Lipschitzness $\forall x,y$

$$L|x-y| \geq |f(x)-f(y)|$$

Smoothness f is β-smooth if ∇f is β-Lipshitz: $\forall x,y$

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x-y\|$$

Ex: $f(x)=x^2$ is 2-smooth over $\mathbb{R}$, $g(x)=x^3$ isn't c-smooth over $\mathbb{R}$ for any constant c.

Convexity $\forall \alpha \in (0,1)$ $\forall x,y \in C$

- $f(\alpha x - (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$

(if differentiable, $\forall x,y$)

$$f(x) \geq f(y) + \langle \nabla f(y), x-y \rangle$$

Goal: Is loss min. a good plan even if we don't have finite VC?

A: Sometimes!

Consider regularized loss minimization:

$$A(S) = \underset{w}{\operatorname{argmin}}\ L_S(w) + R(w)$$

where R is a regularizer of w. We'll focus on ℓ^2 reg, $R(w) = \lambda \|w\|^2$

Example: ℓ^2-regularized logistic regression

$$A(S) = \underset{w}{\operatorname{argmin}} \left[\frac{1}{m} \sum_i \log(1+\exp(-y_i \langle w, x_i \rangle)) + \lambda \|w\|^2 \right]$$

- No closed form solution, but we can solve efficiently.
- This is strictly convex

Theorem. Fix D over $\mathcal{X} \times [-1,1]$, $\mathcal{X} = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$. Let $\mathcal{H} = \{w \in \mathbb{R}^d \mid \|w\| \leq B\}$. For any $\epsilon \in (0,1)$, $m \geq \frac{8B^2}{\epsilon^2}$, $\lambda = \frac{\epsilon}{2B^2}$, ℓ^2-regularized logistic regression satisfies

$$\mathbb{E}_S [L_D(A_{RLM}(S))] \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon$$

[A similar statement about ridge regression ...]

Why? Stability of regularized learning.

$S = (z_1, \dots, z_m)$
$S^{(i)} = (z_1, \dots z_i', \dots z_m)$

what is $\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)$?

(Prolly) ≥ 0

If large, model is likely overfitting on z_i.

Def: On-average replace one stability An alg A is $\epsilon(m)$ on-avg replacement stable if, $\forall D$

$$\mathbb{E}_{S, z', i}[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)] \leq \epsilon(m).$$

Thm Fix $S = (z_1, \dots, z_m)$ & $z_i' \sim D$. Let $U[m]$ be uniform d. over $[m]$. Then $\forall$ Algorithms A,

$$\underbrace{\mathbb{E}_S[L_D(A(S)) - L_S(A(S))]}_{\text{Generalization gap}} \leq \mathbb{E}_{S, z_i', i \sim U[m]}[\ell(A(S^{(i)}), z_i] - \ell(A(S), z_i)]$$

Pf

$$\mathbb{E}_S[L_S(A(S))] = \mathbb{E}_{S,i}[\ell(A(S), z_i)]$$

$$\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_{S,z'}[\ell(A(S), z')] \quad \perp \text{ of } S, z'$$

$$= \mathbb{E}_{S,z'}[\ell(A(S^{(i)}), z')] \quad \forall i \in [m]$$

z_i

$$=$$ since $S^{(i)}$ & z_i, z' are $\perp$

So, On-avg replacement stability $\Rightarrow$ Bd on gen. gap.

We'll show that RLM is OARS if ℓ is convex & smooth or Lipschitz.

Def (Strong Convexity) f is λ-strongly convex if $\forall x, y$ $\alpha \in (0,1)$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\lambda \alpha (1-\alpha)}{2}\|x-y\|^2$$

(λ increasing makes this more difficult to satisfy, $\lambda = 0$ all convex fns satisfy)

Note: if f is convex & g is λ strongly Convex, $f+g$ is λ-SC.

$$A(S) = \underset{w}{\operatorname{argmin}} \underbrace{L(S) + \lambda \|w\|^2}_{f(S)}$$

So, RLM for a convex L is 2λ-SC.

Prop If f is λ-SC, w^* minimizer of f, then $\forall w$

$$f(w) - f(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

Pf:

$$f(w^*) \underset{\text{(minimizer)}}{\leq} f(\alpha w + (1-\alpha) w^*)$$

$$\leq \alpha f(w) + (1-\alpha) f(w^*) - \frac{\lambda \alpha (1-\alpha)}{2} \|w^* - w\|^2$$

(By basic alg $\Rightarrow$ $f(w^*) \leq f(w) - \frac{\lambda}{2} (1-\alpha) \|w^* - w\|^2$

and as $\alpha \to^+ 0$, ✓

□

Let's bound $\| A(S) - A(S^{(i)}) \|$ from above & below

or $\| f_S(A(S)) - f_S(A(S^{(i)})) \|$ regularized loss

LB: Since f_S is 2λ-SC, $f_S(A(S^{(i)})) - f_S(A(S)) \geq \lambda \| A(S^{(i)}) - A(S) \|^2$

$w^* = A(S)$

UB:

$$f_S(A(S^{(i)})) - f_S(A(S)) = L_S(\hat{w}^{(i)}) + \lambda \|\hat{w}^{(i)}\|^2 - \left(L_S(\hat{w}) + \lambda \|\hat{w}\|^2 \right)$$

$$= \underbrace{L_{S^{(i)}}(\hat{w}^{(i)}) + \lambda \|\hat{w}^{(i)}\|^2 - \left[L_{S^{(i)}}(\hat{w}) + \lambda \|\hat{w}\|^2 \right]}_{\leq 0}$$

$$+ \frac{1}{m} \left[\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}^{(i)}, z') + \ell(\hat{w}, z') - \ell(\hat{w}, z_i) \right]$$

$$\leq \frac{1}{m} \left[\ell(A(S^{(i)}), z_i) - \ell(A(S^{(i)}), z') + \ell(A(S), z') - \ell(A(S), z_i) \right] \quad (1)$$

When $\ell(\cdot, z)$ is ρ-Lipschitz $\forall z$,

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \le \rho(\|A(S^{(i)}) - A(S)\|$$

& same for z'. So, along w (1)

$$\lambda \|A(S^{(i)}) - A(S)\|^2 \le \frac{2\rho}{m} \|A(S^{(i)}) - A(S)\|$$

So $$\|A(S^{(i)}) - A(S)\| \le \frac{2\rho}{\lambda m}$$

& $$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \le \frac{2\rho^2}{\lambda m}$$

$\Rightarrow$ when ℓ is convex, ρ-Lipschitz, RLM w. $R(\omega) = \lambda\|\omega\|^2$ is $\frac{2\rho^2}{m}$ - OAROS stable.

So its exp. gen gap is $\le \frac{2\rho^2}{m}$.

[Similar argument when ℓ is β smooth but $\frac{48\beta}{\lambda m}$ UB instead

---

Overall loss:

$$\mathbb{E}_S[L_D(A(S)] = \mathbb{E}_S[L_S(A(S))] + \underbrace{\text{Gen-GAP}}_{\le \text{ stability}}$$

$$\le \mathbb{E}_S[L_S(A(S)) + \lambda\|A(S)\|^2] + \frac{2\rho^2}{\lambda m}$$

for any w^* $$\le \mathbb{E}_S[L_S(w^*) + \lambda\|w^*\|^2] + \underline{\qquad}$$

$$= L_D(w^*) + \lambda \|w^*\| + \frac{2\rho^2}{\lambda m}$$

Corollary If $\ell(\cdot\, z)$ is convex + ρ-lipschitz $\forall z$ & $\|w^*\| \le B$, then $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$

RLM for $R(w) = \lambda \|w\|^2$ satisfies

$$\mathbb{E}_S[L_D(A(S))] \le \min_{w \in \mathcal{H}} L_D(w) + \underbrace{\rho B \sqrt{\frac{8}{m}}}_{\le \epsilon}$$

or, if $m \ge \frac{8\rho^2 B^2}{\epsilon^2}$

$\le \epsilon$

We'll now switch to starting to talking more about knowing how good our predictions are (in a more p'twise sense)

Let's start by recalling squared error of a predictor f (also called the Brier Score)

$$B(f, D) = \mathbb{E}_{(x,y)\sim D}\left[(f(x) - y)^2\right]$$

Let's consider regression, where $f: \mathcal{X} \to [0,1]$

We'd like to learn f^* : $f^*(x) = \mathbb{E}_{y|x}[y]$ (Bayes-OPT)

which we can't do,

---

but maybe we can have marginal mean consistency w. error α:

$$\left| \mathbb{E}_{x\sim D_x}[f(x)] - \mathbb{E}_{(x,y)}[y] \right| \leq \alpha$$

Since it's only a statement abt global f-avgs

If f isn't α-MMC, it's easy to fix: Let

$$\Delta: \mathbb{E}_{(x,y)\sim D}[y] - \mathbb{E}_{x\sim D_x}[f(x)]$$

Then, $\hat{f}(x) = f(x) + \Delta$ is MMC.

But did we make $f \to \hat{f}$ worse in some way?

Lemma: Fix any D, $f: X \to [0,1]$, and $\Delta, \hat{f}$ as above. Then,

$$\ell_2(\hat{f}) = B(\hat{f}, D) = B(f, D) - \Delta^2$$

Eg, the squared loss is lower than for f!

Proof

$$B(\hat{f},D) - B(f,D) = \mathbb{E}[(f(x)-y)^2 - (\hat{f}(x)-y)^2]$$

$$= \mathbb{E}[f^2(x) - 2f(x)y + y^2 - \{\hat{f}^2(x) - 2\hat{f}(x)y + y^2\}]$$

$$= \mathbb{E}[f^2(x) - 2f(x)y - [f(x)+\Delta]^2 + 2[f(x)+\Delta]y]$$

$$= \mathbb{E}[-2f(x)y - 2\Delta f(x) - \Delta^2 + 2f(x)y + 2\Delta y]$$

$$= \mathbb{E}[-2\Delta f(x) - \Delta^2 + 2\Delta y]$$

$$= 2\Delta \underbrace{\mathbb{E}[y - f(x)]}_{\Delta} - \Delta^2$$

$$= \Delta^2$$

Ok, so marginal mean consistency of f isn't hard (or costly in terms of MSE).

What other things about $(x,y) \sim D$ can we find models to satisfy/learn?

What about quantiles? τ is a q-quantile of D if

$$\Pr_{y}[y \leq \tau] = q$$

"q fraction of D is below τ"

It'd be rad to find f that was conditionally (on x) a q-quantile, generally hard/imp. But can ask for it marginally:

Def f has marginal quantile consistency error α for target quantile q if

$$\left| \Pr_{(x,y) \sim D}[y \leq f(x)] - q \right| \leq \alpha$$

Squared error : marginal mean consistency

Pinball loss : quantile

Def: The pinball loss function for quantile q:

$$L_q(\tau, y) = \begin{cases} (y-\tau)q & y > \tau \\ (\tau - y)(1-q) & y \leq \tau \end{cases}$$

For D & f,

$$PB_q(f, D) = \mathbb{E}_{(x,y) \sim D}[L_q(f(x), y)]$$

Lemma: For any continuous distribution every $0 \le q \le 1$

$$\tau_q = \underset{\tau \in [0,1]}{\operatorname{argmin}} \mathbb{E}_y [L_q(\tau, y)]$$

is a q-quantile.

$\int () [(y-\tau) q] \cdot dy$

$\int ( ) [\tau - y)(1-q) \cdot dy$

is continuous

Pf Since y is continuous & actually convex in τ so is min. at a pt where it's subderivative is 0

$$\frac{d \, \mathbb{E}_y [L_q(\tau, y)]}{d\tau} = \mathbb{E}_y [(1-q) \, \mathbb{1}[y \le \tau] - q \, \mathbb{1}[y > \tau]]$$

$$= \mathbb{E}_y [\mathbb{1}[y \le \tau] - q]$$

$$= \Pr[y \le \tau] - q$$

$= 0$ where τ is a q-quantile.

Can we "patch"/fix f to be q-quantile consistent like we did w. means?

Say f is α-violating q-quantile consistency

Define Δ: $\Pr[y \le f(x) + \Delta] = q$ [$\exists$ since y is continuous]

$\hat{f}(x) = f(x) + \Delta$ $\hookrightarrow$ now MQC

{And only has lower pinball loss.}

But its improvement in pinball loss depends on density btwn f & $\hat{f}$

Recall ρ-Lipschitzness applied to cond label distribution $D(x)$ says $\forall 0 \le \tau \le \tau' \le 1$

$$\Pr_{y \sim D(x)}[y \le \tau'] - \Pr_{y \sim D(x)}[y \le \tau] \le \rho[\tau - \tau']$$

The whole of the label dist D is ρ-Lipschitz if $\forall x$, $D(x)$ is ρ-Lipschitz. (we just need marginal Lipschitzness, not $\forall x$)

Lemma: Fix any D which is continuous & ρ-Lipschitz. If f has marginal consistency err α wrt q, Δ as above, $\hat{f} = f + \Delta$ then

$$PB_q(\hat{f}) \le PB_q(f) - \frac{\alpha^2}{2\rho}$$

and

$$PB(f) \le PB(\hat{f}) + |\Delta|\alpha - \frac{\alpha^2}{2\rho}$$

This argument is messier (more calculus / geometry) but morally similar (see p14-15 in Uncertainty notes).

Some overview of argument

$$\frac{d\, PB_q(f(x)+\tau)}{d\,\tau} = \mathbb{E}_x\left[\frac{d\,\mathbb{E}_{y\sim D_x} L_q(f(x)+\tau, y)}{d\tau}\right.$$

$$= \mathbb{E}_x[\Pr[y \le f(x)+\tau] - q]$$

$$= \Pr_{x,y}[y \le f(x)+\tau] - q$$

$$PB_q(f) - PB_q(f') = PB_q(f(x)+\Delta) - PB_q(f(x))$$

$$= \int_0^\Delta \frac{d\, PB_q(f(x)+\tau)}{d\,\tau}\, d\tau$$

$$= \int_0^\Delta \underline{\Pr_{x,y}[(y \le f(x)+\tau] - q}\, d\tau$$

$$\begin{cases} \int_0^\Delta \quad \swarrow \quad]\, d\tau - |\Delta| q & \Delta \ge 0 \\ \int_0^\Delta \Pr(y \le f(x)+\tau]\, d\tau + |\Delta| q & \Delta < 0 \end{cases}$$

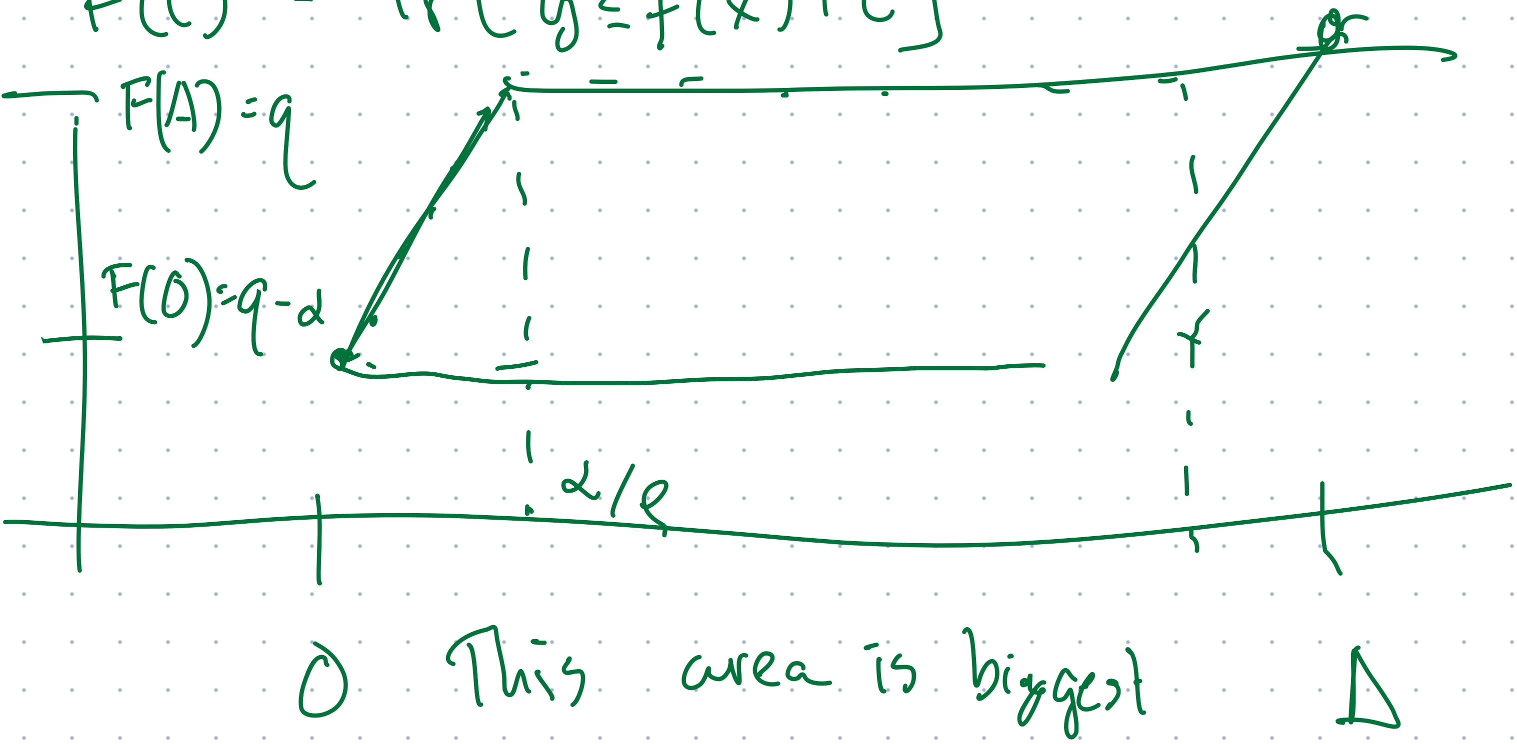
If we want to bound $\mathcal{T}$ It'll be useful to understand that integral.

lets consider $\Delta \geq 0$ (area under curve: between $f(x), f(x)+\Delta$)

f had quantile $q-\alpha$

$\hat{f}$ has quantile q

$F(\tau) = \Pr[y \leq f(x)+\tau]$



This area is biggest if the slopes are as steep as possible

[slope $\leq \rho$!