

# Neural networks

Training neural networks - loss function

# MACHINE LEARNING

**Topics:** stochastic gradient descent (SGD)

- Algorithm that performs updates after each example
  - ▶ initialize  $\boldsymbol{\theta}$  ( $\boldsymbol{\theta} \equiv \{\mathbf{W}^{(1)}, \mathbf{b}^{(1)}, \dots, \mathbf{W}^{(L+1)}, \mathbf{b}^{(L+1)}\}$ )
  - ▶ for N iterations
    - for each training example  $(\mathbf{x}^{(t)}, y^{(t)})$ 

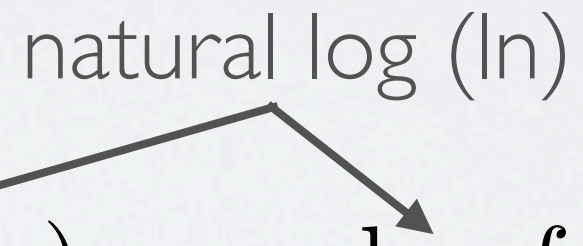
$$\left. \begin{array}{l} \checkmark \Delta = -\nabla_{\boldsymbol{\theta}} l(f(\mathbf{x}^{(t)}; \boldsymbol{\theta}), y^{(t)}) - \lambda \nabla_{\boldsymbol{\theta}} \Omega(\boldsymbol{\theta}) \\ \checkmark \boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \Delta \end{array} \right\} \begin{array}{l} \text{training epoch} \\ = \\ \text{iteration over **all** examples} \end{array}$$
- To apply this algorithm to neural network training, we need
  - ▶ the loss function  $l(\mathbf{f}(\mathbf{x}^{(t)}; \boldsymbol{\theta}), y^{(t)})$
  - ▶ a procedure to compute the parameter gradients  $\nabla_{\boldsymbol{\theta}} l(\mathbf{f}(\mathbf{x}^{(t)}; \boldsymbol{\theta}), y^{(t)})$
  - ▶ the regularizer  $\Omega(\boldsymbol{\theta})$  (and the gradient  $\nabla_{\boldsymbol{\theta}} \Omega(\boldsymbol{\theta})$ )
  - ▶ initialization method



# LOSS FUNCTION

**Topics:** loss function for classification

- Neural network estimates  $f(\mathbf{x})_c = p(y = c|\mathbf{x})$ 
  - we could maximize the probabilities of  $y^{(t)}$  given  $\mathbf{x}^{(t)}$  in the training set
- To frame as minimization, we minimize the negative log-likelihood

$$l(\mathbf{f}(\mathbf{x}), y) = - \sum_c 1_{(y=c)} \log f(\mathbf{x})_c = - \log f(\mathbf{x})_y$$


The diagram shows the text "natural log (ln)" positioned above the equation. Two arrows originate from this text: one points to the  $\log f(\mathbf{x})_c$  term in the first part of the equation, and the other points to the  $\log f(\mathbf{x})_y$  term in the second part of the equation.

- we take the log to simplify for numerical stability and math simplicity
- sometimes referred to as cross-entropy