Collapsed Variational Inference for Sum-Product Networks

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

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 algorithms.

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:

\[ V_i(x | w) = \begin{cases} p(X_i = x_i) & k \text{ is a leaf node over } X_i \\ \sum_{j \in C_i} V_j(x | w) & k \text{ is a product node} \\ \sum_{j \in C_i} w_{ij} V_j(x | w) & k \text{ is a sum node} \end{cases} \]

SPNs as Mixture of Trees:

Let \( \tau_S = V_{root}(1|1) \).

\[ f(w) = V_{root}(x | w) = \sum_{t=1}^{\tau} \prod_{(k,j) \in E} w_{kj} \prod_{i=1}^{n} p_i(X_i = x_i) \]

is a posynomial function of \( w \).

Equivalent Bayesian Networks:

Each SPN \( S \) is equivalent to a Bayesian network \( B \) with bipartite structure.

<table>
<thead>
<tr>
<th>W_1</th>
<th>W_2</th>
<th>W_3</th>
<th>W_m</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_1</td>
<td>H_2</td>
<td>H_3</td>
<td>H_m</td>
</tr>
<tr>
<td>X_1</td>
<td>X_2</td>
<td>X_3</td>
<td>X_m</td>
</tr>
</tbody>
</table>

- Number of observable variables in \( B \) = Number of variables in \( S \).
- Number of sum nodes in \( S \) = Number of hidden variables in \( B \) = \( \Theta(|S|) \).
- \( |S| = O(n|S|) \).
- 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:

\[ p(w | \alpha) = \prod_{i=1}^{n} Dir(w_i | \alpha_i) \]

Exact posterior in computationally intractable:

\[ p(w | \{x_i\}) \propto \prod_{i=1}^{n} Dir(w_i | \alpha_i) \prod_{d=1}^{D} \prod_{j \in T_d} w_{ij} \prod_{i=1}^{m} p(x_d) \]

Standard Variational Bayes Inference:

Mean Field assumption:

\[ q(w, h) = \prod_i q(w_i) \prod_j q(h_j) \]

Evidence Lower Bound (ELBO):

\[ \bar{L} = E_{q(w,h)}[log p(w, h, x)] + \mathbb{E}[q(w, h)] \]

Collapsed Variational Bayes Inference:

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

\[ L = E_{q(w)}[log p(w, x)] + \mathbb{E}[q(w)] \]

Equivalent to marginalizing out local hidden variables:

\[ q(w) = \sum_h q(w, h) \]

before approximating the true marginal posterior distribution.

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

Variational optimization formulation:

\[ \min_{q \in Q} KL[q(w) || p(w)] - E_{q(w)}[log p(x | w)] \]

No closed form solution for \( E_{q(w)}[log p(x | w)] \).

Logarithmic Transformation:

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

\[ \log p(x | w) = \log \left( \sum_{t=1}^{\tau} \prod_{(k,j) \in E} w'_{kj} \prod_{i=1}^{n} p_i(X_i = x_i) \right) \]

a convex function of \( w' \). Apply Jensen’s inequality to obtain further lower bound:

\[ E_{q(w)}[log p(x | w)] = E_{q(w)}[log p(x | w')] \geq \log p(x | E_{q(w')}[w']) \]

Relaxed objective:

\[ \min_{q \in Q} KL[q(w) || p(w)] - \log p(x | E_{q(w')}[w']) \]

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.

- 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.

Conclusion

Carnegie Mellon University