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, $\forall D$

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

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

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

Pf

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

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

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

$$= \qquad z_i \quad \text{since } S^{(i)} + z_i, z' \text{ are } \perp$$

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 $\forall x, y$
$\alpha \in (0,1)$, $\quad f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) - \frac{\lambda \alpha(1-\alpha)}{2}\|x-y\|^2$

(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, $w^* = A(S)$ $f_S(A(S^{(i)})) - f_S(A(S)) \geq \lambda\|A(S^{(i)}) - A(S)\|^2$

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

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

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

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

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. $R(\theta) = \lambda\|w\|^2$ 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^* $$\le \mathbb{E}_S[L_S(w^*) + \lambda\|w^*\|^2] + \underbrace{\qquad}$$

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

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

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^* : $f^*(x) = \mathbb{E}_{y|x}[y]$

which we can't do, (Bayes-OPT)

---

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: