THEORY OF MACHINE LEARNING

LECTURE 15

STRONG CONVEXITY, REGULARIZATION, STABILITY

SUMMARY OF GRADIENT DESCENT

- Argmin f(x), over $x \in D$, where D is a convex domain
- Simple iterative algorithm ("first order")
- Fixed step-size
  - Convergence with error $O(\frac{1}{\sqrt{T}})$ after T steps for any L-Lipschitz function
  - "Noisy gradient oracle" → stochastic gradient descent
  - Error of O(1/T) for "smooth" convex functions (derivative is M-Lipschitz), assuming step size < $\frac{1}{2M}$
  - If function is also *strongly convex* with parameter μ, convergence bound improves to roughly $\exp(-\frac{\mu}{M}T)$ (extends to Polyak-Lojasiewicz)

“OPTIMAL” BOUNDS

- Turns out: under just the Lipschitz assumption, $\frac{1}{\sqrt{T}}$ cannot be improved, at least with “sub-gradient oracle”
- **Smoothness:** purely assuming smoothness, can get rate of 1/T^2 (Nesterov 1983), this is optimal for all “gradient based” methods
- Can we use information *beyond* the gradient?

PRECONDITIONING

- **Hessian** plays informal role in most GD analyses (even 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
- …

STRONG CONVEXITY, MOTIVATION

- Saw that strong convexity leads to “faster” optimization

- Additional benefit – “stability” to small perturbation
  - Example of quadratic

STABILITY OF LOSS MINIMIZATION

- Loss minimization with 'n' examples

- What happens if one example is "replaced"?

STABILITY IMPLIES GENERALIZATION

- Recall the notion of "generalization gap"
  - Can we phrase it in terms of stability?

- Stability versus "utility"!