NN parameter estimation for classification

- Model: $P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta}) \quad where \quad \mathbf{x}, \mathbf{y} \sim P_D \,, \quad \mathbf{y} \in \{0,1\}^N$

one hot encoding

- Optimization:

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{M} -\log P(\mathbf{y}_i\,|\,\mathbf{x}_i;\boldsymbol{\theta}) \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim P_D}\big[\log P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\big] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; \mathbb{E}_{\mathbf{x},\mathbf{y}\sim P_D}\Big[\log \frac{1}{P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})}\Big]
\end{aligned}
$$

Cross entropy between data distribution and the model output distribution, $H\big(P_D(\mathbf{y}\,|\,\mathbf{x}), P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\big)$. Note, this holds for arbitrary PMFs (softmax, Bernoulli, etc…). Because the labels, $\mathbf{y}$, are one hot, they have zero entropy, and thus in this setting minimizing the cross entropy is equivalent to minimizing the KL divergence, $D_{KL}\big(P_D(\mathbf{y}\,|\,\mathbf{x}), P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\big)$.