Group Meeting - August 07, 2020

Paper review & Research progress

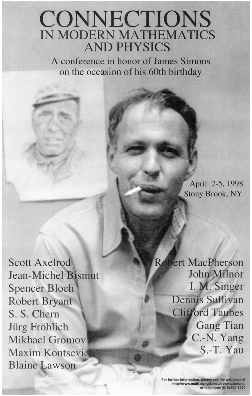
Truong Son Hy *

*Department of Computer Science
The University of Chicago

Ryerson Physical Lab

James Simons

I wasn't the fastest guy in the world. I wouldn't have done well in an Olympiad or a math contest. But I like to ponder. And pondering things, just sort of thinking about it and thinking about it, turns out to be a pretty good approach.



What I realize for myself

1. I love Science and Research.
2. I am no longer into Silicon valley, I don't want to work in the industry.
3. Reading scientific papers is more meaningful than working in a tech company.
4. Prof. Jordan Peterson: 'The pursuit of happiness is a pointless goal. We must instead search for meaning, not for its own sake, but as a defense against the suffering that is intrinsic to our existence.'

Paper

Correlated Variational Auto-Encoders, Da Tang, Dawen Liang, Tony Jebara, Nicholas Ruozzi (ICML 2019)

Standard VAEs (1)

Input data $\boldsymbol{x} = \{\boldsymbol{x}_1, .., \boldsymbol{x}_n\} \subseteq \mathrm{R}^D$. Standard VAEs assume that each data point $\boldsymbol{x}_i$ is generated independently by the following process:

1. Genearte the latent variables $\boldsymbol{z} = \{\boldsymbol{z}_1, .., \boldsymbol{z}_n\} \subseteq \mathrm{R}^d$ $(d << D)$ by drawing i.i.d from the prior distribution (e.g. standard Gaussian distribution) $\boldsymbol{z}_i \sim p_0(\boldsymbol{z}_i)$, for each i.
2. Generate the data points $\boldsymbol{x}_i \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}_i)$ from the model conditional distribution $p_{\boldsymbol{\theta}}$ independently.

Standard VAEs (2)

Optimizing $\boldsymbol{\theta}$ to maximize the likelihood $p_{\boldsymbol{\theta}}$ requires computing **intractable** posterior distribution

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{z}_i|\mathbf{x}_i)$$

VAEs approximates this posterior distribution as $q_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x}) = \prod_{i=1}^{n} q_{\boldsymbol{\lambda}}(\mathbf{z}_i|\mathbf{x}_i)$ via amortized inference and maximize the evidence lower bound (ELBO):

$$L(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \mathrm{E}_{q_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \mathcal{D}(q_{\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{x})||p_0(\mathbf{z}))$$

$$= \sum_{i=1}^{n} \left[\mathrm{E}_{q_{\boldsymbol{\lambda}}(\mathbf{z}_i|\mathbf{x}_i)}[\log p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)] - \mathcal{D}(q_{\boldsymbol{\lambda}}(\mathbf{z}_i|\mathbf{x}_i)||p_0(\mathbf{z}_i)) \right]$$

Correlated priors on acyclic graphs (1)

Input data $\boldsymbol{x} = \{\boldsymbol{x}_1, .., \boldsymbol{x}_n\}$ with correlation structure given by an undirected graph $G = (V, E)$ in which $V = v_1, .., v_n$ is the set of vertices corresponding to each data point, and $(v_i, v_j) \in E$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are correlated. Define prior distribution p_0^{corr} of the latent variables $\boldsymbol{z}_1, .., \boldsymbol{z}_n$ over $(\boldsymbol{z}_1, .., \boldsymbol{z}_n) \in \mathrm{R}^d \times .. \times \mathrm{R}^d$ whose singleton and pairwise marginal distributions satisfying:

$$p_0^{corr}(\boldsymbol{z}_i) = p_0(\boldsymbol{z}_i) \qquad \forall v_i \in V$$

$$p_0^{corr}(\boldsymbol{z}_i, \boldsymbol{z}_j) = p_0(\boldsymbol{z}_i, \boldsymbol{z}_j) \qquad \forall (v_i, v_j) \in E$$

Symmetry and marginalization consistency properties:

$$p_0(\boldsymbol{z}_i, \boldsymbol{z}_j) = p(\boldsymbol{z}_j, \boldsymbol{z}_i) \qquad \forall \boldsymbol{z}_i, \boldsymbol{z}_j \in \mathrm{R}^d$$

$$\int p_0(\boldsymbol{z}_i, \boldsymbol{z}_j) d\boldsymbol{z}_j = p_0(\boldsymbol{z}_i) \qquad \forall \boldsymbol{z}_i \in \mathrm{R}^d$$

Correlated priors on acyclic graphs (2)

The generative process of a CVAE:

1. Sample $\boldsymbol{z}$ from the prior p_0^{corr}.
2. Sample each data point $\boldsymbol{x}_i$ conditionally independently from $\boldsymbol{z}_i$.

Wainwright & Jordan, 2008:

$$p_0^{corr}(\boldsymbol{z}) = \prod_{i=1}^{n} p_0(\boldsymbol{z}_i) \prod_{(v_i,v_j)\in E} \frac{p_0(\boldsymbol{z}_i, \boldsymbol{z}_j)}{p_0(\boldsymbol{z}_i)p_0(\boldsymbol{z}_j)}$$

The prior is factorized as singletons and pairwise marginal distributions.

Variational families

Singleton variational family: Approximate the posterior distribution $p((z)|\boldsymbol{x})$ as

$$q_{\lambda}(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{n} q_{\lambda}(\boldsymbol{z}_i|\boldsymbol{x}_i)$$

Correlated variational family: $q_{\lambda}(\boldsymbol{z}|\boldsymbol{x})$ is factorized as singleton and pairwise marginal distributions

$$q_{\lambda}(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{n} q_{\lambda}(\boldsymbol{z}_i|\boldsymbol{x}_i) \prod_{(v_i,v_j)\in E} \frac{q_{\lambda}(\boldsymbol{z}_i, \boldsymbol{z}_j|\boldsymbol{x}_i, \boldsymbol{x}_j)}{q_{\lambda}(\boldsymbol{z}_i|\boldsymbol{x}_i) q_{\lambda}(\boldsymbol{z}_j|\boldsymbol{x}_j)}$$

On general graphs (1)

Trivial generalization fails

For a general graph G:

$$p_0^{corr}(\boldsymbol{z}) = \prod_{i=1}^{n} p_0(\boldsymbol{z}_i) \prod_{(v_i,v_j)\in E} \frac{p_0(\boldsymbol{z}_i,\boldsymbol{z}_j)}{p_0(\boldsymbol{z}_i)p_0(\boldsymbol{z}_j)}$$

is not guaranteed to be a valid distribution.

On general graphs (1)

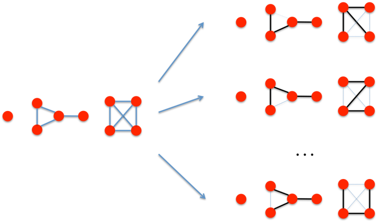
Trivial generalization fails

For a general graph G:

$$p_0^{corr}(\mathbf{z}) = \prod_{i=1}^{n} p_0(\mathbf{z}_i) \prod_{(v_i,v_j)\in E} \frac{p_0(\mathbf{z}_i, \mathbf{z}_j)}{p_0(\mathbf{z}_i)p_0(\mathbf{z}_j)}$$

is not guaranteed to be a valid distribution.

Solution: Factorized a general graph as a set of maximal acylic graphs



On general graphs (2)

Maximal acyclic subgraph

For an undirected graph $G = (V, E)$, a subgraph $G' = (V', E')$ is a maximal acyclic subgraph of G if:

1. G' is acylic.
2. $V' = V$, i.e., G' contains all vertices of G.
3. Adding any edge from E/E' to E' will create a cycle in G'.

On general graphs (2)

Maximal acyclic subgraph

For an undirected graph $G = (V, E)$, a subgraph $G' = (V', E')$ is a maximal acyclic subgraph of G if:

1. G' is acylic.
2. $V' = V$, i.e., G' contains all vertices of G.
3. Adding any edge from E/E' to E' will create a cycle in G'.

- If G is connected, G' is a spanning tree of G.
- Otherwise, a spanning forest.

On general graphs (3)

New prior distribution of $\boldsymbol{z}$ as a uniform mixture over all subgraphs in $\mathcal{A}_G$:

$$p_0^{corr_g} = \frac{1}{|\mathcal{A}_G|} \sum_{G'=(V,E')\in\mathcal{A}_G} p_0^{G'}(\boldsymbol{z})$$

where

$$p_0^{G'}(\boldsymbol{z}) = \prod_{i=1}^{n} p_0(\boldsymbol{z}_i) \prod_{(v_i,v_j)\in E'} \frac{p_0(\boldsymbol{z}_i, \boldsymbol{z}_j)}{p_0(\boldsymbol{z}_i)p_0(\boldsymbol{z}_j)}$$

and $\mathcal{A}_G$ is the set of all spanning forests of G.

On general graphs (4)

Log-likelihood:

$$\log p_{\theta}(\boldsymbol{x}) = \mathrm{E}_{p_0^{corrg}(\boldsymbol{z})}[\log p_{\theta}(\boldsymbol{x}|\boldsymbol{z})] = \frac{1}{|\mathcal{A}_G|}\sum_{G'\in\mathcal{A}_G}\mathrm{E}_{p_0^{G'}(\boldsymbol{z})}[\log p_{\theta}(\boldsymbol{x}|\boldsymbol{z})]$$

$$\geq \frac{1}{|\mathcal{A}_G|}\sum_{G'\in\mathcal{A}_G}\left(\mathrm{E}_{q_{\boldsymbol{\lambda}}^{G'}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\theta}(\boldsymbol{x}|\boldsymbol{z})] - \mathcal{D}(q_{\boldsymbol{\lambda}}^{G'}(\boldsymbol{z}|\boldsymbol{x})||p_0^{G'}(\boldsymbol{z}))\right)$$

where

$$q_{\boldsymbol{\lambda}}^{G'}(\boldsymbol{z}|\boldsymbol{x}) = \prod_{i=1}^{n} q_{\boldsymbol{\lambda}}(\boldsymbol{z}_i|\boldsymbol{x}_i)\prod_{(v_i,v_j)\in E'}\frac{q_{\boldsymbol{\lambda}}(\boldsymbol{z}_i,\boldsymbol{z}_j|\boldsymbol{x}_i,\boldsymbol{x}_j)}{q_{\boldsymbol{\lambda}}(\boldsymbol{z}_i|\boldsymbol{x}_i)q_{\boldsymbol{\lambda}}(\boldsymbol{z}_j|\boldsymbol{x}_j)}$$

On general graphs (5)

1. Singleton terms: all vertices have the same weight.
2. Pairwise terms: an edge e's weight is the fraction of times e appears among all subgraphs in $\mathcal{A}_G$.

$$w_{G,e}^{MAS} = \frac{|\{G' \in \mathcal{A}_G : e \in G'\}|}{|\mathcal{A}_G|}$$

Sum of all edge weights:

$$\sum_{e \in E} w_{G,e}^{MAS} = |V| - |CC(G)|$$

where $CC(G)$ is the set of connected components of G.

For a complete graph K_n, the weight $w_{K_n,e}^{MAS}$ for any edge e of K_n is $2/n$.

On general graphs (6)

New lower bound of the log-likelihood:

$$\sum_{i=1}^{n} \Big(\mathbb{E}_{q_{\boldsymbol{\lambda}}}(\boldsymbol{z}_i|\boldsymbol{x}_i) \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_i|\boldsymbol{z}_i)\right] - \mathrm{KL}(q_{\boldsymbol{\lambda}}(\boldsymbol{z}_i|\boldsymbol{x}_i)||p_0(\boldsymbol{z}_i)) \Big)$$
$$- \sum_{(v_i,v_j)\in E} w^{\mathrm{MAS}}_{G,(v_i,v_j)} \Big(\mathrm{KL}(q_{\boldsymbol{\lambda}}(\boldsymbol{z}_i, \boldsymbol{z}_j|\boldsymbol{x}_i, \boldsymbol{x}_j)||p_0(\boldsymbol{z}_i, \boldsymbol{z}_j))$$
$$- \mathrm{KL}(q_{\boldsymbol{\lambda}}(\boldsymbol{z}_i|\boldsymbol{x}_i)||p_0(\boldsymbol{z}_i)) - \mathrm{KL}(q_{\boldsymbol{\lambda}}(\boldsymbol{z}_j|\boldsymbol{x}_j)||p_0(\boldsymbol{z}_j)) \Big)$$

Some graph theory (1)

Matrix Tree Theorem (Chaiken & Kleitman, 1978)

For an undirected graph $G = (V, E)$, the number of spanning trees of G is the determinant of the sub-matrix of the Laplacian matrix L of G after deleting the i-th row and the i-th column, for any $i = 1, .., n$.

$$L_{i,j} = \begin{cases} \text{degree}(v_i) & \text{if } i = j, \\ -1 & \text{if } (v_i, v_j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Some graph theory (2)

Number of spanning trees containing a particular edge

For an undirected graph $G = (V, E)$ and an edge $(v_i, v_j) \in E$, the number of spanning trees of G containing this edge is the determinant of the sub-matrix of the Laplacian matrix L of G after deleting the i-th, j-th rows and the i-th, j-th columns of it.

Complexity: $O(|E||V|^3)$ that is inefficient!

Some graph theory (3)

Smarter way

For an undirected connected graph $G = (V, E)$ and an edge $e = (v_i, v_j) \in E$, the weight $w_{G,e}^{MAS} = L^+_{i,i} - L^+_{i,j} - L^+_{j,i} + L^+_{j,j}$ where L^+ is the **Moore-Penrose pseudo-inverse** of the Laplacian matrix L of G.

Complexity: $O(|V|^3)$

Algorithm to compute edge weights

Algorithm 1 Computing all weights $w_{G,e}^{\text{MAS}}$

Input: undirected graph $G = (V = \{v_1, \ldots, v_n\}, E)$.
Compute all the connected components $\text{CC}_1, \ldots, \text{CC}_K$ of G using depth-first search or breadth-first search.
for $k = 1$ **to** K **do**
    Compute the Moore-Penrose inverse L_k^+ of the Laplacian matrix L_k of the component CC_k.
    Apply Theorem 3 to compute $w_{G,e}^{\text{MAS}}$ for each edge e in the component CC_k.
end for
Return The weights $w_{G,e}^{\text{MAS}}$ for all $e \in E$.

Baselines and applications

Baselines:

1. Standard VAEs
2. GraphSAGE (Hamilton et al., 2017)

Experiments:

1. Bipartite correlation graph: MovieLens 20M dataset.
2. Perform spectral clustering on a synthetic dataset with a tree-structured latent variable graphical model.
3. Link prediction: Epinions dataset (Massa & Avesani, 2007).

Extension of this work: Markov Random Fields

Hammersley-Clifford theorem

A positive distribution $p(\boldsymbol{z}) > 0$ satisfies the CI (conditional independent) properties of an undirected graph G iff p can be represented as a product of factors, one per **maximal clique**, i.e.,

$$p(\boldsymbol{z}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in \mathcal{C}} \psi_c(\boldsymbol{z}_c|\boldsymbol{\theta}_c)$$

where $\mathcal{C}$ is the set of all the (maximal) cliques of G, and $Z(\boldsymbol{\theta})$ is the **partition function** given by:

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} \prod_{c \in \mathcal{C}} \psi_c(\boldsymbol{z}_c|\boldsymbol{\theta}_c)$$

Note that the partition function is to ensure the overall distribution sums to 1.