Neural networks

Training neural networks - parameter gradient
MACHINE LEARNING

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

- Algorithm that performs updates after each example
  - initialize $\vec{\theta}$ ( $\vec{\theta} \equiv \{\vec{W}^{(1)}, \vec{b}^{(1)}, \ldots, \vec{W}^{(L+1)}, \vec{b}^{(L+1)}\}$
  - for N iterations
    - for each training example $(x^{(t)}, y^{(t)})$
      - $\Delta = -\nabla_{\theta} l(f(x^{(t)}; \theta), y^{(t)}) - \lambda \nabla_{\theta} \Omega(\theta)$
      - $\vec{\theta} \leftarrow \vec{\theta} + \alpha \Delta$
  - training epoch = iteration over all examples

- To apply this algorithm to neural network training, we need
  - the loss function $l(f(x^{(t)}; \theta), y^{(t)})$
  - a procedure to compute the parameter gradients $\nabla_{\theta} l(f(x^{(t)}; \theta), y^{(t)})$
  - the regularizer $\Omega(\theta)$ (and the gradient $\nabla_{\theta} \Omega(\theta)$)
  - initialization method
Topics: loss gradient of parameters

- Partial derivative (weights):

\[
\frac{\partial}{\partial W_{i,j}^{(k)}} - \log f(x)_y \\
= \frac{\partial - \log f(x)_y}{\partial a^{(k)}(x)_i} \frac{\partial a^{(k)}(x)_i}{\partial W_{i,j}^{(k)}} \\
= \frac{\partial - \log f(x)_y}{\partial a^{(k)}(x)_i} h^{(k-1)}_j(x)
\]

REMINDER

\[ a^{(k)}(x)_i = b^{(k)}_i + \sum_j W^{(k)}_{i,j} h^{(k-1)}(x)_j \]
**Topics:** loss gradient of parameters

- Gradient (weights):

\[
\nabla_w f^{(k)} - \log f(x)_y \\
= (\nabla_a^{(k)}(x) - \log f(x)_y) h^{(k-1)}(x)^T
\n\]
Topics: loss gradient of parameters

- Partial derivative (biases):

\[
\frac{\partial}{\partial b_i^{(k)}} - \log f(x)_y \\
= \frac{\partial - \log f(x)_y}{\partial a^{(k)}(x)_i} \frac{\partial a^{(k)}(x)_i}{\partial b_i^{(k)}} \\
= \frac{\partial - \log f(x)_y}{\partial a^{(k)}(x)_i}
\]

REMAINDER

\[ a^{(k)}(x)_i = b_i^{(k)} + \sum_j W_{i,j}^{(k)} h^{(k-1)}(x)_j \]
Topics: loss gradient of parameters

- Gradient (biases):

\[ \nabla b^{(k)} = \log f(x)_y \]
\[ = \nabla a^{(k)}(x) - \log f(x)_y \]

REMINDER

\[ a^{(k)}(x)_i = b^{(k)}_i + \sum_j W^{(k)}_{i,j} h^{(k-1)}(x)_j \]