MATH/COMP 562 LECTURE NOTES
STABILITY THEORY

ADAM M. OBERMAN

Contents

Notes adapted from [MRT18, Chapter 14] and [SSBD14, Ch 13]

1. Convex analysis

1.1. Lipschitz and strong convexity (for functions).

Definition 1.1. The function $f : W \to \mathbb{R}$ is Lipchitz continuous with constant C_f, if

$$|f(w_1) - f(w_2)| \leq C_f \|w_1 - w_2\|, \quad \forall w_1, w_2 \in W$$

Definition 1.2. The function $f : W \subset \mathbb{R}^d \to \mathbb{R}$ is λ-strongly convex (on W), if $f(w) - \lambda\|w\|^2$ is convex on W.

Exercise 1.1. Show that f λ convex on $\mathbb{R}^d$ means $f(w) - \lambda\|w - w_0\|^2$ is convex for any $w_0 \in \mathbb{R}^d$

Lemma 1.3. *Suppose f is λ-strongly convex. Let w^* be a minimizer of f. Then*

$$\text{(SC)} \qquad f(w) - f(w^*) \geq \lambda\|w - w^*\|^2, \qquad \forall w$$

Remark 1.4. Intuition of this: consider $f : \mathbb{R} \to \mathbb{R}$ smooth. Then Taylor expansion around w^*:

$$f(w^* + v) = f(w^*) + \nabla f(w^*) \cdot v + v^2 f''(w^*)/2 + O(v^3)$$

Set $\lambda = f''$ for a local version of the inequality. So the global version of this idea given by strong convexity.

Date: April 12, 2023.

2. Stability Theory Setup

We are considering a supervised learning problem (classification or regression). For now, consider the case of binary classification, or one dimensional regression. We may assume normalized vector data $\mathcal{X} \subset [-1,1]^d \subset \mathbb{R}^d$. We assume labels are either $\mathcal{Y} \subset \mathbb{R}$ or $\mathcal{Y} = \{-1, 1\}$

Definition 2.1. Given the ML setup $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and a class of functions $\mathcal{H}$, $h : \mathcal{X} \to \mathcal{Y}$. A learning algorithm is a operator A which takes a finite subset, $Z \subset \mathcal{Z}$ and returns a function $h = A(Z)$.

Note, in the parametric setting $\mathcal{H} = \{h(x, w) \mid w \in W\}$, so we have $h_w(x) = A(Z)$. We can simply write $w = A(Z)$.

Write $z = (x, y)$ and for the dataset, $S^m = \{(x_i, y_i)\}_{i=1}^m$ write

$$Z = (z_1, \dots, z_m) = ((x_1, y_1), \dots, (x_m, y_m))$$

Definition 2.2 (bounded data)**.** We say the learning problem $(\mathcal{X}, \mathcal{Y})$ is bounded if there exist constants r_x, r_y such that

$$\|x\| \leq r_x, \quad \forall x \in \mathcal{X},$$

and

$$|y| \leq r_y, \quad \forall y \in \mathcal{Y}$$

We are given a loss function

$$\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}^+$$

Example 2.3. In the case of regression,

$$\ell(f, y) = (f - y)^2$$

Example 2.4. For classification, with, $Y = -1, +1$, consider,

$$\ell(f, y) = \begin{cases} -\log(\sigma(f)), & y = +1 \\ -\log(1 - \sigma(f)), & y = -1 \end{cases}$$

3. Stability Theory

Definition 3.1. Define the bounded linear hypotheses

$$\mathcal{H}_{lin,W} = \{h(x, w) = w \cdot x \mid x \in \mathcal{X}, w \in W\}$$

Let $W \subset \mathbb{R}^d$ be a convex and bounded, with radius C_w.

Given a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, define $\ell : W \times Z \to \mathbb{R}$ by

$$\ell(w, z) = \ell(w \cdot x, y)$$

Then, in this overloaded notation,

$$\ell(h_w(x), y) = \ell(w \cdot x, y) = \ell(w, z), \quad \forall h \in \mathcal{H}_{lin}$$

3.1. **Strongly convex losses.** Given a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, define $\ell_w : Z \times \mathbb{R}^+ \to \mathbb{R}$ by

$$\ell_w(z, \lambda) = \ell(w, z, \lambda) = \ell(w \cdot x, y) + \lambda \|w\|^2$$

Lemma 3.2. *Suppose $\ell(s, y)$ is a convex function s, (for all z), and $\lambda \geq 0$. Then $\ell(w, z, \lambda)$ is λ-strongly convex function of w, for all $z = (x, y)$.*

Proof. Class Notes / HW □

We say a loss is Lipschitz continuous if it is Lipschitz continuous as a function of w, independent of z.

Definition 3.3. The loss $\ell : W \times Z \to \mathbb{R}$ is Lipchitz continuous with constant C_ℓ, if

$$|\ell(w_1, z) - \ell(w_2, z)| \leq C_\ell \|w_1 - w_2\|, \quad \forall w_1, w_2 \in W, z \in Z \tag{Lip}$$

Definition 3.4. Given a dataset $Z = \{z_1, \ldots, z_m\}$, where each $z_i \in \mathcal{Z}$. Given a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}^+$ Define for $w \in W$,

$$L(w, Z, \lambda) = \frac{1}{m} \sum_{i=1}^{m} \ell(w, z_i) + \lambda \|w\|^2$$

3.2. **Define stability.**

Remark 3.5. In what follows, we will usually have $\beta(m) = C/m$

Definition 3.6 (Replace one Stability)**.** Given $\ell, \rho(z)$. Let Z_1, Z_2 be two datasets of size m, which differ in exactly one element.

The operator A is replace one stable in w if there exists $C_w > 0$ such that

$$\|A(Z_1) - A(Z_2)\| = \|w_1 - w_2\| \leq \frac{C_w}{m} \tag{S1}$$

The operator A is uniformly replace one stable in ℓ with rate $\beta(m)$, if there exists a function $\beta = \beta(m)$

$$|\ell(w_1, z) - \ell(w_2, z)| \leq \beta(m), \quad \forall z \in \mathcal{Z} \tag{1}$$

The operator A is replace one stable in the expected loss, with rate $\beta(m)$, if there exists a function $\beta = \beta(m)$ such that

$$L(A(Z_1)) - L(A(Z_2)) \leq \beta(m) \tag{S2}$$

for all datasets Z_1, Z_2 of size m which differ by only one element.

3.3. **Stability Theorem.** Define

$$w_i = A(Z_i) = \arg\min_{w \in W} L(w, Z_i, \lambda), \qquad i = 1, 2$$

Lemma 3.7. *Suppose the loss is C_ℓ Lipschitz continuous. Then* (S1) *implies* (1) *and* (S2) *with* $C_L = C_\ell C_w$

Proof.

$$\begin{aligned} \ell(w_1, z) - \ell(w_2, z) &\leq C_\ell \|w_1 - w_2\| && \text{by (Lip)} \\ &\leq C_\ell C_w \frac{1}{m} && \text{by (S1)} \end{aligned}$$

The second result comes from taking expectations of the first inequality.

□

Definition 3.8. Given the loss, ℓ, the dataset, Z, and $\lambda > 0$, define the regularized empirical loss

$$L(w, Z, \lambda) = \frac{1}{m}\sum_{i=1}^{m} \ell(w, z_i) + \lambda \|w\|^2 \tag{REL}$$

The regularized loss minimization problem is to set

$$w = A(Z, \lambda) = \arg\min_w L(w, Z, \lambda) \tag{RLM}$$

Theorem 3.9. *Given the loss $\ell : W \times Z$ which is*

(1) convex in w
(2) Lipschitz continuous in w, as defined by (Lip)

Then the regularized loss minimization problem, (RLM)*, is replace one stable in w with constant $C_w = C_\ell/\lambda$ and replace on stable in the expected loss, with constant $C_L = C_\ell^2/\lambda$.*

Proof. Let S_1, S_2 differ by one point,

$$S_1 = \{z_1, \dots z_{m-1}, z_1'\}, \qquad S_2 = \{z_1, \dots z_{m-1}, z_2'\} \tag{RO}$$

Define w_1, w_2 by (RLM),

$$w_i = w_i(S_i, \lambda) = \arg\min_w L(w, S_i, \lambda), \quad i = 1, 2$$

Since ℓ is convex in w, $L(w, S, \lambda)$ is λ-strongly convex in w. Thus, applying strong convexity of (REL), and the definition of w_1, w_2, we obtain

$$\lambda\|w_1 - w_2\|^2 \le L(w_2, S_1, \lambda) - L(w_1, S_1, \lambda) \qquad \text{by (SC) and (RLM)}$$
$$\lambda\|w_1 - w_2\|^2 \le L(w_1, S_2, \lambda) - L(w_2, S_2, \lambda) \qquad \text{by (SC) and (RLM)}$$

Adding the two inequalities above,

$$2\lambda\|w_1 - w_2\|^2 \le L(w_2, S_1, \lambda) - L(w_1, S_1, \lambda) + L(w_1, S_2, \lambda) - L(w_2, S_2, \lambda)$$

Using the fact that the datasets differ by only one point, we have

$$L(w_2, S_1, \lambda) - L(w_2, S_2, \lambda) = \tfrac{1}{m}\left(\ell(w_2, z_1') - \ell(w_2, z_2')\right) \qquad \text{by (RO)}$$
$$L(w_1, S_2, \lambda) - L(w_1, S_1, \lambda) = \tfrac{1}{m}\left(\ell(w_1, z_2') - \ell(w_1, z_1')\right) \qquad \text{by (RO)}$$

Combining the last three lines,

$$2\lambda\|w_1 - w_2\|^2 \le \frac{1}{m}\left(\ell(w_2, z_1') - \ell(w_2, z_2') + \ell(w_1, z_2') - \ell(w_1, z_1')\right)$$

Now apply the Lipschitz condition to obtain

$$\ell(w_2, z_1') - \ell(w_1, z_1') \le C_\ell\|w_1 - w_2\| \qquad \text{by (Lip) at } z_1'$$
$$-\ell(w_2, z_2') + \ell(w_1, z_2') \le C_\ell\|w_1 - w_2\| \qquad \text{by (Lip) at } z_z'$$

Combining, gives

$$2\lambda\|w_1 - w_2\|^2 \le \frac{2C_\ell}{m}\|w_1 - w_2\|$$

Simplify the last inequality to obtain

$$\|w_1 - w_2\| \le \frac{C_\ell}{\lambda m}$$

as desired. The second result follows directly from Lemma 3.7. □

Remark 3.10. This proof used a symmetric differences technique. We started by fixing the dataset and changing w to get the first inequality. Later we fixed the w and changed the dataset to go from L to ℓ. Finally, we fixed the z and changed the w again to get another inequality.

4. Learning bounds for stable algorithms

Definition 4.1. Define

$$\Phi(S) = L(A(S)) - L_S(A(S)) \tag{2}$$

to be the gap between the expected loss and the *training* loss of an algorithm.

Remark 4.2. The learning bounds for stable algorithms come from applying McDiarmid's inequality to Φ. This is similar to what was done for Rademacher complexity bounds. We need to

(1) Show that Φ satisfies the bounded differences inequality
(2) Apply McDiarmid's inequality to Φ
(3) Bound the expected value of Φ and relate this to the quantity of interest (in this case, stability).

4.1. Bounded differences.

Lemma 4.3. *Let Φ be defined by* (2)*. Let A be a uniformly β-stable algorithm (in expectation). Suppose, in addition, that the loss is bounded by M. Then $\Phi(S)$ satisfies the bounded difference inequality*

$$|\Phi(S_1) - \Phi(S_2)| \leq 2\beta(m) + \frac{M}{m} \tag{3}$$

(where S_1, S_2 differ by one point).

Proof of Lemma 4.3 . Using local notation, since copied from [MRT18].

Let Φ be defined for all samples S by

$$\Phi(S) = R\left(h_S\right) - \widehat{R}_S\left(h_S\right)$$

Let S' be another sample of size m with points drawn i.i.d. according to $\mathcal{D}$ that differs from S by exactly one point. We denote that point by z_m in S, z_m' in S', i.e.,

$$S = (z_1, \ldots, z_{m-1}, z_m) \quad \text{and} \quad S' = (z_1, \ldots, z_{m-1}, z_m')$$

By definition of Φ, the following inequality holds:

$$|\Phi\left(S'\right) - \Phi(S)| \leq |R\left(h_{S'}\right) - R\left(h_S\right)| + \left|\widehat{R}_{S'}\left(h_{S'}\right) - \widehat{R}_S\left(h_S\right)\right|$$

We bound each of these two terms separately.

First, by the β-stability of $\mathcal{A}$, (S2), we have

$$|R\left(h_S\right) - R\left(h_{S'}\right)| = \left|\mathbb{E}_z\left[L_z\left(h_S\right)\right] - \mathbb{E}_z\left[L_z\left(h_{S'}\right)\right]\right| \leq \mathbb{E}_z\left[\left|L_z\left(h_S\right) - L_z\left(h_{S'}\right)\right|\right] \leq \beta$$

$$\begin{aligned}\left|\widehat{R}_S\left(h_S\right) - \widehat{R}_{S'}\left(h_{S'}\right)\right| &= \frac{1}{m}\left|\left(\sum_{i=1}^{m-1} L_{z_i}\left(h_S\right) - L_{z_i}\left(h_{S'}\right)\right) + L_{z_m}\left(h_S\right) - L_{z_m'}\left(h_{S'}\right)\right| \\ &\leq \frac{1}{m}\left[\left(\sum_{i=1}^{m-1}\left|L_{z_i}\left(h_S\right) - L_{z_i}\left(h_{S'}\right)\right|\right) + \left|L_{z_m}\left(h_S\right) - L_{z_m'}\left(h_{S'}\right)\right|\right]\end{aligned}$$

Using the uniform β-stability of $\mathcal{A}$, (1), for the first terms, along with boundedness of L, for the last term, we have

$$\left|\widehat{R}_S\left(h_S\right)-\widehat{R}_{S'}\left(h_{S'}\right)\right| \leq \frac{m-1}{m}\beta(m)+\frac{M}{m} \leq \beta(m)+\frac{M}{m}$$

Thus, Φ satisfies (3). □

In the next result, we take expectations of Φ.

Lemma 4.4. *Let* Φ *be defined by* (2). *Then*

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\Phi(S) \leq \underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}\left[\left|L_z\left(h_S\right)-L_z\left(h_{S'}\right)\right|\right]$$

Proof. Local notation scope: we are copying from Mohri, so using his notation.

Rewrite

$$\Phi(S)=R(h_S)-\widehat{R}_S(h_S)$$

We now bound the expectation term, first noting that by linearity of expectation

$$\mathbb{E}_S[\Phi(S)]=\mathbb{E}_S\left[R\left(h_S\right)\right]-\mathbb{E}_S\left[\widehat{R}_S\left(h_S\right)\right]$$

By definition of the generalization error,

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[R\left(h_S\right)\right]=\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[\underset{z\sim\mathcal{D}}{\mathbb{E}}\left[L_z\left(h_S\right)\right]\right]=\underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}\left[L_z\left(h_S\right)\right]$$

By the linearity of expectation,

$$\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[\widehat{R}_S\left(h_S\right)\right]=\frac{1}{m}\sum_{i=1}^{m}\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[L_{z_i}\left(h_S\right)\right]=\underset{S\sim\mathcal{D}^m}{\mathbb{E}}\left[L_{z_1}\left(h_S\right)\right]$$

where the second equality follows from the fact that the z_i are drawn i.i.d. and thus the expectations $\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{z_i}\left(h_S\right)\right], i\in[m]$, are all equal. The last expression in the equation above is the expected loss of a hypothesis on one of its training points. We can rewrite it as

$$\mathbb{E}_{S\sim\mathcal{D}^m}\left[L_{z_1}\left(h_S\right)\right]=\mathbb{E}_{S,z\sim\mathcal{D}^{m+1}}\left[L_z\left(h_{S'}\right)\right]$$

where S' is a sample of m points containing z extracted from the $m+1$ points formed by S and z. Thus,

$$\begin{aligned}\left|\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[\Phi(S)]\right| &= \left|\underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}\left[L_z\left(h_S\right)\right]-\underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}\left[L_z\left(h_{S'}\right)\right]\right| \\ &\leq \underset{S,z\sim\mathcal{D}^{m+1}}{\mathbb{E}}\left[\left|L_z\left(h_S\right)-L_z\left(h_{S'}\right)\right|\right]\end{aligned}$$

as desired. □

Combining the lemma and the theorem, it means we can apply McDiarmid's inequality to $\Phi(S)$. Thus we have

Theorem 4.5. *Assume that the loss function* L *is bounded by* $M\geq 0$. *Let* $\mathcal{A}$ *be a uniformly* β*-stable learning algorithm. Let* S *be a sample of* m *points drawn i.i.d. according to distribution* $\mathcal{D}$. *Then, with probability at least* $1-\delta$ *over the sample* S *drawn, the following holds:*

$$R\left(h_S\right) \leq \widehat{R}_S\left(h_S\right)+\beta+(2m\beta+M)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

In particular, in the case $\beta = C/m$, we have

$$R(h_S) \leq \widehat{R}_S(h_S) + \frac{C}{m} + (2C + M)\sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

Proof. Use the two previous results, to apply McD inequality.
Also apply the uniform stable definition. □

References

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.