Other Differentiation Algorithms

Sargur N. Srihari
srihari@buffalo.edu

Topics (Deep Feedforward Networks)

- Overview

1. Example: Learning XOR
2. Gradient-Based Learning
3. Hidden Units
4. Architecture Design
5. Backpropagation and Other Differentiation Algorithms
6. Historical Notes

Topics in Backpropagation

- Forward and Backward Propagation

1. Computational Graphs
2. Chain Rule of Calculus
3. Recursively applying the chain rule to obtain backprop
4. Backpropagation computation in fully-connected MLP
5. Symbol-to-symbol derivatives
6. General backpropagation
7. Ex: backpropagation for MLP training
8. Complications
9. Differentiation outside the deep learning community
10. Higher-order derivatives

9. Differentiation outside the Deep Learning Community

Automatic Differentiation

- Deep learning community has been outside the CS community dealing with automatic differentiation
- The back-propagation algorithm is only one approach to automatic differentiation
- It is a special case of a broader class of techniques called *reverse mode accumulation*

Computational Complexity

- In general, determining the order of evaluation that results in the lowest computational cost is a difficult problem
- Finding the optimal sequence of operations to compute the gradient is NP-complete (Naumann, 2008)
  - in the sense that it may require simplifying algebraic expressions into their least expensive form

Algebraic substitution

- If p_i are probabilities and z_i are unnormalized log probabilities. Suppose $\boxed{q_i = \frac{\exp(z_i)}{\sum_i \exp(z_i)}}$
  - where we build the softmax function out of exponentiation, summation and division, and construct a cross-entropy loss $J = -\Sigma_i\, p_i \log q_i$
- A human mathematician can observe that the derivative of J wrt z_i takes a simple form: $q_i - p_i$
  - whereas backprop propagates gradients through log and exp operations through the original graph
- Theano performs some algebraic substitution to improve over graph proposed by pure backprop

Future differentiation technology

- Backprop is not the only- or optimal-way of computing the gradient, but a practical method for deep learning
- In the future, differentiation technology for deep networks may improve with advances in the broader field of automatic differentiation