Collapsed Variational Inference for Sum-Product Networks

Han Zhao[1], Tameem Adel[2], Geoff Gordon[1] and Brandon Amos[1]
[1]{han.zhao, ggordon, bamos}@cs.cmu.edu, [2]T.M.A.A.Hesham@uva.nl
[1]Carnegie Mellon University, [2]University of Amsterdam;Radboud University

MACHINE LEARNING DEPARTMENT

Introduction

- Sum-Product Networks (SPNs) are probabilistic inference machines that admit exact inference in linear time in the size of the network.
- We develop a deterministic collapsed variational inference algorithm for SPNs that is both computationally and statistically efficient.
- The proposed algorithm can be easily adapted to stochastic and distributed settings.
- The proposed algorithm has a linear reduction in both time and space complexity compared with standard variational inference algorithm.

Background

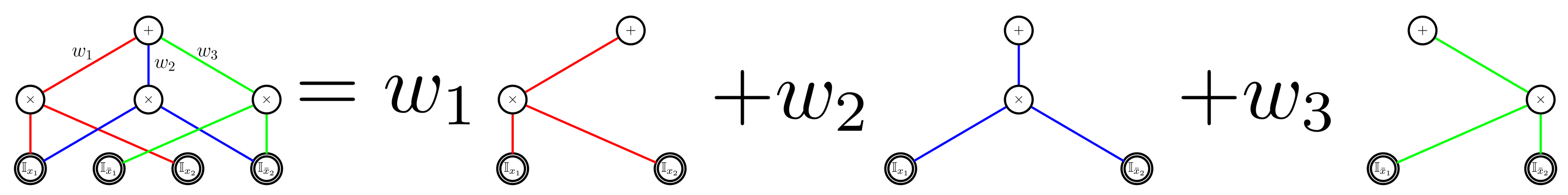
Sum-Product Networks (SPNs):

- 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.

Recursive computation of the network:

$$V_k(\mathbf{x} \mid \mathbf{w}) = \begin{cases} p(X_i = \mathbf{x}_i) & k \text{ is a leaf node over } X_i \\ \prod_{j \in \text{Ch}(k)} V_j(\mathbf{x} \mid \mathbf{w}) & k \text{ is a product node} \\ \sum_{j \in \text{Ch}(k)} w_{kj} V_j(\mathbf{x} \mid \mathbf{w}) & k \text{ is a sum node} \end{cases}$$

SPNs as Mixture of Trees:



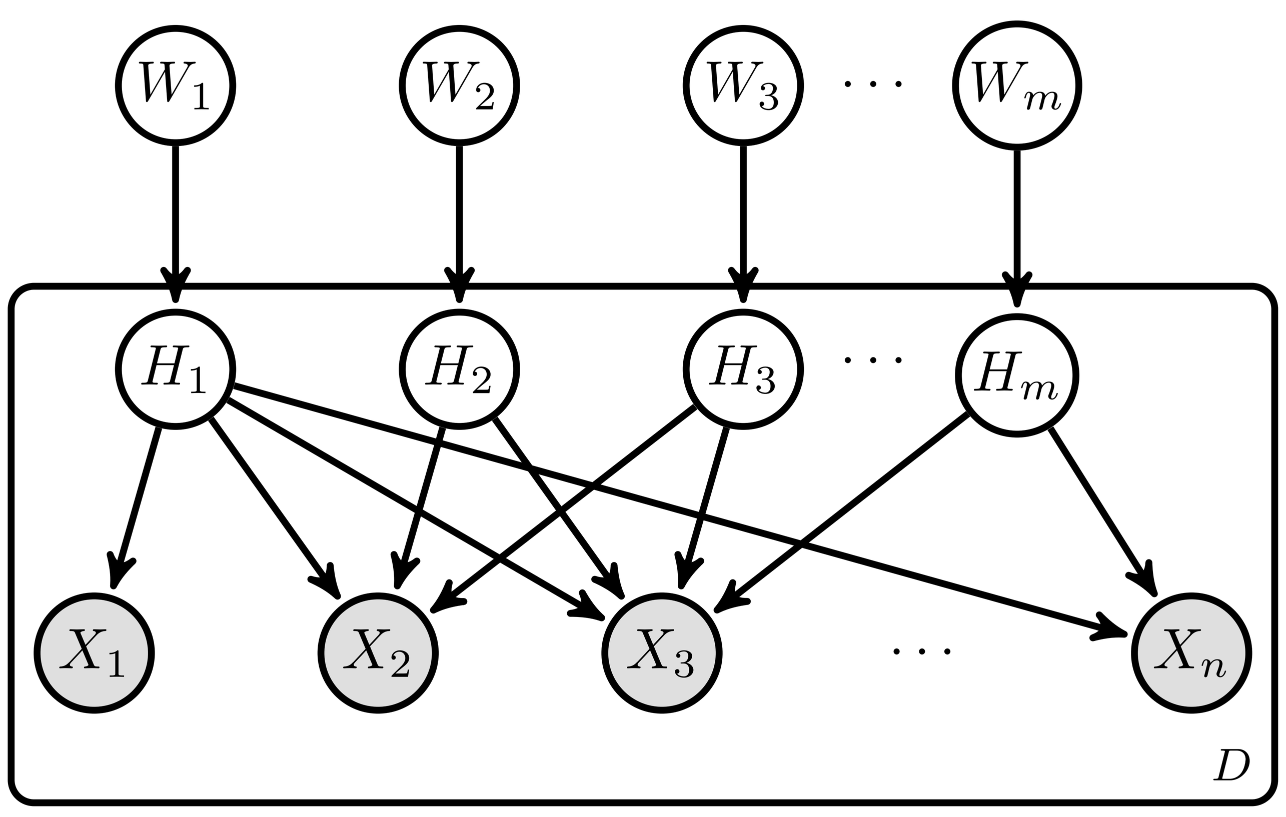
Let $\tau_{\mathcal{S}} = V_{\text{root}}(\mathbf{1}|\mathbf{1})$.

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

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

Equivalent Bayesian Networks:

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



- Number of observable variables in $\mathcal{B}$ = Number of variables in $\mathcal{S}$.
- Number of sum nodes in $\mathcal{S}$ = Number of hidden variables in $\mathcal{B} = \Theta(|\mathcal{S}|)$. $|\mathcal{B}| = O(n|\mathcal{S}|)$.
- Typically number of hidden variables $\gg$ number of observable variables, i.e., $m \gg n$.
- H_j are local hidden variables. W_j are global hidden variables.

Collapsed Variational Inference

Prior distribution over model parameters: $p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{k=1}^{m} p(w_k|\alpha_k) = \prod_{k=1}^{m} \text{Dir}(w_k|\alpha_k)$. Exact posterior in computationally intractable:

$$p(\mathbf{w}|\{\mathbf{x}_d\}_{d=1}^{D}, \boldsymbol{\alpha}) \propto \prod_{k=1}^{m} \text{Dir}(w_k|\alpha_k) \prod_{d=1}^{D} \sum_{t=1}^{\tau_{\mathcal{S}}} \prod_{(k,j) \in \mathcal{T}_{tE}} w_{kj} \prod_{i=1}^{n} p_t(x_{di})$$

Standard Variational Bayes Inference:

Mean Field assumption:

$$q(\mathbf{w}, \mathbf{h}) = \prod_i q(w_i) \prod_j q(h_j)$$

Evidence Lower BOund (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 Bayes Inference:

Using exact conditional distribution $q(\mathbf{h}|\mathbf{w})$, leading to the new ELBO:

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

Equivalent to marginalizing out local hidden variables: $q(\mathbf{w}) = \sum_{\mathbf{h}} q(\mathbf{w}, \mathbf{h})$ before approximating the true marginal posterior distribution.

- A better lower bound: $\log p(\mathbf{x}|\boldsymbol{\alpha}) \geq \mathcal{L} \geq \widehat{\mathcal{L}}$.
- Reduced space complexity: $O(D|\mathcal{S}|) \Rightarrow O(|\mathcal{S}|)$.
- Reduced time complexity: $O(nD|\mathcal{S}|) \Rightarrow O(D|\mathcal{S}|)$.

Variational optimization formulation:

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

No closed form solution for $\mathbb{E}_{q(\mathbf{w})}[\log p(\mathbf{x}|\mathbf{w})]$ due to the non-conjugacy between $q(\mathbf{w})$ and $p(\mathbf{x}|\mathbf{w})$.

Logarithmic Transformation:

Bijective mapping (change of variable) $\mathbf{w}' = \log(\mathbf{w})$, leading to:

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

a convex function of $\mathbf{w}'$. Apply Jensen's inequality to obtain further lower bound:

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

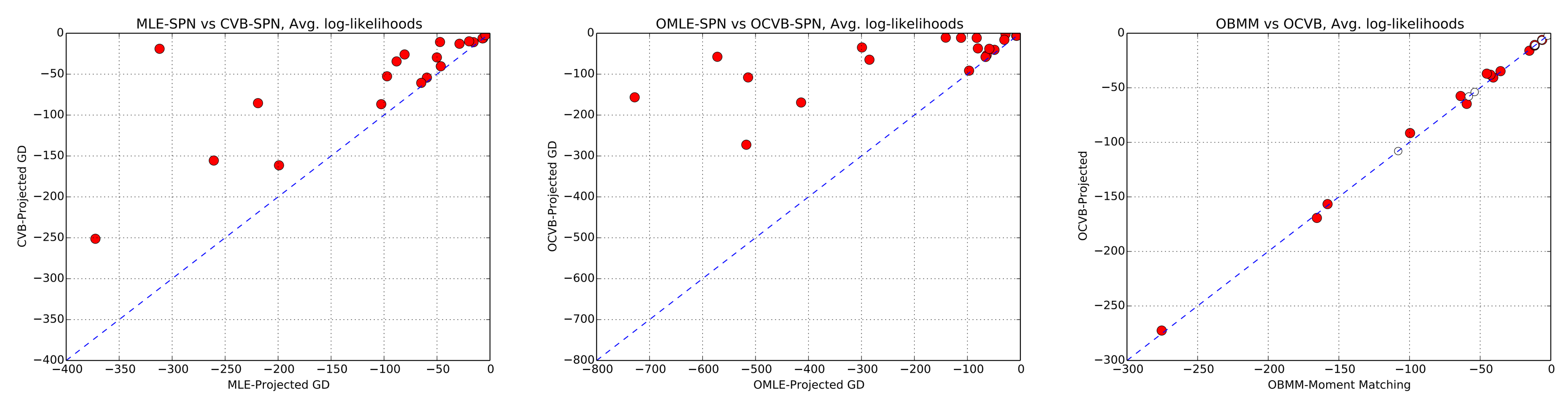
Relaxed objective:

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

Optimized by Projected gradient descent. Easily extended to stochastic and distributed settings.

Experiments

Compared with (O)MLE-SPN, OBMM on 20 benchmark data sets. Measuring average log-likelihoods on test data.



Conclusion

- CVB-SPN maintains a variational posterior distribution over global hidden variables by marginalizing out all the local hidden variables.
- CVB-SPN is both computationally and statistically efficient.

Carnegie Mellon University