MIT Department of Brain and Cognitive Sciences
9.641J, Spring 2005 - Introduction to Neural Networks
Instructor: Professor Sebastian Seung

Backprop for recurrent networks

Steady state

- Reward is an explicit function of *x.*
- The steady state of a recurrent network.

$$x_i = f\left(\sum_j W_{ij} x_j + b_i\right)$$

$$\max_{W,b} R(x)$$

Recurrent backpropagation

- Find steady state $x = f(Wx + b)$
- Calculate slopes $D = \mathrm{diag}\{f'(Wx + b)\}$
- Solve for *s* $\left(D^{-1} - W^T\right)s = \frac{\partial R}{\partial x}$
- Weight update $\Delta W = \eta s x^T$

Sensitivity lemma

$$\frac{\partial R}{\partial W_{ij}} = \frac{\partial R}{\partial b_i} x_j$$

Input as a function of output

- What input *b* is required to make *x* a steady state?

$$b_i = f^{-1}(x_i) - \sum_j W_{ij} x_j$$

- This is unique, even when output is not a unique function of the input!

Jacobian matrix

$$b_i = f^{-1}(x_i) - \sum_j W_{ij} x_j$$

$$\frac{\partial b_i}{\partial x_j} = f^{-1\prime}(x_i)\delta_{ij} - W_{ij}$$

$$= \left(D^{-1} - W\right)_{ij}$$

Chain rule

$$\frac{\partial R}{\partial x_j} = \sum_i \frac{\partial R}{\partial b_i} \frac{\partial b_i}{\partial x_j}$$

$$= \sum_i \frac{\partial R}{\partial b_i} \left(D^{-1} - W\right)_{ij}$$

$$\frac{\partial R}{\partial x} = \left(D^{-1} - W^T\right) \frac{\partial R}{\partial b}$$

Trajectories

- Initialize at *x*(0)
- Iterate for *T* time steps

$$x_i(t) = f\left(\sum_j W_{ij} x_j(t-1) + b_i\right)$$

$$\max_{W,b} R(x(1),\ldots,x(T))$$

Unfold time into space

- Multilayer perceptron
  - Same number of neurons in each layer
  - Same weights and biases in each layer (weight-sharing)

$$x(0) \xrightarrow{W,b} x(1) \xrightarrow{W,b} \cdots \xrightarrow{W,b} x(T)$$

Backpropagation through time

- Initial condition x(0)

$$x(t) = f\left(Wx(t-1) + b(t)\right)$$

- Compute $R / \partial x(t)$
- Final condition s(T+1)=0

$$s(t) = D(t)W^T s(t+1) + D(t)\frac{\partial R}{\partial x(t)}$$

$$\Delta W = \eta \sum_t s(t) x(t-1)^T \qquad \Delta b = \eta \sum_t s(t)$$

Input as a function of output

$$x(t) = f\left(Wx(t-1) + b(t)\right)$$

$$b(t) = f^{-1}\left(x(t)\right) - Wx(t-1)$$

$$x(1), x(2), \ldots, x(T-1), x(T)$$

$$b(1), b(2), \ldots, b(T-1), b(T)$$

Jacobian matrix

$$b(t) = f^{-1}(x(t)) - Wx(t-1)$$

$$\frac{\partial b_i(t)}{\partial x_j(t')} = \delta_{tt'}\left(D^{-1}(t)\right)_{ij} - W_{ij}\delta_{t-1,t'}$$

$$D(t) = \mathrm{diag}\{f(Wx(t-1)+b(t))\}$$

Chain rule

$$\frac{\partial R}{\partial x_j(t')} = \sum_{i,t'} \frac{\partial R}{\partial b_i(t)} \frac{\partial b_i(t)}{\partial x_j(t')}$$

$$= \sum_i s_i(t')\left(D^{-1}(t')\right)_{ij} - \sum_i s_i(t'+1)W_{ij}$$

$$\frac{\partial R}{\partial x(t)} = D^{-1}(t)s(t) - W^T s(t+1)$$