Margin Natarajan dimension of Multi-Layer Perceptrons

Tom Masini

LORIA - UL

Day of the department

July 11, 2022

Agnostic Learning (Kearns et al., 1994) for Pattern Classification

Problem characterization

1. Link between descriptions $x \in \mathcal{X}$ and their categories $y \in \mathcal{Y} = [\![1; C]\!]$
2. Existence of a random pair (X, Y) taking values in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, distributed according to a probability measure P
3. The joint distribution of (X, Y) is unknown.

What is available

1. $\mathbf{Z}_m = ((X_i, Y_i))_{1 \leqslant i \leqslant m}$: m-sample made up of independent copies of (X, Y)
2. $\{\mathcal{G}_k : 1 \leqslant k \leqslant C\}$: set of classes of functions from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]$ with $M_{\mathcal{G}} \geqslant 1$ which are uniform Glivenko-Cantelli (uGC) classes
3. $\mathcal{G} \subset \prod_{k=1}^{C} \mathcal{G}_k$: margin classifier, with the decision rule dr which maps every function $g \in \mathcal{G}$ to $\mathrm{dr}_g \in (\mathcal{Y} \bigcup \{*\})^{\mathcal{X}}$. For every pair $(g, x) \in \mathcal{G} \times \mathcal{X}$, $\mathrm{dr}_g(x)$ is either the index of the component function of g taking the highest value at x, or the dummy category $*$ in case of ex æquo.

Majoration bound theory

Objective

Find a Law of large numbers that uniformly upper bounds the probability of error as a function of the frequency of error and a confidence interval depending on the basic parameters : m, C and γ.

Margin operator

Definition 1 (Margin operator ρ)

Let $\mathcal{G}$ be a function class defined as above. Define ρ as an operator on $\mathcal{G}$ such that:

$$\begin{array}{rccc} \rho : & \mathcal{G} & \longrightarrow & \rho_{\mathcal{G}} \\ & g & \mapsto & \rho_g \end{array}$$

$$\forall (x, k) \in \mathcal{Z}, \ \ \rho_g (x, k) = \frac{1}{2} \left(g_k (x) - \max_{l \neq k} g_l (x) \right).$$

The function ρ_g is the margin function associated with g.

Risks

Definition 2 (Margin loss functions)

A class of margin loss functions ϕ_γ *parameterized by* $\gamma \in (0,1]$ *is a class of nonincreasing functions from* $\mathbb{R}$ *into* $[0,1]$ *satisfying:*

$$\begin{cases} \forall \gamma \in (0,1],\ \phi_\gamma(0) = 1 \text{ and } \phi_\gamma(\gamma) = 0 \\ \forall (\gamma, \gamma') \in (0,1]^2,\ \gamma < \gamma' \implies \phi_{\gamma'} \text{ majorizes } \phi_\gamma \end{cases} .$$

Definition 3 (Squashing operator π_γ)

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{T}}$. *For* $\gamma \in (0,1]$, *define the piecewise-linear squashing operator* π_γ *as:*

$$\begin{array}{cccc} \pi_\gamma : & \mathcal{F} & \longrightarrow & \mathcal{F}_\gamma \\ & f & \mapsto & f_\gamma \end{array}$$

$$\forall t \in \mathbb{R}, \quad f_\gamma(t) = f(t)\,\mathbb{1}_{\{f(t) \in (0,\gamma]\}} + \gamma \mathbb{1}_{\{f(t) > \gamma\}}.$$

Risks

The function class whose behaviour characterizes the generalization performance is

$$\rho_{\mathcal{G},\gamma} = \left\{ \rho_{g,\gamma} = \pi_\gamma \circ \rho_g : \; g \in \mathcal{G} \right\}.$$

Definition 4 (Risks)

Let $\mathcal{G}$ be a function class defined as above and ϕ_γ a margin loss function. Let P_m be the empirical measure supported on $\mathbf{Z}_m$.

$$\begin{cases} L(g) = \mathbb{E}_{Z \sim P}\left[\mathbb{1}_{\{\rho_g(Z) \leqslant 0\}}\right] = P\left(dr_g(X) \neq Y\right) & \textit{(risk)} \\ L_\gamma(g) = \mathbb{E}_{Z \sim P}\left[\phi_\gamma \circ \rho_g(Z)\right] = \mathbb{E}_{Z \sim P}\left[\phi_\gamma \circ \rho_{g,\gamma}(Z)\right] & \textit{(margin risk)} \\ L_m(g) = \mathbb{E}_{Z \sim P_m}\left[\mathbb{1}_{\{\rho_g(Z) \leqslant 0\}}\right] & \textit{(empirical risk)} \\ L_{\gamma,m}(g) = \mathbb{E}_{Z \sim P_m}\left[\phi_\gamma \circ \rho_g(Z)\right] & \textit{(empirical margin risk)} \end{cases}$$

Guaranteed Risks

Starting point

Theorem 1 (Basic supremum inequalities)

Let $\mathcal{G}$ be a function class defined as above. For $\gamma \in (0,1]$ and $\delta \in (0,1)$,

$$P^m \left\{ \sup_{g \in \mathcal{G}} \left(L_* \left(g \right) - L_{\gamma,m} \left(g \right) \right) > F_i \left(m, \gamma, \delta, cap \left(\rho_{\mathcal{G},\gamma} \right) \right) \right\} \leqslant \delta,$$

where L_* is either L or L_γ and F_i is the confidence interval, with $cap\left(\rho_{\mathcal{G},\gamma}\right)$ standing for the capacity of $\rho_{\mathcal{G},\gamma}$.

Objective

Theorem 2 (Guaranteed risks)

Let $\mathcal{G}$ be a function class defined as above. For $\gamma \in (0,1]$ and $\delta \in (0,1)$,

$$P^m \left\{ \sup_{g \in \mathcal{G}} \left(L_* \left(g \right) - L_{\gamma,m} \left(g \right) \right) > F_f \left(m, C, \gamma, \delta \right) \right\} \leqslant \delta.$$

Capacity Measures - Combinatorial Dimension

Definition 5 (γ-dimension, Kearns and Schapire, 1994)

Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For $\gamma \in \mathbb{R}_+^*$, $s_{\mathcal{T}^n} = \{t_i : 1 \leqslant i \leqslant n\} \subset \mathcal{T}$ is said to be γ-shattered by $\mathcal{F}$ if there is a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ such that, for every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1, 1\}^n$, there is a function $f_{\mathbf{s}_n} \in \mathcal{F}$ satisfying

$$\forall i \in [\![1; n]\!], \; s_i \left(f_{\mathbf{s}_n}(t_i) - b_i \right) \geqslant \gamma.$$

The γ-dimension of $\mathcal{F}$, $\gamma\text{-dim}(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{T}$ γ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite γ-dimension.

$$\mathcal{F} \subset \{-1, 1\}^{\mathcal{T}} \Longrightarrow 1\text{-dim}(\mathcal{F}) = \text{VC-dim}(\mathcal{F})$$

Canonical Scheme of Derivation of the Guaranteed Risks

$$\begin{tikzcd}
& F_i\left(m, \gamma, \delta, \mathrm{cap}\left(\rho_{\mathcal{G},\gamma}\right)\right) \arrow[dl] \arrow[d] & & & \\
R_m\left(\rho_{\mathcal{G},\gamma}\right) \arrow[r, "\text{chaining}"] \arrow[d] & \mathcal{N}_p^{\text{int}}\left(\epsilon, \rho_{\mathcal{G},\gamma}, m'\right) \arrow[d] & \leqslant & \mathcal{M}_p\left(\epsilon, \rho_{\mathcal{G},\gamma}, m'\right) \arrow[r, "\text{combinatorial res.}"] \arrow[d, phantom, "\text{structural result}"] & \epsilon'\text{-dim}\left(\rho_{\mathcal{G},\gamma}\right) \arrow[d] \\
R_m\left(\mathcal{G}_0\right) \arrow[r, "\text{chaining}"] \arrow[drr] & \mathcal{N}_p^{\text{int}}\left(\epsilon'', \mathcal{G}_0, m'\right) \arrow[dr] & \leqslant & \mathcal{M}_p\left(\epsilon'', \mathcal{G}_0, m'\right) \arrow[r, "\text{combinatorial res.}"] \arrow[d, phantom, "\text{direct computation}"] & \epsilon'''\text{-dim}\left(\mathcal{G}_0\right) \arrow[dll] \\
& & F_f\left(m, C, \gamma, \delta\right) & &
\end{tikzcd}$$

Figure: Graph of the transitions from a function F_i to a function F_f, where $\mathcal{G}_0 = \bigcup_{k=1}^{C} \mathcal{G}_k$.

Structural Result for the γ-dimension

Lemma 1 (After Lemma 6.2 in Duan, 2012)

Let $\mathcal{G}$ be a function class defined as above. For every $\gamma \in (0,1]$ and $\epsilon \in \left(0, \frac{\gamma}{2}\right]$,

$$\begin{aligned}
\epsilon\text{-}dim\left(\rho_{\mathcal{G},\gamma}\right) &\leqslant \epsilon\text{-}dim\left(\rho_{\mathcal{G}}\right) \\
&\leqslant 320 \log_2\left(\frac{24 M_{\mathcal{G}}\sqrt{C}}{\epsilon}\right) \sum_{k=1}^{C} \left(\frac{\epsilon}{96\sqrt{C}}\right)\text{-}dim\left(\mathcal{G}_k\right) \\
&\leqslant 320 \log_2\left(\frac{24 M_{\mathcal{G}}\sqrt{C}}{\epsilon}\right) C \left(\frac{\epsilon}{96\sqrt{C}}\right)\text{-}dim\left(\mathcal{G}_0\right).
\end{aligned}$$

Derivation of the Structural Result

$$F_i\left(m, \gamma, \delta, \mathrm{cap}\left(\rho_{\mathcal{G},\gamma}\right)\right)$$

$$\begin{tikzcd}
R_m\left(\rho_{\mathcal{G},\gamma}\right) & \mathcal{N}_2^{\text{int}}\left(\epsilon', \rho_{\mathcal{G},\gamma}, m'\right) \arrow[d] & \geqslant & \mathcal{M}_2\left(\epsilon, \rho_{\mathcal{G},\gamma}, m'\right) & \epsilon\text{-dim}\left(\rho_{\mathcal{G},\gamma}\right) \arrow[l, "\text{Talagrand (2003)}"'] \\
 & & & \text{Lemma 5} & \\
R_m\left(\mathcal{G}_0\right) & \mathcal{N}_2^{\text{int}}\left(\epsilon'', \mathcal{G}_0, m'\right) & \leqslant & \mathcal{M}_2\left(\epsilon'', \mathcal{G}_0, m'\right) \arrow[r, "\text{combinatorial result}"] & \epsilon'''\text{-dim}\left(\mathcal{G}_0\right)
\end{tikzcd}$$

$$F_f\left(m, C, \gamma, \delta\right)$$

Figure: Transitions from ϵ-dim $\left(\rho_{\mathcal{G},\gamma}\right)$ to ϵ'''-dim $\left(\mathcal{G}_0\right)$.

Discussion

1. Lemma 1 is useless.
2. Can a change of combinatorial dimension bring an improvement?

The fat-shattering dimension of $\rho_{\mathcal{G}}$ can be replaced with γ-Ψ-dimensions.

Derivation of Guaranteed Risks Involving γ-Ψ-dimensions

$$
\begin{tikzcd}
 & F_i\left(m, \gamma, \delta, \mathrm{cap}\left(\rho_{\mathcal{G},\gamma}\right)\right) \arrow[dl] \arrow[dr] & & & \\
R_m\left(\rho_{\mathcal{G},\gamma}\right) \arrow[rr, "\text{chaining}"] \arrow[d] & & \mathcal{N}_p^{\mathrm{int}}\left(\epsilon, \rho_{\mathcal{G},\gamma}, m'\right) \arrow[d] & \leqslant \mathcal{M}_p\left(\epsilon, \rho_{\mathcal{G},\gamma}, m'\right) \arrow[r, "\text{combinatorial res.}"] \arrow[d, phantom, "\text{structural result}"] & \epsilon'\text{-}\Psi\text{-dim}\left(\rho_{\mathcal{G}}\right) \arrow[d] \\
R_m\left(\mathcal{G}_0\right) \arrow[rr, "\text{chaining}"] \arrow[drr] & & \mathcal{N}_p^{\mathrm{int}}\left(\epsilon'', \mathcal{G}_0, m'\right) \arrow[d] & \leqslant \mathcal{M}_p\left(\epsilon'', \mathcal{G}_0, m'\right) \arrow[r, "\text{combinatorial res.}"] \arrow[d, phantom, "\text{direct computation}"] & \epsilon'''\text{-dim}\left(\mathcal{G}_0\right) \arrow[dll] \\
 & & F_f\left(m, C, \gamma, \delta\right) & &
\end{tikzcd}
$$

Figure: Paths from F_i to F_f involving γ-Ψ-dimensions of the class $\rho_{\mathcal{G}}$.

γ-Ψ-dimensions

Definition 6 (γ-Ψ-dimensions, Definition 28 in Guermeur, 2007)

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{Z}}$ be such that:

$$\forall f \in \mathcal{F},\ \forall x \in \mathcal{X},\ \max_{1 \leqslant k < l \leqslant C} \{f(x,k) + f(x,l)\} = 0.$$

Let Ψ be a family of mappings from $\mathcal{Y}$ into $\{-1,0,1\}$. For $\gamma \in \mathbb{R}_+^*$, a subset $s_{\mathcal{Z}^n} = \{z_i = (x_i, y_i) :\ 1 \leqslant i \leqslant n\}$ of $\mathcal{Z}$ is said to be γ-Ψ-shattered by $\mathcal{F}$ if there is a vector $\boldsymbol{\psi}_n = \left(\psi^{(i)}\right)_{1 \leqslant i \leqslant n} \in \Psi^n$ satisfying $\left(\psi^{(i)}(y_i)\right)_{1 \leqslant i \leqslant n} = \mathbf{1}_n$, and a vector $\mathbf{b}_n = (b_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}_+^n$ such that, for every vector $\mathbf{s}_n = (s_i)_{1 \leqslant i \leqslant n} \in \{-1,1\}^n$, there is a function $f_{\mathbf{s}_n} \in \mathcal{F}$ satisfying

$$\forall i \in [\![1;n]\!],\quad s_i \left(s_i \max_{\left\{k:\ \psi^{(i)}(k) = s_i\right\}} f_{\mathbf{s}_n}(x_i, k) - b_i \right) \geqslant \gamma. \tag{1}$$

The γ-Ψ-dimension of $\mathcal{F}$, denoted by γ-Ψ-dim$(\mathcal{F})$, is the maximal cardinality of a subset of $\mathcal{Z}$ γ-Ψ-shattered by $\mathcal{F}$, if such maximum exists. Otherwise, $\mathcal{F}$ is said to have infinite γ-Ψ-dimension.

Margin Natarajan Dimension

Definition 7 (Margin Natarajan dimension)

Let $\mathcal{F}$ be a function class defined as in Definition 6 and let $\gamma \in \mathbb{R}_+^*$.
The Natarajan dimension with margin γ of $\mathcal{F}$, denoted by γ-N-dim$(\mathcal{F})$, is the γ-Ψ-dimension of $\mathcal{F}$ corresponding to the following choice for Ψ:

$$\Psi_N = \left\{ \left(\psi_{k,l} : y \mapsto \mathbb{1}_{\{y=k\}} - \mathbb{1}_{\{y=l\}} \right) : \{k, l\} \subset \mathcal{Y} \right\}.$$

Remark 1

For the instantiation of (1) associated with the margin Natarajan dimension, choosing $\boldsymbol{\psi}_n$ is equivalent to choosing a vector $\mathbf{c}_n = (c_i)_{1 \leqslant i \leqslant n} \in \mathcal{Y}^n$ (satisfying for every $i \in [\![1; n]\!]$, $c_i \neq y_i$). Then, $\boldsymbol{\psi}_n$ is set equal to $(\psi_{y_i, c_i})_{1 \leqslant i \leqslant n}$, so that (1) becomes

$$\forall i \in [\![1; n]\!], \begin{cases} \text{if } s_i = 1, \ f_{\mathbf{s}_n}(x_i, y_i) - b_i \geqslant \gamma \\ \text{if } s_i = -1, \ f_{\mathbf{s}_n}(x_i, c_i) + b_i \geqslant \gamma \end{cases}.$$

Structural Results - Margin Natarajan Dimension

Lemma 2

Let $\mathcal{G}$ be a function class defined as above and let $\mathcal{D}_{\mathcal{G}}$ be the function class $\left\{\frac{1}{2}(g_k - g_l) : g \in \mathcal{G}, 1 \leqslant k < l \leqslant C\right\}$. Then for every value of γ in $(0, M_{\mathcal{G}}]$,

$$\gamma\text{-}N\text{-}dim(\rho_{\mathcal{G}}) \leqslant \binom{C}{2} \cdot \gamma\text{-}dim(\mathcal{D}_{\mathcal{G}})$$

and

$$\gamma\text{-}N\text{-}dim(\rho_{\mathcal{G}}) \leqslant 384 \binom{C}{2} \log_2\left(\frac{20 M_{\mathcal{G}}}{\gamma}\right)\left(\frac{\gamma}{48}\right)\text{-}dim(\mathcal{G}_0).$$

Many popular classifiers have the closure property $\mathcal{D}_{\mathcal{G}} \subset \mathcal{G}_0$.

Structural Results - Margin Natarajan Dimension

Corollary 1

Let $\mathcal{H}^{(1)}$ be a class of functions from $\mathcal{X}$ into a Hilbert space $(\mathbf{H}, \langle \cdot, \cdot \rangle_{\mathbf{H}})$ and $(\Lambda_1, \Lambda_2) \in \left(\mathbb{R}_+^*\right)^2$. Let $\mathcal{H}^{(2)}$ be the class of functions $h^{(2)}$ from $\mathcal{X}$ into $[-\Lambda_1\Lambda_2, \Lambda_1\Lambda_2]^C$ of the form:

$$\forall x \in \mathcal{X}, \quad h^{(2)}(x) = \left(\left\langle \mathbf{w}_k, h^{(1)}(x) \right\rangle_{\mathbf{H}} \right)_{1 \leqslant k \leqslant C},$$

where $h^{(1)} \in \mathcal{H}^{(1)}$ satisfies $\sup_{x \in \mathcal{X}} \left\| h^{(1)}(x) \right\|_{\mathbf{H}} \leqslant \Lambda_1$ and the vector $(\mathbf{w}_k)_{1 \leqslant k \leqslant C} \in \mathbf{H}^C$ satisfies $\max_{1 \leqslant k \leqslant C} \|\mathbf{w}_k\|_{\mathbf{H}} \leqslant \Lambda_2$. Let $\mathcal{H}_0^{(2)}$ be the class of all the component functions of the functions in $\mathcal{H}^{(2)}$. Then,

$$\forall \gamma \in (0, \Lambda_1\Lambda_2], \quad \gamma\text{-}N\text{-}dim\left(\rho_{\mathcal{H}^{(2)}}\right) \leqslant \binom{C}{2} \cdot \gamma\text{-}dim\left(\mathcal{H}_0^{(2)}\right).$$

Corollary 1 applies to both C-category SVMs and C-category MLPs.

Multi-layer Perceptrons

Definition 8 (MLPs)

Let $\mathcal{F}$ be a function class from $\mathbb{R}^{d_0}$ to $[-M/2,\ M/2]$, with $d_0 \in \mathbb{N}^*$ and $M > 0$. For $\Lambda \geqslant 0$, we define the class $\mathcal{H}$ of vector-valued functions h parameterised by $d_1 \in \mathbb{N}^*$, $(f_j)_{1 \leqslant j \leqslant d_1} \in \mathcal{F}^{d_1}$ and the matrix $A = (a_{kj}) \in \mathcal{M}_{C,d_1}(\mathbb{R})$, such that :

$$\mathcal{H} = \left\{ h = A\,(f_j)_{1 \leqslant j \leqslant d_1} :\ \|A\|_1 \leqslant \Lambda \right\}.$$

Multi-layer Perceptrons

Lemma 3

Let $\mathcal{H}_0$ be the class of binary MLPs and $\gamma \in (0, \Lambda M/2]$. Then,

$$\gamma\text{-}dim\left(\mathcal{H}_0\right) \leqslant \frac{KM^2\Lambda^2}{\gamma^2} \log_2\left(\frac{25M\Lambda}{\gamma}\right)\left(\frac{\gamma}{192\Lambda}\right)\text{-}dim\left(\mathcal{F}\right),$$

where $K = 2560$.

Corollary 2

Let $\mathcal{H}$ be a function class defined as Definition 8 and $\gamma \in (0, \Lambda M/2]$. Then,

$$\gamma\text{-}N\text{-}dim\left(\rho_{\mathcal{H}}\right) \leqslant \binom{C}{2}\frac{KM^2\Lambda^2}{\gamma^2} \log_2\left(\frac{25M\Lambda}{\gamma}\right)\left(\frac{\gamma}{192\Lambda}\right)\text{-}dim\left(\mathcal{F}\right),$$

where $K = 2560$.

Conclusion and Future work

Conclusion
First workable upper bound on a combinatorial dimension of MLPs

Ongoing research
Derivation of a sharper bound for a more general model of MLPs

Future work
Include this contribution in the general study on phase transitions

Capacity Measures - Rademacher Complexity

Definition 9 (Rademacher complexity, Bartlett and Mendelson, 2002)

Let $(\mathcal{T}, \mathcal{A}_{\mathcal{T}}, P_{\mathcal{T}})$ *be a probability space and let* T *be a random variable distributed according to* $P_{\mathcal{T}}$*. For* $n \in \mathbb{N}^*$*, let* $\mathbf{T}_n = (T_i)_{1 \leqslant i \leqslant n}$ *be an n-sample made up of independent copies of* T *and let* $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leqslant i \leqslant n}$ *be a Rademacher sequence. Let* $\mathcal{F}$ *be a class of real-valued functions with domain* $\mathcal{T}$*.*
The empirical Rademacher complexity of $\mathcal{F}$ *given* $\mathbf{T}_n$ *is*

$$\hat{R}_n(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}_n \sim \{\pm 1\}^n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(T_i) \,\middle|\, \mathbf{T}_n \right].$$

The Rademacher complexity of $\mathcal{F}$ *is*

$$R_n(\mathcal{F}) = \mathbb{E}_{\mathbf{T}_n \sim P_{\mathcal{T}}^n} \left[\hat{R}_n(\mathcal{F}) \right].$$

Capacity Measures - Covering and Packing Numbers

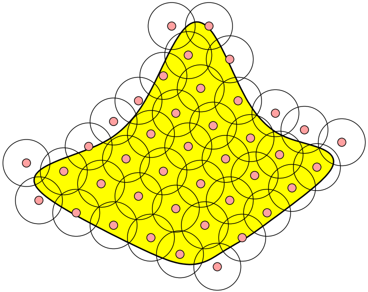

Figure: ϵ-net and ϵ-cover of a set $\mathcal{E}'$ in a pseudo-metric space $(\mathcal{E}, \rho)$

Definition 10 (Covering numbers, Kolmogorov and Tihomirov, 1961)

$\mathcal{N}(\epsilon, \mathcal{E}', \rho)$: *minimal number of open balls of radius* ϵ *needed to cover* $\mathcal{E}'$ *(or* $+\infty$*)*
$\mathcal{N}^{int}(\epsilon, \mathcal{E}', \rho)$: *the* ϵ*-nets considered are included in* $\mathcal{E}'$ *(proper to* $\mathcal{E}'$*)*

Definition 11 (Packing numbers, Kolmogorov and Tihomirov, 1961)

$\mathcal{E}' \subset \mathcal{E}$ *is* *ϵ-separated* $\Longleftrightarrow \forall \{e, e'\} \subset \mathcal{E}', \rho(e, e') \geqslant \epsilon$
$\mathcal{M}(\epsilon, \mathcal{E}', \rho)$: *maximal cardinality of an* ϵ*-separated subset of* $\mathcal{E}'$ *(or* $+\infty$*)*

Structural Results - Rademacher Complexity

Lemma 4 (After Theorem 9.2 in Mohri et al., 2018)

Let $\mathcal{G}$ be a function class defined as above. Then,

$$\forall \gamma \in (0,1]\,,\ \forall n \in \mathbb{N}^*,\ R_n\left(\rho_{\mathcal{G},\gamma}\right) \leqslant \min\left\{R_n\left(\rho_{\mathcal{G}}\right), CR_n\left(\mathcal{G}_0\right)\right\}.$$

Structural Results - Covering Numbers

Lemma 5 (Lemma 1 in Guermeur, 2017)

Let $\mathcal{G}$ be a function class defined as above. For every $\gamma \in (0,1]$, $\epsilon \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$, $p \in [1,+\infty]$, and $\mathbf{z}_n = ((x_i, y_i))_{1 \leqslant i \leqslant n} \in \mathcal{Z}^n$,

$$\mathcal{N}^{int}\left(\epsilon, \rho_{\mathcal{G},\gamma}, d_{p,\mathbf{z}_n}\right) \leqslant \mathcal{N}^{int}\left(\epsilon, \rho_{\mathcal{G}}, d_{p,\mathbf{z}_n}\right) \leqslant \prod_{k=1}^{C} \mathcal{N}^{int}\left(C^{-\frac{1}{p}}\epsilon, \mathcal{G}_k, d_{p,\mathbf{x}_n}\right)$$
$$\leqslant \left(\mathcal{N}^{int}\left(C^{-\frac{1}{p}}\epsilon, \mathcal{G}_0, d_{p,\mathbf{x}_n}\right)\right)^C,$$

where $\mathbf{x}_n = (x_i)_{1 \leqslant i \leqslant n}$.

Combinatorial Results - Margin Natarajan Dimension

Lemma 6

Let $\mathcal{F}$ be a function class defined as in Definition 6. For $\epsilon \in \mathbb{R}_+^*$, let $d_N(\epsilon) = \epsilon$-N-dim$(\mathcal{F})$. Then for every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$ such that $n \geqslant d_N\left(\frac{\epsilon}{4}\right)$,

$$\mathcal{M}_\infty(\epsilon, \mathcal{F}_\gamma, n) \leqslant \left(\frac{6\gamma\sqrt{C-1}n}{\epsilon}\right)^{d_N\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{2\gamma(C-1)en}{d_N\left(\frac{\epsilon}{4}\right)\epsilon}\right)}.$$

Lemma 7

Let $\mathcal{F}$ be a function class defined as in Definition 6. For $\epsilon \in \mathbb{R}_+^*$, let $d_N(\epsilon) = \epsilon$-N-dim$(\mathcal{F})$. Then for every $\gamma \in (0,1]$, $\epsilon \in (0,\gamma]$ and $n \in \mathbb{N}^*$,

$$\mathcal{M}_2(\epsilon, \mathcal{F}_\gamma, n) \leqslant \left((C-1)\left(\frac{4\gamma}{\epsilon}\right)^5\right)^{\frac{3}{2}\log_2\left(2\left(\frac{14\gamma}{\epsilon}\right)^2(C-1)\right)d_N\left(\frac{\epsilon}{28}\right)}.$$