THEORY OF MACHINE LEARNING

LECTURE 13

GRADIENT DESCENT, THOUGHTS

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

BASIC THEOREM

- Assume f is L Lipschitz, domain is all of R^d, $|w_0 - w^*| \leq B$
- **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
- Proof works even if functions at different time steps were different!

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

ANALYSIS

- 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 $f(w^*)$
- $\Phi_t - \Phi_{t+1} \geq 2\eta \left(f(w_t) - f(w^*)\right) - \eta^2 |\nabla f(w_t)|^2$
- Summing over t gives the bound
- Applications to online convex optimization, SGD

STOCHASTIC GRADIENT DESCENT

- Consider the setting where the function f can be decomposed as

$$f(w) = \frac{1}{m} \sum_i g_i(w)$$

- In iteration t, pick index i_t uniformly at random and take a gradient step, i.e., $w_{t+1} = w_t - \eta \nabla g_{i_t}(w_t)$
- Now w_t is a random variable, and we need to argue about $E[\, f(w_t)]$
- We still have that the expected potential drop at step t (conditioned on any trajectory so far) is $2\eta\big(f(w_t) - f(w^*)\big) - \eta^2 L^2$
- Earlier bound holds <u>in expectation</u>

NOISY GRADIENT DESCENT

- Consider the setting where we perform gradient descent on the function f using a "noisy gradient oracle"
- Given a point w, suppose we get "noisy gradient"
- Same bound holds

ADDITIONAL STRUCTURE ON FUNCTIONS

- **Smoothness** – function is M smooth if *gradient* is M-Lipschitz
- **Key observation:** in this case, <u>every iteration</u> yields drop in function value!

- After T steps, $\sum_t |\nabla f(w_t)|^2$ is bounded by $4M\,(f(w_0) - f(w^*))$
- Convergence rate of 1/T

NONCONVEX (SMOOTH) FUNCTIONS

ADDITIONAL STRUCTURE ON FUNCTIONS

- **Smoothness**: function is M smooth if *gradient* is M-Lipschitz
- **Strongly convex:** function is m-strongly convex if we have a "lower bound" via a parabola

IMPROVEMENTS, GENERALIZATIONS

- **Polyak-Lojasiewicz inequality**: suppose f satisfies:
  $|\nabla f(w)|^2 \geq c(f(w) - f(w^*))$ for all w
- "Global" condition, but can be satisfied for non-convex f

- Polyak's "heavy ball" method (momentum)
- AdaGrad and related methods
- Second order (Newton) methods
- …