Regularization & Stability

§1 RLM rule

Def 1 Regularized Loss Minimization (RLM) is a learning rule of the form

$$\operatorname{argmin}_w \left(L_S(w) + R(w)\right), \text{ with}$$

$R: \mathbb{R}^d \to \mathbb{R}$ a regularization function.
The case $R(w) = \lambda \|w\|_2^2$, $\lambda > 0$ is called Tikhonov regularization.

§2 Stable rules & Overfitting

Notations

$U(m)$: Uniform distribution on $[m]$
A : learning Algorithm
$S = (z_1, \dots, z_m)$
$A(S)$: output of A
$S^{(i)} = (z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_m)$, for another example z'.

Def 2

Let $\varepsilon: \mathbb{N} \to \mathbb{R}$ be a monotonically dec. function.
A is on-average-replace-one-stable with rate $\varepsilon(m)$ if for every distr. $\mathcal{D}$:

$$\mathop{\mathbb{E}}_{(S,z')\sim \mathcal{D}^{m+1}}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\right] \leq \varepsilon(m)$$

Theorem 1

For any learning Algorith A, we have:

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}(A(S)) - L_S(A(S))\right] = \mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1}, i\in U(m)}\left[\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)\right]$$

Proof

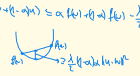
$$\mathop{\mathbb{E}}_S\left[L_\mathcal{D}(A(S))\right] \underset{\text{Def } L_\mathcal{D}}{=} \mathop{\mathbb{E}}_{S,z'}\left[\ell(A(S), z')\right] \underset{\text{i.i.d.}}{=} \mathop{\mathbb{E}}_{S,z'}\left[\ell(A(S^{(i)}), z')\right] \quad (\text{for all } i \in [m])$$

Also: $\mathop{\mathbb{E}}_S\left[L_S(A(S))\right] = \mathop{\mathbb{E}}_{S,i}\left[\ell(A(S), z_i)\right]$ □

§3 Strong Convexivity

Def 3 A function f is λ-strongly-convex if for w, u, $\alpha \in (0,1)$, we have:

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



Lemma 1

1. $f(w) = \lambda\|w\|^2$ is 2λ strongly convex
2. if f is λ-strongly convex and g is convex then $f+g$ is λ-strongly convex
3. if f is λ strongly convex, u is minimizer of f, then for every w

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

Proof 3

Divide def of s.c. by α and rearrange.

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

Let $g(\alpha) = f(u + \alpha(w-u))$. Take limit $\alpha \to 0$

$$0 = g'(0) \leq f(w) - f(u) - \frac{\lambda}{2}\|w-u\|^2$$ □

§4 Tikhonov Regularization as a stabilizer

Assumption Loss function is convex.
Goal bound for $\|A(S^{(i)}) - A(S)\|$ for tikhonov reg.

Define $f_S(w) = L_S(w) + \lambda\|w\|^2$, $A(S) = \operatorname{argmin}_w f_S(w)$
By Lemma 1, part 1,2 f_S is 2λ strongly convex
Now for any v by Part 3 of Lemma 1:

$$f_S(v) - f_S(A(S)) \geq \lambda\|v - A(S)\|^2 \qquad (2)$$

Also for any u, v, i:

$$\begin{aligned} f_S(v) - f_S(u) &= L_S(v) + \lambda\|v\|^2 - (L_S(u) + \lambda\|u\|^2) \\ &= L_{S^{(i)}}(v) + \lambda\|v\|^2 - (L_{S^{(i)}}(u) + \lambda\|u\|^2) \\ &\quad + \frac{\ell(v, z_i) - \ell(u, z_i)}{m} + \frac{\ell(u, z') - \ell(v, z')}{m} \end{aligned}$$

For $v = A(S^{(i)})$, $u = A(S)$ we obtain (v is minimizer of $L_{S^{(i)}}(v) + \lambda\|v\|^2$)

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

by (2):

$$\lambda\|A(S^{(i)}) - A(S)\|^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m} \qquad (3)$$

Theorem 2

Assume a convex, ρ-Lipschitz loss function. Then the RLM rule with $\lambda\|w\|^2$ regularization is on-average-replace-one-stable with rate $\frac{2\rho^2}{\lambda m}$
This implies (by theorem 1)

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[L_\mathcal{D}(A(S)) - L_S(A(S))\right] \leq \frac{2\rho^2}{\lambda m}$$

Proof

$\ell(\cdot, z_i)$ ρ-Lipschitz

$$\ell(A(S), z_i) - \ell(A(S^{(i)}), z_i) \leq \rho\|A(S^{(i)}) - A(S)\| \qquad (4)$$

$$\ell(A(S^{(i)}), z') - \ell(A(S), z') \leq \rho\|A(S^{(i)}) - A(S)\|$$

By (3)

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

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

by (4)

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

Def

A differentiable function $f: \mathbb{R}^d \to \mathbb{R}$ is β smooth if ∇f is β Lipschitz.

Corollary 1

For a convex-lipschitz-bounded problem ($\|w\| \leq B$ for $w \in \mathcal{H}$)

$$\mathbb{E}\left[L_\mathcal{D}(A(S))\right] \leq \min L_\mathcal{D}(w) + \rho B\sqrt{\frac{8}{m}}$$

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

Remark This implies PAC learnability because

If A guarantees that for $m \geq m_H(\varepsilon)$

$$\mathbb{E}\left[L_\mathcal{D}(A(S))\right] \leq \min_{h\in H} L_\mathcal{D}(h) + \varepsilon$$

then it is apac learnable.

Proof

Let $\delta \in (0,1)$, $m \geq m(\varepsilon\delta)$
$X := L_\mathcal{D}(A(S)) - \min_{h\in H} L_\mathcal{D}(h)$
$\Rightarrow X \geq 0$ $\mathbb{E}[X] \leq \varepsilon\delta$

$$P[X \geq \varepsilon] \leq \frac{\mathbb{E}[X]}{\varepsilon} = \frac{\varepsilon\delta}{\varepsilon} = \delta$$

$\|$

$$\mathbb{P}\left[L_\mathcal{D}(A(S)) \geq \min_{h\in H} L_\mathcal{D}(h) + \varepsilon\right]$$ □

$|\langle w, (x+\varepsilon)\rangle - y|$
$= |wx + w\varepsilon - y|$
$\leq |wx - y| + |w\varepsilon|$
$\leq |wx - y| + \|w\|\|\varepsilon\|$

$$\mathop{\mathbb{E}}_S\left[L_\mathcal{D}(A(S))\right] \leq L_\mathcal{D}(w^*) + \lambda\|w^*\|^2 + \frac{2\rho^2}{\lambda m}$$

$\|w\| \leq B$
$\|w - \alpha\| \leq B$