THEORY OF MACHINE LEARNING

LECTURE 10

CONVEX OPTIMIZATION, GRADIENT DESCENT

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
- Many candidate loss functions

RECAP: CONVEX OPTIMIZATION

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

- Local opt = global opt (just due to convexity)
- Question: how to find a "locally better" point? (assume f is continuous, differentiable)

TAYLOR APPROXIMATION

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

GRADIENT DESCENT ALGORITHM

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

NATURAL ISSUES

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

GRADIENT DESCENT – 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

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

ALTERNATE DEFINITION OF CONVEXITY

- Function lies “above” the tangent plane *at any point* !
- Related to the definition of convex functions/bodies via half spaces

GRADIENT DESCENT – 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

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

ANALYSIS VIA POTENTIAL FUNCTIONS

DEALING WITH THE DOMAIN – PROJECTED GD