THEORY OF MACHINE LEARNING

LECTURE 21

GRADIENT DESCENT FOR NN: NEURAL TANGENT KERNEL

ANNOUNCEMENTS

- **Homework 3 due on Monday April 11**
- Ideas
  - Implicit regularization
  - Stability in optimization
  - Depth vs width (converse)
  - Learning parities – earliest lower bound for deep learning
- Last lecture's discussion on feature learning

NEURAL NETWORK TRAINING

- **Question (supervised learning):** given data $(x_1, y_1), (x_2, y_2), \ldots$ from some distribution D, find *h* (with given "architecture") that minimizes the risk
  - Really hard theoretically (even if inputs Gaussian and risk zero is achievable)
  - In practice, solved via gradient descent
    [fast implementation (backprop) by Rumelhart, Hinton, Williams]

- Question for today:
  - How can one analyze dynamics of gradient descent?
  - Are there cases where we can reason about resulting solution?

OVER-PARAMETRIZATION

- **Observation:** modern deep nets pretty overparametrized, but they still don't overfit
- Question out of desperation: Is GD easier to analyze when network is "heavily" overparametrized?

Theorem. [Jacot, Gabriel, Hongler 18] [Arora, et al. 2019] A width ~ n^3 network (any number of layers) trained via GD from random initialization achieves zero training error. Moreover, the final solution is equivalent to solving a "Kernel regression" problem with a specific kernel.

ASIDE: KERNEL REGRESSION

- Ubiquitous motivation: function value known at a bunch of points, "interpolate" to rest of space
  - One way to think of all of ML!
- Suppose K defines "point similarity" K(x,y)
- Consider interpolation via functions of specific form…

NEURAL NET TRAINING REVISITED

KERNEL IN THE INFINITE WIDTH LIMIT