mathscr{A} \quad \mu(f)(a) := \sum_{b \leq a} \mu(a, b) f(b)$$

Combinatorial Loss (9)

Proposed global loss that we call ‘Combinatorial loss’:

for a functor F from $\mathscr{A}^{op}$ (the poset with inverse relation) to vector spaces, and $u = (u_a \in F(a), a \in \mathscr{A})$:

$$l(u) = \sum_{a\in\mathscr{A}} \sum_{b\leq a} \mu(a,b) l_b(u_b) \qquad \text{(CLoss)}$$

Optimization problem **Solve:**

$$\min_{u\in \lim F} l(u)$$

Combinatorial Loss (10)

The Combinatorial loss can be rewritten as,

$$l(u) = \sum_{a \in \mathscr{A}} c(a) l_a(u_a)$$

where $c(a) = \sum_{b \geq a} \mu(b, a)$.

In the inclusion-exclusion principle for two sets A, B, $c(A) = 1$, $c(B) = 1$, $c(A \cap B) = -1$.

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Critical points of Regionalized loss

When G is a functor from $\mathscr{A}$ to vector spaces, the collection of dual maps

$$G_a^{b^*} : G(a)^* \to G(b)^*$$

defines a functor from $\mathscr{A}^{op}$ to vector spaces denoted as G^*

Theorem (GSP)

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

$$d_u l_a = \sum_{b\leq a} F_b^{a*} \left(\sum_{c\leq b} F_c^{b^*} m_{b\to c} - \sum_{c\geq b} m_{c\to b} \right) \tag{CP}$$

Message passing algorithms (1)

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.

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

Messages:

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

Auxiliary variables,

$$A(t) \in \bigoplus_{a \in \mathscr{A}} F(a)^*$$

Message passing algorithms (2)

For any $a, b \in \mathscr{A}$ such that $b \leq a$, the update rule is given by,

$$A_a(t) = \sum_{b:b\leq a} \sum_{c:b\geq c} F_c^{a*} m_{b\to c}(t) - \sum_{b:b\leq a} \sum_{c:c\geq b} F_b^{a*} m_{c\to b}(t)$$

$$m_{a\to b}(t+1) = m_{a\to b}(t) + F_b^a g_a(A_a(t)) - g_b(A_b(t)) \tag{MSP}$$

Fix points of MSP $\leftrightarrow$ Critical points CP

Theorem (GSP)

Fix points of message passing algorithm (MSP) are critical points of 'global' Combinatorial loss: if $MSP(m^*) = m^*$, then let $\forall a \in \mathscr{A}$,

$$u_a^* = g_a\left[\sum_{b\leq a} F_b^{a*}\left(\sum_{c\leq b} F_c^{b*} m_{b\to c}^* - \sum_{c\geq b} m_{c\to b}^*\right)\right]$$

Then u^* satisfies (CP).

Extends previous result of Yedidia, Freeman, Weiss, Peltre (Theorem 5 [YFW05], Theorem 5.15 [Pel20]) stating that:

Fix points of General Belief Propagation $\leftrightarrow$ critical points of Region based approximation of free energy.

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

To go further (2)

<u>Understanding expression of critical points:</u>

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

$$\mu_{F^*} d_u I|_{\lim F} = 0$$

To go further (3)

Understanding expression of critical points:

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

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 $\ker \delta = \lim F$.

To go further (4)

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 $a \in \mathscr{A}$,

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

To go further (5)

Rewriting condition on fix points:

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

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,

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

To go further (6)

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 $b \leq a$,

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

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

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

To go further (7)

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!

Example of applications

- Extension of General Belief Propagation to noisy channel networks
- PCA for filtered data like time series

Thank you very much for your attention

Thank you very much for your attention!

References I

- Kanchan Jha, Sriparna Saha, and Hiteshi Singh, *Prediction of protein–protein interaction using graph neural networks*, Scientific Reports (2022).
- Olivier Peltre, *Homology of Message-Passing Algorithms*, `http://opeltre.github.io`, 2020, Ph.D. thesis (preprint).
- Gian-Carlo Rota, *On the foundations of combinatorial theory I. Theory of Möbius functions*, Probability theory and related fields **2** (1964), no. 4, 340–368.
- Grégoire Sergeant-Perthuis, *Intersection property and interaction decomposition*, arXiv:1904.09017v2, 2019.
- Grégoire Sergeant-Perthuis, *Intersection property, interaction decomposition, regionalized optimization and applications*, Ph.D. thesis, Université de Paris, 2021.

References II

- Grégoire Sergeant-Perthuis, *Regionalized optimization*, arXiv:2201.11876v1, 2022.
- Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu, *A comprehensive survey on graph neural networks*, IEEE Transactions on Neural Networks and Learning Systems **32** (2021), no. 1, 4–24.
- Jonathan S Yedidia, William T Freeman, and Yair Weiss, *Constructing free-energy approximations and generalized belief propagation algorithms*, IEEE Transactions on information theory **51** (2005), no. 7, 2282–2312.