Collapsed Variational Inference for Sum-Product Networks

Han Zhao[1], Tameem Adel[2], Geoff Gordon[1], Brandon Amos[1]
Presented by: Han Zhao

Carnegie Mellon University[1], University of Amsterdam[2]

June. 20th, 2016

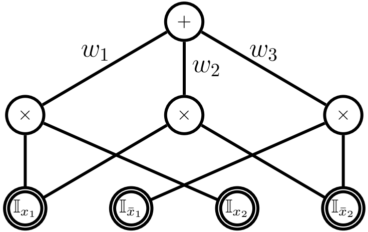
Outline

Sum-Product Networks

Definition

A Sum-Product Network (SPN) is a

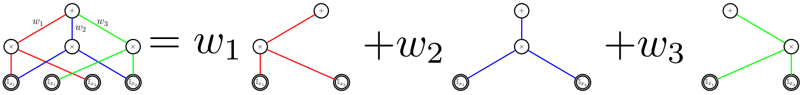
- Rooted directed acyclic graph of univariate distributions, sum nodes and product nodes.
- Value of a product node is the product of its children.
- Value of a sum node is the weighted sum of its children, where the weights are nonnegative.
- Value of the network is the value at the root.



Sum-Product Networks

Mixture of Trees

Each SPN can be decomposed as a mixture of trees:



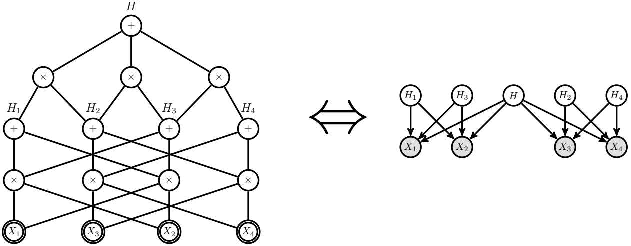
- Each tree is a product of univariate distributions.
- Number of mixture components is $\Omega(2^{\text{Depth}})$.
- Each network computes a positive polynomial (posynomial) function of model parameters:

$$V_{\text{root}}(\mathbf{x} \mid \mathbf{w}) = \sum_{t=1}^{\tau_{\mathcal{S}}} \prod_{(k,j)\in\mathcal{T}_{tE}} w_{kj} \prod_{i=1}^{n} p_t(X_i = \mathbf{x}_i)$$

Sum-Product Networks

Bayesian Network

Alternatively, each SPN $\mathcal{S}$ is equivalent to a Bayesian network $\mathcal{B}$ with bipartite structure.



- Number of sum nodes in $\mathcal{S}$ = Number of hidden variables in $\mathcal{B} = \Theta(|\mathcal{S}|)$. $|\mathcal{B}| = O(n|\mathcal{S}|)$
- Number of observable variables in $\mathcal{B}$ = Number of variables modeled by $\mathcal{S}$.
- Typically number of hidden variables $\gg$ number of observable variables.

Bayesian Inference:

$$\underbrace{p(\mathbf{w} \mid \mathbf{x})}_{\text{posterior}} \propto \underbrace{p(\mathbf{w})}_{\text{prior}} \; \underbrace{p(\mathbf{x} \mid \mathbf{w})}_{\text{likelihood}}$$

Often intractable because of:

- No analytical solution.
- Expensive numerical integration.

General idea: find the best approximation in a tractable family of distributions Q:

$$\text{minimize}_{q \in Q} \quad \mathrm{KL}[q(\mathbf{w}) \mid\mid p(\mathbf{w} \mid \mathbf{x})]$$

Typical choice of approximation families: Mean-field, structured mean-field, etc.

Variational method: Optimization-based, deterministic approach for approximate Bayesian inference.

$$\inf_{q \in Q} \mathrm{KL}[q(\mathbf{w}) \,||\, p(\mathbf{w} \mid \mathbf{x})] \Leftrightarrow \sup_{q \in Q} \mathbb{E}_q[\log p(\mathbf{w}, \mathbf{x})] + \mathbb{H}[q]$$

Evidence Lower Bound $\widehat{\mathcal{L}}$:

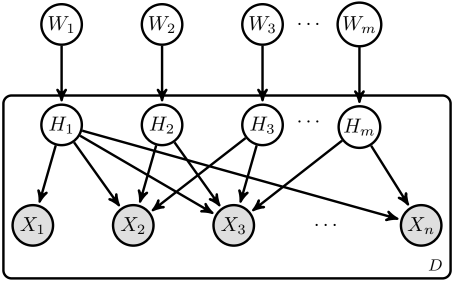
$$\log p(\mathbf{x}) \geq \sup_{q \in Q} \mathbb{E}_q[\log p(\mathbf{w}, \mathbf{x})] + \mathbb{H}[q] =: \widehat{\mathcal{L}}$$

Bayesian inference algorithms for SPNs:

- Flexible at incorporating prior knowledge about the structure of SPNs.
- More robust to overfitting.

Collapsed Variational Inference

Motivations and Challenges



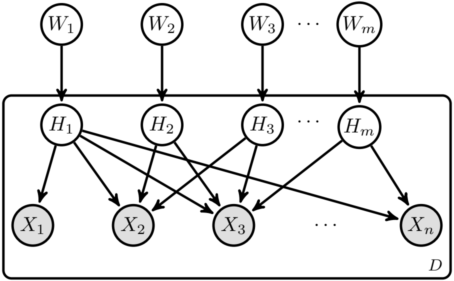
- W – Model parameters, global hidden variables.
- H – Assignments of sum nodes, local hidden variables.
- X – Observable variables.
- D – Number of instances.

Challenges for standard VB:

- Large number of local hidden variables: number of local hidden variables = Number of sum nodes = $\Theta(|\mathcal{S}|)$.

Collapsed Variational Inference

Motivations and Challenges



- W – Model parameters, global hidden variables.
- H – Assignments of sum nodes, local hidden variables.
- X – Observable variables.
- D – Number of instances.

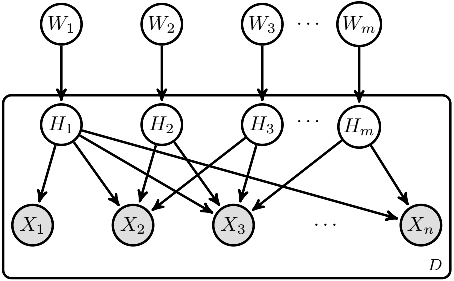
Challenges for standard VB:

- Large number of local hidden variables: number of local hidden variables = Number of sum nodes = $\Theta(|\mathcal{S}|)$.
- Memory overhead: space complexity $O(D|\mathcal{S}|)$.

Carnegie Mellon University

Collapsed Variational Inference

Motivations and Challenges



- W – Model parameters, global hidden variables.
- H – Assignments of sum nodes, local hidden variables.
- X – Observable variables.
- D – Number of instances.

Challenges for standard VB:

- Large number of local hidden variables: number of local hidden variables = Number of sum nodes = $\Theta(|\mathcal{S}|)$.
- Memory overhead: space complexity $O(D|\mathcal{S}|)$.
- Time complexity: $O(nD|\mathcal{S}|)$.

Our contributions:

- We obtain better ELBO $\mathcal{L}$ to optimize than $\widehat{\mathcal{L}}$, the one obtained by mean-field.
- Reduced space complexity: $O(D|\mathcal{S}|) \Rightarrow O(|\mathcal{S}|)$, space complexity is independent of training size.
- Reduced time complexity: $O(nD|\mathcal{S}|) \Rightarrow O(D|\mathcal{S}|)$, removing the explicit dependency on the dimension.

Collapsed Variational Inference

Efficient Marginalization

Recall ELBO in standard VI:

$$\widehat{\mathcal{L}} := \mathbb{E}_{q(\mathbf{w},\mathbf{h})}[\log p(\mathbf{w},\mathbf{h},\mathbf{x})] + \mathbb{H}[q(\mathbf{w},\mathbf{h})]$$

Consider the new ELBO in Collapsed VI:

$$\begin{aligned}\mathcal{L} :=& \mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{w},\mathbf{x})] + \mathbb{H}[q(\mathbf{w})] \\ =& \mathbb{E}_{q(\mathbf{w})}[\log \sum_{\mathbf{h}} p(\mathbf{w},\mathbf{h},\mathbf{x})] + \mathbb{H}[q(\mathbf{w})]\end{aligned}$$

We can establish the following inequality:

$$\log p(\mathbf{x}) \geq \mathcal{L} \geq \widehat{\mathcal{L}}$$

The new ELBO in Collapsed VI leads to a better lower bound than the one used in standard VI!

Standard Variational Inference

Mean-field assumption: $q(\mathbf{w}, \mathbf{h}) = \prod_i q(w_i) \prod_j q(h_j)$

ELBO: $\widehat{\mathcal{L}} := \mathbb{E}_{q(\mathbf{w},\mathbf{h})}[\log p(\mathbf{w}, \mathbf{h}, \mathbf{x})] + \mathbb{H}[q(\mathbf{w}, \mathbf{h})]$

Collapsed Variational Inference for LDA, HDP

Collapsed out global hidden variables: $q(\mathbf{h}) = \int_{\mathbf{w}} q(\mathbf{w}, \mathbf{h}) \; d\mathbf{w}$

ELBO: $\mathcal{L}_{\mathbf{h}} := \mathbb{E}_{q(\mathbf{h})}[\log p(\mathbf{h}, \mathbf{x})] + \mathbb{H}[q(\mathbf{h})]$

Better lower bound: $\mathcal{L}_{\mathbf{h}} \geq \widehat{\mathcal{L}}$

Collapsed Variational Inference for SPN

Collapsed out local hidden variables: $q(\mathbf{w}) = \sum_{\mathbf{h}} q(\mathbf{w}, \mathbf{h})$

ELBO: $\mathcal{L}_{\mathbf{w}} := \mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{w}, \mathbf{x})] + \mathbb{H}[q(\mathbf{w})]$

Better lower bound: $\mathcal{L}_{\mathbf{w}} \geq \widehat{\mathcal{L}}$

Time complexity of the exact marginalization incurred in computing $\sum_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}, \mathbf{x})$:

- Time complexity of marginalization in graphical model $\mathcal{G}$: $O(D \cdot 2^{\mathrm{tw}(\mathcal{G})})$.

Space complexity reduction:

Collapsed Variational Inference

Efficient Marginalization

Time complexity of the exact marginalization incurred in computing $\sum_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}, \mathbf{x})$:

- Time complexity of marginalization in graphical model $\mathcal{G}$: $O(D \cdot 2^{\text{tw}(\mathcal{G})})$.
- Exact marginalization in BN $\mathcal{B}$ with algebraic decision diagram as local factors: $O(D|\mathcal{B}|) = O(nD|\mathcal{S}|)$.

Space complexity reduction:

Collapsed Variational Inference

Efficient Marginalization

Time complexity of the exact marginalization incurred in computing $\sum_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}, \mathbf{x})$:

- Time complexity of marginalization in graphical model $\mathcal{G}$: $O(D \cdot 2^{\mathrm{tw}(\mathcal{G})})$.
- Exact marginalization in BN $\mathcal{B}$ with algebraic decision diagram as local factors: $O(D|\mathcal{B}|) = O(nD|\mathcal{S}|)$.
- Exact marginalization in SPN $\mathcal{S}$: $O(D|\mathcal{S}|)$.

Space complexity reduction:

Time complexity of the exact marginalization incurred in computing $\sum_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}, \mathbf{x})$:

- Time complexity of marginalization in graphical model $\mathcal{G}$: $O(D \cdot 2^{\text{tw}(\mathcal{G})})$.
- Exact marginalization in BN $\mathcal{B}$ with algebraic decision diagram as local factors: $O(D|\mathcal{B}|) = O(nD|\mathcal{S}|)$.
- Exact marginalization in SPN $\mathcal{S}$: $O(D|\mathcal{S}|)$.

Space complexity reduction:

- No posterior over $\mathbf{h}$ to approximate anymore.

Collapsed Variational Inference

Efficient Marginalization

Time complexity of the exact marginalization incurred in computing $\sum_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}, \mathbf{x})$:

- Time complexity of marginalization in graphical model $\mathcal{G}$: $O(D \cdot 2^{\mathrm{tw}(\mathcal{G})})$.
- Exact marginalization in BN $\mathcal{B}$ with algebraic decision diagram as local factors: $O(D|\mathcal{B}|) = O(nD|\mathcal{S}|)$.
- Exact marginalization in SPN $\mathcal{S}$: $O(D|\mathcal{S}|)$.

Space complexity reduction:

- No posterior over $\mathbf{h}$ to approximate anymore.
- No variational variables over $\mathbf{h}$ needed: $O(D|\mathcal{S}|) \Rightarrow O(|\mathcal{S}|)$.

Collapsed Variational Inference

Logarithmic Transformation

New optimization objective:

$$\text{maximize}_{q \in Q} \quad \mathbb{E}_{q(\mathbf{w})}[\log \sum_{\mathbf{h}} p(\mathbf{w}, \mathbf{h}, \mathbf{x})] + \mathbb{H}[q(\mathbf{w})]$$

which is equivalent to

$$\text{minimize}_{q \in Q} \quad \mathrm{KL}[q(\mathbf{w}) \mid\mid p(\mathbf{w})] - \mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{x} \mid \mathbf{w})]$$

- $p(\mathbf{w})$ – prior distribution over $\mathbf{w}$, product of Dirichlets.
- $q(\mathbf{w})$ – variational posterior over $\mathbf{w}$, product of Dirichlets.
- $p(\mathbf{x} \mid \mathbf{w})$ – likelihood, not multinomial anymore after marginalization.

Non-conjugate $q(\mathbf{w})$ and $p(\mathbf{x} \mid \mathbf{w})$, no analytical solution for $\mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{x} \mid \mathbf{w})]$.

Key observation:

$$p(\mathbf{x} \mid \mathbf{w}) = V_{\text{root}}(\mathbf{x} \mid \mathbf{w}) = \sum_{t=1}^{\tau_{\mathcal{S}}} \prod_{(k,j)\in\mathcal{T}_{tE}} w_{kj} \prod_{i=1}^{n} p_t(X_i = \mathbf{x}_i)$$

is a posynomial function of $\mathbf{w}$.

Make a bijective mapping (change of variable): $\mathbf{w}' = \log(\mathbf{w})$.

- Dates back to the literature of geometric programming.
- The new objective after transformation is convex in $\mathbf{w}'$.

$$\log p(\mathbf{x} \mid \mathbf{w}) = \log \left(\sum_{t=1}^{\tau_{\mathcal{S}}} \exp \left(c_t + \sum_{(k,j)\in\mathcal{T}_{tE}} w'_{kj} \right) \right)$$

Jensen's inequality to obtain further lower bound.

Collapsed Variational Inference

Logarithmic Transformation

Further lower bound:

$$\mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{x} \mid \mathbf{w})] = \mathbb{E}_{q(\mathbf{w}')}[\log p(\mathbf{x} \mid \mathbf{w}')] \geq \log p(\mathbf{x} \mid \mathbb{E}_{q'(\mathbf{w}')}[\mathbf{w}'])$$

Relaxed objective:

$$\text{minimize}_{q \in Q} \quad \underbrace{\mathrm{KL}[q(\mathbf{w}) \mid\mid p(\mathbf{w})]}_{\text{Regularity}} \underbrace{- \log p(\mathbf{x} \mid \mathbb{E}_{q'(\mathbf{w}')}[\mathbf{w}'])}_{\text{Data fitting}}$$

Roughly, $\log p(\mathbf{x} \mid \mathbb{E}_{q'(\mathbf{w}')}[\mathbf{w}'])$ corresponds the log-likelihood by setting the weights of SPN as the posterior mean of $q(\mathbf{w})$.

Optimized by projected GD.

Collapsed Variational Inference

Algorithm

Algorithm 1 CVB-SPN

Input: Initial $\boldsymbol{\beta}$, prior hyperparameter $\boldsymbol{\alpha}$, training instances $\{\mathbf{x}_d\}_{d=1}^{D}$.

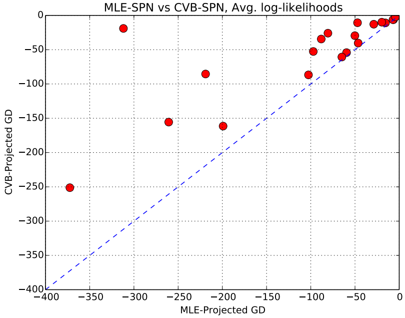
Output: Locally optimal $\boldsymbol{\beta}^*$.

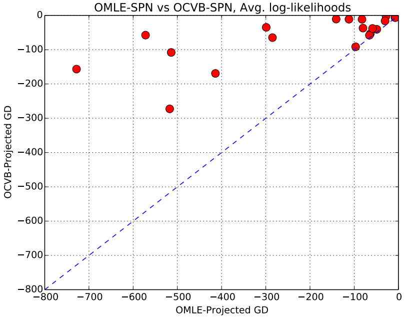
```
1:  while not converged do
2:      Update w = exp(E_{q'(w'|β)}[w']) with Eq. 10.
3:      Set ∇_β = 0.
4:      for d = 1 to D do
5:          Bottom-up evaluation of log p(x_d|w).
6:          Top-down differentiation of ∂/∂w log p(x_d|w).
7:          Update ∇_β based on x_d.
8:      end for
9:      Update ∇_β based on KL(q(w|β) || p(w|α)).
10:     Update β with projected GD.
11: end while
```

- Line 4 – 8 easily parallelizable, distributed version.
- Sample minibatch in Line 4 – 8, stochastic version.

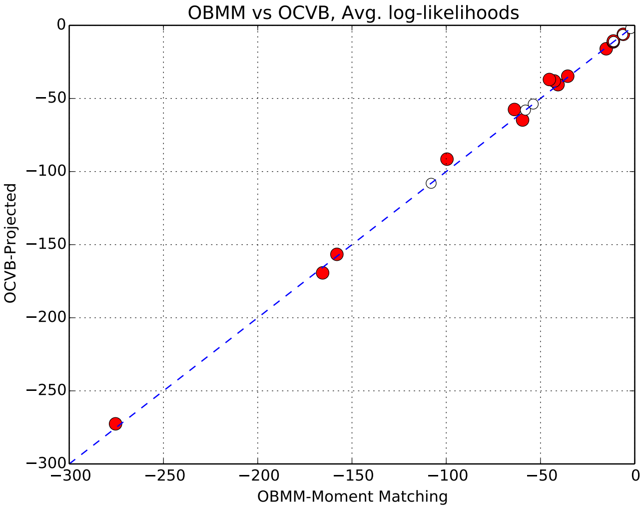
Experiments

- Experiments on 20 data sets, report average log-likelihoods, Wilcoxon ranked test.
- Compared with (O)MLE-SPN and OBMM-SPN.





Experiments



Summary

Thanks

Q & A