THEORY OF MACHINE LEARNING

LECTURE 16

REGULARIZATION, STABILITY

SUMMARY OF GRADIENT DESCENT

- 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)
- Nesterov's "acceleration", preconditioning via the Hessian, or by using first order proxies (AdaGrad), momentum

IMPROVEMENTS, GENERALIZATIONS

- Polyak's "heavy ball" method (momentum)
  - Originally designed for strongly convex functions – achieves $\sqrt{\kappa}$ in exponent
- Second order methods, first order "proxies" (AdaGrad)
- Theme: avoid "slow" convergence – take large steps when possible
  - Non-convex functions – "slip out" of local minima
  - Perturbed gradient descent -- if you're not moving much via gradient descent, just make a "random jump" to a point in a neighborhood
  - Can prove formally that you get out of "bad saddles"

MANY VARIANTS OF GD

ML Hipster
@ML_Hipster

"Oh sure, going in that direction will totally minimize the objective function" —Sarcastic Gradient Descent.

CHOOSING LOSS FUNCTIONS

- Saw that "smoother" loss functions lead to "faster" optimization
- Utility versus niceness

- Today's topic
  - "Nice" loss functions come with added benefit: "stability" to input changes
  - Example of quadratic
  - Stability is a form of "simplicity" => generalization

STABILITY OF A LOSS MINIMIZATION ALGORITHM

- Given examples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, do loss minimization
- Can be viewed as map from examples -> parameters w
- How does changing a single (x_i, y_i) change the w?

UNDERSTANDING STABILITY -- LINEAR FUNCTIONS

- Suppose we are optimizing over $w \in [-1,1]$
- Consider sequence of functions:

  w, -2w, 2w, -2w, 2w, …

STABILITY IMPLIES GENERALIZATION

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

CONCENTRATION BOUNDS AND STABILITY

- [Talagrand '80s], [Boucheron, Lugosi, Massart], [Efron-Stein 60s]