Rademacher Complexity and Lipschitz Functions

Amiran George Ambroladze & John Shawe-Taylor

University of Southampton, UK

Definition 1 (Rademacher complexity) *Let X be an input space, D be a distribution on X, and F be a real-valued function class defined on X. Let $S = \{x_1, \cdots, x_l\}$ be a random sample generated (independently) by D. The* empirical Rademacher complexity *of F for the given sample S is the following random variable:*

$$\hat{R}_l(F) = \mathbb{E}_r \left[\sup_{f \in F} \frac{2}{l} \left| \sum_{i=1}^{l} r_i f(x_i) \right| \right],$$

where $r = \{r_1, \cdots r_l\}$ are iid $\{\pm 1\}$-valued random variables with equal probabilities for $+1$ and -1 and the expectation is taken with respect to r.

The Rademacher complexity *of F is*

$$R_l(F) = \mathbb{E}_S \left[\hat{R}_l(F) \right].$$

Theorem 1 *Fix $\delta \in (0,1)$ and let H be a class of functions mapping from $Z = X \times \{1, -1\}$ to $[0,1]$. Let $z_1, \cdots, z_l$ be drawn independently according to a probability distribution D. Then with probability at least $1-\delta$ over random draws of samples of size l, every $h \in H$ satisfies:*

$$\begin{aligned}\mathbb{E}_D[h(z)] &\leq \hat{\mathbb{E}}[h(z)] + R_l(H) + \sqrt{\frac{\ln(2/\delta)}{2l}} \\ &\leq \hat{\mathbb{E}}[h(z)] + \hat{R}_l(H) + 3\sqrt{\frac{\ln(2/\delta)}{2l}},\end{aligned}$$

where $\mathbb{E}_D[h(z)]$ is the true expectation of $h(z)$ and $\hat{\mathbb{E}}[h(z)]$ is the corresponding empirical one.

Theorem 2 *Let $F, F_1, \cdots, F_n$ and G be classes of real functions. Then:*

(1) $If\ \ F \subseteq G,\ then\ \hat{R}_l(F) \leq \hat{R}_l(G).$
(2) $\hat{R}_l(F) = \hat{R}_l(\mathrm{conv} F).$
(3) $\hat{R}_l(cF) = |c|\hat{R}_l(F)\ \ for\ every\ c \in \mathbb{R}.$
(4) $If\ \ A : \mathbb{R} \longrightarrow \mathbb{R}\ is\ a\ Lipschitz\ with\ constant\ L\ and\ satisfies\ A(0) = 0,$ $then$ $\mathbf{\hat{R}_l(A \circ F) \leq 2L\hat{R}_l(F)}.$
(5) $\hat{R}_l(F + h) \leq \hat{R}_l(F) + 2\sqrt{\hat{\mathbb{E}}[h^2]/l}\ \ for\ any\ h.$
(6) $\hat{R}_l(\{|F - h|^q\}) \leq 2q\left(\hat{R}_l(F) + 2\sqrt{\hat{\mathbb{E}}[h^2]/l}\right)$ $if\ 1 \leq q \leq \infty\ and\ ||f - h||_\infty \leq 1.$
(7) $\hat{R}_l(\sum_{i=1}^n F_i) \leq \sum_{i=1}^n \hat{R}_l(F_i).$

Lipschitz property for general complexity

Theorem 3 *Let μ be an arbitrary distribution on $\mathbb{R}$ with zero mean. Let $\hat{C}_l$ denote the corresponding empirical complexity for this distribution. Let $A : \mathbb{R} \longrightarrow \mathbb{R}$ be a Lipschitz function with constant L and satisfy $A(0) = 0$. Then for any real-valued function class F we have:*

$$\hat{C}_l(A \circ F) \leq 2L\hat{C}_l(F).$$

Theorem 4 *If the function A in Theorem 3 is an odd function ($A(-t) = -A(t)$) then*

$$\hat{C}_l(A \circ F) \leq L\hat{C}_l(F).$$

Proofs

Denote $f(x_i) = f_i$. We have to prove

$$\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} r_i A(f_i) \right| \right] \leq 2\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} r_i f_i \right| \right].$$

Rademacher $\Longrightarrow$ Gauss:

$$\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} \sum_{j=1}^{m} r_{ij} A(f_i) \right| \right] \leq 2\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} \sum_{j=1}^{m} r_{ij} f_i \right| \right].$$

$$\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} g_i A(f_i) \right| \right] \leq 2\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} g_i f_i \right| \right].$$

$$\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} r_i A(f_i) \right| \right] \leq 2\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} r_i f_i \right| \right].$$

Assume $\equiv 0 \in F$.

Assume $f \in F \Longrightarrow -f \in F$.

Denote $A^+(t) = A(t)$, $A^-(t) = -A(-t)$ and $A^\pm = \{A^+,\ A^-\}$.

$$\mathbb{E}_r \left[\sup_{f \in F, A \in A^\pm} \left| \sum_{i=1}^{l} r_i A(f_i) \right| \right] \leq 2\mathbb{E}_r \left[\sup_{f \in F} \left| \sum_{i=1}^{l} r_i f_i \right| \right]$$

$$\mathbb{E}_r \left[\sup_{f \in F, A \in A^\pm} \sum_{i=1}^{l} r_i A(f_i) \right] \leq 2\mathbb{E}_r \left[\sup_{f \in F} \sum_{i=1}^{l} r_i f_i \right]$$

$$\sup_{f \in F, A \in A^\pm} \sum_{i=1}^{l} r_i A(f_i) \leq$$

$$\sup_{f \in F} \sum_{i=1}^{l} r_i A^+(f_i) + \sup_{f \in F} \sum_{i=1}^{l} r_i A^-(f_i).$$

$$\mathbb{E}_r\left[\sup_{f\in F}\sum_{i=1}^{l} r_i A(f_i)\right] \leq \mathbb{E}_r\left[\sup_{f\in F}\sum_{i=1}^{l} r_i f_i\right].$$

More general:

$$\mathbb{E}_r\left[\sup_{f\in F}\sum_{i=1}^{l} r_i A_i(f_i)\right] \leq \mathbb{E}_r\left[\sup_{f\in F}\sum_{i=1}^{l} r_i f_i\right].$$

Remove all the A_i but step-by-step, one at a time. First step: Prove that

$$\mathbb{E}_r\left[\sup_{f\in F}(r_1 A_1(f_1) + r_2 A_2(f_2) + \cdots r_l A_l(f_l))\right] \leq$$

$$\mathbb{E}_r\left[\sup_{f\in F}(r_1 f_1 + r_2 A_2(f_2) + \cdots r_l A_l(f_l))\right].$$

A first naive attempt fails:

$$\left[\sup_{f\in F}(r_1 A_1(f_1) + r_2 A_2(f_2) + \cdots r_l A_l(f_l))\right] \not\leq$$

$$\left[\sup_{f\in F}(r_1 f_1 + r_2 A_2(f_2) + \cdots r_l A_l(f_l))\right].$$

Next attemt: Group $(r_1, r_2, \cdots, r_l)$, $r_1 \geq 0$, with $(-r_1, r_2, \cdots, r_l)$; then we can assert:

$$\sup_{f\in F}(r_1 A_1(f_1)+r_2 A_2(f_2)+\cdots+r_l A_l(f_l))+ \sup_{f\in F}(-r_1 A_1(f_1)+r_2 A_2(f_2)+\cdots+r_l A_l(f_l)) \leq$$

$$\sup_{f\in F}(r_1 \cdot f_1+r_2 A_2(f_2)+\cdots+r_l A_l(f_l))+ \sup_{f\in F}(-r_1 \cdot f_1+r_2 A_2(f_2)+\cdots+r_l A_l(f_l)).$$

To prove the last inequality it suffices to show that for each couple of functions $\{f^+, f^-\} \subset F$ there is another couple of functions $\{g^+, g^-\} \subset F$ such that

$$(r_1 \cdot A(f_1^+)+r_2 A_2(f_2^+)+\cdots+r_l A_l(f_l^+))+ (-r_1 \cdot A(f_1^-)+r_2 A_2(f_2^-)+\cdots+r_l A_l(f_l^-)) \leq$$

$$(r_1 \cdot g_1^+ + r_2 A_2(g_2^+)+\cdots+r_l A_l(g_l^+))+ (-r_1 \cdot g_1^- + r_2 A_2(g_2^-)+\cdots+r_l A_l(g_l^-)).$$

The choice $g^+ = f^+$, $g^- = f^-$ gives

$$A(f_1^+) - A(f_1^-) \leq f_1^+ - f_1^-.$$

The choice $g^+ = f^-$, $g^- = f^+$ gives

$$A(f_1^+) - A(f_1^-) \leq f_1^- - f_1^+.$$

Question: Is the factor 2 optimal **???**

$$\hat{C}_l(A \circ F) \leq \mathbf{2} \cdot L \cdot \hat{C}_l(F).$$

In simple simulations

$$\hat{C}_l(A \circ F) \leq \frac{\mathbf{4}}{\mathbf{3}} \cdot L \cdot \hat{C}_l(F).$$

Theorem 5 *Let H be the class of Lipschitz functions with Lipschitz constants at most L on the interval $\Delta = [0,1]$ and vanishing at some point of this interval. Then for any set of points $\{x_1, \cdots, x_l\} \subset \Delta$ we have*

$$\hat{R}_l(H) \leq 2L\hat{R}_l(\mathbf{1_\Delta}),$$

where $\mathbf{1_\Delta}$ is the function identically equal to 1 on Δ.

If we consider the class of functions vanishing at the origin we gain the factor 2:

$$\hat{R}_l(H) \leq L\hat{R}_l(\mathbf{1_\Delta}).$$

Question: Is it possible to drop the factor **2** in Theorem 5?

The factor 2 in Theorem 5 can not be made smaller then 1. (Take $h(x) = Lx$, $x_1 = x_2 = \cdots = x_l = 1$.)

We will prove that

$$\hat{R}_l(H) \le \hat{R}_l(\mathbf{1}_{\boldsymbol{\Delta}})$$

for the class of contractions vanishing at the origin.

Fix $\{x_1, \cdots, x_l\} \subset \Delta = [0,1],\ 0 \le x_1 \le x_2 \le \cdots \le x_l \le 1$. Fix $r = (r_1, \cdots, r_l),\ r_i = \pm 1,\ i = 1, \cdots, l$.

we can assume that

$$\sup_{f \in H} \left| \sum_{i=1}^{l} r_i f(x_i) \right| = \sum_{i=1}^{l} r_i h(x_i)$$

for some $h \in H$.

Denote $d_1 = x_1 - 0,\ d_2 = x_2 - x_1, \cdots, d_l = x_l - x_{l-1}$. We have

$$d_i \ge 0,\ \sum_{i=1}^{l} d_i \le 1. \qquad (1)$$

$h(0) = 0 \Rightarrow |h(x_1)| \le d_1$.

If $\mathrm{sgn}(\mathrm{r}_1 + \cdots + \mathrm{r}_l) > 0$, then we must have $h(x_1) = d_1$.

$$h(x_1) = d_1 \mathrm{sgn}(\mathrm{r}_1 + \cdots + \mathrm{r}_\mathrm{l}).$$

Now having fixed $h(x_1)$ we show in the same way that

$$h(x_2) = h(x_1) + d_2 \mathrm{sgn}(\mathrm{r}_2 + \cdots + \mathrm{r}_\mathrm{l}) = d_1 \mathrm{sgn}(\mathrm{r}_1 + \cdots + \mathrm{r}_\mathrm{l}) + \mathrm{d}_2 \mathrm{sgn}(\mathrm{r}_2 + \cdots + \mathrm{r}_\mathrm{l}).$$

In general for $i = 1, \cdots, l$ we have

$$h(x_i) = d_1 \mathrm{sgn}(\mathrm{r}_1 + \cdots + \mathrm{r}_\mathrm{l}) + \cdots + \mathrm{d}_\mathrm{i} \mathrm{sgn}(\mathrm{r}_\mathrm{i} + \cdots + \mathrm{r}_\mathrm{l}).$$

The last equality gives an expression for $\sum_{i=1}^{l} r_i h(x_i)$ only in terms of $r = (r_1, \cdots, r_l)$ (recall that $d_1, \cdots, d_l$ are fixed):

$$\sum_{i=1}^{l} r_i h(x_i) =$$
$$r_1[d_1 \mathrm{sgn}(\mathrm{r}_1 + \cdots + \mathrm{r}_\mathrm{l})] +$$
$$r_2[d_1 \mathrm{sgn}(\mathrm{r}_1 + \cdots + \mathrm{r}_\mathrm{l}) + \mathrm{d}_2 \mathrm{sgn}(\mathrm{r}_2 + \cdots + \mathrm{r}_\mathrm{l})] +$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$r_l[d_1 \mathrm{sgn}(\mathrm{r}_1 + \cdots + \mathrm{r}_\mathrm{l}) + \mathrm{d}_2 \mathrm{sgn}(\mathrm{r}_2 + \cdots + \mathrm{r}_\mathrm{l}) + \cdots + \mathrm{d}_\mathrm{l} \mathrm{sgn}(\mathrm{r}_\mathrm{l})].$$

The expectation of the last expression is exactly the empirical Rademacher complexity. In order to estimate this expectation we denote

$$m_{l-i+1} := \mathbb{E}_r[r_i \cdot \mathrm{sgn(r_i + \cdots + r_l)}].$$

Evidently it depends only on the index $l - i + 1$. Then for the Rademacher complexity we get from the last equality that (now we write h_r instead of h to indicate the dependence of h on r):

$$\mathbb{E}_r \left[\sum_{i=1}^{l} r_i h_r(x_i)\right] =$$
$$d_1 m_l +$$
$$[d_1 m_l + d_2 m_{l-1}] +$$
$$\cdots\cdots\cdots\cdots$$
$$\cdots\cdots\cdots\cdots$$
$$[d_1 m_l + d_2 m_{l-1} + \cdots + d_l m_1]$$

This gives

$$\hat{R}_l(H) =$$
$$d_1[l \cdot m_l] + d_2[(l-1) \cdot m_{l-1}] + \cdots + d_l[1 \cdot m_1].$$

We show that $m_1, \cdots, m_l$ constitute the central (middle) elements in the Pascal triangle made of binomial coefficients (here each line should be divided 2 powered by the index of the line):

$$
\begin{array}{ccccccccccc}
 & & & & & \boxed{1} & & & & & \\
 & & & & \boxed{1} & & 1 & & & & \\
 & & & 1 & & \boxed{2} & & 1 & & & \\
 & & 1 & & \boxed{3} & & 3 & & 1 & & \\
 & 1 & & 4 & & \boxed{6} & & 4 & & 1 & \\
1 & & 5 & & \boxed{10} & & 10 & & 5 & & 1 \\
. & . & . & . & . & . & . & . & . & &
\end{array}
$$

Now we can prove that

$$lm_l \geq (l-1)m_{l-1} \geq \cdots \geq 1 \cdot m_1.$$

The last inequality together with (1) show that $\hat{R}_l(H)$ will achieve its maximum if we take d_1 as big as possible, namely if we take $d_1 = 1$, which gives that $x_1 =$

$x_2 = \cdots = x_l = 1$. And the Rademacher complexity in this case will be maximal if $|h(1)|$ is as big as possible. Due to the Lipschitz condition (with constant $L = 1$) the maximal value for $|h(1)|$ is 1. We can take $h(1) = 1$ for all $r = (r_1, \cdots, r_l)$. Evidently the Rademacher complexity in this case ($x_1 = x_2 = \cdots = x_l = 1$) is the same as the Rademacher complexity of the identical one function $\mathbf{1}_{[\mathbf{0},\mathbf{1}]}$ (for arbitrary choice of $\{x_1, \cdots, x_l\} \subset [0, 1]$.

Theorem 5 is proved.