



## Why

We can approximate a density with a tree density similar to how we can approximate a distribution with a tree distribution.

## Definition

We use the differential relative entropy as a criterion of approximation. An *optimal tree approximator* of density for a tree is a density which factors according to a tree and minimizes its differential relative entropy with the given density.

## Notation

Let  $g : \mathbf{R}^n \rightarrow \mathbf{R}$  be a density and  $T$  be a tree on  $\{1, \dots, n\}$ . An optimal tree approximator of  $g$  for  $T$  is a density  $f$  that factors according to  $T$  and minimizes  $d(g, f)$ . In other words, given  $g$  and  $T$  we want to find  $f$  to

$$\begin{aligned} & \text{minimize} && d(g, f) \\ & \text{subject to} && f \text{ factors according to } T. \end{aligned}$$

## Result

**Proposition 1.** *Let  $g : \mathbf{R}^n \rightarrow \mathbf{R}$  be a density and  $T$  be a tree on  $\{1, \dots, n\}$ . The density  $f_T^* : \mathbf{R}^d \rightarrow \mathbf{R}$  defined by*

$$f_T^* = g \prod_{i \neq \text{pa } i} g_{\text{pa } i}$$

*minimizes the differential relative entropy with  $g$  among all densities on  $\mathbf{R}^n$  which factor according to  $T$  ( $\text{pa } i$  is the parent of  $i$  in  $T$  rooted at vertex 1,  $i = 2, \dots, n$ ).*

*Proof.* Let  $f : \mathbf{R}^d \rightarrow \mathbf{R}$  be a density factoring according to  $T$ . First, express

$$f = f_1 \prod_{i=2}^n f_{i|\text{pa } i}.$$

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

Third, express

$$\begin{aligned} h(g, f) &= - \int_{\mathbf{R}^d} g \log f \\ &= - \int_{\mathbf{R}^d} g(x) \left( \log f_i(x_i) + \sum_{i \neq 1} \log f_{i|\text{pa } i}(x_i, x_{\text{pa } i}) \right) dx \\ &= h(g_1, f_1) + \sum_{i \neq 1} \left( \int_{\mathbf{R}} g_{\text{pa } i}(\xi) h(g_{i|\text{pa } i}(\cdot, \xi), f_{i|\text{pa } i}(\cdot, \xi)) d\xi \right) \end{aligned}$$

which separates across  $f_1$  and  $f_{i|\text{pa } i}(\cdot, \xi)$  for  $i = 1, \dots, n$  and  $\xi \in \mathbf{R}$ . In particular, since  $g_{\text{pa } i} \geq 0$ , we can minimize the integrand pointwise.

Fourth, recall  $h(\phi, \psi) \geq 0$  for densities  $\phi, \psi$  of any dimension, and zero if  $\phi = \psi$ . So  $f_1 = g_1$  and  $f_{i|\text{pa } i} = g_{i|\text{pa } i}$  are solutions.  $\square$

