Rademacher Complexity of Margin Multi-category Classifiers

Yann Guermeur

LORIA - CNRS

WSOM+ 2017

June 30, 2017

Agnostic Learning (Kearns et al., 1994)

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. For $k \in [\![1, C]\!]$, $\mathcal{G}_k$: class of functions from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]$ with $M_{\mathcal{G}} \geqslant 1$ which is a uniform Glivenko-Cantelli class

Uniform Glivenko-Cantelli class

Definition 1 (Dudley et al., 1991)

Let $(\mathcal{T}, \mathcal{A}_{\mathcal{T}})$ be a measurable space and let T be a random variable with values in $\mathcal{T}$, distributed according to a probability measure P_T on $(\mathcal{T}, \mathcal{A}_{\mathcal{T}})$. 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. Let $\mathcal{F}$ be a class of measurable functions on $\mathcal{T}$. Then $\mathcal{F}$ is a uniform Glivenko-Cantelli class if for every $\epsilon \in \mathbb{R}_+^*$,

$$\lim_{n \longrightarrow +\infty} \sup_{P_T} \mathbb{P}\left(\sup_{n' \geqslant n} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{T' \sim P_{\mathbf{T}_{n'}}} \left[f\left(T'\right) \right] - \mathbb{E}_{T \sim P_T} \left[f\left(T\right) \right] \right| > \epsilon \right) = 0,$$

where $\mathbb{P}$ denotes the infinite product measure P_T^{∞} and $P_{\mathbf{T}_{n'}}$ denotes the empirical measure supported on $\mathbf{T}_{n'}$.

Margin Classifier

Pattern Classification

1. $\mathcal{G} = \prod_{k=1}^{C} \mathcal{G}_k$: class of functions $g = (g_k)_{1 \leqslant k \leqslant C}$, from $\mathcal{X}$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]^C$
2. Decision rule dr: operator from $\mathcal{G}$ into $\left(\mathcal{Y} \bigcup \{*\}\right)^{\mathcal{X}}$ mapping g to dr_g

$$\forall g \in \mathcal{G}, \forall x \in \mathcal{X}, \begin{cases} \left|\operatorname{argmax}_{1 \leqslant k \leqslant C} g_k(x)\right| = 1 \Longrightarrow \mathrm{dr}_g(x) = \operatorname{argmax}_{1 \leqslant k \leqslant C} g_k(x) \\ \left|\operatorname{argmax}_{1 \leqslant k \leqslant C} g_k(x)\right| > 1 \Longrightarrow \mathrm{dr}_g(x) = * \end{cases}$$

where $|\cdot|$ returns the cardinality of its argument and $*$ stands for a dummy category

Function Selection

Minimization over $\mathcal{G}$ of the <u>risk</u> $L(g) = \mathbb{E}_{Z \sim P}\left[\mathbb{1}_{\{\mathrm{dr}_g(X) \neq Y\}}\right] = P\left(\mathrm{dr}_g(X) \neq Y\right)$

Margin

Definition 2 (Class of functions $\mathcal{F}_{\mathcal{G}}$)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. For every $g \in \mathcal{G}$, the function f_g from $\mathcal{X} \times [\![1, C]\!]$ into $[-M_{\mathcal{G}}, M_{\mathcal{G}}]$ is defined by:

$$\forall (x, k) \in \mathcal{X} \times [\![1, C]\!], \quad f_g(x, k) = \frac{1}{2}\left(g_k(x) - \max_{l \neq k} g_l(x)\right).$$

Then, the class $\mathcal{F}_{\mathcal{G}}$ is defined as follows: $\mathcal{F}_{\mathcal{G}} = \{f_g : g \in \mathcal{G}\}$.

Definition 3 (Margin)

Let g be a function computed by a margin multi-category classifier. The margin of g on (x, y) is defined as $f_g(x, y)$.

1. $L(g) = \mathbb{E}_{Z \sim P}\left[\mathbb{1}_{\{f_g(Z) \leqslant 0\}}\right]$
2. The margin bears useful information on the generalization performance.
3. Its exploitation calls for the implementation of a scale-sensitive approach.

Margin Risks

Definition 4 (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:*

1. $\forall \gamma \in (0, 1], \ \ \phi_\gamma(0) = 1 \wedge \phi_\gamma(\gamma) = 0$;
2. $\forall (\gamma, \gamma') \in (0, 1]^2, \ \ \gamma < \gamma' \Longrightarrow \forall t \in (0, \gamma), \ \ \phi_\gamma(t) \leqslant \phi_{\gamma'}(t)$.

Definition 5 (Margin risk)

Let $\mathcal{G}$ *be the class of functions computed by a margin multi-category classifier and* ϕ_γ *a margin loss function. The risk with margin* γ *of* $g \in \mathcal{G}$ *is defined as:*

$$L_\gamma(g) = \mathbb{E}_{Z \sim P}\left[\phi_\gamma \circ f_g(Z)\right].$$

$L_{\gamma,m}(g)$ *designates the corresponding empirical risk, measured on* $\mathbf{Z}_m$.

Guaranteed Risks

Definition 6 (Piecewise-linear squashing function π_γ)

For $\gamma \in (0,1]$, the piecewise-linear squashing function π_γ is defined by:

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

Definition 7 (Class of functions $\mathcal{F}_{\mathcal{G},\gamma}$)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. $\forall \gamma \in (0,1]$, the class $\mathcal{F}_{\mathcal{G},\gamma}$ is defined as follows: $\mathcal{F}_{\mathcal{G},\gamma} = \{f_{g,\gamma} = \pi_\gamma \circ f_g : g \in \mathcal{G}\}$.

Theorem 1 (Guaranteed risk)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. Let $\gamma \in (0,1]$ and $\delta \in (0,1)$. With P^m-probability at least $1-\delta$,

$$\sup_{g \in \mathcal{G}} \left(L(g) - L_{\gamma,m}(g)\right) \leqslant F\left(C, m, \gamma, \delta, d\left(\mathcal{F}_{\mathcal{G},\gamma}\right)\right)$$

where $d\left(\mathcal{F}_{\mathcal{G},\gamma}\right)$ is a scale-sensitive measure of the capacity of $\mathcal{F}_{\mathcal{G},\gamma}$.

Rademacher Complexity

Definition 8 (Rademacher complexity)

Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{T}_n = (T_i)_{1 \leqslant i \leqslant n}$ be a sequence of n i.i.d. random variables taking values in $\mathcal{T}$ and let $\boldsymbol{\sigma}_n = (\sigma_i)_{1 \leqslant i \leqslant n}$ be a Rademacher sequence. The <u>empirical Rademacher complexity</u> *of $\mathcal{F}$ given $\mathbf{T}_n$ is*

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

The <u>Rademacher complexity</u> *of $\mathcal{F}$ is*

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

Covering Numbers

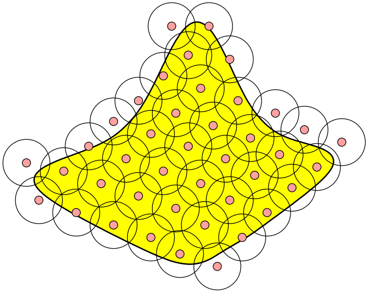

Figure : ϵ-net and ϵ-cover of a set $\mathcal{E}'$ in a pseudo-metric space $(\mathcal{E}, \rho)$

Definition 9 (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 $\mathcal{E}'$)*

Packing Numbers

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

Let $(\mathcal{E}, \rho)$ be a pseudo-metric space and $\epsilon \in \mathbb{R}_+^*$. A set $\mathcal{E}' \subset \mathcal{E}$ is ϵ-separated if, for any distinct points e and e' in $\mathcal{E}'$, $\rho(e, e') \geqslant \epsilon$. The ϵ-packing number of $\mathcal{E}'' \subset \mathcal{E}$, $\mathcal{M}(\epsilon, \mathcal{E}'', \rho)$, is the maximal cardinality of an ϵ-separated subset of $\mathcal{E}''$, if such maximum exists. Otherwise, the ϵ-packing number of $\mathcal{E}''$ is defined to be infinite.

Lemma 1 (After Theorem IV in Kolmogorov and Tihomirov, 1961)

Let $(\mathcal{E}, \rho)$ be a pseudo-metric space. For every totally bounded set $\mathcal{E}' \subset \mathcal{E}$ and $\epsilon \in \mathbb{R}_+^*$,

$$\mathcal{N}^{int}(\epsilon, \mathcal{E}', \rho) \leqslant \mathcal{M}(\epsilon, \mathcal{E}', \rho) \leqslant \mathcal{N}^{int}\left(\frac{\epsilon}{2}, \mathcal{E}', \rho\right).$$

Fat-Shattering Dimension

Definition 11 (Fat-shattering 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

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

The fat-shattering dimension with margin γ of the class $\mathcal{F}$, γ-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 fat-shattering dimension with margin γ.

Empirical Pseudo-metrics

Definition 12 (Pseudo-distance $d_{p,\mathbf{t}_n}$)

Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$ and $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$. Then,

$$\begin{cases} \forall p \in [1, +\infty), \forall (f, f') \in \mathcal{F}^2, \;\; d_{p,\mathbf{t}_n}(f, f') = \left(\frac{1}{n} \sum_{i=1}^{n} |f(t_i) - f'(t_i)|^p \right)^{\frac{1}{p}} \\ \forall (f, f') \in \mathcal{F}^2, \;\; d_{\infty,\mathbf{t}_n}(f, f') = \max_{1 \leqslant i \leqslant n} |f(t_i) - f'(t_i)| \end{cases} .$$

Definition 13 (Uniform covering and packing numbers)

Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{T}$ and $\bar{\mathcal{F}} \subset \mathcal{F}$. For $p \in [1, +\infty]$, $\epsilon \in \mathbb{R}_+^*$ and $n \in \mathbb{N}^*$, the uniform covering number $\mathcal{N}_p\left(\epsilon, \bar{\mathcal{F}}, n\right)$ and the uniform packing number $\mathcal{M}_p\left(\epsilon, \bar{\mathcal{F}}, n\right)$ are defined as follows:

$$\begin{cases} \mathcal{N}_p\left(\epsilon, \bar{\mathcal{F}}, n\right) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{N}\left(\epsilon, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right) \\ \mathcal{M}_p\left(\epsilon, \bar{\mathcal{F}}, n\right) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{M}\left(\epsilon, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right) \end{cases} .$$

Accordingly,

$$\mathcal{N}_p^{int}\left(\epsilon, \bar{\mathcal{F}}, n\right) = \sup_{\mathbf{t}_n \in \mathcal{T}^n} \mathcal{N}^{int}\left(\epsilon, \bar{\mathcal{F}}, d_{p,\mathbf{t}_n}\right).$$

Transitions Between Capacity Measures

"Complete" Pathway

$$R(\mathcal{F}) \xrightarrow{\text{chaining}} \mathcal{N}_p^{\text{int}}(\epsilon, \mathcal{F}, n) \xrightarrow{\leqslant} \mathcal{M}_p(\epsilon, \mathcal{F}, n) \xrightarrow{\text{Sauer-Shelah lemma}} \gamma\text{-dim}(\mathcal{F})$$

"Partial" Pathways

Depend on the choice of the norm, the classifier...

Guaranteed Risk Based on the L_∞-norm

Margin Loss Function

$$\forall \gamma \in (0,1]\,,\ \forall t \in \mathbb{R},\ \ \phi_{\infty,\gamma}(t) = \mathbb{1}_{\{t<\gamma\}}$$

Basic Supremum Inequality

Theorem 2 (After Theorem 22 in Guermeur, 2007)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. Let $\gamma \in (0,1]$ and $\delta \in (0,1)$. With P^m-probability at least $1-\delta$,

$$\sup_{g \in \mathcal{G}} \left(L(g) - L_{\gamma,m}(g) \right) \leqslant \sqrt{\frac{2}{m}\left(\ln\left(\mathcal{N}_\infty^{int}\left(\frac{\gamma}{2}, \mathcal{F}_{\mathcal{G},\gamma}, 2m\right)\right) + \ln\left(\frac{2}{\delta}\right)\right)} + \frac{1}{m}.$$

Guaranteed Risk Based on the L_∞-norm

Decomposition Lemma

Lemma 2 (Lemma 1 in Guermeur, 2017)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. Then for $\gamma \in (0, 1]$, $\epsilon \in \mathbb{R}_+^*$, $m \in \mathbb{N}^*$ and $\mathbf{z}_m = ((x_i, y_i))_{1 \leqslant i \leqslant m} \in \mathcal{Z}^m$,

$$\forall p \in [1, +\infty], \;\; \mathcal{N}^{int}\left(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, d_{p,\mathbf{z}_m}\right) \leqslant \prod_{k=1}^{C} \mathcal{N}^{int}\left(C^{-\frac{1}{p}}\epsilon, \mathcal{G}_k, d_{p,\mathbf{x}_m}\right),$$

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

Generalized Sauer-Shelah Lemma

Lemma 3 (After Lemma 3.5 in Alon et al., 1997)

Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F})$. Then for $\epsilon \in (0, 2M_{\mathcal{F}}]$ and $n \in \mathbb{N}^*$,

$$\mathcal{M}_\infty(\epsilon, \mathcal{F}, n) < 2\left(\frac{16 M_{\mathcal{F}}^2 n}{\epsilon^2}\right)^{d\left(\frac{\epsilon}{4}\right)\log_2\left(\frac{4 M_{\mathcal{F}} e n}{d\left(\frac{\epsilon}{4}\right)\epsilon}\right)}.$$

Guaranteed Risk Based on the L_∞-norm

Theorem 3 (Theorem 3 in Guermeur, 2017)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. For $\epsilon \in (0, M_{\mathcal{G}}]$, let $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim(\mathcal{G}_k)$. Let $\gamma \in (0,1]$ and $\delta \in (0,1)$. With P^m-probability at least $1-\delta$,

$$\sup_{g \in \mathcal{G}} \left(L(g) - L_{\gamma,m}(g)\right) \leqslant \sqrt{\frac{2}{m}\left(3Cd\left(\frac{\gamma}{8}\right)\ln^2\left(\frac{128 M_{\mathcal{G}}^2 m}{\gamma^2}\right) + \ln\left(\frac{2}{\delta}\right)\right)} + \frac{1}{m}.$$

Guaranteed Risk Based on the L_2-norm

Margin Loss Function

$$\forall \gamma \in (0,1]\,,\ \forall t \in \mathbb{R},\ \ \phi_{2,\gamma}(t) = \mathbb{1}_{\{t \leqslant 0\}} + \left(1 - \frac{t}{\gamma}\right) \mathbb{1}_{\{t \in (0,\gamma]\}}.$$

Basic Supremum Inequality

Theorem 4 (After Theorem 8.1 in Mohri et al., 2012)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. Let $\gamma \in (0,1]$ and $\delta \in (0,1)$. With P^m-probability at least $1-\delta$,

$$\sup_{g \in \mathcal{G}} \left(L_\gamma(g) - L_{\gamma,m}(g)\right) \leqslant \frac{2}{\gamma} R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}.$$

Guaranteed Risk Based on the L_2-norm

Decomposition Lemma

Lemma 4 (Kuznetsov et al., 2014)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. For $\gamma \in (0, 1]$,

$$R_m \left(\mathcal{F}_{\mathcal{G},\gamma} \right) \leqslant C R_m \left(\bigcup_{k=1}^{C} \mathcal{G}_k \right).$$

Theorem 5 (After Theorem 3 in Kuznetsov et al., 2014)

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. Let $\gamma \in (0, 1]$ and $\delta \in (0, 1)$. With P^m-probability at least $1 - \delta$,

$$\sup_{g \in \mathcal{G}} \left(L_\gamma(g) - L_{\gamma,m}(g) \right) \leqslant \frac{2C}{\gamma} R_m \left(\bigcup_{k=1}^{C} \mathcal{G}_k \right) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2m}}.$$

State-of-the-Art L_p-norm Sauer-Shelah Lemma

Theorem 6 (After Theorem 3.2 in Mendelson, 2002)

Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon$-dim$(\mathcal{F})$. Then for $\epsilon \in (0, 2M_{\mathcal{F}}]$ and $n \in \mathbb{N}^*$,

$$\forall p \in [1, +\infty), \;\; \ln\left(\mathcal{M}_p\left(\epsilon, \mathcal{F}, n\right)\right) \leqslant K_p \cdot d\left(\frac{\epsilon}{8}\right) \ln^2\left(\frac{2 \cdot d\left(\frac{\epsilon}{8}\right)}{\epsilon}\right),$$

where K_p is a constant depending only on p.

Main Lemmas in the Proof

Probabilistic Extraction Result

Lemma 5 (After Lemma 3.1 in Mendelson, 2002)

Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[M_{\mathcal{F}-}, M_{\mathcal{F}+}]$. For $n \in \mathbb{N}^*$, $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$, $p \in [1, +\infty)$ and $\epsilon \in [0, M_{\mathcal{F}+} - M_{\mathcal{F}-}]$, assume that $\mathcal{M}\left(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}\right) > 1$. Then there exists a subvector $\mathbf{t}_q$ of $\mathbf{t}_n$ of size q satisfying $q \leqslant K_p \left(\frac{M_{\mathcal{F}+} - M_{\mathcal{F}-}}{\epsilon}\right)^p \ln\left(\mathcal{M}\left(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}\right)\right)$ such that

$$\mathcal{M}\left(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}\right) \leqslant \mathcal{M}\left(\frac{\epsilon}{2}, \mathcal{F}, d_{\infty,\mathbf{t}_q}\right),$$

where K_p is a constant depending only on p.

Sauer-Shelah Lemma of Alon and co-authors (Lemma 3)

State-of-the-Art L_2-norm Sauer-Shelah Lemma

Lemma 6 (After Theorem 1 in Mendelson and Vershynin, 2003)

Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon\text{-dim}(\mathcal{F})$. Then for $\epsilon \in (0, 2M_{\mathcal{F}}]$ and $n \in \mathbb{N}^*$,

$$\mathcal{M}_2(\epsilon, \mathcal{F}, n) \leqslant \left(K \left(\frac{2M_{\mathcal{F}}}{\epsilon} \right)^5 \right)^{4d\left(\frac{\epsilon}{96}\right)}$$

where $K = 3584e$.

Lemma 7

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. Then, for $\epsilon \in (0, \gamma]$,

$$\ln\left(\mathcal{N}_2^{int}(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m)\right) \leqslant 20Cd\left(\frac{\epsilon}{96\sqrt{C}}\right) \ln\left(\frac{14M_{\mathcal{G}}\sqrt{C}}{\epsilon}\right)$$

where $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-dim}(\mathcal{G}_k)$.

Key Lemma in the Proof

Probabilistic Extraction Result

Lemma 8 (After Lemma 13 in Mendelson and Vershynin, 2003)

Let $\mathcal{T} = \{t_i : 1 \leqslant i \leqslant n\}$ be a finite set and $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n}$. Let $\mathcal{F}$ be a finite class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$. Assume that for some $\epsilon \in (0, 2M_{\mathcal{F}}]$, $\mathcal{F}$ is ϵ-separated with respect to the pseudo-metric $d_{2,\mathbf{t}_n}$. If $r \in [1, n]$ is such that $|\mathcal{F}| \leqslant \exp\left(K_e r \epsilon^4\right)$ with

$$K_e = \frac{3}{112\left(2M_{\mathcal{F}}\right)^4},$$

then there exists a subvector $\mathbf{t}_q$ of $\mathbf{t}_n$ of size $q \leqslant r$ such that $\mathcal{F}$ is $\frac{\epsilon}{2}$-separated with respect to the pseudo-metric $d_{2,\mathbf{t}_q}$.

L_p-norm Sauer-Shelah Lemma

Lemma 9 (Lemma 2 in Guermeur, 2017)

Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $[-M_{\mathcal{F}}, M_{\mathcal{F}}]$. For $\epsilon \in (0, M_{\mathcal{F}}]$, let $d(\epsilon) = \epsilon$-dim$(\mathcal{F})$. Then for $\epsilon \in (0, 2M_{\mathcal{F}}]$ and $n \in \mathbb{N}^*$,

$$\forall p \in [1, +\infty), \ \ \mathcal{M}_p(\epsilon, \mathcal{F}, n) \leqslant 4^{K_\epsilon(p)+1} \left(\frac{6272eK_\epsilon(p)}{3} \left(\frac{2M_{\mathcal{F}}}{\epsilon} \right)^{2p+1} \right)^{2K_\epsilon(p) d\left(\frac{\epsilon}{45}\right)},$$

where $K_\epsilon(p) = \log_2 \left(\left\lceil \left\lceil \frac{112 M_{\mathcal{F}}}{\epsilon} \right\rceil^{p+2} \right\rceil \right)$.

A logarithmic factor is gained compared to Mendelson's lemma (Theorem 6).

Key Lemmas in the Proof

Main Combinatorial Result

Lemma 10 (L_p-norm extension of Lemma 8 in Bartlett and Long, 1995)

Let $\mathcal{T} = \{t_i : 1 \leqslant i \leqslant n\}$ be a finite set and $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n}$. Let $\mathcal{F}$ be a class of functions from $\mathcal{T}$ into $\mathcal{S} = \left\{ 2M_{\mathcal{F}} \frac{j}{N} : 0 \leqslant j \leqslant N \right\}$ with $M_{\mathcal{F}} \in \mathbb{R}_+^*$ and $N \in \mathbb{N} \setminus [\![0, 3]\!]$. For $\epsilon \in \left(\frac{6M_{\mathcal{F}}}{N}, 2M_{\mathcal{F}} \right]$, let $d = \left(\frac{\epsilon}{2} - \frac{3M_{\mathcal{F}}}{N} \right)$-dim$(\mathcal{F})$. Then

$$\forall p \in [1, +\infty), \quad \mathcal{M}\left(\epsilon, \mathcal{F}, d_{p,\mathbf{t}_n}\right) < 2 \left\lceil N^{p+2} \right\rceil \left(\frac{e(N-1)n}{d} \right)^{\log_2\left(\left\lceil N^{p+2} \right\rceil\right) d}.$$

Probabilistic Extraction Result

L_p-norm extension of Lemma 13 in Mendelson and Vershynin (2003)

L_2-norm Sauer-Shelah Lemma

Lemma 11

Let $\mathcal{G}$ be the class of functions computed by a margin multi-category classifier. Then for $\epsilon \in (0, \gamma]$,

$$\ln\left(\mathcal{M}_2\left(\epsilon, \mathcal{F}_{\mathcal{G},\gamma}, m\right)\right) \leqslant 3Cd\left(\frac{\epsilon}{16}\right)\ln^2\left(\frac{K\left(2M_{\mathcal{G}}\right)^3\gamma^2 C^{\frac{3}{2}} d\left(\frac{\epsilon}{192\sqrt{C}}\right)}{\epsilon^5}\right)$$

where $d(\epsilon) = \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim\left(\mathcal{G}_k\right)$ and $K = \frac{143360}{3}$.

Idea: Use the L_2-norm to get a dimension-free result <u>and</u> the L_∞-norm to optimize the dependency on C (get the best of two worlds)

Chaining Method

Theorem 7 (Dudley's metric entropy bound)

Let $\mathcal{F}$ be a class of bounded real-valued functions on $\mathcal{T}$. For $n \in \mathbb{N}^*$, let $\mathbf{t}_n = (t_i)_{1 \leqslant i \leqslant n} \in \mathcal{T}^n$ and let $diam(\mathcal{F}) = \sup_{(f,f') \in \mathcal{F}^2} d_{2,\mathbf{t}_n}(f, f')$. Let h be a positive and decreasing function on $\mathbb{N}$ such that $h(0) \geqslant diam(\mathcal{F})$. Then for $N \in \mathbb{N}^*$,

$$\hat{R}_n(\mathcal{F}) \leqslant h(N) + 2 \sum_{j=1}^{N} (h(j) + h(j-1)) \sqrt{\frac{\ln\left(\mathcal{N}^{int}(h(j), \mathcal{F}, d_{2,\mathbf{t}_n})\right)}{n}}$$

and

$$\hat{R}_n(\mathcal{F}) \leqslant 12 \int_0^{\frac{1}{2} \cdot diam(\mathcal{F})} \sqrt{\frac{\ln\left(\mathcal{N}^{int}(\epsilon, \mathcal{F}, d_{2,\mathbf{t}_n})\right)}{n}}\, d\epsilon.$$

Polynomial Growth of the γ-dimension

Hypothesis 1

We consider margin multi-category classifiers for which there exists a pair $(d_{\mathcal{G}}, K_{\mathcal{G}}) \in \left(\mathbb{R}_+^*\right)^2$ *such that*

$$\forall \epsilon \in (0, M_{\mathcal{G}}], \quad \max_{1 \leqslant k \leqslant C} \epsilon\text{-}dim\left(\mathcal{G}_k\right) \leqslant K_{\mathcal{G}} \epsilon^{-d_{\mathcal{G}}}.$$

Classifier	$d_{\mathcal{G}}$	Reference
MLP	4	(Bartlett, 1998)
M-SVM	2	(Bartlett and Shawe-Taylor, 1999)
LVQ	?	

Table : Characterization of γ-dimensions

Bound Based on the Lemma of Mendelson and Vershynin

Theorem 8 (Theorem 7 in Guermeur, 2017)

Let $\mathcal{G}$ be a class of functions satisfying Hypothesis 1 and $\gamma \in (0, 1]$.

$d_{\mathcal{G}}$	*Bound on* $R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right)$
< 2	$8\frac{1+2^{\frac{2}{2-d_{\mathcal{G}}}}}{\sqrt{2(2-d_{\mathcal{G}})}}\gamma^{1-\frac{d_{\mathcal{G}}}{2}}\sqrt{\frac{5 \cdot 96^{d_{\mathcal{G}}} K_{\mathcal{G}}}{m}}C^{\frac{d_{\mathcal{G}}+2}{4}}\left\{\sqrt{\ln\left(F\left(C\right)\right)}+\sqrt{\frac{1}{4\ln(F(C))}}\right\}$
2	$\frac{\gamma C^{\frac{3}{4}}}{\sqrt{m}} + 1152\sqrt{\frac{5K_{\mathcal{G}}}{m}}C\left\lceil \frac{1}{2}\log_2\left(\frac{m}{C}\right)\right\rceil\sqrt{\ln\left(\frac{14M_{\mathcal{G}}\sqrt{m}}{\gamma C^{\frac{1}{4}}}\right)}$
> 2	$\gamma\sqrt{C}\left(\frac{C}{m}\right)^{\frac{1}{d_{\mathcal{G}}}}\left(1+8\left(1+2^{\frac{2}{d_{\mathcal{G}}-2}}\right)\gamma^{-\frac{d_{\mathcal{G}}}{2}}\sqrt{5 \cdot 96^{d_{\mathcal{G}}}K_{\mathcal{G}}}\sqrt{\ln\left(\frac{14M_{\mathcal{G}}}{\gamma}\left(\frac{m}{C}\right)^{\frac{1}{d_{\mathcal{G}}}}\right)}\right)$

where

$$F\left(C\right) = 2\left(\frac{14M_{\mathcal{G}}\sqrt{C}}{\gamma}\right)^{\frac{2-d_{\mathcal{G}}}{2}}.$$

Bound Based on the New L_2-norm Lemma

Theorem 9

Let $\mathcal{G}$ be a class of functions satisfying Hypothesis 1 and $\gamma \in (0, 1]$. Then

$$R_m\left(\mathcal{F}_{\mathcal{G},\gamma}\right) \leqslant K\left(M_{\mathcal{G}}, \gamma, d_{\mathcal{G}}\right) F\left(m, C\right)$$

with

$d_{\mathcal{G}}$	$F\left(m, C\right)$ *(Guermeur, 2017)*	$F\left(m, C\right)$ *Present study*
< 2	$\sqrt{\frac{C^{\frac{d_{\mathcal{G}}+2}{2}} \ln(C)}{m}}$	$\frac{\sqrt{C}\ln(C)}{\sqrt{m}}$
2	$\frac{C \ln^{\frac{3}{2}}\left(\frac{m}{C}\right)}{\sqrt{m}}$	$\frac{\sqrt{C}\ln(Cm)\ln(m)}{\sqrt{m}}$
> 2	$\sqrt{C}\left(\frac{C}{m}\right)^{\frac{1}{d_{\mathcal{G}}}} \sqrt{\ln\left(\frac{m}{C}\right)}$	$\frac{\sqrt{C}\ln(Cm)}{m^{\frac{1}{d_{\mathcal{G}}}}}$

Rademacher Complexity of a Linear Separator

Theorem 10 (Theorem 4.3 in Mohri et al., 2012)

Let $\mathcal{H} = \{x \mapsto w \cdot x\}$ *with* $\|x\| \leqslant \Lambda_x$ *and* $\|w\| \leqslant \Lambda_w$. *Then,*

$$R_m(\mathcal{H}) \leqslant \frac{\Lambda_w \Lambda_x}{\sqrt{m}}.$$

Covering Numbers of a Linear Separator

Theorem 11 (Theorem 4 in Zhang, 2002)

Let $\mathcal{H} = \{x \mapsto w \cdot x\}$ *with* $\|x\| \leqslant \Lambda_x$ *and* $\|w\| \leqslant \Lambda_w$. *Then,*

$$\ln\left(\mathcal{N}_{\infty}^{int}(\epsilon, \mathcal{H}, m)\right) \leqslant 36 \left(\frac{\Lambda_w \Lambda_x}{\epsilon}\right)^2 \ln\left(2 \left\lceil \frac{4\Lambda_w \Lambda_x}{\epsilon} + 2 \right\rceil m + 1\right).$$

Conclusions and Ongoing Research

Conclusions

1. The control terms of our guaranteed risks grow sublinearly with C.
2. An optimal trade-off between this dependency and the convergence rate is to be looked for.

Ongoing research

1. Application to LVQ
2. Derivation of lower bounds
3. Characterization of the phase transitions