



# Optimal Tree Density Approximators

## 1 Why

Which is the optimal tree to use for tree density approximation?

## 2 Definition

We want to choose the tree whose corresponding approximator for the given density achieves minimum relative entropy with the given density among all tree density approximators. We call such a density an *optimal tree approximator* of the given density. We call a tree according to which an optimal tree approximator factors and optimal approximator tree.

## 3 Result

**Proposition 1.** *Let  $g : \mathbf{R}^n \rightarrow [0, 1]$  be a density. A tree  $T$  on  $\{1, \dots, n\}$  is an optimal approximator tree if and only if it is a maximal spanning tree of the differential mutual information graph of  $g$ .*

*Proof.* First, denote the optimal approximator of  $g$  for tree  $T$  by  $f_T^*$ . Recall

$$f_T^* = f_1 \prod_{i \neq 1} f_{i|\mathbf{pa}_i}$$

Second, recall  $d(g, f) = H(g, f) - H(g)$ . Since  $H(g)$  does not depend on  $f$ ,  $f$  is a minimizer of  $d(g, f)$  if and only if it is a minimizer of  $H(g, f)$ .

Third, express the cross entropy of  $f_T^*$  relative to  $g$  as

$$\begin{aligned}
H(q, p_T^*) &= h(q_1) - \sum_{j \neq i} \left( \int_{\mathbf{R}^d} g(x) \log g_{i|pai}(x_i, x_{\mathbf{pa}_i}) dx \right) \\
&= H(q_1) - \sum_{i \neq 1} \sum_{a \in A} q(a) (\log q_{i, \mathbf{pa}_i}(a_i, a_{\mathbf{pa}_i}) - \log q_{\mathbf{pa}_i}(a_{\mathbf{pa}_i})) \\
&= H(q_1) - \sum_{i \neq 1} \sum_{a \in A} q(a) (\log q_{i, \mathbf{pa}_i}(a_i, a_{\mathbf{pa}_i}) - \log q_{\mathbf{pa}_i}(a_{\mathbf{pa}_i}) - \log q_i(a_i) + \log q_i(a_i)) \\
&= \sum_{i=1}^n H(q_i) - \sum_{i \neq 1} I(q_i, q_{\mathbf{pa}_i}) \\
&= \sum_{i=1}^n H(q_i) - \sum_{\{i,j\} \in T} I(q_i, q_j)
\end{aligned}$$

where  $\mathbf{pa}_i$  denotes the parent of vertex  $i$  in  $T$  ( $i = 2, \dots, n$ ).  $H(g_i)$  does not depend on the choice of tree. Choosing a tree to minimize the second term in the final expression above is equivalent to choosing a maximal spanning tree from the weighted graph with differential mutual information edge weights; namely, the mutual information graph of  $g$ .

□