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", <u>preconditioning</u> via the Hessian, or by using first order proxies (AdaGrad), momentum

IMPROVEMENTS, GENERALIZATIONS

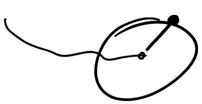
$\exp\left(-\frac{\mu}{M}\cdot T\right)$ $1/\kappa$

$w_{t+1} = w_t - \eta \nabla f(w_t) - \delta(w_t - w_{t+1})$

$\exp\left(-\frac{1}{\sqrt{\kappa}}\cdot T\right)$

- 1960s. 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

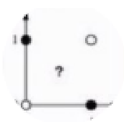
  - Perturbed gradient descent -- if you're not moving much via gradient descent, just make a "random jump" to a point in a neighborhood
  - Can prove formally that you get out of "bad saddles" [Chi Jin, Rong Ge, M. Jordan '17.]



You end up at "approximately" local minima.

$\nabla^2 f \succeq 0$. $\succeq -\delta I$.

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

earlier: binary loss $\to$ logistic loss.

(HW)

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

  Fit

  dataset: $(x_1, y_1), (x_2, y_2) \ldots, (x_m, y_m)$

  $$\text{Loss}(\omega) = \left(\frac{1}{m} \cdot \sum_{i=1}^{m} \ell(\omega, x_i, y_i)\right) + \theta \|\omega\|^2$$

  $$\ell'(\omega, x_i, y_i) = \ell + \theta \|\omega\|^2$$

- 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 $\Leftarrow$ Concentration).

Common Regularizers:
- ℓ_p regulari[zers].
- entropy. reg.
- Log-barrier.
- "Self-concordant" fns.

STABILITY OF A LOSS MINIMIZATION ALGORITHM

(deterministic training procedures).

- Given examples $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$, do loss minimization

$$\underline{L(w)} = \frac{1}{m} \cdot \sum_{i=1}^{m} \ell(w; x_i, y_i) \longrightarrow \arg\min_{w} L(w) = w^*.$$

- Can be viewed as map from examples -> parameters w

$$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\} \longrightarrow w^*(S)$$

- How does changing a single $\underline{(x_i, y_i)}$ change the w^*?

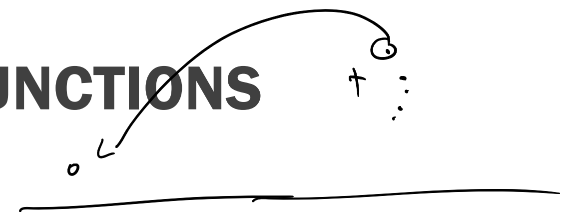
$$S' = \{(x_1, y_1), (x_2, y_2), \dots, (x_{m-1}, y_{m-1}), (x'_m, y'_m)\} \longrightarrow w^*(S')$$

Def. '<u>Loss function is $\delta^{(m)}$-stable</u> (for training sets of size m) if <u>$\forall\, x, y, x'_m, y'_m$ as above</u>,

$$\|w^*(S') - w^*(S)\| \leq \delta(m).$$

$$|L(w^*(S')) - L(w^*(m))| \leq \frac{1}{m}.$$

UNDERSTANDING STABILITY -- LINEAR FUNCTIONS



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

w, -2w, 2w, -2w, 2w, …

loss fn for example 3, . . .

$-2w$.

loss fn for example 2

$l_2(w)$

$l_1(w)$

$$L(w) = \frac{1}{m} \cdot \sum_{i=1}^{m} l_i(w)$$

$$\underset{w}{\mathrm{argmin}}\ L(w) = \begin{cases} -1 & \text{if } m \text{ is even} \\ 1 & \text{if } m \text{ is odd.} \end{cases}$$

In general, changing one of the l_i's can significantly change the w^*.

$\boxed{\text{Theorem:}}$ Suppose $L(w)$ is $\boxed{\alpha\text{-strongly convex}}$ and we replace $l_i(w)$ with $l_i'(w)$, such that $\|\nabla(l_i' - l_i)\| \le G$, say. Then $|w^*(S') - w^*(S)| \le \frac{G}{m\alpha}$

STABILITY IMPLIES GENERALIZATION

- Suppose our loss fn $\ell(w; x)$ is $\delta(m)$-stable; then generalization error is $\leq \delta(m)$.

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

data $S \equiv x_1 \quad x_2 \quad \dots \quad x_{m-1} \quad x_m \xrightarrow{\text{train}} w^*(S)$

data $S' \equiv x_1 \quad x_2 \quad x_{m-1} \quad x_m' \xrightarrow{\text{train}} w^*(S')$

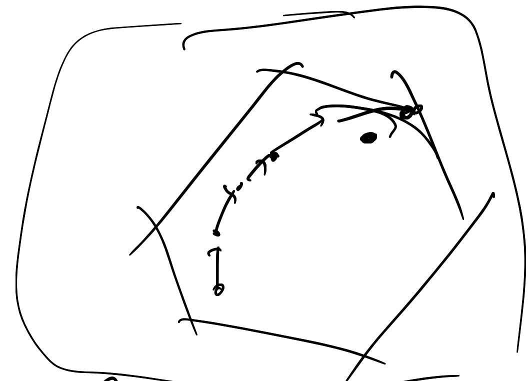
Generalization: (data is from distn $\mathcal{D}$)

$|\text{Loss on test} - \text{Loss on training}|$

loss on test of $w^*(S)$ $:= \mathbb{E}_{x_m' \sim \mathcal{D}} \, \ell(w^*(S), x_m')$

$\mathbb{E}_{i \sim [m]} \, \ell(w^*(S), x_i) \rightsquigarrow \text{I}$.

$$\mathrm{I} \equiv \mathbb{E}_{i \sim [m]} \, \ell\big(w^*(S), x_i\big) \approx \underbrace{\mathbb{E}_{i', i \sim [m]} \, \ell\big(w^*(S \setminus x_i + x_i'), x_i\big)}_{\delta(m)}$$



Regularization:

(1) Improves generalization gap. ("~~fa~~ fake win" sometimes, because you might have high error before reg.).

(2) Improves rate of convergence.

CONCENTRATION BOUNDS AND STABILITY

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

Chernoff: $X_1, X_2, \ldots, X_n \sim$ iid from $\mathcal{D}$ with support $[a,b]$.
& mean μ

$$\Pr\left[\left|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\right| > t\right] \le e^{-t^2 n/(a-b)^2}$$

$$\frac{|a-b|}{n}.$$

Talagrand: Let f be any fn which is δ-stable, ie,

$$|f(X_1, X_2, \ldots X_n) - f(X_1, X_2, \ldots, X_i', X_{i+1}, \ldots X_n)| \le \delta.$$

Then if $X_1, \ldots, X_n$ are iid r.v.s; then

$$\Pr\left[|f(X_1, \ldots, X_n) - \text{median}| \geq t\right] \le e^{-t^2 n/\delta^2}$$