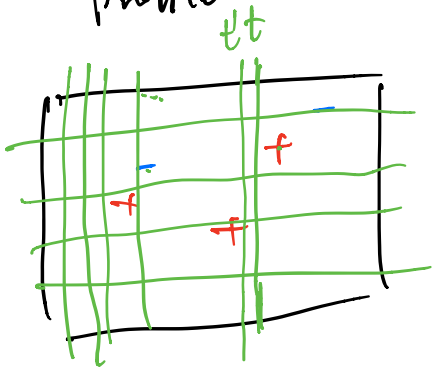
1. Project proposal due today on gradescope

2 HW. 2 Problem 4

(n) = 1.2 in hint

m

HW. 2 Problem 1



$B = \{ s \cdot (2I(x_i \geq t) - 1) :$

$s \in \{\pm 1\}$

$i \in \{1 \dots d\}$

$t \in \mathbb{R} \}$

$s=1$

$i=1$

$t \in \{ x_i : (x,y) \text{ in training set} \}$

Stability: another view of generalization.

$\hat{w} \leftarrow \underset{w}{\operatorname{argmin}} \; R(w) + \mathbb{E}_S \ell(w)$

intuition: alg stable if small change in input dataset does not change the output too much.

(1) Setting:

training set $S = (z_1 \dots z_m) \overset{iid}{\sim} D$.

learning model: parametrized by w.

loss fn: $\ell(w, z) \in \mathbb{R}$ $\quad z = (x,y)$ $\quad$ 0-1 loss

$\ell(w,(x,y)) = I(y\langle w,x\rangle \leq 0)$.

hinge loss

$= \max(0, 1 - y\langle w,x\rangle)$.

$L_D(w) = \mathbb{E}_{z \sim D} \ell(w,z)$. : generalization loss

$L_S(w) = \mathbb{E}_S \ell(w,z) = \frac{1}{|S|} \sum_{z \in S} \ell(w,z)$: empirical loss.

A : learning algorithm $\quad A(S) = \hat{w}$.

want to bound : $\mathop{\mathbb{E}}_{S\sim D^m}\left[L_D(\hat{w}) - L_S(\hat{w}) \right]$.

generalization gap.

(previously want to give a high prob. upper bound of gen gap).

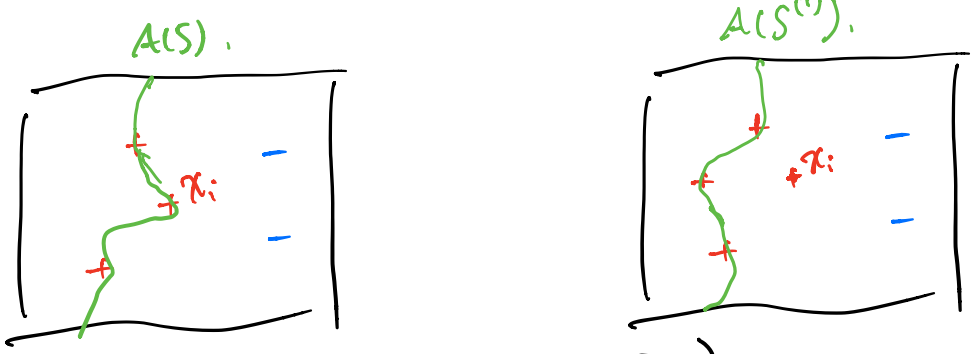
② S : $z_1, z_2, \dots, z_i, \dots, z_m$ $\xrightarrow{A}$ $A(S)$.

$i \in \{1, \dots, m\}$.

$S^{(i)}$: $z_1, z_2, \dots, z', \dots, z_m$ $\xrightarrow{A}$ $A(S^{(i)})$.

$z' \sim D$. indept from S.

$\ell(A(S), z_i)$ << vs. $\ell(A(S^{(i)}), z_i)$ when A is overfitting on data set S.



(OARO)

Def : learning alg A is on-average-replace-one stable with rate fn $g : \mathbb{N} \to \mathbb{R}$ if $\forall$ distn D. $\forall m$ (sample size).

$$\mathop{\mathbb{E}}_{(S,z')\sim D^{m+1},\, i \sim \mathrm{Unif}(\{1,\dots,m\})} \left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \right] \le g(m).$$

$\hat{w}^{(i)}$ $\qquad$ $\hat{w}$

(*)

(If $A(S^{(i)}) \approx A(S)$, then this will be small).

Thm: If A is OARO-stable w/ rate g, then

$$\underset{S\sim D^m}{\mathbb{E}}\left[L_D(A(S)) - L_S(A(S))\right] \le g(m).$$

(1)

pf: show (1) = (2)

First term in (*):

$$\underset{S\sim D^m,\, z'\sim D}{\mathbb{E}} \ell(A(S), z') = \underset{S\sim D^m}{\mathbb{E}} L_D(A(S))$$

$$= \underset{(S,z)\sim D^{m+1},\, i\sim U(\{1,\dots,m\})}{\mathbb{E}} \ell(A(S^{(i)}), z_i)$$

observe: for fixed i, $(S^{(i)}, z_i) \overset{d}{=} (S, z') \sim D^{m+1}$.

Second term in (*):

$$\underset{(S,z')\sim D^{m+1},\, i\sim U(\{1,\dots,m\})}{\mathbb{E}} \ell(A(S), z_i) = \underset{S\sim D^m}{\mathbb{E}} L_S(A(S)).$$

$$\frac{1}{m}\sum_{i=1}^{m} \ell(A(S), z_i) = L_S(A(S)).$$

④ ℓ_2-regularization gives stability.

Assume: ① $\ell(w,z)$ is ρ-Lipschitz wrt w. for any z

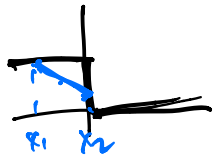
$$\forall w_1, w_2 \quad |\ell(w_1,z) - \ell(w_2,z)| \le \rho \cdot |w_1 - w_2|.$$

(if f differentiable $\max_z |f'(z)|$. $\quad f(z_1) - f(z_2) \overset{MVT}{=} f'(\theta)(z_1 - z_2)$, $\theta \in [z_1, z_2]$)

② $\ell(w,z)$ is convex wrt w, for any z.

(hinge loss. logistic loss. exponential loss. etc)

$y = \mathbb{1}_{x>0}$.



$I(y\langle w,x\rangle \le 0)$ not cvx.

③ $\hat{w} = A(S) = \underset{w}{\operatorname{argmin}} \left(\frac{\lambda}{2}\|w\|_2^2 + L_S(w) \right)$.

we will show A is $g(m) = \frac{2\rho^2}{\lambda m}$ – OARO stable.

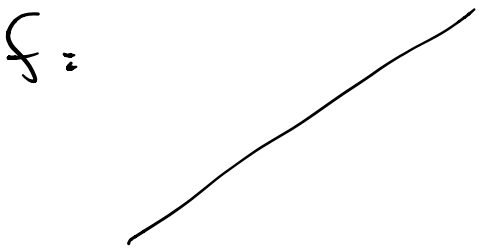
key tool: strong convexity

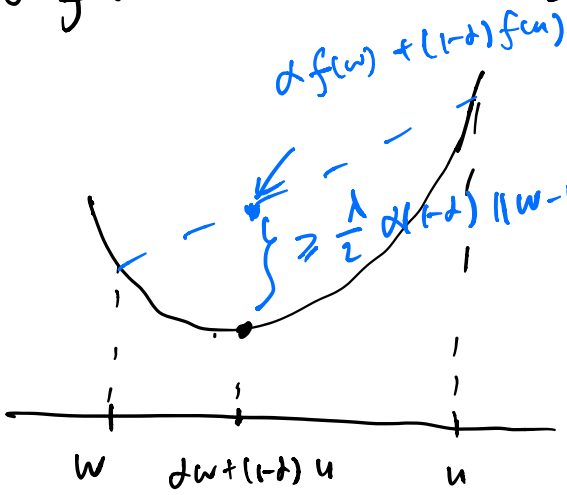
Def: fn f is λ-strongly cvx (λ-sc) if $\forall\, w, u$.

$\alpha \in (0,1)$,

$$f(\alpha w + (1-\alpha)u) \le \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2.$$

(0-sc $\Leftrightarrow$ cvx).





- λ-sc for $\lambda > 0$?

$\lambda \uparrow$

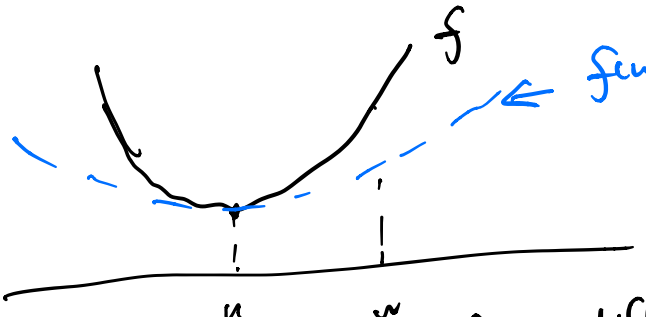
key properties of sc. fns:

1. $f(w) = \frac{\lambda}{2}\|w\|^2$ is λ-sc. (Exercise)
2. if f is λ-sc, g is cvx, then $h = f + g$ is λ-sc.

(write down the λ-sc. cvx defns of f & g & and sum).

3. If f is λ-sc & $u = \underset{w}{\operatorname{argmin}} f(w)$. then.

$\forall w. \quad f(w) - f(u) \geq \frac{\lambda}{2}\|w-u\|^2$.



Pf: focus on the case when f is differentiable. (for general f, we need the defn of subgradient).

$$\nabla f(u) = 0.$$

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

$g(\alpha) - g(0)$

taking limit w/ $\alpha \to 0^+$.

$$0 \leq f(w) - f(u) - \frac{\lambda}{2}\|w-u\|^2.$$

$g(\alpha) = f(u+\alpha(w-u))$.

LHS $\to g'(0) = 0$.

$g'(\alpha) = \langle \nabla f(u+\alpha(w-u)), w-u \rangle$

$\overset{\alpha=0}{=} 0$

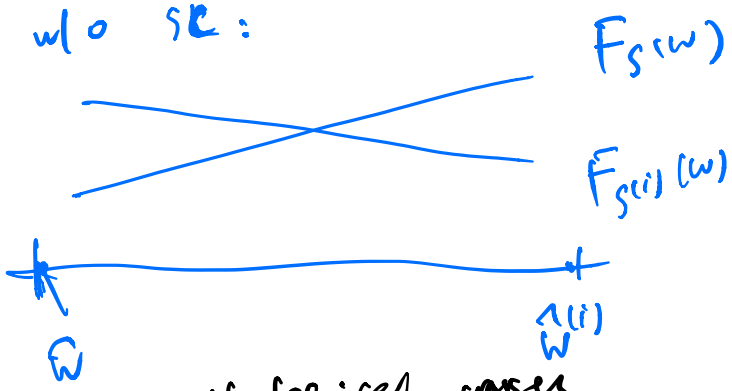
Back to stability:

$$\hat{w} = \arg\min_w \underbrace{\left(L_S(w) + \frac{\lambda}{2}\|w\|^2 \right)}_{= F_S(w)} \quad \because \lambda\text{-SC.}$$

$$\hat{w}^{(i)} = \arg\min_w \underbrace{\left(L_{S^{(i)}}(w) + \frac{\lambda}{2}\|w\|^2 \right)}_{= F_{S^{(i)}}(w)} \quad \because \lambda\text{-SC.}$$

want to show $\hat{w}$ is close to $\hat{w}^{(i)}$.

caveat: w/o SC:



w/ SC: can rule out such pathological cases.

property 3: $F_S(\hat{w}^{(i)}) - F_S(\hat{w}) \geq \frac{\lambda}{2}\|\hat{w}^{(i)} - \hat{w}\|^2$

$F_{S^{(i)}}(\hat{w}) - F_{S^{(i)}}(\hat{w}^{(i)}) \geq \frac{\lambda}{2}\|\hat{w}^{(i)} - \hat{w}\|^2$.

summing up:

$$\left(F_S(\hat{w}^{(i)}) - F_{S^{(i)}}(\hat{w}^{(i)})\right) - \left(F_S(\hat{w}) - F_{S^{(i)}}(\hat{w})\right) \geq \lambda\|\hat{w}^{(i)} - \hat{w}\|^2$$

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

$\leq \rho \cdot \|\hat{w}^{(i)} - \hat{w}\|$ ‖ regroup wrt z_i or z'

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

‖

$$\frac{2\rho}{m}\|\hat{w} - \hat{w}^{(i)}\|.$$

$$\Rightarrow \|\hat{w}^{(i)} - \hat{w}\| \leq \frac{2\rho}{m \cdot \lambda}.$$

$$\Rightarrow \ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i) \leq \rho \cdot \|\hat{w}^{(i)} - \hat{w}\|$$

$$\leq \frac{2\rho^2}{m\lambda}.$$

$\Rightarrow$ A is $g(m) = \frac{2\rho^2}{m\lambda}$ – OARO – stable.

Next class:

- Application of stability argument to Regularized loss minimization w/ the goal of minimizing final generalization loss
- online learning