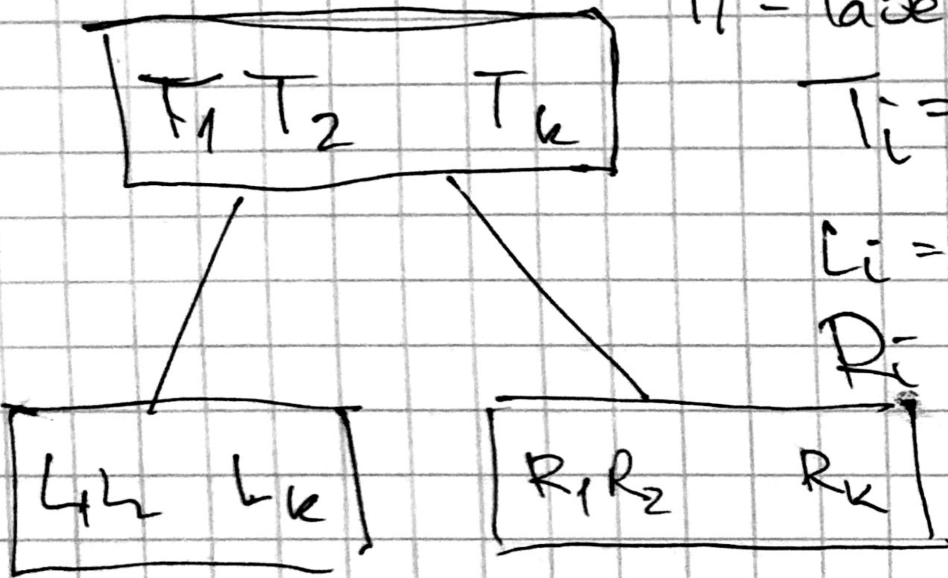


T_i = label i count before split; K labels

$T_i = L_i + R_i$

L_i = label i count on left branch

R_i = label i count on right branch

$\sum T_i = T$; $\sum L_i = L$; $\sum R_i = R$

$$H_T = H_{before} = \sum \frac{T_i}{T} \log\left(\frac{T}{T_i}\right)$$

$$H_S = H_{after} = \frac{L}{T} \sum \frac{L_i}{L} \log\left(\frac{L}{L_i}\right) + \frac{R}{T} \sum \frac{R_i}{R} \log\left(\frac{R}{R_i}\right)$$

want to prove that $H_{before} - H_{after} \leq 1$

$$\sum_{i=1}^{K} \frac{T_i}{T}\log\frac{T}{T_i} - \sum_{i=1}^{K}\frac{L_i}{T}\log\frac{L}{L_i} - \sum_{i=1}^{K}\frac{R_i}{T}\log\frac{R}{R_i} \leq 1$$

$$\sum_{i} \frac{T_i}{T}\left(\log\frac{T}{T_i} - \frac{L_i}{T_i}\log\frac{L}{L_i} - \frac{R_i}{T_i}\log\frac{R}{R_i}\right) \leq 1$$

$$\sum_{i} \frac{T_i}{T}\left[\log\frac{T}{T_i} + \frac{L_i}{T_i}\log\frac{L_i}{L} + \frac{R_i}{T_i}\log\frac{R_i}{R}\right] \leq 1$$

log concave $\Rightarrow$ $p\log x + (1-p)\log y \le \log(px + (1-p)y)$
if $p \in (0,1)$

So $\frac{L_i}{T_i} \log \frac{L_i}{L} + \frac{R_i}{T_i} \log \frac{R_i}{R} \le \log\left(\frac{L_i}{T_i} \cdot \frac{L_i}{L} + \frac{R_i}{T_i} \cdot \frac{R_i}{R}\right)$

Want: $\sum_i \frac{T_i}{T}\left[\log \frac{T}{T_i} + \log\left(\frac{L_i}{T_i} \cdot \frac{L_i}{L} + \frac{R_i}{T_i} \cdot \frac{R_i}{R}\right)\right] \le 1$

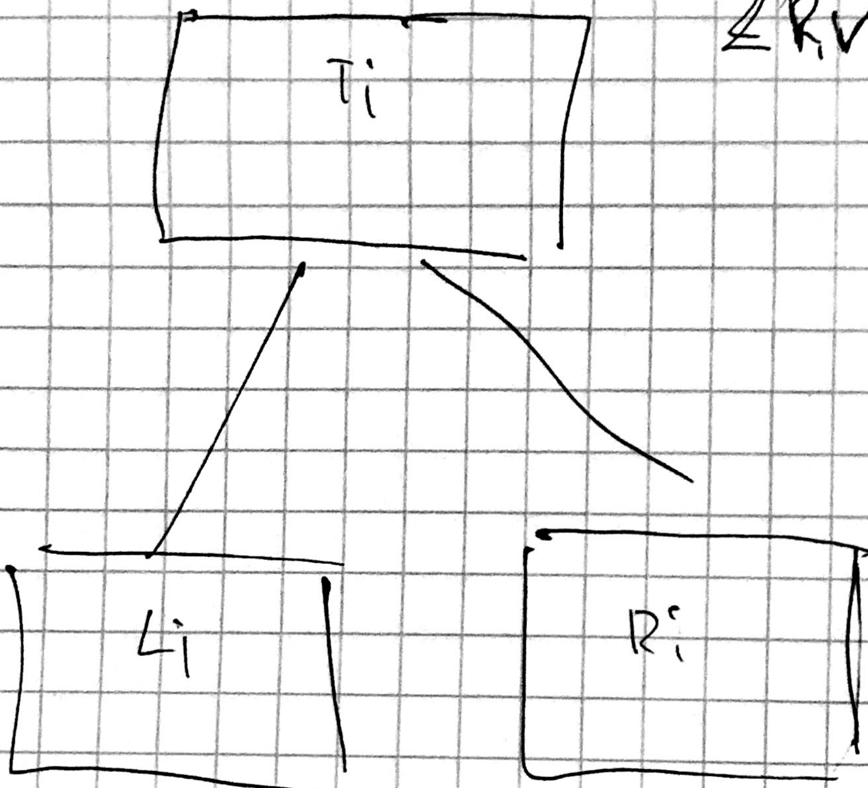
$$\sum \frac{T_i}{T} \log\left[\frac{T}{T_i} \cdot \left(\frac{L_i}{T_i} \cdot \frac{L_i}{L} + \frac{R_i}{T_i} \cdot \frac{R_i}{R}\right)\right] \le 1$$

log concave $\Rightarrow \sum p_i \log x_i \le \log(\sum p_i x_i)$ if $\sum p_i = 1$

$$\sum \frac{T_i}{T} \log\left(\frac{T}{T_i}\left(\frac{L_i}{T_i} \cdot \frac{L_i}{L} + \frac{R_i}{T_i} \cdot \frac{R_i}{R}\right)\right) \le \log\left[\sum\left(\frac{L_i}{T_i} \cdot \frac{L_i}{L} + \frac{R_i}{T_i} \cdot \frac{R_i}{R}\right)\right]$$

Want: $\log\left(\sum\left(\frac{L_i}{T_i} \cdot \frac{L_i}{L} + \frac{R_i}{T_i} \cdot \frac{R_i}{R}\right)\right) \le 1$

$$\sum\left(\frac{L_i}{T_i} \cdot \frac{L_i}{L} + \frac{R_i}{T_i} \frac{R_i}{R}\right) \le 2$$



Z RV $\quad Z = \log \frac{1}{T_i}$

$$2^{-Z} = 2^{-\log\frac{1}{T_i}} = 2^{\log T_i} = T_i = L_i + R_i$$

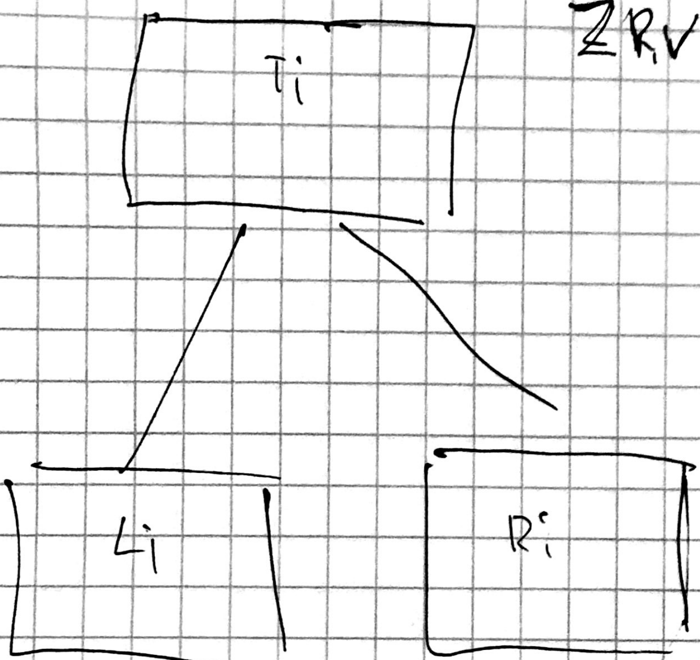
$$= 2^{-X} + 2^{-Y}$$

X RV $\cancel{=}$ $\quad$ Y RV

$X = \log \frac{1}{L_i}$ $\quad$ $Y = \log \frac{1}{R_i}$

Info gain = $H_{before} - H_{after} = \sum \frac{T_i}{T} \log \frac{T}{T_i} -$

$$- \frac{L}{T} \sum \frac{L_i}{L} \log \frac{L}{L_i} - \frac{R}{T} \sum \frac{R_i}{R} \log \frac{R}{R_i} =$$



Z RV $\quad Z = \log \frac{1}{T_i}$

$$2^{-Z} = 2^{-\log \frac{1}{T_i}} = 2^{\log T_i} = T_i = L_i + R_i$$

$$= 2^{-X} + 2^{-Y}$$

X RV $\quad X = \log \frac{1}{L_i}$

Y RV $\quad Y = \log \frac{1}{R_i}$

$$\text{Info gain} = H_{\text{before}} - H_{\text{after}} = \sum \frac{T_i}{T} \log \frac{T}{T_i} -$$

$$- \frac{L}{T} \sum \frac{L_i}{L} \log \frac{L}{L_i} - \frac{R}{T} \sum \frac{R_i}{R} \log \frac{R}{R_i} =$$

$$= \log T - \sum \frac{T_i}{T} \log \frac{1}{T_i} - \frac{L}{T} \log L - \frac{L}{T} \sum \frac{L_i}{L} \log \frac{1}{L_i} - \frac{R}{T} \log R - \frac{R}{T} \sum \frac{R_i}{R} \log \frac{1}{R_i}$$

$$= \log T + E[Z] - \frac{L}{T} \log L - \frac{L}{T} E[X] - \frac{R}{T} \log R - \frac{R}{T} E[Y]$$

$$= \log T - \frac{L}{T} \log L - \frac{R}{T} \log R + \frac{L}{T} E[Z-X] + \frac{R}{T} E[Z-Y]$$

$$= \log T + \underbrace{\frac{L}{T} \log \frac{1}{L} + \frac{R}{T} \log \frac{1}{R}}_{\leq \log \frac{2}{T}} + \frac{L}{T} \underbrace{E[Z-X]}_{\leq 0} + \frac{R}{T} \underbrace{E[Z-Y]}_{\leq 0}$$

$$\leq \log T + \log \frac{2}{T} = \log 2 = 1$$