Lecture 17. Logistics — Project Schedule Out; HW3 due Thursday

Recap. [lipschitzness]. a function $f(\cdot)$ is β-lipschitz if

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

measures how fast the function changes over the input.

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: **regularization** makes learning algorithms more **stable**

Definition: regularized loss minimization

For a regularizer $R: \mathbb{R}^d \to \mathbb{R}$, RLM outputs

$$A(S) \leftarrow \underset{w \in \mathbb{R}^d}{\arg\min} \left(L_S(w) + R(w) \right)$$

- the regularizer $R(w)$ is a **measure of complexity** of hypothesis w
- we use $R(w) = \lambda \cdot \|w\|_2^2$, also called **Tikhonov Regularizer**

Example: Ridge Regression from CSE 446/546

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

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

Our goal today is to show the following

[Theorem] D is a distribution over $\mathcal{X} \times [-1,1]$, where $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \le 1\}$

$\mathcal{H} = \{w \in \mathbb{R}^d : \|w\| \le B\}$, loss is ρ-Lipschitz.

For any $\varepsilon \in (0,1)$, if $m \ge \frac{8\rho^2 B^2}{\varepsilon^2}$ then

Ridge Regression with $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ satisfies

$$\mathbb{E}_S[\, L_D(A(S)) \,] \le \min_{w \in \mathcal{H}_B} L_D(w) + \varepsilon$$

($A(S)$: RLM; S: Random)

Remarks:

- both $\mathcal{X}$ and $\mathcal{H}$ are **bounded**, which is important.
- relatedly, **regularization** is important.
- we are bounding the **expectation**, and not the $L_D(A(S))$ with high probability, because expectation is directly related to **stability**.

Definition:

A learning algorithm is **stable** if "a small change in the input" does not "change the output much". We make this formal

- small change in input: $S = (z_1, z_2, \dots, z_{i-1}, \boxed{z_i}, z_{i+1}, \dots, z_m)$
  $S^{(i)} = (z_1, z_2, \dots, z_{i-1}, \boxed{z_i'}, z_{i+1}, \dots, z_m)$
- measure of effect: loss on $z_i = (x_i, y_i)$, i.e.,
  $$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)$$

* intuitively, this is non-negative, because an algorithm that saw z_i in training will likely have less error on z_i.
* 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

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

$S\sim D^m$, $\tilde{z}'\sim D$, $i\sim U[m]$ (uniform)

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

*this notion of stability is further justified by the following, which shows stable algorithms generalize.

[Theorem 13.2] $S=(z_1,\dots,z_m)$ i.i.d. from D, and z' is another iid sample, let $U[m]$ be uniform distribution over $[m]$. then for any algorithm A,

$$\mathbb{E}_S\left[\underbrace{L_D(A(S))-L_S(A(S))}_{\text{generalization error PAC.}}\right]=\mathbb{E}_{S,\tilde{z}',i}\left[\underbrace{\ell(A(S^{(i)}),\tilde{z}_i)-\ell(A(S),\tilde{z}_i)}_{\text{bound by one-replace-stable }\varepsilon(m)}\right]$$

$S\sim D^m$, $\tilde{z}'\sim D$, $i\sim U[m]$

Proof) by definition,

$$\text{2nd term}=\mathbb{E}_S[L_S(A(S))]=\mathbb{E}_S\left[\frac{1}{m}\sum_{i=1}^{m}\ell(A(\underbrace{S}_{(z_1,\dots,z_m)}),z_i)\right]$$

* this is indep z_i'
* note that this holds even if A treats $(z_1,\dots,z_m)$ differently, for example train on every even index data

$$=\mathbb{E}_{S,I\sim U[m]}\left[\ell(A(S),z_I)\right]$$

$$\text{1st term}=\mathbb{E}_S[L_D(A(S))]=\mathbb{E}_{S,z_i'}\left[\ell(A(S),\underbrace{z_i'}_{\text{indep.}})\right]$$

$$=\mathbb{E}_{S,\tilde{z}',i}\left[\ell(A(\{z_1,\dots,z_{i-1},\tilde{z}_i,z_{i+1},\dots z_m\}),\tilde{z}_i)\right]$$

$$=\mathbb{E}_{z_1\dots z_m,\tilde{z}_i}\left[\ell(A(S^{(i)}),z_i)\right]$$

* 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 $\forall u, v$ & $\alpha \in [0,1]$,

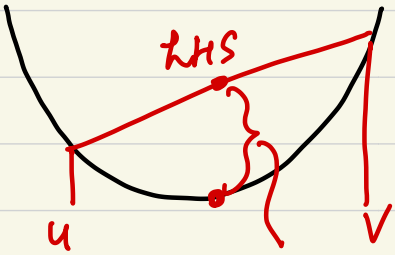
$$\alpha f(u) + (1-\alpha) f(v) \geq f(\alpha u + (1-\alpha) v) + \frac{\lambda}{2} \alpha (1-\alpha) \cdot \|u - v\|^2$$

example >

piecewise linear.

0-strongly convex

$f(u) = \lambda \cdot \|u\|^2$ is 2λ-strongly convex,



$\geq \lambda \alpha (1-\alpha) \cdot \|u - v\|^2$

- smallest eigenvalue of Hessian.

* in particular, we use the fact that

① f is convex & g is λ-strongly convex $\rightarrow$ $f + g$ is λ-strongly convex.

[Proposition]

② if f is λ-strongly convex, and w^* is minimizer of f.

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

we assume $\begin{cases} \ell(w, z) \text{ is Convex in } w \in \mathbb{R}^d \\ R(w) = \lambda \|w\|^2, \text{ which is } 2\lambda\text{-strongly Convex.} \end{cases}$

Recall that RLM outputs.

$$A(S) \leftarrow \arg\min_w \left[f_S(w) \right] \triangleq \hat{L}_S(w) + \lambda \|w\|^2$$

by ①, $f_S(w)$ is 2λ-strongly Convex

let $\hat{w}^{(i)} = A(S^{(i)})$, and $\hat{w} = A(S)$, then

Lower Bound: by ② and 2λ-strong Convexity, and the fact that $\hat{w}$ is minimizer of $f_S(w)$.

$$f_S(\hat{w}^{(i)}) - f_S(\hat{w}) \underset{\text{strong Convexity due to regularization.}}{\geq} \lambda \cdot \|\hat{w}^{(i)} - \hat{w}\|_2^2 \underset{\text{Convexity.}}{\geq} 0$$

Upper Bound:

$$f_S(\hat{w}^{(i)}) - f_S(\hat{w}) = \underbrace{\hat{L}_S(\hat{w}^{(i)}) + \lambda \|\hat{w}^{(i)}\|^2}_{\hat{L}_{S^{(i)}}(\hat{w}^{(i)}) + \frac{1}{m}\left(\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}^{(i)}, z_i')\right)} - \underbrace{\hat{L}_S(\hat{w}) - \lambda \|\hat{w}\|^2}_{\hat{L}_S^{(i)}(\hat{w}) + \frac{1}{m}\left(\ell(\hat{w}, z_i) - \ell(\hat{w}, z_i')\right)}$$

$$= \underbrace{\left(\hat{L}_{S^{(i)}}(\hat{w}^{(i)}) + \lambda \|\hat{w}^{(i)}\|^2\right)}_{f_{S^{(i)}}(\hat{w}^{(i)})} - \underbrace{\left(\hat{L}_S^{(i)}(\hat{w}) + \lambda \|\hat{w}\|^2\right)}_{f_{S^{(i)}}(\hat{w})} + \frac{\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i)}{m} + \frac{\ell(\hat{w}, z_i') - \ell(\hat{w}^{(i)}, z_i')}{m}$$

$\hat{w}^{(i)}$ minimizer of $f_{S^{(i)}}$, so $f_{S^{(i)}}(\hat{w}^{(i)}) - f_{S^{(i)}}(\hat{w}) \leq 0$

$$\leq \underbrace{\frac{\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i)}{m}}_{\text{How fast } \ell \text{ changes as we change } \hat{w}^{(i)}, \text{ related to Lipschitzness of } \ell.} + \frac{\ell(\hat{w}, z_i') - \ell(\hat{w}^{(i)}, z_i')}{m}$$

* Putting the Upper Bound and Lower Bound together, we get

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

This holds for any convex RLM.

we use this to show that if the loss $\ell(\cdot, z_i)$ is well-behaved (Lipschitz, smooth) then RLM is stable.

Case 1: ρ-Lipschitz loss $\ell(\cdot, z_i)$ → $\ell(w_1, z_i) - \ell(w_2, z_i) \leq \rho \|w_1 - w_2\|$.

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

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

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

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

Recall Stability is a bound on

Stability: $\mathbb{E}\left[\underbrace{\ell(\hat{w}^{(i)}, z_i) - \ell(\hat{w}, z_i)}_{\leq \frac{2\rho^2}{\lambda m}}\right] \leq \frac{2\rho^2}{\lambda m}$

more Lipschitz, steeper loss.

more sample, less influence by 1

$\lambda \uparrow$, more strictly convex, more stable

Thm 13.2.

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

Generalization: $\mathbb{E}[L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda m}$

Case 2. Similarly, for β-smooth loss and $\lambda \geq \frac{2\beta}{m}$,

$$\mathbb{E}_S\left[\ell_D(A(S)) - \ell_S(A(S))\right] \leq \frac{48\beta}{\lambda m}\,\mathbb{E}\left[\ell_S(A(S))\right]$$

---

* Fitting - Stability Tradeoff

$$\underbrace{\mathbb{E}_S\left[\ell_D(A(S))\right]}_{\text{what we care about, Test loss}} = \underbrace{\mathbb{E}_S\left[\ell_S(A(S))\right]}_{\text{Train loss: how well you fit Train data}} + \underbrace{\left\{\mathbb{E}_S\left[\ell_D(A(S)) - \ell_S(A(S))\right]\right\}}_{\approx \text{ Stability: Corollary 13.6} \leq \frac{2\beta^2}{\lambda m}}$$

Let's expand:

$$\leq \mathbb{E}_S\left[\underbrace{\ell_S(A(S)) + \lambda\|A(S)\|^2}_{\text{minimizer RLM}}\right]$$

$$\leq \mathbb{E}_S\left[\ell_S(w^*) + \lambda\|w^*\|^2\right]$$

$$\leq \underbrace{\ell_D(w^*)}_{\text{irreducible error}} + \underbrace{\lambda\|w^*\|^2}_{\text{mis fit due to reg.}} + \underbrace{\frac{2\beta^2}{\lambda m}}_{\text{stability due to reg.}}$$

Controlled by λ.

[Corollary 13.9]. (Convex-Lipschitz-Bounded Risk bound)
If ℓ is β-Lipschitz, $\|w^*\| \leq B$, then $\lambda = \sqrt{\frac{2\beta^2}{B^2 m}}$ achieves

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

($A(S)$: RLM)

this proves the [Theorem].