8 Combinatorial Measures of Complexity

By a careful inspection of the Massart's Finite Lemma and its proof, we notice that the functions that belong to our function class $\mathcal{F}$ enter the bounds only through their evaluation at the data points. That is, if the functions have bounded second moment under the empirical distribution over the data set, then Rademacher complexity decays with n. We will make use of this observation throughout this section.

8.1 Shattering Coefficient

Above, in our proof of Massart's Lemma, we flagged the inequality in (7.2), the step $(*)$ as the point at which we appealed to $|\mathcal{F}| < \infty$ to convert $\sup_{f\in\mathcal{F}}$ to a summation $\sum_{f\in\mathcal{F}}$.

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] \leq \mathbb{E}\left[\sum_{f\in\mathcal{F}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}\Big|z_{1:n}\right]. \tag{8.1}$$

We notice that $f \in \mathcal{F}$ enters this bound only through $f(z)$ for $z \in \mathcal{Z}$. As in our next example, $\mathcal{F}$ can be infinitely large as long as it has finite behavior over $\mathcal{Z}$.

Example. Let's assume that we have integer data points, $z \in \mathcal{Z} \subset \mathbb{Z}$, and the function class is given as $\mathcal{F} = \{z \to \sin(z\pi k), k \in \mathbb{N}\}$. We notice that even though $|\mathcal{F}| = \infty$, clearly $f(z) = 0$ for $\forall f \in \mathcal{F}$ and $z \in \mathcal{Z}$.

This behavior is not at all uncommon. Especially when we are working with loss functions with finite range, we always have finitely many function behavior over data. An example to this case is the 0-1 loss.

Example. Let's assume we are working with 0-1 loss function, and the loss class that enter the Rademacher complexity-based generalization bound is given as $\mathcal{G} = \{z \to \ell(z, f), f \in \mathcal{F}\}$. Then over data $z_1, \ldots, z_n$ we can have at most $|\mathcal{G}| = 2^n$ different assignments for the vectors $[f(z_1), ..., f(z_n)]$.

The above argument is not enough. Assume that we can replace $|\mathcal{F}|$ in the Massart's Finite Lemma, with the above exponential number 2^n. The bound on Rademacher complexity becomes $\mathcal{O}(1)$, which doesn't yield generalization. Of course, we would require sub-exponential behaviour over the data to have a useful bound which in turn yields generalization.

Let's modify the inequality (8.1) a little bit. We write

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}\Big|z_{1:n}\right] = \mathbb{E}\left[\sup_{[f_1\ldots f_n]\in\{[f(z_1)\ldots f(z_n)]:f\in\mathcal{F}\}} \exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i\right\}\Big|z_{1:n}\right]. \tag{8.2}$$

We define the shattering coefficient as follows.

Definition 31 (Shattering Coefficient). *For $\mathcal{F} = \{f : \mathcal{Z} \to \mathcal{Y}\}$, define*

$$s(\mathcal{F}, n) = \max_{z_1,\ldots,z_n\in\mathcal{Z}} \Big|\Big\{\big[f(z_1)\cdots f(z_n)\big] : f \in \mathcal{F}\Big\}\Big|.$$

The term inside our set $\{\cdot\}$ is counting how many different configurations of the vector $[f(z_1)\ldots f(z_n)]$ are possible.

We pick up from the inequality (8.2) and write

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}}\exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(z_i)\right\}|z_{1:n}\right] &= \mathbb{E}\left[\sup_{[f_1\dots f_n]\in\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}}\exp\left\{t\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i\right\}|z_{1:n}\right] \\
&\leq \sum_{[f_1\dots f_n]\in\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}}\mathbb{E}\left[\exp\left\{t\cdot\frac{1}{n}\sum_{i=1}^{n}\sigma_i f_i\right\}\right] \\
&= \sum_{[f_1\dots f_n]\in\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}}\prod_{i=1}^{n}\underbrace{M_\sigma\left(\tfrac{tf_i}{n}\right)}_{\leq\exp(t^2f_i^2/(2n^2))} \\
&\leq |\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}|\exp\left\{\frac{t^2\kappa^2}{2n}\right\} \\
&\leq \max_{z_1,\dots,z_n}|\{[f(z_1)\dots f(z_n)]:f\in\mathcal{F}\}|\exp\left\{\frac{t^2\kappa^2}{2n}\right\} \\
&= s(\mathcal{F},n)\exp\left\{\frac{t^2\kappa^2}{2n}\right\}.
\end{aligned}$$

These steps are exactly the same as before. The only difference is that instead of summing over the entire $\mathcal{F}$, this time we sum over different function evaluations. The max argument was applied over the data points to remove their dependence so we take an expectation on the empirical Rademacher complexity which would give us (population) Rademacher complexity.

We can write the following upgraded version of Massart's Finite Lemma.

Lemma 32 (**Modified Massart's Lemma**). *If* $\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}f(z_i)^2\leq\kappa^2$, *then*

$$\hat{\mathfrak{R}}_n(\mathcal{F})\leq\kappa\sqrt{\frac{2\log s(\mathcal{F},n)}{n}}.$$

Remark. Compared to Massart's Finite Lemma, $|\mathcal{F}|$ is replaced by $s(\mathcal{F},n)$.

For the 0-1 loss, we obtained a bound on shattering coefficient that grows exponentially in n, i.e. $s(\mathcal{F},n)=2^n$. Notice that exponentially growing shattering coefficient doesn't yield generalization based on the above theorem. We need at most sub-exponential growth to achieve generalization. To make this more concrete, we define the notion of a "shattered" set next.

In the sequel, we only consider Boolean functions, $f:\mathcal{Z}\to\{0,1\}$**.**

Definition 33. *Let* $\mathcal{F}$ *be a class of Boolean functions on a domain* $\mathcal{Z}$*. We say that* $\mathcal{F}$ *shatters a subset* $\mathcal{D}\subset\mathcal{Z}$ *if any function* $g:\mathcal{D}\to\{0,1\}$ *can be obtained by restricting some function* $f\in\mathcal{F}$ *to* $\mathcal{D}$*.*

Example. For the data $\mathcal{D}=\{z_1,...,z_n\}$, and $f\in\mathcal{F}$, consider the n-dimensional vectors $[f(z_1),...,f(z_n)]$. These are Boolean vectors, and if we can get every possible 2^n Boolean vectors by varying $f\in\mathcal{F}$, then $\mathcal{F}$ shatters $\mathcal{D}$. Notice that, here $\mathcal{D}$ is fixed and the Boolean vectors are changing since we change f.

For Boolean functions, if the shattering coefficient satisfies $s(\mathcal{F},n)=2^n$, this ultimately means that $\exists\mathcal{D}\subset\mathcal{Z}$ such that $\mathcal{F}$ shatters $\mathcal{D}$. It is worth restating that whenever this happens, Massart's Lemma doesn't yield generalization.

Similar to the previous section, we first justify the move from the Rademacher complexity of the loss class $\mathfrak{R}_n(\mathcal{G})$ to that of hypothesis class $\mathfrak{R}_n(\mathcal{F})$. Our next example serves as a demonstration for this.

Example. Assume that we use 0-1 loss $\ell((y,x),f) = \mathbb{1}_{\{y\neq f(x)\}} \in \{0,1\}$ and let $y \in \{\pm 1\}$ and $f : \mathcal{X} \to \{\pm 1\}$. This is not Boolean, but mapping to that case is trivial. The loss class in this case is given as $\mathcal{G} = \{(y,x) \to \mathbb{1}_{\{y\neq f(x)\}}\}$. Let (y_i, x_i) for $i = 1, 2, ..., n$ denote the samples in the data. Then notice that there is a bijection from the set of vectors $\{[f(x_1), ..., f(x_n)] : \ f \in \mathcal{F}\}$ to $\{[\ell((y_1,x_1),f), ..., \ell((y_n,x_n),f)], \ f \in \mathcal{F}\}$. This can be seen by considering the mapping from $f(x_i) \to (1 - y_i f(x_i))/2 = \ell((y_i, x_i), f)$.

Next example is a demonstration to how we calculate shattering coefficient for simple function classes.

Example. [Indicators of rays] Let's consider the function class $\mathcal{F} = \{z \to \mathbb{1}_{\{z\geq t\}} \ t \in \mathbb{R}\}$. Clearly, this function class has $|F| = |\mathbb{R}|$. But we can easily verify that $s(\mathcal{F}, n) = n + 1$. This shattering coefficient is sub-exponential and thus, the Massart's lemma will provide us with generalization. It is also worth noting that $s(\mathcal{F}, n) = 2^n$ only if $2^n = n + 1$; therefore, for $n > 1$ $\mathcal{F}$ cannot shatter any subset of size n.

8.2 Vapnik-Chervonenkis Dimension

Definition 34 (**VC-dimension of a boolean $\mathcal{F}$**). *VC dimension of $\mathcal{F}$, denoted by VC($\mathcal{F}$), is the largest cardinality of a subset $\mathcal{D} \subset \mathcal{Z}$ that can be shattered by $\mathcal{F}$.*

Remark. Notice that since we are concerned only with Boolean function classes, we can equivalently write $\mathrm{VC}(\mathcal{F})$ as

$$\mathrm{VC}(\mathcal{F}) = \sup\{n : s(\mathcal{F}, n) = 2^n\}.$$

If the VC dimension of a function class $\mathcal{F}$ is d, i.e. $\mathrm{VC}(\mathcal{F}) = d$, this means that there exists $\mathcal{D} \subset \mathcal{Z}$ with $|\mathcal{D}| = d$ such that $\mathcal{F}$ shatters $\mathcal{D}$, i.e. $s(\mathcal{F}, d) = 2^d$, and no subset $\mathcal{D} \subset \mathcal{Z}$ of size $|\mathcal{D}| > d$ can be shattered by $\mathcal{F}$, i.e. $s(\mathcal{F}, d+1) < 2^{d+1}$.

Example. If we revisit the example for the indicators of rays, we found that $s(\mathcal{F}, n) = n + 1$ for every n. Therefore, in order to get $s(\mathcal{F}, n) = 2^n$, we need $n = 1$. Also, for any $n > 1$, we have $s(\mathcal{F}, n) < 2^n$ which proves that $\mathrm{VC}(\mathcal{F}) = 1$.

Example. [Indicators of closed intervals] Consider the following boolean function class

$$\mathcal{F} = \{z \to \mathbb{1}_{\{z\in[a,b]\}}, a < b, \ a, b \in \mathbb{R}\}.$$

We can show that for $n = 1, 2$, we have $s(\mathcal{F}, n) = 2^n$. This can be done by considering every possible 2^n cases. However for $n = 3$, for z_1, z_2, z_3, we cannot obtain $[f(z_1), f(z_2), f(z_3)] = [1, 0, 1]$ using the above function class. In fact, any other configuration is achievable which makes the shattering coefficient $s(\mathcal{F}, 3) = 7$. Therefore we conclude that $\mathrm{VC}(\mathcal{F}) = 2$.

So far we consider simple function classes where it is simple to reason about their shattering coefficient. The following lemma however, can be used together with Massart's lemma and yield a generalization bound directly related to the VC-dimension of the function class $\mathcal{F}$.

Lemma 35 (**Sauer-Shelah's Lemma**). *If VC($\mathcal{F}$) = d, then*

$$s(\mathcal{F}, n) \leq \begin{cases} 2^n & \text{if } n \leq d, \\ \left(\frac{en}{d}\right)^d & \text{if } n > d. \end{cases}$$

Remark. $\mathrm{VC}(\mathcal{F})$ is the n at which the shattering coefficient stops being exponential and starts becoming polynomial (and useful for generalization). In fact, whenever $n > \mathrm{VC}(\mathcal{F})$, by the Massart's and Sauer-Shelah's lemmas, we can write

$$\begin{aligned}\mathfrak{R}_n(\mathcal{F}) \leq & \sqrt{\frac{2\log s(\mathcal{F}, n)}{n}} \leq \sqrt{\frac{2\mathrm{VC}(\mathcal{F})\log(en/\mathrm{VC}(\mathcal{F}))}{n}} \\ \leq & \sqrt{\frac{3\mathrm{VC}(\mathcal{F})\log(n)}{n}}.\end{aligned}$$

We have also seen examples that the Rademacher complexity of loss class $\mathcal{G}$, can be upper bounded by that of function class $\mathcal{F}$. Plugging this into the generalization bound obtained through Rademacher complexity (7.1), we get with probability at least $1-\delta$

$$R(\hat{f}) - R(f_*) \leq 4\sqrt{\frac{3\mathrm{VC}(\mathcal{F})\log(n)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Note that $\mathrm{VC}(\mathcal{G})$ in the bound can be replaced with $\mathrm{VC}(\mathcal{F})$ for binary classification problems (See the previous example).

Proof. Let $\mathcal{Z}^* = \{z_1^*, z_2^*, ..., z_n^*\}$ be such that $s(\mathcal{F}, n) = |\{[f(z_1^*), ..., f(z_n^*)] \ : \ f \in \mathcal{F}\}|$, restrict $\mathcal{F}$ onto $\mathcal{Z}^*$ and call it $\mathcal{F}^*$. We notice that $\mathcal{F}^*$ is finite and its size is equal to $s(\mathcal{F}, n)$ by construction, i.e. $|\mathcal{F}^*| = s(\mathcal{F}, n)$. We state the following lemma due to Pajor.

Lemma 36 (Pajor's lemma)**.** *If $\mathcal{F}^*$ is a class of Boolean functions on a finite domain $\mathcal{Z}^*$, then*

$$|\mathcal{F}^*| \leq |\{\Lambda \subset \mathcal{Z}^* \ : \ \Lambda \textit{ is shattered by } \mathcal{F}^*\}|.$$

We prove the above lemma in the homework. Now, let $d^* = \mathrm{VC}(\mathcal{F}^*)$, and by Pajor's lemma, we obtain

$$s(\mathcal{F}, n) \leq \sum_{i=0}^{d^*} \binom{n}{i} \tag{8.3}$$

where there right hand side above is the number of subsets of $\mathcal{Z}^*$ of size at most d^*.

But if $\Lambda \subset \mathcal{Z}^* \subset \mathcal{Z}$ is shattered by $\mathcal{F}^*$, it is also shattered by $\mathcal{F}$ since former is a restriction of the latter. Therefore, $\mathrm{VC}(\mathcal{F}^*) \leq \mathrm{VC}(\mathcal{F})$.

Now, if $d \geq n$, the right hand side of (8.3) is easily bounded by 2^n since $\mathcal{F}^*$ is class from domain of size n and it can shatter at most a set of size n. If $d < n$, then we get

$$\begin{aligned}s(\mathcal{F}, n) \leq & \sum_{i=0}^{d} \binom{n}{i}, \\ = & \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^i \left(\frac{d}{n}\right)^i, \\ \leq & \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i, \\ \leq & \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n, \\ \leq & \left(\frac{en}{d}\right)^d,\end{aligned}$$

which concludes the proof.

□