THEORY OF MACHINE LEARNING

LECTURE 14

GRADIENT DESCENT – SMOOTH, STRONGLY CONVEX

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
- 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 $\Phi_t = |w_t - w^*|^2$

NOISY GRADIENT DESCENT (GENERALIZES SGD)

- Doing gradient descent on f using a "noisy gradient oracle"
- Given a point w, suppose we get "noisy gradient"
- Same bound holds assuming noise is unbiased, and has low variance

ADDITIONAL STRUCTURE: SMOOTHNESS

- **Smoothness** – function is M smooth if *gradient* is M-Lipschitz
- **Key observation:** in this case, every iteration yields drop in function value (first order approx. is accurate in ball of radius < 1/2M)

- After T steps, $\sum_t |\nabla f(w_t)|^2$ is bounded by $4M\ (f(w_0) - f(w^*))$
- Convergence rate of 1/T
- GD on smooth non-convex functions converges to "approximately singular" points

CAN WE GO BEYOND 1/T CONVERGENCE?

- **Smoothness**: function is M smooth if *gradient* is M-Lipschitz
- Purely assuming smoothness, can get rate of 1/T^2 (Nesterov 1983)

[Optimal for all "gradient based" methods]

STRONG CONVEXITY

- **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

PRECONDITIONING

- **Hessian** plays informal role in most GD analyses (M-smooth)

- “Directions” of Hessian can matter
- Optimal movement using second order information

IMPROVEMENTS, GENERALIZATIONS

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