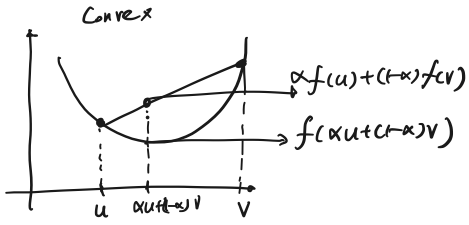
chapter 12. Convex Learning Problems.

Definition [Convex Set] A set C in a vector space is convex if for any two vectors u, v in C, the line segment between u and v is contained in C. That is, for any $\alpha \in [0,1]$, we have

$$\alpha \cdot u + (1-\alpha) \cdot v \in C, \quad \forall u, v \in C$$

Definition [Convex function] Let C be a convex set. A function $f: C \to \mathbb{R}$ is convex if for every $u, v \in C$, and $\alpha \in [0,1]$

$$f(\alpha \cdot u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v).$$

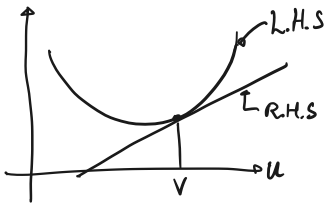


- we will be using equivalent definitions of convexity:
  - for convex differentiable f, $\forall u$

$$\boxed{f(u) \geq f(v) + \langle \nabla f(v), u - v \rangle}$$

$\nabla f(v)$: gradient, $\nabla f(v) \in \mathbb{R}^d$ if $v \in \mathbb{R}^d$, $\nabla f(v)_i = \frac{\partial f(v)}{\partial v_i}$

$\langle \cdot, \cdot \rangle$: inner product, $\langle a, b \rangle = \sum_i a_i \cdot b_i$



Definition [Lipschitzness] Let $C \subset \mathbb{R}^d$. A function $f: \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if for every $w_1, w_2 \in C$ we have

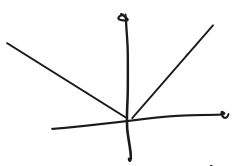
$$\| f(w_1) - f(w_2) \| \leq \rho \| w_1 - w_2 \|$$

- Lipschitz constant ρ measures how fast the function is changing.
- it follows that if $\| \nabla f(w) \|$ is bounded, then $f(\cdot)$ is Lipschitz.

ex> $f(x) = |x|$ is 1-Lipschitz.

$$\begin{aligned} |x_1| - |x_2| &= |x_1 - x_2 + x_2| - |x_2| \\ &\leq |x_1 - x_2| + |x_2| - |x_2| \\ &= |x_1 - x_2|. \end{aligned}$$

by symmetry $|x_1| - |x_2| \leq |x_1 - x_2|$



ex> $f(x)=x^2$ is not Lipschitz

Let $x_1=0, x_2=1+\rho$, $\rho>0$

$(1+\rho)^2-(0)^2 > \rho^2+\rho = \rho\cdot(1+\rho) = \rho|x_2-x_1|$,

taking ρ as large as we want, we can make this arbitrarily large

---

Definition [Smoothness] A differentiable function $f:\mathbb{R}^d\to\mathbb{R}$ is β-smooth if its gradient is β-Lipschitz; $\forall v,w$ we have

$$\|\nabla f(v)-\nabla f(w)\| \le \beta\cdot\|v-w\|$$

$$\nabla f(v) = \Big(\underbrace{\frac{\partial f(v)}{\partial v_1}}_{\mathbb{R}^d}, \frac{\partial f(v)}{\partial v_2}, \dots, \frac{\partial f(v)}{\partial v_d}\Big)$$

- this measures how smoothly the gradient is changing.
- we will be using the fact that, for smooth $f(\cdot)$,

$$\boxed{f(v) \le f(w) + \langle \nabla f(w), v-w\rangle + \frac{\beta}{2}\|v-w\|^2}$$

ex> $f(x)=x^2$ is 2-smooth.

$f'(x)=2x$ is 2-Lipschitz

Chapter 13.

We will learn a new learning paradigm: Regularized Loss Minimization (RLM) and show that convex-Lipschitz-bounded and convex-smooth-bounded families of learning problems are learnable. Key insight is that regularizers make learning algorithm more stable.

- Regularized Loss Minimization

  for a regularizer is $R:\mathbb{R}^d\to\mathbb{R}$, the RLM outputs

$$A(S) \leftarrow \operatorname*{arg\,min}_{w\in\mathbb{R}^d} \big(L_S(w) + R(w)\big)$$

- the value of the regularization measures the complexity of the hypothesis, w
- we use $R(w)=\lambda\cdot\|w\|^2$, also called Tikhonov regularization

example> Ridge Regression from CSE 446/546

$$A(S) \leftarrow \operatorname*{arg\,min}_{w\in\mathbb{R}^d} \Big(\underbrace{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\big(\langle w,x_i\rangle-y_i\big)^2}_{\text{linear regression loss}} + \underbrace{\lambda\cdot\|w\|_2^2}_{\text{regularizer}}\Big)$$

- this has a closed form solution, which we get by setting the gradient to zero.

- Our goal today is to show the following

[Theorem] $\mathcal{D}$ is a distribution over $\mathcal{X}\times[-1,1]$, where $\mathcal{X}=\{x\in\mathbb{R}^d:\|x\|\le 1\}$ and $\mathcal{H}=\{w\in\mathbb{R}^d:\|w\|\le B\}$, ρ-Lipschitz loss, For any $\varepsilon\in(0,1)$, if $m\ge\frac{8\rho^2B^2}{\varepsilon^2}$ then ridge regression with $\lambda=\sqrt{\frac{2\rho^2}{B^2m}}$ satisfies

$$\mathbb{E}_S\big[L_{\mathcal{D}}(\underbrace{A(S)}_{\text{RLM}})\big] \le \min_{w\in\mathcal{H}} L_{\mathcal{D}}(w) + \varepsilon.$$

Remarks:

- Both $\mathcal{X}$ and $\mathcal{H}$ are bounded, and we will learn that this is important.
- relatedly, regularization is important
- We are proving a bound on <u>Expected</u> test error (or Risk) as opposed to the usual high probability bound on the test error. Note that test error is random because hypothesis $A(S)$ is random.
- Bounded expected risk implies agnostic PAC learnability, which we will not cover in this lecture.
- We use expected Risk because it is related to stability.

A learning algorithm is <u>stable</u> if "a small change in the input" does not

"change the output much". we make this formal.

- $S = (z_1, z_2, \cdots, z_{i-1}, z_i, z_{i+1}, \cdots, z_n)$
- $S^{(i)} = (z_1, z_2, \cdots, z_{i-1}, z', z_{i+1}, \cdots, z_n)$

"small change in the input"

- measure of the effect: loss on $z_i = (x_i, y_i)$

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

- intuitively this is non-negative because an algorithm that saw z_i in training will likely have less error on z_i, and a stable algorithm will make the above difference small.

Definition. [on-average-replace-one-stable]
We say an algorithm A is on-average-replace-one-stable with rate $\varepsilon(m)$ if

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

for all D, for some monotonically decreasing $\varepsilon(\cdot)$.

- this is further justified by the following, which shows stable algorithms generalize.

[Theorem 13.2], $S = (z_1, \cdots, z_m)$ iid from D and z' is another iid sample.
let $U(m)$ be the uniform distribution over $[m]$. Then for any Algorithm, A,

$$\mathbb{E}_S\left[L_D(A(S)) - L_S(A(S))\right] = \mathop{\mathbb{E}}_{S, z', i \sim U(m)}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\right]$$

proof> by definition

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

$$\mathbb{E}_S[L_D(A(S))] = \mathbb{E}_{S,z'}\left[\ell(A(S), z')\right]$$

(z': independent)

$$= \mathbb{E}_{S,z'}\left[\ell(A(S^{(i)}), z_i)\right] \quad \text{for all } i.$$

(z_i: independent)

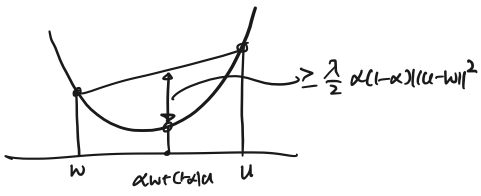
- Assuming convex loss function with Lipschitzness or Smoothness, we show that RLM is stable, because it is strongly convex

Definition [strong convexity]
A function is λ-strongly convex if for all w, u and $\alpha \in (0,1)$ we have

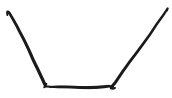
$$f(\alpha w + (1-\alpha)w) \le \alpha f(w) + (1-\alpha) f(u) - \frac{\lambda}{2}\alpha(1-\alpha)\|w-u\|^2$$

(how large can this λ be.)

example> $f(w) = \lambda\|w\|^2$ is 2λ-strongly convex



0-strongly convex
= convex



- in particular, we use the fact that
  - f is convex and g is λ-strongly convex $\longrightarrow$ $f+g$ is λ strongly convex
  - if f is λ-strongly convex, then

$$f(w) - f(u) \ge \frac{\lambda}{2}\|w-u\|^2 \qquad (*)$$

Recall, for RLM

$$A(S) \leftarrow \mathop{\text{argmin}}_w \Big(\underbrace{L_S(w) + \lambda\|w\|^2}_{f_S(w)\ \text{— } (2\lambda)\text{-strongly convex.}} \Big)$$

lower bound: $f_S(A(S^{(i)})) - f_S(A(S)) \ge \lambda \cdot \|A(S^{(i)}) - A(S)\|^2$ ← (*)

upper bound:

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

($f_S(A(S^{(i)}))$: $\hat{w}^{(i)}$, $f_S(A(S))$: $\hat{w}$)

$$= \underbrace{\left\{ L_{S^{(i)}}(\hat{w}^{(i)}) + \lambda\|\hat{w}^{(i)}\|^2 - \left(L_{S^{(i)}}(\hat{w}) + \lambda\|\hat{w}\|^2 \right) \right\}}_{\leq 0 \text{ because } \hat{w}^{(i)} \text{ is optimal solution for } f_{S^{(i)}}(\cdot)} + \frac{\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i)}{m} + \frac{\ell(\hat{w}, z') - \ell(\hat{w}^{(i)}, z')}{m}$$

Together,

$$\lambda\|\hat{w}^{(i)} - \hat{w}\|^2 \leq \frac{\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i)}{m} - \frac{\ell(\hat{w}, z') - \ell(\hat{w}^{(i)}, z')}{m} \quad (**)$$

---

we use (**) to show that RLM is stable.

Case 1. Lipschitz loss

If $\ell(\cdot, z_i)$ is ρ-Lipschitz, then

$$\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i) \leq \rho \cdot \|\hat{w}^{(i)} - \hat{w}\|, \text{ and} \quad (***)$$
$$\ell(\hat{w}, z') - \ell(\hat{w}^{(i)}, z') \leq \rho \cdot \|\hat{w}^{(i)} - \hat{w}\|, \text{ plugging into (**)},$$

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

$$\rightarrow \quad \|\hat{w}^{(i)} - \hat{w}\| \leq \frac{2\rho}{\lambda m}, \text{ plugging into (***)},$$

$$\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i) \leq \frac{2\rho^2}{\lambda m}. \leftarrow \text{which is monotonically decreasing in } m.$$

This implies

[Corollary 13.6] for ρ-Lipschitz loss function, RLM with $\lambda\|w\|^2$ regularization achieves

$$\mathbb{E}_S\left[L_D(A(S)) - L_S(A(S))\right] \leq \frac{2\rho^2}{\lambda m}.$$

$\lambda\uparrow$ → more strongly convex → better generalization
$\rho\downarrow$ → more smooth loss → better generalization.

- we skip the proof for smooth losses, but for β-smooth losses,

for $\lambda \geq \frac{2\beta}{m}$, $\quad \mathbb{E}_S\left[L_D(A(S)) - L_S(A(S))\right] \leq \frac{48\beta}{\lambda m} \cdot \mathbb{E}\left[L_S(A(S))\right]$

---

* Fitting-Stability Tradeoff

$$\mathbb{E}_S\left[L_D(A(S))\right] = \underbrace{\mathbb{E}_S\left[L_S(A(S))\right]}_{\text{fit train data}} + \underbrace{\left\{ \mathbb{E}_S\left[L_D(A(S)) - L_S(A(S))\right] \right\}}_{\approx \text{stability : Corollary 13.6},\ \leq \frac{2\rho^2}{\lambda m}}$$

$$\leq \mathbb{E}_S\left[L_S(A(S)) + \lambda\|A(S)\|^2\right]$$

$A(S)$ is optimal solution to RLM → $\leq \mathbb{E}_S\left[L_S(w^*) + \lambda\|w^*\|^2\right]$

$$= L_D(w^*) + \lambda\|w^*\|^2, \ \forall w^*$$

This gives $\quad \mathbb{E}_S\left[L_D(A(S))\right] \leq L_D(w^*) + \lambda\|w^*\|^2 + \frac{2\rho^2}{\lambda m}$

(λ: Trade off by λ.)

[Corollary 13.9] if ℓ is ρ-Lipschitz, and $\|w^*\| \leq B$, then

$\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ achieves

$$\mathbb{E}_D\left(L_D(A(S))\right) \leq \min_{w \in H} L_D(w) + \rho B\sqrt{\frac{8}{m}}$$

($A(S)$: RLM)

This implies the [Theorem].
