THEORY OF MACHINE LEARNING

LECTURE 9

INTRODUCTION TO OPTIMIZATION

SO FAR IN THE COURSE

- Reasoning about learning
  - Learning and generalization (distribution D)
  - Inductive bias and necessity of choosing a hypothesis class (H)
  - PAC learnability of a class (given examples, find hypothesis that is as good as the best in class); agnostic, improper
- Growth function of a class $\tau_H(m)$; small growth function => random sample is representative => class is learnable
- Fundamental theorem: VC dimension "captures" learnability
- Homework 1

FUNDAMENTAL THEOREM OF (STAT) LEARNING THEORY

- Theorem: The following statements are *equivalent:*
  - Class H is PAC learnable
  - Class H is *agnostically* PAC learnable
  - Class H has finite VC dimension

RECALL: IMPLICATIONS

- If H has infinite VC dimension, it is **not** PAC learnable! (same proof as no-free-lunch theorem – homework)
- Agnostic case usually as hard as *realizable* case
- **Caveat**. Generalization guarantees only apply to ERM, not (say) to an improper learner
- ERM is all you need, assuming you have enough samples
  - Doing ERM efficiently is a challenge
  - Loss + optimization is the standard approach

OPTIMIZATION

SOLVING ERM EFFICIENTLY

RECAP: ERM IS OFTEN HARD WITH FINITE CLASSES

- Finding best linear classifier (fewest mistakes) is NP hard even to approximate!

- Common remedy: loss functions
- Discrete optimization (mistake bounds) to continuous

LOSS MINIMIZATION

- Given training examples $(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)$, find hypothesis h to minimize "loss"

- Examples: square loss, l1 loss, logistic loss, …
- Parametrize h
- Not all loss functions can be minimized efficiently!

CONVEX OPTIMIZATION

- **Problem.** Given a convex function defined over a convex domain, find the minimizer (or min value).

TAYLOR APPROXIMATION

- Functions over R^d, gradients, Hessian
- First order approximation

GRADIENT DESCENT ALGORITHM

- Very general approach

NATURAL ISSUES

- "Learning rate"
- Staying in the domain

GRADIENT DESCENT – VANILLA ANALYSIS

- Consider 'f' that is L-Lipschitz, fixed learning rate, domain all of R^d
- Suppose OPT was distance B away from initial point