Challenges in Neural Network Optimization

Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics

- Importance of Optimization in machine learning
- How learning differs from optimization
- Challenges in neural network optimization
    1. Ill-conditioning
    2. Local minima
    3. Plateaus, saddle points and other flat regions
    4. Cliffs and exploding gradients
    5. Long-term dependencies
    6. Inexact gradients
    7. Poor correspondence between local & global structure
    8. Theoretical limits of optimization
- Basic Algorithms
- Parameter initialization strategies

Optimization is a difficult task

- Traditionally ML has avoided difficulty of general optimization by carefully designing the objective function and constraints to ensure that optimization problem is convex
- When training neural networks, we must confront the nonconvex case
- We summarize challenges in optimization for training deep models

1. Ill-conditioning of the Hessian

- Even when optimizing convex functions one problem is an ill conditioned Hessian matrix, H
  - Very general problem in optimization, convex or not
- Causes SGD to be stuck: even very small steps cause increase in cost function

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}^{(0)}) + (\boldsymbol{x} - \boldsymbol{x}^{(0)})^T \boldsymbol{g} + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(0)})^T H (\boldsymbol{x} - \boldsymbol{x}^{(0)})$$

Substituting $\boldsymbol{x} = \boldsymbol{x}^{(0)} - \varepsilon\, \boldsymbol{g}$

$$f(\boldsymbol{x}^{(0)} - \varepsilon \boldsymbol{g}) \approx f(\boldsymbol{x}^{(0)}) - \varepsilon \boldsymbol{g}^T \boldsymbol{g} + \frac{1}{2}\varepsilon^2 g^T H \boldsymbol{g}$$

  - Gradient descent step of $-\varepsilon\, \boldsymbol{g}$ will add to the cost $-\varepsilon \boldsymbol{g}^T \boldsymbol{g} + \frac{1}{2}\varepsilon^2 g^T H \boldsymbol{g}$
  - Ill conditioning becomes a problem when $\frac{1}{2}\varepsilon^2 g^T H \boldsymbol{g} > \varepsilon \boldsymbol{g}^T \boldsymbol{g}$
  - To determine whether ill-conditioning is detrimental monitor $g^T g$ and $g^T H g$ terms
    - Gradient norm doesn't shrink but $g^T H g$ grows order of magnitude
  - Learning becomes very slow despite a strong gradient

2. Local Minima

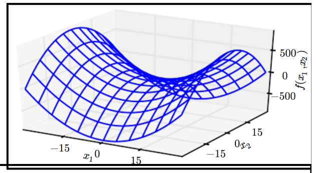
- In convex optimization, problem is one of finding a local minimum
- Some convex functions have a flat region rather than a global minimum point
- Any point within the flat region is acceptable
- With non-convexity of neural nets many local minima are possible
- Many deep models are guaranteed to have an extremely large no. of local minima
- This is not necessarily a major problem

Model Identifiability

- Model is identifiable if large training sample set can rule out all but one setting of parameters
  - Models with latent variables are not identifiable
    - Because we can exchange latent variables
      - If we have m layers with n units each there are $n!^m$ ways of arranging the hidden units
    - This non-identifiability is *weight space symmetry*
  - Another is scaling incoming weights and biases
    - By a factor α and scale outgoing weights by $1/\alpha$
- Even if a neural net has uncountable no. of minima, they are equivalent in cost
  - So not a problematic form of non-convexity

3. Plateaus, Saddle Points etc

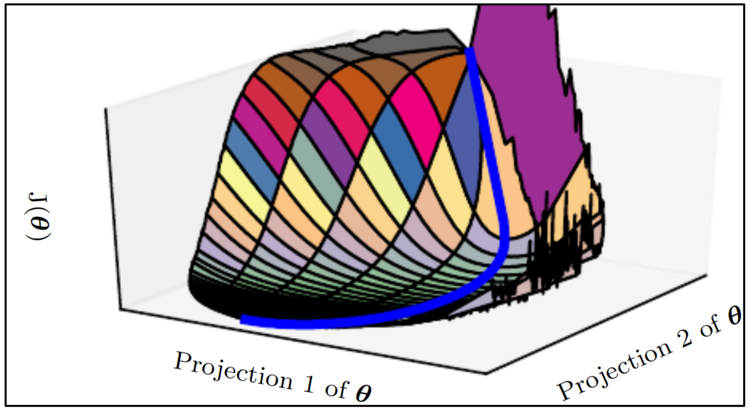
- More common than local minima/maxima are:
  - Another kind of zero gradient points: saddle points
    - At saddle, Hessian has both positive and negative values
      - Positive: cost greater than saddle point
      - Negative values have lower value
    - In low dimensions:
      - Local minima are more common
    - In high dimensions:
      - Local minima are rare, saddle points more common
- For Newton's saddle points pose a problem
  - Explains why second-order methods have not replaced gradient descent



Contains both positive and negative curvature
Function is $f(\boldsymbol{x})=x_1^2-x_2^2$

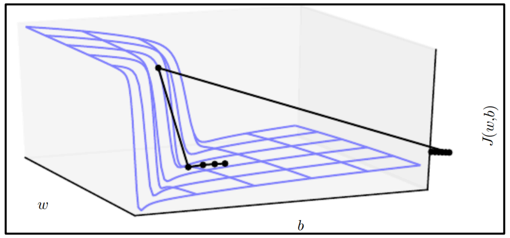
Cost Function of Neural Network

- Visualizations are similar for
  - Feedforward networks
  - Convolutional networks
  - Recurrent networks



- Applied to object recognition and NLP tasks
- Primary obstacle is not multiple minima but saddle points
- Most of training time spent on traversing flat valley of the Hessian matrix or circumnavigating tall "mountain" via an indirect arcing path

4.Cliffs and Exploding Gradients



- Neural networks with many layers
  - Have steep regions resembling cliffs
    - Result from multiplying several large weights
    - E.g., RNNs with many factors at each time step
- Gradient update step can move parameters extremely far, jumping off cliff altogether
- Cliffs dangerous from either direction
- *Gradient clipping* heuristics can be used

5. Long-Term Dependencies

- When computational graphs become extremely deep, as with
  - feed-forward networks with many layers
  - RNNs which construct deep computational graphs by repeatedly applying the same operation at each time step
- Repeated application of same parameters gives rise to difficulties
- Discussed further with RNNs in 10.7

6. Inexact Gradients

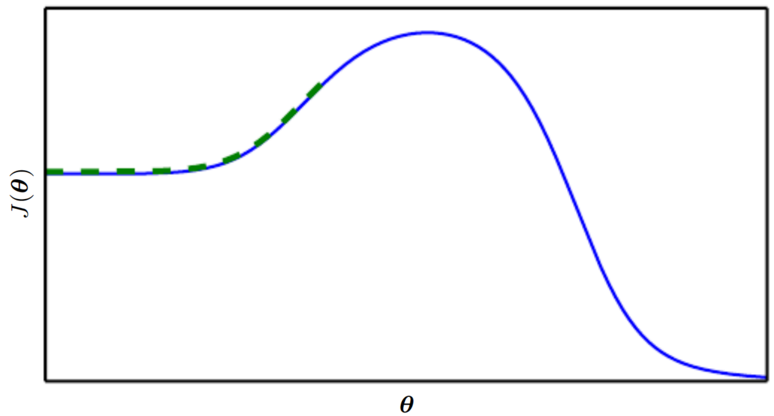
- Optimization algorithms assume we have access to exact gradient or Hessian matrix
- In practice we have a noisy or biased estimate
  - Every deep learning algorithm relies on sampling-based estimates
    - In using minibatch of training examples
  - In other case, objective function is intractable
    - In which case gradient is intractable as well
    - Contrastive Divergence gives a technique for approximating the gradient of the intractable log-likelihood of a Boltzmann machine

7. Poor Correspondence between Local and Global Structure

- It can be difficult to make a single step if:
  - $J(\boldsymbol{\theta})$ is poorly conditioned at the current point $\boldsymbol{\theta}$
  - $\boldsymbol{\theta}$ lies on a cliff
  - $\boldsymbol{\theta}$ is a saddle point hiding the opportunity to make progress downhill from the gradient
- It is possible to overcome all these problems and still perform poorly
  - if the direction that makes most improvement locally does not point towards distant regions of much lower cost

Need for good initial points

- Optimization based on local downhill moves can fail if local surface does not point towards the global solution
- Research directions are aimed at finding good initial points for problems with a difficult global structure
  - Ex: no saddle points or local minima
    - Trajectory of circumventing such mountains may be long and result in excessive training time



8. Theoretical Limits of Optimization

- There are limits on the performance of any optimization algorithm we might design for neural networks
- These results have little bearing on the use of neural networks in practice
  - Some apply only to networks that output discrete values
    - Most neural networks output smoothly increasing values
  - Some show that there exist problem classes that are intractable
    - But difficult to tell whether problem falls in thet class