

# Encoder-Decoder: Training Objective

For output word  $y_i$  we have:

- estimated conditional distribution  $\hat{p}_i = \frac{\exp t_i}{\sum \exp t_j}$  (softmax function)
- unknown true distribution  $p_i$ , we lay  $p_i \equiv \mathbf{1}[y_i]$

Cross entropy  $\approx$  distance of  $\hat{p}$  and  $p$ :

$$\mathcal{L} = H(\hat{p}, p) = \mathbf{E}_p(-\log \hat{p}) = - \sum_{v \in V} p(v) \log \hat{p}(v) = -\log \hat{p}(y_i)$$

...computing  $\frac{\partial \mathcal{L}}{\partial t_i}$  is quite simple

See <https://eli.thegreenplace.net/2016/the-softmax-function-and-its-derivative/>

**...but we expect the model to produce  
the exact word at the exact position!**