We want $\hat{p}_t$ to be close to $p = 0.05$.

Alternatives:

1) add $(\hat{p}_t - p)^2$ to the loss function f.

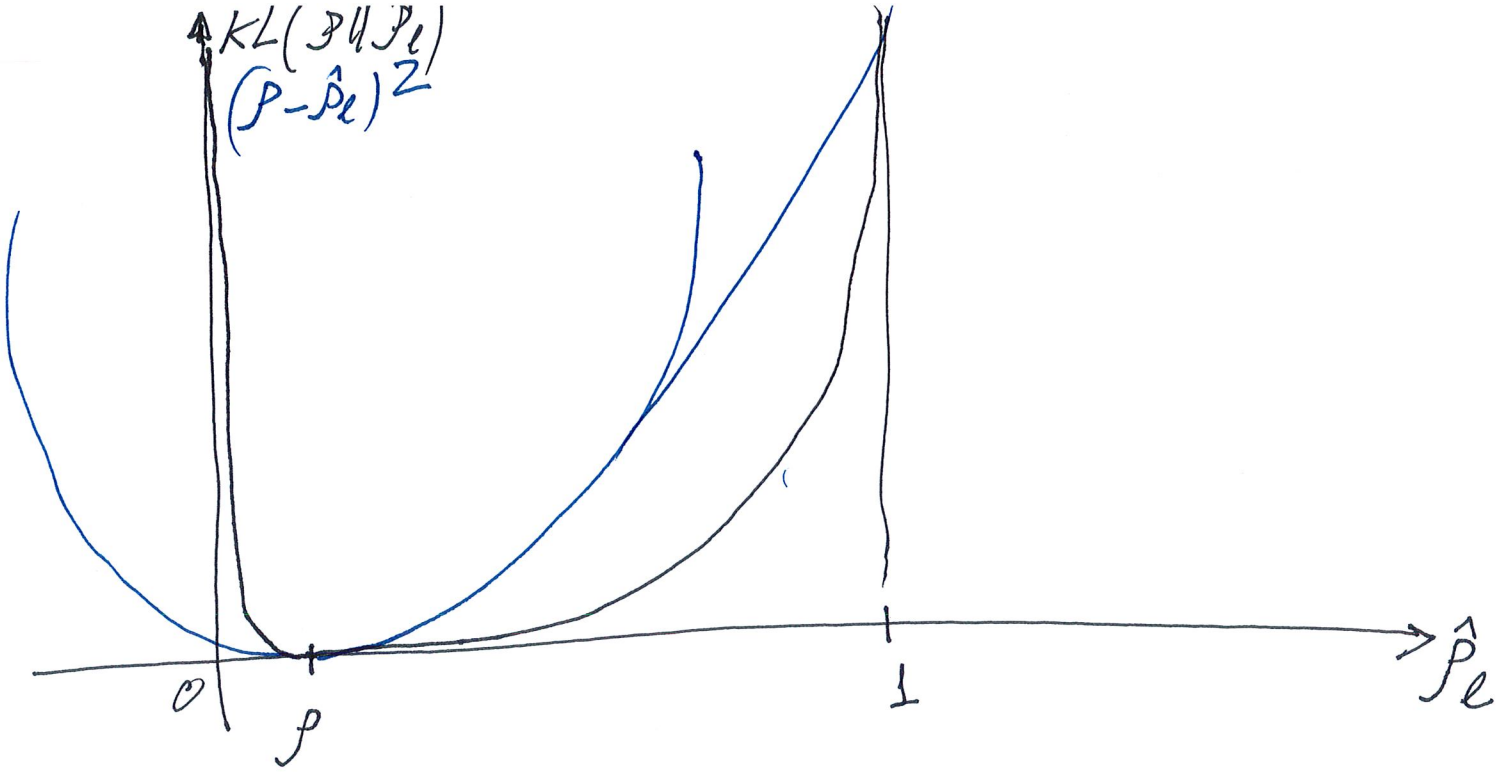
2) $\boxed{KL(p \,\|\, \hat{p}_t)} \neq -p \log \frac{\hat{p}_t}{p} - (1-p) \log \frac{1-\hat{p}_t}{1-p}$

$$= +p \log \frac{p}{\hat{p}_t} + (1-p) \log \frac{1-p}{1-\hat{p}_t}$$

$$KL(p \,\|\, q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

→ KL divergence between $p \equiv$ Bernoulli distrib w/ param p

$q \equiv$ — " — w/ param $\hat{p}_t$



input $a^{(1)} = x$ to $a^{(2)}$ to $a^{(3)} = h(x)$

$W_{ij}^{(1)}$ $W^{(2)}$

$$\hat{\rho}_\ell = \frac{1}{m} \sum_{k=1}^{m} a^{(2)}(x^{(k)})$$

$$J = \frac{1}{2} \| a^{(3)} - x \|^2 + \frac{\lambda}{2} \| W \|^2 + \beta \sum_{\ell=1}^{s_2} KL(\rho \| \hat{\rho}_\ell)$$

Need to compute gradient of new (sparsity) term $\beta \sum_{\ell=1}^{s} KL(\rho \| \hat{\rho}_\ell)$ w.r.t. $W_{ij}^{(1)}$

$$KL(\rho, \hat{\rho}_\ell) = \rho \log \frac{\rho}{\hat{\rho}_\ell} + (1-\rho) \log \frac{(1-\rho)}{(1-\hat{\rho}_\ell)}$$

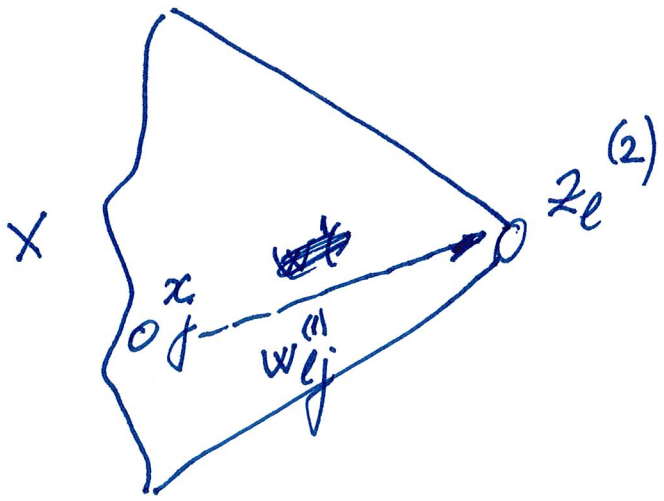
$$\hat{\rho}_\ell = \frac{1}{m} \sum_{k=1}^{m} a_\ell^{(2)}(x^{(k)}) = \frac{1}{m} \sum_{k=1}^{m} f\left(\sum_{u=1}^{|x^{(k)}|} W_{\ell u}^{(1)} \cdot x_u^{(k)} + b_\ell^{(1)} \right)$$

$$\frac{\partial KL(\hat{\rho}_\ell)}{\partial W_{ij}^{(1)}} = \frac{\partial KL(\hat{\rho}_\ell)}{\partial \hat{\rho}_\ell} \cdot \frac{\partial \hat{\rho}_\ell}{\partial W_{ij}^{(1)}}$$

$$KL(\rho, \hat{\rho}_\ell) = -\rho \log \frac{\hat{\rho}_\ell}{\rho} - (1-\rho) \log \frac{1-\hat{\rho}_\ell}{1-\rho}$$

$$\Rightarrow \frac{\partial KL}{\partial \hat{\rho}_\ell} = -\rho \cdot \frac{\rho}{\hat{\rho}_\ell} \cdot \frac{1}{\rho} - (1-\rho) \cdot \frac{1-\rho}{1-\hat{\rho}_\ell} \cdot -\frac{1}{1-\rho}$$

$$= -\frac{\rho}{\hat{\rho}_\ell} + \frac{1-\rho}{1-\hat{\rho}_\ell}$$



$$\frac{\partial z_l}{\partial w_{ij}^{(1)}} = \begin{cases} x_j \ (\text{also } a_j^{(1)}) & \text{if } l = i \\ 0, & \text{if } l \neq i \end{cases}$$