NN parameter estimation for classification

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

  one hot encoding

- Optimization:

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

↓

Cross entropy between data distribution and the model output distribution, $H\left(P_D(\mathbf{y} \mid \mathbf{x}), P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})\right)$. 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, $D_{KL}\left(P_D(\mathbf{y} \mid \mathbf{x}), P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta})\right)$.