THEORY OF MACHINE LEARNING

LECTURE 12

GRADIENT DESCENT VARIANTS

RECAP: CONVEX OPTIMIZATION

- **Problem.** Given a convex function defined over a convex domain, find the minimizer (or min value).
- Gradient descent – inspired by Taylor approximation
  - Start with some feasible w_0
  - For t = 0, 1, …, T-1, set $w_{t+1} = w_t - \eta \nabla f(w_t)$
- How do you set/"tune" the learning rate?
- Staying feasible

GRADIENT DESCENT ANALYSIS

- Assume f is L Lipschitz, domain is all of R^d, $|w_0 - w^*| \leq B$
- Use "basic inequality" about convex functions, for any t,

$$f(w^*) \geq f(w_t) + \langle w^* - w_t, \nabla f(w_t) \rangle$$

- Use the potential function $\Phi_t = |w_t - w^*|^2$
- Note that $\Phi_t - \Phi_{t+1}$ (potential drop) is <u>lower bounded</u> by how far $f(w_t)$ is from $f(w^*)$

BASIC THEOREM

- **Theorem.** Consider running T steps of gradient descent with a fixed learning rate η. Then we have

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

- Same proof works if we had a constrained domain. Simply project iterates to feasible set
- Proof works even if functions at different time steps were different!

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

APPLICATIONS

- Online convex optimization
- Stochastic gradient descent

EXTENSIONS – MORE STRUCTURE ON FUNCTION

- What if function is "smooth"? Get improved rate
- What if function is "strongly convex"?