

Coordinate Descent

Sargur N. Srihari

srihari@cedar.buffalo.edu

Topics in Optimization for Deep Models

- Importance of Optimization in machine learning
- How learning differs from optimization
- Challenges in neural network optimization
- Basic Optimization Algorithms
- Parameter initialization strategies
- Algorithms with adaptive learning rates
- Approximate second-order methods
- Optimization strategies and meta-algorithms

Topics in Optimization Strategies and Meta-Algorithms

1. Batch Normalization
2. Coordinate Descent
3. Polyak Averaging
4. Supervised Pretraining
5. Designing Models to Aid Optimization
6. Continuation Methods and Curriculum Learning

Solving pieces independently

- It may be possible to solve an optimization problem quickly by breaking it into separate pieces
- Minimize $f(\mathbf{x})$ wrt one variable x_i , then wrt another variable x_j , and so on, repeatedly cycling through all variables, we are guaranteed to arrive at a local minimum
- This is called *coordinate descent*
 - *Block coordinate descent* refers to minimizing wrt a subset of variables

When to use coordinate descent?

- When the different variables can be separated into groups that play relatively isolated roles
- Or when optimization wrt a subset of variables is significantly more efficient than optimization wrt all variables
 - Sparse coding is an example (see next)

Sparse Coding

- Consider cost function

$$J(\mathbf{H}, \mathbf{W}) = \sum_{i,j} |H_{i,j}| + \sum_{i,j} \left(\mathbf{X} - \mathbf{W}^\top \mathbf{H} \right)_{i,j}^2$$

- Goal is to find a weight matrix \mathbf{W} that can linearly decode a matrix of activation values \mathbf{H} to reconstruct the training set \mathbf{X}
- Most applications of sparse coding also involve weight decay or a constraint on the norms of the columns of \mathbf{W}
 - in order to prevent the pathological solution with extremely small \mathbf{H} and large \mathbf{W}

Example of Sparse Coding

- The function J is not convex. However, divide the training inputs into two sets:
 - Dictionary parameters W , Code representations H .
- Minimizing J wrt either one of these sets of variables is a convex problem.
- Thus coordinate descent allows us to use efficient convex optimization algorithms
 - By alternating between optimizing W with H fixed, then optimizing H with W fixed.