CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 18: November 2

Lecturer: Csaba Szepesvári *Scribes: Vedant Vyas*

Note: *LaTeX template courtesy of UC Berkeley EECS dept. (link to directory)*
Disclaimer: *These notes have **not** been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

18.1 Leave One Out (L.O.O) Stability

Definition 18.1. $k : \overset{\text{target}}{Z^n} \cup \overset{\text{auxillary}}{Z^{n+1}} \to \mathcal{M}_1(\mathbb{R}^{\mathbb{Z}})$ is ε-LOO (Leave One Out) stable with $\varepsilon : \mathbb{Z}^{n+1} \cup \mathbb{Z} \to [0, \infty)$ if $\forall z_{1:n+1} \in Z^{n+1} \mu_k(z_{1:n}, z_{n+1}) \le \mu_k(z_{1:n+1}, z_{n+1}) + \varepsilon(z_{1:n+1}, z_{n+1})$

Proposition 18.2. *Assume* $\mu_k(z_{1:n+1}^{i \leftrightarrow n+1}, z_{n+1}) = \mu_k(z_{1:n+1}, z_{n+1})$ *and* $\varepsilon(z_{1:n+1}^{i \leftrightarrow n+1}, z_{n+1}) = \varepsilon(z_{1:n+1}, z_{n+1})$ *& K is* ε*-LOO stable then*

$$\mathbb{E}\left[\mu_k\left(z_{1:n}, z_{n+1}\right)\right] \le \mathbb{E}P_{n+1}\mu_k\left(z_{1:n+1}\right) + \mathbb{E}P_{n+1}\varepsilon\left(z_{1:n+1}\right)$$

Note: From our assumptions on μ_k we can get complete symmetry as to swap i $\leftrightarrow$ j, we can do $i \leftrightarrow n+1$, and then $n+1 \leftrightarrow j$

Proof.

$$\begin{aligned}
z_{1:n+1}^{i \leftrightarrow n+1} &= (z_1, , , , z_{i-1}, z_{n+1}, \dots . z_i) \\
(z_{1:n+1}^{i \leftrightarrow n+1}, z_i) &\overset{P}{=} (z_{1:n+1}, z_{n+1}) \qquad \text{(Same joint distribution)} \\
(\mu_k + \varepsilon)(z_{1:n+1}^{i \leftrightarrow n+1}, z_i) &\sim (\mu_k + \varepsilon)(z_{1:n+1}, z_{n+1}) \\
\text{So, } \mu_k(z_{1:n}, z_{n+1}) &\le (\mu_k + \varepsilon)(z_{1:n+1}, z_{n+1}) \\
&\overset{P}{=} (\mu_k + \varepsilon)(z_{1:n+1}^{i \leftrightarrow n+1}, z_i)
\end{aligned}$$

Taking expectation on both sides, and averaging over i would lead to the desired result □

18.2 First Order Optimality

Lemma 18.3. $f : C \to \mathbb{R}, C \subseteq \mathbb{R}^d$ *closed, convex,* $C \ne \varnothing$

$$x^* \in \underset{x \in C}{\text{argmin}} f(x).$$

(1.) $\exists \theta \in \partial f\left(x^*\right)$ *s.t.* $\theta^T\left(x - x^*\right) \ge 0.$
(2.) Assume f *is* $\lambda - SOC, \tilde{\tau} \in \partial f(x), \tilde{\theta}^\top\left(x^* - x\right) \ge -g\left\|x - x^*\right\|$ *for some* $g > 0 \Rightarrow \left\|x - x^*\right\| \le \frac{g}{\lambda}$.

Proof.

$$f'(x^*; x - x^*) = \theta^\top(x - x^*) \text{ for some } \theta \in \partial f(x)$$

$$\begin{aligned} f(x^*) &\geqslant f(x) + \tilde{\theta}^\top (x^* - x) + \frac{\lambda}{2}\|x^* - x\|^2 \\ &\geqslant f(x) - g\|x - x^*\| + \frac{\lambda}{2}\|x^* - x\|^2 \end{aligned}$$

$$\begin{aligned} f(x) &\geq f(x^*) + \theta^\top (x - x^*) + \frac{\lambda}{2}\|x - x^*\|^2 \\ &\geqslant f(x^*) + \frac{\lambda}{2}\|x - x^*\|^2 \qquad \text{(Using 1)} \end{aligned}$$

$$g\|x - x_*\| \geqslant \lambda\|x - x^*\|^2$$

Q.E.D.

□

Theorem 18.4. $G = \{g_w : w \in C\}, C \subseteq \mathbb{R}^d$ *closed and convex.* $w \mapsto g_w(z) \quad (W_1, \|\cdot\|_2) \to (\mathbb{R}, |\cdot|) \quad G(z)$*-Lipschitz*
$h : \mathbb{R}^d \to \mathbb{R}; \quad \bar{g}_w(z) = g_w(z) + h(w)$.
Assume $\omega \mapsto P_{z_{1:n}}\bar{g}_w$ *is* λ*-SOC (* λ *Strongly Convex) ;* $\forall z_{1:n} \in Z^n$

$$\begin{aligned} \mathcal{A}(z_{1:n}) &:= \underset{w \in C}{\operatorname{argmin}} P_{z_{1:n}} \bar{g}_w = \underset{w \in C}{\operatorname{argmin}} P_{z_{1:n}} g_w + h(w) \\ \mathcal{A}(z_{1:n+1}) &= \underset{w \in C}{\operatorname{argmin}} P_{z_{1:n+1}} g_w + \frac{n}{n+1} h(w) \end{aligned}$$

Then: 1.) $\mathcal{A}$ *is* $\varepsilon\left(z_{1:n+1}, z'_{n+1}\right) = \frac{G\left(z'_n+1\right)^2}{\lambda(n+1)}$*−LOO Stable*
2.) $\mathcal{A}$ *is* $\frac{2\|G\|_\infty^2}{\lambda n}$ *uniformly stable*
3.) $z_{1:n} \sim p^{\otimes n}$*. Then, for* $w_n = \mathcal{A}(z_{1:n})$*,*

$$\mathbb{E}Pg_{w_n} \leqslant \operatorname{iuf}_{\omega \in C}\left(Pg_w + h(\omega)\right) + \frac{\mathbb{E}G^2(z_1)}{\lambda(n+1)}$$

Proof.
Part 1: L.O.O Stability

$$\begin{aligned} \text{Let } Z_{1:n+1} &\sim P^{\otimes n+1} \\ L_n(w) &= P_{z_{1:n}} \bar{g}_w \\ w_n &= \mathcal{A}(z_{1:n}) = \operatorname{argmin}_w \; L_n(w) \\ L_{n+1}(w) &= P_{z_{1:n+1}} g + \frac{n}{n+1} h(w) \\ w_{n+1} &= \mathcal{A}(z_{1:n+1}) = \operatorname{argmin}_w \; L_{n+1}(w) \end{aligned}$$

Assume L_n, L_{n+1} are differentiable. By F.O.O. Lemma (18.4): $\theta_n := \nabla L_n(w_n)$ s.t.

$$\theta_n^\top (w_{n+1} - w_n) \geqslant 0. \quad (*)$$

$$\text{Now, } L_{n+1}(w) = \frac{n}{n+1} L_n(w) + \frac{1}{n+1} g_w(z_{n+1}),$$
$$\nabla L_{n+1}(w_n) = \frac{n}{n+1} \nabla L_n(w_n) + \frac{1}{n+1} \nabla g_{w_n}(z_{n+1}),$$
$$\Rightarrow \nabla L_{n+1}(w_n)^T (w_{n+1} - w_n) \overset{(*)}{\geqslant} \frac{1}{n+1} \nabla g_{w_n}(z_{n+1})^\top (w_{n+1} - w_n),$$
$$\geqslant -\frac{G(z_{n+1})}{n+1} \|w_{n+1} - w_n\|,$$
$$\downarrow$$
$$\|\nabla g_w\| \leqslant G, \qquad \text{(Cauchy Schwartz).}$$

Now by using F.O.O Lemma (Part 2), we get

$$\Rightarrow \|W_{n+1} - W_n\| \leq \frac{G(z_{n+1})}{\lambda(n+1)}$$
$$\Rightarrow g_{w_n}(z_{n+1}) - g_{w_{n+1}}(z_{n+1}) \leq G(z_{n+1}) \|w_n - w_{n+1}\|$$
$$\leq \frac{G^2(z_{n+1})}{\lambda(n+1)}$$
$$\Rightarrow \mathcal{A} \text{ is } \frac{Q^2}{\lambda(t+1)} - L \cdot O.O \text{ stable}$$

Part 2: Uniform Stability Proof for Uniform stability follows similary but with the use of $\|\cdot\|_\infty$

Part 3: Symmetry

$$\mathcal{A}\left(z_{1:n+1}^{i \leftrightarrow n+1}\right) = \mathcal{A}(z_{1:n+1}),$$
$$\varepsilon\left(z_{1:n+1}^{i \oplus n+1}\right) = \varepsilon(z_1 : n-1).$$

LOO Theorem: $Z_{1:n} \sim P^{\otimes n}; W_n : A(Z_{1:n}), W_{n+1} = A(Z_{1:n+1})$.

$$\begin{aligned}
\mathbb{E} P g_{w_n} &\leq \mathbb{E} P_{n+1} g_{w_{n+1}} + \mathbb{E} P_{n+1} \varepsilon(z_{n:n+1}) \\
&\leq \mathbb{E} L_{n+1}(w_{n+1}) + \frac{\mathbb{E} G^2(z_1)}{\lambda(n+1)} \\
&\leq \mathbb{E} L_{n+1}(\omega) + \frac{1}{n+1} h(\omega) + \frac{\mathbb{E} G^2(z_1)}{\lambda(n+1)} \\
&= P g_w + h(\omega) + \frac{\mathbb{E} G^2(z_1)}{\lambda(n+1)}
\end{aligned}$$

$\square$

Example)

$$\overset{hingeloss}{l\left(f_1(x,y)\right)} = \max(1 - f(\tilde{x})y, 0), y \in \pm 1$$
$$f_w(\alpha) = w^\top \psi(x), w \in \mathbb{R}^d$$
$$g(w) = \frac{\lambda}{2}\|w\|^2; \quad h = 0.$$

$$\text{Then } \mathbb{E}\left[P\ell \circ f_{w_n}\right] \leq \inf_w Pl \circ f_w + \frac{\lambda}{2}\|w\|^2 + \frac{\mathbb{E}\left(\|\psi(X)\| + \sqrt{2\lambda}\right)^2}{\lambda(n+1)}$$

Proof. Consider the function defined by

$$g_\omega(x,y) = \ell\left(f_{w_1}(x,y)\right) + \frac{\lambda}{2}\|\omega\|^2. \tag{18.1}$$

The mapping $\omega \mapsto g_w(z)$ induces a gradient

$$\nabla g_{w(1)} = \underbrace{l'(fw,z)}_{\in\{0,1\}}\psi + \underbrace{\lambda\omega}_{\text{unbounded}}. \tag{18.2}$$

Given $z_{1:n} \in Z^n$ and $w_n \in \operatorname{argmin} L_n(w)$, we have

$$\frac{\lambda}{2}\left\|w_n\right\|^2 \leq L_n\left(w_n\right) \leq L_n(0) \leq 1 \Rightarrow \|w_n\| \leq \sqrt{\frac{2}{\lambda}}, \tag{18.3}$$

where

$$\mathcal{A}\left(z_{1:n}\right) = \operatorname{argmin}_{w\in C} L_n(w) \quad \text{and} \quad C = \left\{w : \|w\| \leq \sqrt{\frac{2}{\lambda}}\right\}. \tag{18.4}$$

□

Theorem 18.5. *Assume $\forall z \in Z$, the mapping $w \mapsto g_w(z)$ is λ-strongly convex and L-smooth. Then the following properties hold:*

$$\mathcal{A}\left(z_{1:n}\right) = \operatorname*{argmin}_w P_{z_{1:n}} g_w, \tag{18.5}$$
$$\mathcal{A}\left(z_{1:n+1}\right) = \operatorname{argmin}_w P_{z_{1:n+1}} g_w, \tag{18.6}$$
$$\varepsilon\left(z_{1:n+1}, z'_{n+1}\right) = \left(1 + \frac{L}{2\lambda n}\right)\frac{\|\nabla g_{\mathcal{A}(z_{1:m})}\left(z'_{n+1}\right)\|_2^2}{\lambda n}. \tag{18.7}$$

1. *$\mathcal{A}$ is ε-L.O.O stable.*

2. *If $L \leq 0.2\lambda n$, then*

$$\mathbb{E}Pg_{\omega_n} \leq \inf_\omega\left(Pg_\omega + \frac{2.2}{\lambda n}P\|\nabla g_\omega\|_2^2\right). \tag{18.8}$$

Example: let the loss function l and regularization term g be defined as:

$$l\left(f_1(x,y)\right) = (f(x) - y)^2,$$
$$g(w) = \frac{\lambda}{2}\|w\|_2^2.$$

Then, the model f parameterized by weights w, and the composite objective function g_ω can be expressed as:

$$\begin{aligned} f &= f_w = w^\top \psi, \\ g_\omega(z) &= l\left(f_w, z\right) + g(\omega), \text{ implying strong convexity } (\lambda - \text{SOC}). \\ \nabla g_w(z) &= 2\left(f_w(x) - y\right)\psi(x) + \lambda\omega, \\ \nabla^2 g_w(z) &= 2\psi(x)^\top \psi(x) + \lambda I, \\ \lambda_{\max}\left(\nabla^2 g_\omega(z)\right) &\leq 2\|\psi(x)\|^2 + \lambda. \end{aligned}$$

We define the Lipschitz constant L as:

$$L := \sup_x \left(2\|\psi(x)\|^2 + \lambda\right).$$

If $L \leqslant 0.2\lambda n$, it follows that:

$$\mathbb{E}Pg_{w_n} \leq \inf_\omega \left(Pg_\omega + \frac{8.8}{\lambda n} \times \mathbb{E}\|\psi(X)^2\|(f_w(x) - y)^2\right).$$

Assuming there exists a w_* such that:

$$\mathbb{E}\left[(f_{w_*}(X) - Y)^2 \mid X\right] \leq 0^2 \text{ almost surely (a.s.)},$$

We can deduce that:

$$\mathbb{E}Pg_{w_n} \leq \sigma^2 + \frac{\lambda}{2}\|w_*\|_2^2 + \frac{8.8\sigma^2}{\lambda n}\mathbb{E}\|\psi(x)\|^2.$$

H.W: Compare to result that does not use smoothness!

18.3 Bibliography