THEORY OF MACHINE LEARNING

LECTURE 11

CONVEX OPTIMIZATION, GRADIENT DESCENT

RECAP: LOSS MINIMIZATION

- ERM is hard, so we consider minimization of loss
- General problem

- Optimization can be hard in general, we study "easy" case of convex optimization
- Min f(x) over D, where f is convex, domain D is convex
- Minimization is important (max can be hard)

RECAP: CONVEX OPTIMIZATION

- **Problem.** Given a convex function defined over a convex domain, find the minimizer (or min value).
- $f(tx + (1-t)y) \le t\,f(x) + (1-t)f(y)$ for all $t \in (0,1)$ and $x, y \in D$

- Local opt = global opt (just due to convexity)
- <u>Question:</u> how to find a "locally better" point? (assume f is continuous, differentiable)
- Gradient descent – inspired by Taylor approximation

GRADIENT DESCENT ALGORITHM

- Generally applicable – even to non-convex functions
  (in which case you only find local opt)

- Choosing how much to move! (aka *learning rate*)
- Staying in the domain

VANILLA ANALYSIS

- Suppose f is L-Lipschitz, and domain $D = R^d$
- Suppose OPT was distance B away from initial point
- **Theorem.** Consider running T steps of gradient descent with a fixed learning rate η. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w) \leq \frac{B^2}{2\eta T} + \frac{\rho^2 \eta}{2}$$

Proof uses "basic inequality" of convexity

BASIC INEQUALITY, POTENTIAL FUNCTION ANALYSIS

DEALING WITH THE DOMAIN – PROJECTED GD

EXTENSIONS

- What if function is "smooth"? Get improved 'rate'
- What if function is "strongly convex"?
- What if functions at different steps are different? (!)