1 Review

- last time we introduced neural network equations
  \[ \tau \dot{x}_i + x_i = f \left( \sum_j W_{ij} x_j + b_i \right) \]

- good approximation to the dynamics of spiking neurons, assuming
  - point neuron
  - slow synapses of a single type
  - all synaptic time constants and other biophysical properties are homogeneous

2 What can a perceptron compute?

- perceptron
  \[ f \left( \sum_j w_j x_j - \theta \right) \]
  - intuitive idea: convergence of synaptic connections determines selectivity
  - let \( f \) be the step function for simplicity, so the output is binary.
  - threshold \( \theta \) (notational difference from bias \( b \))
  - static mapping, need not consider dynamics, as there are no feedback loops.

- geometry of perceptron
  - weight vector \( w \)
  - input vector \( x \)
  - angle \( \phi \) between these two vectors
  \[ \cos \phi = \frac{w \cdot x}{|w||x|} \]
\[ w \cdot x \] measures the similarity of the vectors

- \( w \cdot x = \theta \) defines a separating hyperplane.

- examples: some logical functions can be represented, but others cannot.
  - AND
  - OR
  - XOR

linear separability has both geometric and algebraic representations.

- by putting perceptrons together, we can compute anything. any digital circuit, since we can construct basic logic gates.

3 Example: handwritten digit recognition

- database collected by NIST. segmented and size normalized by AT&T group

- let’s make a ”two” detector
  - let a training set be given
  - take the mean of all the ”two”s
  - compute two histograms
    - \( w \cdot x \) for all ”two”s
    - \( w \cdot x \) for all other classes
  - set threshold in the middle
  - nonzero training error: distribution tails
    - false positives
    - false negatives
  - performance on test set

- important distinction: training and generalization error
  - dangerous to test on your training data
  - performance on training data may be unrealistically good

4 Perceptron learning algorithm

- suppose we have a set of examples \( x \), and labels \( y \)

- forget about bias (like an input clamped at -1)

- loop over examples. if there is an error
- positive example: \( w = w + x \)
- negative example: \( w = w - x \)

- perceptron convergence theorem. This procedure converges if there is a separating hyperplane
  - Make all examples positive by reflection.
  - Define the margin \( D(w) \)
    \[
    D(w) = \min_{\mu} \frac{w \cdot x^\mu}{|w|}
    \]
  - Write \( w = \sum_{\mu} m^\mu x^\mu \), where \( m^\mu \) is the number of errors that were made on example \( \mu \).
  - Let \( w^* \) define some separating hyperplane.
    \[
    w \cdot w^* = \sum_{\mu} m^\mu x^\mu \cdot w^* \geq |w^*| D(w^*) M
    \]
    where \( M = \sum_{\mu} m_\mu \) is the total number of errors.
  - Now bound the magnitude of \( w \)
    \[
    \Delta |w|^2 = |w + x^\alpha|^2 - |w|^2 \geq |x^\alpha|^2 + 2 w \cdot x^\alpha \leq X_{max}^2
    \]
    Therefore \( |w|^2 \leq MX_{max}^2 \)
  - Put the two inequalities together to bound the angle between \( w \) and \( w^* \)
    \[
    1 \geq \frac{w \cdot w^*}{|w||w^*|} \geq \frac{D(w^*)}{X_{max}} \]
    - This yields an upper bound on the number of weight updates that can happen
      \[
      M \leq \left[ \frac{X_{max}}{D(w^*)} \right]^2
      \]
- no guarantees if examples are not linearly separable.

5 Learning by gradient descent

- formulate learning as an optimization problem, with a squared error cost function
  \[
  E(w) = \frac{1}{2} \sum_{\mu} [y^\mu - f(w \cdot x^\mu)]^2
  \]
• use sigmoid rather than step function for differentiability
• gradient descent (batch update)

\[
\Delta w = -\eta \frac{\partial E}{\partial w}
\]
\[
= \eta \sum_{\mu} [y^\mu - f(w \cdot x^\mu)] f'(w \cdot x^\mu) x^\mu
\]

• online update (similar to perceptron algorithm)

\[
\Delta w^\mu = \eta [y^\mu - f(w \cdot x^\mu)] f'(w \cdot x^\mu) x^\mu
\]

• problems with gradient descent
  – good if cost function is isotropic
  – is not optimal, just convenient

6 Problem of credit assignment

• many local elements contributed to the global outcome
• how should credit or blame be assigned?
• both temporal and spatial aspects
• most fundamental problem in learning