Neurocomputing 505 (2022) 125–141

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

The benefits of adversarial defense in generalization

Luca Oneto [a,*], Sandro Ridella [a], Davide Anguita [a]

[a] University of Genoa – Via Opera Pia 11a, 16145 Genova, Italy

ARTICLE INFO

Article history:
Received 8 February 2022
Revised 19 June 2022
Accepted 12 July 2022
Available online 16 July 2022
Communicated by Zidong Wang

Keywords:
Adversarial machine learning
Evasion attacks
Adversarial defense
Statistical learning theory
Generalization
(Local) Vapnik–Chervonenkis theory
(Local) Rademacher complexity theory

ABSTRACT

Recent research has shown that models induced by machine learning, and in particular by deep learning, can be easily fooled by an adversary who carefully crafts imperceptible, at least from the human perspective, or physically plausible modifications of the input data. This discovery gave birth to a new field of research, the adversarial machine learning, where new methods of attacks and defense are developed continuously, mimicking what is happening from a long time in cybersecurity. In this paper we will show that the drawbacks of inducing models from data less prone to be misled can actually provide some benefits when it comes to assessing their generalization abilities. We will show these benefits both from a theoretical perspective, using state-of-the-art statistical learning theory, and both with practical examples.

1. Introduction

In the last decades, Artificial Intelligence, and in particular Machine Learning, has become pervasive in all aspects of our lives experiencing a fast process of commodification and reaching the society at large. From self-driving cars to smart IoT devices, almost every consumer application now leverages such technologies to make sense and get insights from the vast amount of data collected and stored by these devices. In some tasks (e.g., medicine, vision, and games) recent deep-learning algorithms have shown superhuman performance [1–4] or are expected to do so in the near future [5]. For this reason, it has been extremely surprising to discover that such algorithms can be easily fooled by an adversary who carefully crafts imperceptible, at least from the human perspective, or physically plausible modifications of the input data forcing models to perceiving things that are not there [6–9]. Intrigued by this discovery, and worried about its potential impact on the field, a large number of researchers and stakeholders started to study, understand, and address this problem developing proper mitigation strategies. Despite such large interest, this challenging problem is still far from being solved. In fact new methods of attacks (i.e., adversarial attacks) and mitigation strategies (i.e., adversarial defense) are developed continuously [7,10–14], mimicking what is happening from a long time in cybersecurity, giving birth to an entire new field of research: the adversarial machine learning.

In this setting, one of the main issues, similar to the classical learning setting, is how to estimate the generalization ability of the defense strategies, i.e., how robust will be the protected model on data that have not been observed during the learning phase. Some recent works [15–17] have focused on this problem using the Rademacher Complexity based bounds but without concentrating on the benefits of adversarial defense in generalization.

For this reason, in this work, we propose a change of perspective. Instead of focusing on the challenges posed by the tension between adversarial attackers and defenders we focus our attention on its potential benefits. In particular, we will study what happens when we try to estimate the generalization capabilities of a model learned in the classical setting, where no adversary is present (Non-Adversarial Setting), against the ones of a model designed to be less prone to attacks and then less exposed to adversaries (Adversarial Setting). Exploiting the two well know pillars of statistical learning theory, i.e., the (Local) Vapnik–Chervonenkis [18,19] and the (Local) Rademacher Complexity [20–24], we will show that the introduction of a mechanism of defense in the learning phase of a model can actually improve our ability to accurately estimate its generalization performance (i.e., the tightness of the generalization bound). Moreover, we will show that these theoretical results can be also observed in practical cases. Note that, the proposed generalization bounds for the Adversarial

* Corresponding author.
E-mail addresses: luca.oneto@unige.it (L. Oneto), sandro.ridella@unige.it (S. Ridella), davide.anguita@unige.it (D. Anguita).

https://doi.org/10.1016/j.neucom.2022.07.010
0925-2312/

Setting based on the (Local) Vapnik–Chervonenkis and on the Local Rademacher Complexity are novel while the ones based on the Rademacher Complexity have already been studied in [17]. We will also perform a study on the connection between the (Local) Vapnik–Chervonenkis and on the (Local) Rademacher Complexity in the Adversarial Setting which is again new. Finally, the theoretical and practical analysis of the behaviour of the (Local) Vapnik–Chervonenkis and the (Local) Rademacher Complexity based bound in the Adversarial Setting when the perturbation domain changes in size is new and shed new light on a previously unknown phenomenon: increasing the size of the perturbation domain can increase the tightness of the generalization error bounds.

The rest of the paper is organized as follows. Section 2 recall some background notions. The theoretical and practical analysis will be performed respectively in Sections 3 and 4. Finally, Section 5 concludes the paper.

2. Preliminaries

Let us consider the binary classification problem[1] [18] under evasion attach [6]. Based on a random observation of $X \in \mathcal{X}$ one has to estimate $Y \in \mathcal{Y} \subseteq \{\pm 1\}$ by choosing a suitable hypothesis (function) $h : \mathcal{X} \to \hat{\mathcal{Y}}$, with $\hat{\mathcal{Y}} \in \mathbb{R}$, in a set of possible ones $\mathcal{H}$. Note that, in the forward phase, in order to take a decision, we assume to apply $\text{sgn}(h(X))$ where

$$\text{sgn}[x] = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ +1 & \text{if } x > 0 \end{cases}. \tag{1}$$

Note that choosing the right $\mathcal{H}$ is the so-called model selection problem [25] which is out of the scope of this paper. The hypothesis h is subject to an adversary which tries to fool the model into mistakes by modifying the observation X according to a set of possible modifications $\mathcal{B}(X) \subseteq \mathcal{X}$ such that $X \in \mathcal{B}(X)$, namely

$$X^*_{\mathcal{B}} : \arg \sup_{\widetilde{X} \in \mathcal{B}(X)} \left[\text{sgn}\left[h\left(\widetilde{X}\right)\right] \neq \text{sgn}[h(X)] \right], \tag{2}$$

where the Iverson bracket notation is exploited. A learning algorithm selects $h \in \mathcal{H}$ by exploiting a set of n labeled samples

$$\mathcal{D} = \{Z_1, \cdots, Z_n\} = \{(X_1, Y_1), \cdots, (X_n, Y_n)\}, \tag{3}$$

where $Z \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\mathcal{D}$ consists of a sequence of independent samples distributed according to μ over $\mathcal{Z}$ (i.e., i.i.d. samples). The generalization error (i.e., the risk)

$$\mathsf{L}^Y_\ell(h) = \mathbb{E}_Z \ell(h, Z), \tag{4}$$

associated to an hypothesis $h \in \mathcal{H}$, is defined through a loss function $\ell : \mathcal{H} \times \mathcal{Z} \to [0, 1]$. As μ is unknown, $\mathsf{L}^Y_\ell(h)$ cannot be explicitly computed, but we can compute the empirical error (i.e., the empirical risk) namely the empirical estimator of the generalization error

$$\hat{\mathsf{L}}^Y_\ell(h) = \frac{1}{n} \sum_{Z \in \mathcal{D}} \ell(h, Z). \tag{5}$$

The purpose of any learning procedure is to find the minimizer h^* of the generalization error $\mathsf{L}^Y_\ell(h)$

$$h^*_\ell = \arg \inf_{h \in \mathcal{H}} \mathsf{L}^Y_\ell(h), \tag{6}$$

but since $\mathsf{L}^Y_\ell(h)$ is unknown we have to estimate h^*_ℓ exploiting an empirical estimator $\hat{h}_\ell$ which is the empirical risk minimizer

$$\hat{h}_\ell = \arg \inf_{h \in \mathcal{H}} \hat{\mathsf{L}}^Y_\ell(h). \tag{7}$$

$\hat{h}_\ell$ is effective when $\mathcal{H}$ is carefully tuned [25] (e.g., via Structural Risk Minimization [18] or Invariant Risk Minimization [26]). Nevertheless, in our case, we have a further level of complexity because of the adversary which tries to fool the learned model. For this reason, we have to make the learned model robust to adversarial perturbation using the now-classical approach of Adversarial Defense (also called Adversarial Training) [6,27]. The idea is that the attack of Eq. (2) can be reformulated as

$$X^*_{\mathcal{B}} = X^*_{\ell_{\mathcal{B}}} = \arg \sup_{\widetilde{X} \in \mathcal{B}(X)} \ell\left(h, \left(\widetilde{X}, Y\right)\right), \tag{8}$$

and then we can consider the now-classical problem of Adversarial Defense [6]

$$h^*_{\tilde{\ell}_{\mathcal{B}}} = \arg \inf_{h \in \mathcal{H}} \mathsf{L}^Y_{\tilde{\ell}_{\mathcal{B}}}(h), \tag{9}$$

where

$$\mathsf{L}^Y_{\tilde{\ell}_{\mathcal{B}}}(h) = \mathbb{E}_Z \sup_{\widetilde{X} \in \mathcal{B}(X)} \ell\left(h, \left(\widetilde{X}, Y\right)\right) = \mathbb{E}_Z \tilde{\ell}_{\mathcal{B}}(h, Z), \tag{10}$$

is the generalization error in the Adversarial Setting and its empirical estimator is

$$\hat{h}_{\tilde{\ell}_{\mathcal{B}}} : \arg \inf_{h \in \mathcal{F}} \hat{\mathsf{L}}^Y_{\tilde{\ell}_{\mathcal{B}}}(h), \tag{11}$$

where

$$\hat{\mathsf{L}}^Y_{\tilde{\ell}_{\mathcal{B}}}(h) = \frac{1}{n} \sum_{Z \in \mathcal{D}} \tilde{\ell}_{\mathcal{B}}(h, Z). \tag{12}$$

Note that when $\mathcal{B}(X) = X$ (i.e., $\mathcal{B}$ is the identity $\mathcal{I}$) we have that

$$\begin{aligned} &\mathsf{L}^Y_{\tilde{\ell}_{\mathcal{B}}}(h) = \mathsf{L}^Y_{\tilde{\ell}_{\mathcal{I}}}(h) = \mathsf{L}^Y_\ell(h), \quad h^*_{\tilde{\ell}_{\mathcal{B}}} = h^*_{\tilde{\ell}_{\mathcal{I}}} = h^*_\ell, \\ &\hat{\mathsf{L}}^Y_{\tilde{\ell}_{\mathcal{B}}}(h) = \hat{\mathsf{L}}^Y_{\tilde{\ell}_{\mathcal{I}}}(h) = \hat{\mathsf{L}}^Y_\ell(h), \quad \hat{h}_{\tilde{\ell}_{\mathcal{B}}} = \hat{h}_{\tilde{\ell}_{\mathcal{I}}} = \hat{h}_\ell. \end{aligned} \tag{13}$$

3. Theoretical analysis of generalization

In this section we will study the problem of estimating the generalization ability of $\hat{h}_{\tilde{\ell}_{\mathcal{B}}}$ using two different powerful theories inside the statistical learning theory, the (Local) Vapnik–Chervonenkis [18,19] and the (Local) Rademacher Complexity [20,24], showing the possible benefits related to the Adversarial Setting with respect to dealing with the classical model $\hat{h}_\ell$ in the Non-Adversarial Setting.

3.1. (Local) Vapnik–Chervonenkis Theory

In this section we will first study the classical Non-Adversarial Setting (Section 3.1.1), then the Adversarial Setting (Section 3.1.2), and finally we will compare the two settings (Section 3.1.3) using and extending the (Local) Vapnik–Chervonenkis Theory.

In particular, in this section, we will consider the case in which a $\{0, 1\}$ values loss is used $\ell(h, Z) \in \{0, 1\}$, (i.e., the Hard loss function $\ell(h, Z) = [Yh(X) \leqslant 0]$ in the Non-Adversarial Setting) [18].

[1] Everything we will present can be generalized with some technical steps to the whole supervised learning framework but, for simplicity and clarity of the notation and since this extension does not add much to the content of the paper, we will restrict the presentation to the binary classification framework.

3.1.1. Non-adversarial setting

Let us consider the Non-Adversarial setting in which one has learned $\hat{h}_\ell$ and has to bound its performance, namely estimate value of $\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right)$ just based on empirical quantities.

In this setting we can define the following set

$$\mathscr{L}_{\ell,\mathscr{D}} = \{[\ell(h,Z_1),\cdots,\ell(h,Z_n)] : h \in \mathscr{H}\},$$

which is the set of possible distinct vectors of configuration of the $\{0,1\}$-errors distinguishable within $\mathscr{H}$ with respect to the dataset $\mathscr{D}$. Finding all these vectors can be computational expensive to compute but we can resort to an estimation [19], via Monte Carlo methods.

Then the empirical Vapnik–Chervonenkis Entropy (VCE) can be defined as follows[2] [18]

$$\hat{\mathsf{V}}_\ell(\mathscr{H}) = \ln\left(\max\left[1,\left|\mathscr{L}_{\ell,\mathscr{D}}\right|\right]\right), \tag{15}$$

namely, the VCE is the number of possible distinct vectors of configuration of the $\{0,1\}$-errors distinguishable within $\mathscr{H}$ with respect to the dataset $\mathscr{D}$.

In general it is possible to prove that [18,19]

$$\mathbb{P}\left\{\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right) \leqslant \hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right) + 3\sqrt{\frac{\hat{\mathsf{V}}_\ell(\mathscr{H})}{n}} + 3\sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{n}}\right\} \geqslant 1-\delta, \tag{16}$$

with $\delta \in (0,1)$. The bound of Eq. (16) is a fully empirical bound, namely all the quantities can be estimated from the data [19].

Note that the bound of Eq. (16) can be summarized as follows: the generalization error $\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right)$ of the empirical minimizer[3] $\hat{h}_\ell$ is bounded, with probability at least $(1-\delta)$, by its empirical error $\hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right)$, plus a complexity term $\hat{\mathsf{C}}_\ell(\mathscr{H})$ which measures the risk due to the size of $\mathscr{H}$ (i.e., the larger is $\mathscr{H}$ the larger is the risk), plus a confidence term $\phi(\delta)$ which measures the risk associated to the sample (i.e., we have a risk due to inferring something about the population with a finite number of samples). In other words

$$\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right) \overset{(1-\delta)}{\leqslant} \hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right) + \hat{\mathsf{C}}_\ell(\mathscr{H}) + \phi(\delta), \tag{17}$$

where we did not explicitly specify the dependency from n since it was kept constant in the work. For the bound of Eq. (16) obviously $\hat{\mathsf{C}}_\ell(\mathscr{H}) = 3\sqrt{\frac{\hat{\mathsf{V}}_\ell(\mathscr{H})}{n}}$ and $\phi(\delta) = 3\sqrt{\frac{2\ln\left(\frac{4}{\delta}\right)}{n}}$.

Note also that the bound of Eq. (16) can be improved both in the constants (e.g., using stronger concentration results [19]) and both in the rate of convergence (e.g., we can obtain fast rates for $\hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right) = 0$ [18]) but for the purpose of this paper and for simplicity we prefer not to further weigh down the presentation.

In fact, as we will discuss from now on, the discussion is centered on the empirical risk and on the complexity term. In fact, for the purpose of our discussion, $\phi(\delta)$ is constant (see later) and it can be also disregarded assuming that the sample $\mathscr{D}$ well represents the population.

The complexity of Eq. (15) and the bound of Eq. (16) are also called Global VCE (GVCE) and GVCE based bound respectively since all the functions in $\mathscr{H}$ contribute to $\hat{\mathsf{C}}_\ell(\mathscr{H})$, even the ones that will be never chosen by the algorithm, namely the one characterized by high error. For this reason a Local VCE (LVCE), and the corresponding LVCE based bound have been proposed [19]. The latter are able to not take into account functions with high error resulting in tighter bounds.

Let us then present the LVCE and the LVCE based bounds. Let us first localize the set of functions defined in Eq. (14) by introducing a constraint on the error, controlled by a parameter $r \in [0,1]$

$$\begin{aligned}\mathscr{L}_{\ell,\mathscr{D},r} &= \left\{[\ell(h,Z_1),\cdots,\ell(h,Z_n)] : h \in \mathscr{H}, \hat{\mathsf{L}}_\ell^Y(h) \leqslant r\right\} \\ &= \left\{[\ell(h,Z_1),\cdots,\ell(h,Z_n)] : h \in \mathscr{H}, \sum_{i=1}^n \ell(h,Z_i) \leqslant nr\right\},\end{aligned} \tag{18}$$

which is the $\mathscr{L}_{\ell,\mathscr{D}}$ limited at the vectors of configuration of the $\{0,1\}$-errors such that the number of ones in this vector is small.

Then, the empirical LVCE can be defined as

$$\hat{\mathsf{V}}_\ell(\mathscr{H},r) = \ln\left(\max\left[1,\left|\mathscr{L}_{\ell,\mathscr{D},r}\right|\right]\right), \tag{19}$$

which is the GVCE, limited to the vectors of configuration of the $\{0,1\}$-errors with a small number of ones.

Note that $\hat{\mathsf{V}}_\ell(\mathscr{H}) = \hat{\mathsf{V}}_\ell(\mathscr{H},1)$ namely the LVCE degenerates in the GVCE. Moreover $\hat{\mathsf{V}}_\ell(\mathscr{H},r)$ is obviously monotonically increasing in r, namely if $0 \leqslant r_1 \leqslant r_2 \leqslant 1$ then $\hat{\mathsf{V}}_\ell(\mathscr{H},r_1) \leqslant \hat{\mathsf{V}}_\ell(\mathscr{H},r_2)$.

In this setting it is possible to prove that [19]

$$\begin{aligned}&\mathbb{P}\left\{\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right) \leqslant \min_{K\in(1,\infty)} \tfrac{K}{K-1}\hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right) + \tfrac{r}{K}\right\} \geqslant 1-\delta, \\ &\text{s.t.} \begin{cases} \sup\limits_{\alpha\in(0,1]} 6\sqrt{\frac{r\alpha\left[\mathsf{T}(r,\alpha)+2\ln\left(\frac{9}{\delta}\right)\right]}{n}} \leqslant \frac{r}{K} \\ r > 0 \\ \mathsf{T}(r,\alpha) \leqslant \hat{\mathsf{V}}_\ell\left(\mathscr{H}, \frac{r}{\alpha} + 3\sqrt{\frac{\mathsf{T}(r,\alpha)+2\ln\left(\frac{9}{\delta}\right)}{n}}\right) \\ \mathscr{T}(r,\alpha) = \hat{\mathsf{V}}_\ell\left(\left\{h : h \in \mathscr{H},\ \mathsf{L}_\ell^Y(h) \leqslant \frac{r}{\alpha}\right\}\right) \end{cases}.\end{aligned} \tag{20}$$

This bound is then able to discard functions with high error thanks to the fact that the functions with high error are not contemplated in the complexity term.

The same comments made for GVCE based bound of Eq. (16) holds also for the LVCE based bound of Eq. (20) (namely its localized version). In fact the latter is fully empirical, can be improved in both constants and rate of convergence, actually holds for any hypothesis chosen in $\mathscr{H}$ according to $\mathscr{D}$ if $\mathscr{H}$ is chosen before observing $\mathscr{D}$, and can be simplified as follows

$$\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right) \overset{(1-\delta)}{\leqslant} c_1\hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right) + \hat{\mathsf{C}}_\ell(\mathscr{H},c_2) + \phi(\delta), \tag{21}$$

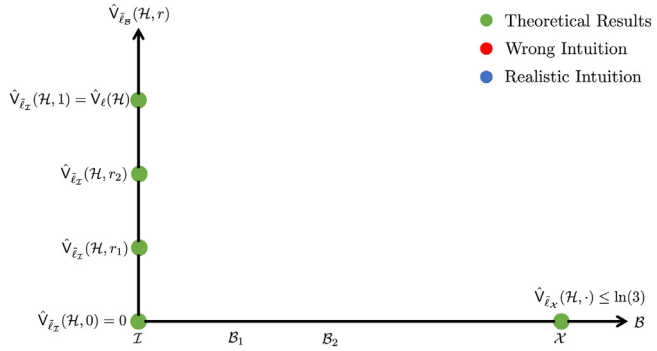
where $c_1 \in (0,\infty)$ and $c_2 \in (0,\infty)$ are quantities that can be computed from the data and $\hat{\mathsf{C}}_\ell(\mathscr{H},c_2)$ is a complexity term which measures the risk due to the size of $\mathscr{H}$ taking into account just the functions with empirical error smaller then c_3. This simplification can be made apparent by noting in the LVCE based bound of Eq. (20) that $\hat{\mathsf{V}}_\ell\left(\left\{h : h \in \mathscr{H},\ \mathsf{L}_\ell^Y(h) \leqslant \frac{r}{\alpha}\right\}\right) \leqslant \hat{\mathsf{V}}_\ell(\mathscr{H})$ and then with probability at least $(1-\delta)$ we have that

$$\begin{aligned}&\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right) \leqslant \min_{K\in(1,\infty)} \tfrac{K}{K-1}\hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right) + \tfrac{r}{K}, \\ &\text{s.t.} \sup_{\alpha\in(0,1],\ r>0} 6\sqrt{\frac{r\alpha\hat{\mathsf{V}}_\ell\left(\mathscr{H}, \frac{r}{\alpha}+3\sqrt{\frac{\hat{\mathsf{V}}_\ell(\mathscr{H})+2\ln\left(\frac{9}{\delta}\right)}{n}}\right)}{n}} + 6\sqrt{\frac{r\alpha 2\ln\left(\frac{9}{\delta}\right)}{n}} \leqslant \frac{r}{K}.\end{aligned} \tag{22}$$

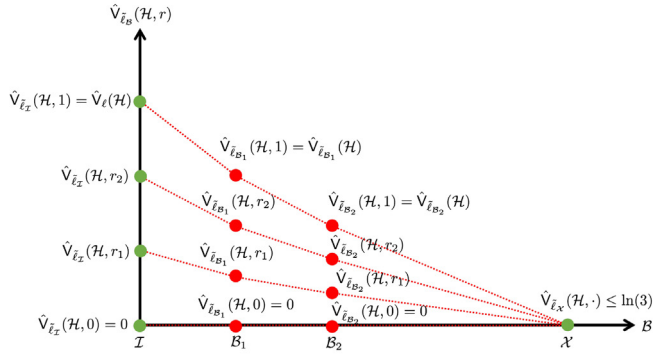
Assuming then that we have solved the problem of Eq. (22) (which counts just quantities that can be computed from the data) and then found K^*, α^*, and r^* we have that $c_1 = \frac{K^*}{K^*-1}$, $\hat{\mathsf{C}}_\ell(\mathscr{H},c_2) = 6\sqrt{\frac{r^*\alpha^*\hat{\mathsf{V}}_\ell(\mathscr{H},c_2)}{n}}$, $c_2 = \frac{r^*}{\alpha^*} + 3\sqrt{\frac{\hat{\mathsf{V}}_\ell(\mathscr{H})+2\ln\left(\frac{9}{\delta}\right)}{n}}$, and $\phi(\delta) = 6\sqrt{\frac{r^*\alpha^* 2\ln\left(\frac{9}{\delta}\right)}{n}}$.

[2] The operator $\max[1,\cdot]$ is needed to deal with empty set namely the case $\ln(0)$.

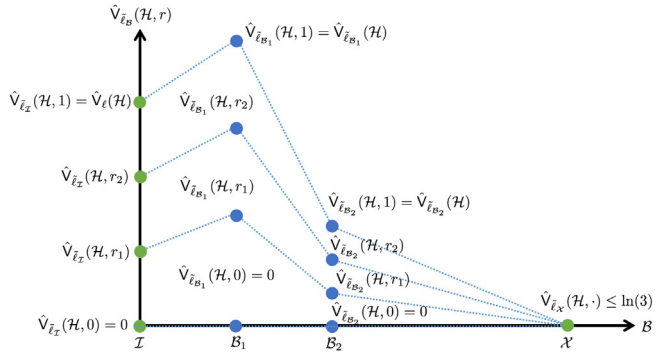
[3] Actually the bound holds for any hypothesis chosen in $\mathscr{H}$ according to $\mathscr{D}$ if $\mathscr{H}$ is chosen before observing $\mathscr{D}$.



(a) Properties of Eqns. (32) and (33)



(b) Wrong intuition from the properties of Eqns. (32) and (33) to fill the properties of Eq. (34)



(c) Plausible intuition from the properties of Eqns. (32) and (33) to fill the properties of Eq. (34) exploiting the properties derived from the Toy sample of Figure 2

Fig. 1. Qualitative analysis of the properties of the (A) GVCE and the (A) LVCE.

3.1.2. Adversarial Setting

Let us consider the Adversarial Setting in which one has learned $\hat{h}_{\tilde{\ell}_{\mathscr{B}}}$ and has to bound its performance, namely estimate value of $\mathsf{L}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ just based on empirical quantities. Our adversary on $h \in \mathscr{H}$ is modeled according to Eq. (2) or, equivalently, according to Eq. (8) using the loss function. In other words, a model needs to label a point X and all the points inside $\mathscr{B}(X)$ with the same (possibly correct) label in order to be robust to attacks and not to make mistakes.

In order the extend the notion of GVCE to the Adversarial Setting, defining the Adversarial GVCE (AGVCE), we need to note that the bounds of the previous sections hold for every loss such that $\ell(h,Z) \in \{0,1\}$ [18]. In the Adversarial Setting our loss becomes

$$\tilde{\ell}_{\mathscr{B}}(h,Z) = \sup_{\widetilde{X} \in \mathscr{B}(X)} \ell\left(h,\left(\widetilde{X},Y\right)\right) \in \{0,1\}, \tag{23}$$

and consequently all the bounds presented in the previous section still hold with a simple series of substitutions and redefinitions. In fact, using the Hard Loss function of the Non-Adversarial setting we have that

$$\tilde{\ell}_{\mathscr{B}}(h,Z) = \sup_{\widetilde{X} \in \mathscr{B}(X)} \left[Yh\left(\widetilde{X}\right) \leqslant 0\right] \in \{0,1\}. \tag{24}$$

Consequently, at least formally, the switch is painless. Nevertheless, as we will see deeper in Section 3.1.3 but that we start to discuss here, these substitutions and redefinitions imply rather counterintuitive results.

Let us start with the definition of AGVCE. For this purpose let us define the counterpart of the set of Eq. (14) in the Adversarial Setting as

$$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}},\mathscr{D}} = \left\{\left[\tilde{\ell}_{\mathscr{B}}(h,Z_1),\cdots,\tilde{\ell}_{\mathscr{B}}(h,Z_n)\right] : h \in \mathscr{H}\right\}, \tag{25}$$

and then the empirical AGVCE can be defined as follows

$$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) = \ln\left(\max\left[1,\left|\mathscr{L}_{\tilde{\ell}_{\mathscr{B}},\mathscr{D}}\right|\right]\right), \tag{26}$$

Thanks to this definition we can state the counterpart of the bound of Eq. (17) for the Adversarial Setting substituting $\mathsf{L}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\mathsf{L}^{Y}_{\ell}\left(\hat{h}_{\ell}\right), \hat{\mathsf{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\hat{\mathsf{L}}^{Y}_{\ell}\left(\hat{h}_{\ell}\right), \hat{\mathsf{C}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) = 3\sqrt{\frac{\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H})}{n}}$ to $\hat{\mathsf{C}}_{\ell}(\mathscr{H})$ and with the same $\phi(\delta)$ of Eq. (17)

$$\mathsf{L}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right) \overset{(1-\delta)}{\leqslant} \hat{\mathsf{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right) + \hat{\mathsf{C}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) + \phi(\delta). \tag{27}$$

Analogously it is possible to define a ALVCE by first localize the set of functions defined in Eq. (25), analogously to what has been done for Eq. (14) in Eq. (18), by introducing a constraint on the error, controlled by a parameter $r \in [0,1]$

$$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}},\mathscr{D},r} = \left\{\left[\tilde{\ell}_{\mathscr{B}}(h,Z_1),\cdots,\tilde{\ell}_{\mathscr{B}}(h,Z_n)\right] : h \in \mathscr{H}, \hat{\mathsf{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}(h) \leqslant r\right\}, \tag{28}$$

then, the empirical ALVCE can be defined as

$$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H},r) = \ln\left(\max\left[1,\left|\mathscr{L}_{\tilde{\ell}_{\mathscr{B}},\mathscr{D},r}\right|\right]\right). \tag{29}$$

Thanks to this definition we can state the counterpart of the bound of Eq. (21) for the Adversarial Setting substituting $\mathsf{L}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\mathsf{L}^{Y}_{\ell}\left(\hat{h}_{\ell}\right), \hat{\mathsf{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\hat{\mathsf{L}}^{Y}_{\ell}\left(\hat{h}_{\ell}\right), c_1 = \frac{K^*}{K^*-1}, \hat{\mathsf{C}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H},c_2) = 6\sqrt{\frac{r^*\alpha^*\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H},c_2)}{n}}, c_2 = \frac{r^*}{\alpha^*} + 3\sqrt{\frac{\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H})+2\ln\left(\frac{9}{\delta}\right)}{n}}$, and $\phi(\delta) = 6\sqrt{\frac{r^*\alpha^*2\ln\left(\frac{9}{\delta}\right)}{n}}$

$$\mathsf{L}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right) \overset{(1-\delta)}{\leqslant} c_1\hat{\mathsf{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right) + \hat{\mathsf{C}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H},c_2) + \phi(\delta), \tag{30}$$

where K^*, α^*, and r^* are found by solving Eq. (22) substituting $\mathsf{L}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\mathsf{L}^{Y}_{\ell}\left(\hat{h}_{\ell}\right), \hat{\mathsf{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\hat{\mathsf{L}}^{Y}_{\ell}\left(\hat{h}_{\ell}\right), \hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H})$ to $\hat{\mathsf{V}}_{\ell}(\mathscr{H})$, and $\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H},\cdot)$ to $\hat{\mathsf{V}}_{\ell}(\mathscr{H},\cdot)$. In fact the proof of the LVCE based bound is based on the results one the GVCE plus some technical steps that hold also for any $\{0,1\}$-valued loss like the one we are using in the Adversarial Setting [19] and by simply redefining the concept of GVCE and LVCE in the AGVCE and ALVCE respectively as we did before.

3.1.3. Non-adversarial and adversarial settings: a comparison

Let us now consider the two settings described in Sections 3.1.1 and 3.1.2 and let us observe the (A) GVCE based bounds of Eqns. (17) and (27) and the (A) LVCE based bounds of Eqns. (21) and (30). By considering these bounds we can immediately observe, by definition, a series of properties.

First, let us define some quantities. Let us consider a perturbation domain $\mathscr{B}$ such that $\mathscr{I} \subset \mathscr{B} \subset \mathscr{X}$. Let us also consider two perturbations domains $\mathscr{B}_1$ and $\mathscr{B}_2$ such that $\mathscr{I} \subset \mathscr{B}_1 \subset \mathscr{B}_2 \subset \mathscr{X}$. Note that, by definition of perturbation domain, $X \in \mathscr{I}(X), \mathscr{B}(X), \mathscr{B}_1(X), \mathscr{B}_2(X), \mathscr{X}(X)$. Finally let us consider three variables r, r_1, and r_2 such that $0 \leqslant r \leqslant 1$ and $0 < r_1 < r_2 < 1$.

Then let us consider the behavior of the generalization error and the empirical error in both the Non-Adversarial and Adversarial Settings

$$\begin{aligned} &\mathsf{L}_\ell^Y(h) = \mathsf{L}_{\hat{\ell}_{\mathscr{I}}}^Y(h) \leqslant \mathsf{L}_{\hat{\ell}_{\mathscr{B}_1}}^Y(h) \leqslant \mathsf{L}_{\hat{\ell}_{\mathscr{B}_2}}^Y(h) \leqslant \mathsf{L}_{\hat{\ell}_{\mathscr{X}}}^Y(h) = 1, \\ &\hat{\mathsf{L}}_\ell^Y(h) = \hat{\mathsf{L}}_{\hat{\ell}_{\mathscr{I}}}^Y(h) \leqslant \hat{\mathsf{L}}_{\hat{\ell}_{\mathscr{B}_1}}^Y(h) \leqslant \hat{\mathsf{L}}_{\hat{\ell}_{\mathscr{B}_2}}^Y(h) \leqslant \hat{\mathsf{L}}_{\hat{\ell}_{\mathscr{X}}}^Y(h) = 1, \quad \forall h \in \mathscr{H} \end{aligned} \tag{31}$$

which follows directly by the definition of $\mathsf{L}_\ell^Y, \hat{\mathsf{L}}_\ell^Y, \mathsf{L}_{\hat{\ell}_{\mathscr{B}}}^Y$, and $\hat{\mathsf{L}}_{\hat{\ell}_{\mathscr{B}}}^Y$. In fact, the larger the perturbation that we can apply to the samples in the dataset the higher is the probability to find a perturbation that induces a hypothesis $h \in \mathscr{H}$ into a mistake. Note that, there can be a set of perturbation $\mathscr{B}$ (in some cases also large) that does not change the actual meaning of X and so, if $\mathscr{H}$ is carefully tuned, we should be still able to keep $\mathsf{L}_{\hat{\ell}_{\mathscr{B}}}^Y$ and $\hat{\mathsf{L}}_{\hat{\ell}_{\mathscr{B}}}^Y$ small [6,28–31,10].

Then let us consider the complexity term. Also in this case there are many properties that we can state that follows directly from their definitions. In particular, we start by analyzing the (A) GVCE and (A) LVCE. In this case we can state that

$$\begin{aligned} &0 = \hat{\mathsf{V}}_\ell(\mathscr{H}, 0) \leqslant \hat{\mathsf{V}}_\ell(\mathscr{H}, r_1) \leqslant \hat{\mathsf{V}}_\ell(\mathscr{H}, r_2) \leqslant \hat{\mathsf{V}}_\ell(\mathscr{H}, 1) = \hat{\mathsf{V}}_\ell(\mathscr{H}) \leqslant n \ln(2), \\ &0 = \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}, 0) \leqslant \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}, r_1) \leqslant \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}, r_2) \leqslant \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}, 1) = \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}) \leqslant n \ln(2), \end{aligned} \tag{32}$$

which means that the (A) LVCE is always smaller than its Global counterpart (A) GVCE, that the (A) LVCE degenerates in the (A) GVCE respectively, and that for $r = 0$ the (A) LVCE leave just one possibility namely the functions with all zero errors[4].

Moreover we can state some relations between the GVCE and LVCE and the AGVCE and the ALVCE respectively that follow again from their definitions. In fact

$$\begin{aligned} &\lim_{\mathscr{B} \to \mathscr{X}} \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}) \leqslant \ln(3), \\ &\hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{X}}}(\mathscr{H}) \leqslant \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}), \quad \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{X}}}(\mathscr{H}) \leqslant \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{I}}}(\mathscr{H}) = \hat{\mathsf{V}}_\ell(\mathscr{H}), \\ &\lim_{\mathscr{B} \to \mathscr{X}} \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}, \cdot) \leqslant \ln(3), \\ &\hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{X}}}(\mathscr{H}, r) \leqslant \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}}}(\mathscr{H}, r), \quad \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{X}}}(\mathscr{H}, r) \leqslant \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{I}}}(\mathscr{H}, r) = \hat{\mathsf{V}}_\ell(\mathscr{H}, r), \end{aligned} \tag{33}$$

namely no matter what $\mathscr{D}$ and $\mathscr{H}$ there are only three possible ways of configuring the vectors $\{0,1\}$-errors distinguishable within $\mathscr{H}$ with respect to the dataset $\mathscr{D}$: correctly label all the $Y_i = +1$ with $h(X) = +1$, correctly label all the $Y_i = -1$ with $h(X) = -1$, and make mistakes on all points with any other $h \in \mathscr{H}$.

These properties then tell us different things. For $\mathscr{B} = \mathscr{I}$ the AGVCE degenerates in the GVCE and the ALVCE degenerates in the LVCE. Moreover for $\mathscr{B} = \mathscr{X}$ we have the smallest possible AGVCE and ALVCE.

The properties of Eqns. (32) and (33) are graphically represented in Fig. 1(a).

What is not easy to prove is which operator $\circ \in \{\leqslant, \geqslant, =\}$ can be inserted in the following relations

$$\begin{aligned} &\hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{X}}}(\mathscr{H}) \circ \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}_2}}(\mathscr{H}) \circ \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}_1}}(\mathscr{H}) \circ \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{I}}}(\mathscr{H}), \\ &\hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{X}}}(\mathscr{H}, r) \circ \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}_2}}(\mathscr{H}, r) \circ \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{B}_1}}(\mathscr{H}, r) \circ \hat{\mathsf{V}}_{\hat{\ell}_{\mathscr{I}}}(\mathscr{H}, r). \end{aligned} \tag{34}$$

The first idea would be to put $\circ \rightarrow \leqslant$ because of the properties of Eq. (33) (see Fig. 1(b)).

As we will show in a toy example (see Fig. 2(a)), this intuition is wrong. Let us consider the following toy example: $\mathscr{X} = [-3,3] \times [-2,2] \subset \mathbb{R} \times \mathscr{R}, \mathscr{D} = \{Z_1, Z_2, Z_3\} = \{([-1,0], +1), ([0,0], +1), ([+1,0], -1)\}, \mathscr{B}_1(X) = \|\widetilde{X} : \|\widetilde{X} - \widetilde{X}\|_2 \leqslant .2\|, \mathscr{B}_2(X) = \|\widetilde{X} : \|\widetilde{X} - \widetilde{X}\|_2 \leqslant 1.1\|$, and $\mathscr{H}$ are all the linear separators in $\mathscr{H}$. In this setting we can study the (A) GVCE and the (A) LVCE varying $\mathscr{B} \in \{\mathscr{I}, \mathscr{B}_1, \mathscr{B}_2, \mathscr{X}\}$ and $r \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ (all the possible values since the empirical error can assume only these four values). For this purpose let us observe Table 1 where we report all the quantities related to the (A) GVCE and the (A) LVCE and we also visualize the same quantities in Fig. 2.

Thanks to the toy sample represented and studied in Fig. 2 and Table 1, we can now state that the intuition to put $\circ \rightarrow \leqslant$ in the properties of Eq. (34) see Fig. 1(b)) is wrong since we found a counterexample (our toy example) that contradicts this intuition.

A more realistic intuition, that comes from the behavior observed in our toy example, is the one of Fig. 1(c), namely for small $\mathscr{B}(X)$ the (A) GVCE and the (A) LVCE may increase but as $\mathscr{B}(X)$ increases they should then start to decrease. We will test this realistic intuition in the experimental section later (see Section 4).

Thanks to the properties that we observed in this section we argue that there could be a benefit when estimating the generalization ability of a model in the Adversarial Setting with respect to the Non-Adversarial Setting. In fact, as we said before, for reasonably large perturbation the empirical error can still be small (i.e., the best model in the hypothesis space is still able to perform well on the true labels) while the complexity can remarkably decrease (i.e., the models in the hypothesis are not able to generate too many possible distinct vectors of configuration of the $\{0,1\}$-errors) creating an optimal perturbation size that is able to keep the empirical error small while decreasing the complexity resulting in tighter generalization bound. We will challenge this argument in the experimental section later (see Section 4).

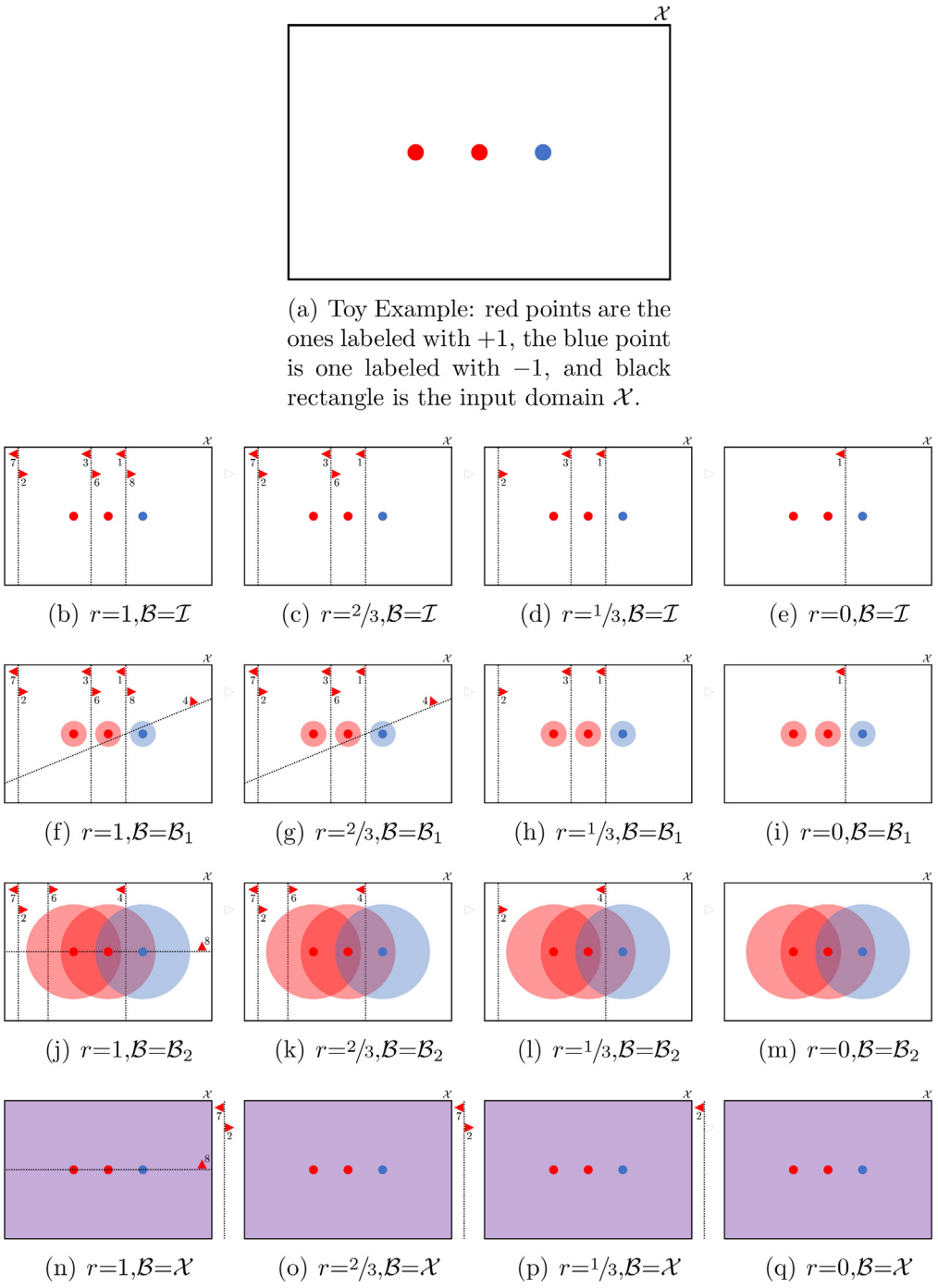
Note also another important fact. From Eq. (31) we can state that the generalization error of a model in the Non-Adversarial Setting is bounded by the generalization error of the same model in the Adversarial Setting with any possible perturbation $\mathscr{B}$. This means that if we can find a perturbation $\mathscr{B}$ such that the estimated generalization error for a particular model in the Adversarial Setting is smaller than the one estimated in the Non-Adversarial Setting we can use the estimated generalization in the Adversarial Setting to get a tighter bound also for the error in the Non-Adversarial Setting. This can open a new field of research: find perturbations $\mathscr{B}$ able to minimize the estimated generalization error. Note that this is not a trivial task since the perturbation needs to be designed before seeing the data.

3.2. (Local) Rademacher complexity theory

In this section, as we did for the (A) GVCE and for the (A) LVCE in Section 3.1, we will first study the classical Non-Adversarial Setting (Section 3.2.1), then the Adversarial Setting (Section 3.2.1), and finally we will compare the two settings (Section 3.2.3) using and extending the (Local) Rademacher Complexity theory.

In particular, in this section, we will consider the case in which $\ell(h, Z) \in [0,1]$, namely whatever $[0,1]$-bounded loss (of course it is reasonable to assume that $\ell(h, Z) = 0$ if $h(X) = Y$ and that $\ell(h, Z) = 1$ if $h(X) = -Y$).

[4] Note that this function may not exist so, in practice, there is a minimum value for r, which is $\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right)$ for the LVCE and $\mathsf{L}_{\hat{\ell}_{\mathscr{B}}}^Y\left(\hat{h}_{\hat{\ell}_{\mathscr{B}}}\right)$ for the ALVCE, below which LVCE and ALVCE are zero.



(a) Toy Example: red points are the ones labeled with $+1$, the blue point is one labeled with -1, and black rectangle is the input domain $\mathcal{X}$.

(b) $r{=}1,\mathcal{B}{=}\mathcal{I}$ (c) $r{=}2/3,\mathcal{B}{=}\mathcal{I}$ (d) $r{=}1/3,\mathcal{B}{=}\mathcal{I}$ (e) $r{=}0,\mathcal{B}{=}\mathcal{I}$

(f) $r{=}1,\mathcal{B}{=}\mathcal{B}_1$ (g) $r{=}2/3,\mathcal{B}{=}\mathcal{B}_1$ (h) $r{=}1/3,\mathcal{B}{=}\mathcal{B}_1$ (i) $r{=}0,\mathcal{B}{=}\mathcal{B}_1$

(j) $r{=}1,\mathcal{B}{=}\mathcal{B}_2$ (k) $r{=}2/3,\mathcal{B}{=}\mathcal{B}_2$ (l) $r{=}1/3,\mathcal{B}{=}\mathcal{B}_2$ (m) $r{=}0,\mathcal{B}{=}\mathcal{B}_2$

(n) $r{=}1,\mathcal{B}{=}\mathcal{X}$ (o) $r{=}2/3,\mathcal{B}{=}\mathcal{X}$ (p) $r{=}1/3,\mathcal{B}{=}\mathcal{X}$ (q) $r{=}0,\mathcal{B}{=}\mathcal{X}$

Fig. 2. Toy example for studying (A) GVCE and the (A) LVCE. Dotted lines are the functions h numbered in Table 1 were the red triangle points to the semi space with label $+1$ and the semi-transparent circles represent the $\mathscr{B}(X)$.

3.2.1. Non-adversarial setting

Let us consider, again, the Non-Adversarial Setting in which one has learned $\hat{h}_\ell$ and has to bound its performance, namely estimate value of $\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right)$ just based on empirical quantities.

Let us define n random variables $\mathscr{S} = \{\sigma_1, \cdots, \sigma_n\}$ with $\sigma_i \in \{\pm 1\}$ such that $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = \frac{1}{2}, \forall i \in \{1, \cdots, n\}$. Then we can define the empirical Rademacher Complexity (RC) as [20,21]

$$\hat{\mathsf{R}}_\ell(\mathscr{H}) = \mathbb{E}_{\mathscr{S}} \sup_{h \in \mathscr{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell(h, Z_i). \tag{35}$$

The RC measures the ability of the space of functions to fit noise (i.e., the less is the ability of the space of functions to fit the noise, the smaller is RC). This interpretation of RC can be made apparent by reformulating RC assuming the loss function to be symmetric[5], namely $\ell(h,(X,-Y)) = 1 - \ell(h,(X,Y))$, and defining $\mathscr{S}^+(\mathscr{S}) = \{i : i \in \{1,\cdots,n\}, \sigma_i = +1\}$ and $\mathscr{S}^-(\mathscr{S}) = \{i : i \in \{1,\cdots,n\}, \sigma_i = -1\}$. In this setting we can note that

[5] Note that this property is satisfied, for example, by the Hard loss function [18] $\ell(h,Z) = [Yh(X) \leqslant 0]$ and the truncated Hinge (or Ramp or Soft) loss function [32] $\ell(h,Z) = \frac{1}{2}\min[2, \max[0, 1 - Yh(X)]]$.

Table 1
Studying the (A) GVCE and the (A) LVCE varying $\mathscr{B} \in \{\mathscr{I}, \mathscr{B}_1, \mathscr{B}_2, \mathscr{X}\}$ and $r \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$ for the toy example of Fig. 2(a). Please refer to Fig. 2 to observe the visualization of this table.

h	1	2	3	4	5	6	7	8		
$\ell(h,Z_1)$, $\tilde{\ell}_{\mathscr{B}}(h,Z_1)$	0	0	0	0	1	1	1	1		
$\ell(h,Z_2)$, $\tilde{\ell}_{\mathscr{B}}(h,Z_2)$	0	0	1	1	0	0	1	1		
$\ell(h,Z_3)$, $\tilde{\ell}_{\mathscr{B}}(h,Z_3)$	0	1	0	1	0	1	0	1		
$\hat{\mathsf{L}}^Y_\ell(h)$, $\hat{\mathsf{L}}^Y_{\tilde{\ell}_{\mathscr{B}}}(h)$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{2}{3}$	1		
$\mathscr{L}_{\tilde{\ell}_{\mathscr{I}},\mathscr{D},1} = \mathscr{L}_{\ell,\mathscr{D},1}$ $= \mathscr{L}_{\tilde{\ell}_{\mathscr{X}},\mathscr{D}} = \mathscr{L}_{\ell,\mathscr{D}}$	✓	✓	✓			✓	✓	✓	→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H},1) = \hat{\mathsf{V}}_{\ell}(\mathscr{H},1)$ $= \hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H}) = \hat{\mathsf{V}}_{\ell}(\mathscr{H}) = \ln(6)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_1},\mathscr{D},1} = \mathscr{L}_{\tilde{\ell}_{\mathscr{B}_1},\mathscr{D}}$	✓	✓	✓	✓		✓	✓	✓	→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_1}}(\mathscr{H},1) = \hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_1}}(\mathscr{H}) = \ln(7)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_2},\mathscr{D},1} = \mathscr{L}_{\tilde{\ell}_{\mathscr{B}_2},\mathscr{D}}$		✓		✓		✓	✓	✓	→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_2}}(\mathscr{H},1) = \hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_2}}(\mathscr{H}) = \ln(5)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{X}},\mathscr{D},1} = \mathscr{L}_{\tilde{\ell}_{\mathscr{X}},\mathscr{D}}$		✓					✓	✓	→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H},1) = \hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H}) = \ln(3)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{I}},\mathscr{D},\frac{2}{3}} = \mathscr{L}_{\ell,\mathscr{D},\frac{2}{3}}$	✓	✓	✓			✓	✓		→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H},\frac{2}{3}) = \hat{\mathsf{V}}_{\ell}(\mathscr{H},\frac{2}{3}) = \ln(5)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_1},\mathscr{D},\frac{2}{3}}$	✓	✓	✓	✓		✓	✓		→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_1}}(\mathscr{H},\frac{2}{3}) = \ln(6)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_2},\mathscr{D},\frac{2}{3}}$		✓		✓		✓	✓		→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_2}}(\mathscr{H},\frac{2}{3}) = \ln(4)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{X}},\mathscr{D},\frac{2}{3}}$		✓					✓		→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H},\frac{2}{3}) = \ln(2)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{I}},\mathscr{D},\frac{1}{3}} = \mathscr{L}_{\ell,\mathscr{D},\frac{1}{3}}$	✓	✓	✓						→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H},\frac{1}{3}) = \hat{\mathsf{V}}_{\ell}(\mathscr{H},\frac{1}{3}) = \ln(3)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_1},\mathscr{D},\frac{1}{3}}$	✓	✓	✓						→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_1}}(\mathscr{H},\frac{1}{3}) = \ln(3)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_2},\mathscr{D},\frac{1}{3}}$		✓							→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_2}}(\mathscr{H},\frac{1}{3}) = \ln(1)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{X}},\mathscr{D},\frac{1}{3}}$		✓							→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H},\frac{1}{3}) = \ln(1)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{I}},\mathscr{D},0} = \mathscr{L}_{\ell,\mathscr{D},0}$	✓								→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H},0) = \hat{\mathsf{V}}_{\ell}(\mathscr{H},0) = \ln(1)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_1},\mathscr{D},0}$	✓								→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_1}}(\mathscr{H},0) = \ln(1)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{B}_2},\mathscr{D},0}$									→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{B}_2}}(\mathscr{H},0) = \ln(1)$
$\mathscr{L}_{\tilde{\ell}_{\mathscr{X}},\mathscr{D},0}$									→	$\hat{\mathsf{V}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H},0) = \ln(1)$

$$
\begin{aligned}
\hat{\mathsf{R}}_\ell(\mathscr{H}) &= \mathbb{E}_{\mathscr{S}} \sup_{h\in\mathscr{H}} \frac{2}{n}\sum_{i=1}^{n}\sigma_i\ell(h,Z_i) \\
&= 1 + \mathbb{E}_{\mathscr{S}} \sup_{h\in\mathscr{H}} \frac{2}{n}\left\{\sum_{i\in\mathscr{S}^+(\mathscr{S})}[\ell(h,Z_i)-1] + \sum_{i\in\mathscr{S}^-(\mathscr{S})} -\ell(h,Z_i)\right\} \\
&= 1 - \mathbb{E}_{\mathscr{S}} \inf_{h\in\mathscr{H}} \frac{2}{n}\left[\sum_{i\in\mathscr{S}^+(\mathscr{S})}\ell(h,(X_i,-Y_i)) + \sum_{i\in\mathscr{S}^-(\mathscr{S})}\ell(h,(X_i,Y_i))\right] \\
&= 1 - 2\mathbb{E}_{\mathscr{S}} \inf_{h\in\mathscr{H}} \frac{1}{n}\sum_{i=1}^{n}\ell(h,(X_i,\sigma_i)) \\
&= 1 - 2\mathbb{E}_{\mathscr{S}} \inf_{h\in\mathscr{H}} \hat{\mathsf{L}}^{\sigma}_{\ell}(h) \in [0,1],
\end{aligned}
\tag{36}
$$

which allows us to state that the RC is the average maximum accuracy on random labels [33].

Note that the RC has a strong connection with the VCE [33] when we use the Hard loss function $\ell(h,Z) = [Yh(X) \leqslant 0]$. In fact, in this case, we can observe by definition that [18,33]

$$
\hat{\mathsf{R}}_\ell(\mathscr{H}) = \mathbb{E}_{\mathscr{S}} \sup_{h\in\mathscr{H}} \frac{2}{n}\sum_{i=1}^{n}\sigma_i\ell(h,Z_i) = \mathbb{E}_{\mathscr{S}} \sup_{[\ell_1,\cdots,\ell_n]\in\mathscr{L}_{\ell,\mathscr{D}}} \frac{2}{n}\sum_{i=1}^{n}\sigma_i\ell_i, \tag{37}
$$

which means that the RC is the ability of the distinct vectors of configuration of the $\{0,1\}$-errors distinguishable within $\mathscr{H}$ with respect to the dataset $\mathscr{D}$ (the VCE) to be aligned with the 2^n possible configurations of the σ.

In the general setting of a $[0,1]$-bounded loss it is possible to prove a bound in the form of Eq. (17) by setting $\hat{\mathsf{C}}_\ell(\mathscr{H}) = \hat{\mathsf{R}}_\ell(\mathscr{H})$ and $\phi(\delta) = 3\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}$ [20,21].

The bound of Eq. (17) is, also in this case, a fully empirical bound, namely all the quantities can be estimated from the data, the bound actually holds for any hypothesis chosen in $\mathscr{H}$ according to $\mathscr{D}$ if $\mathscr{H}$ is chosen before observing $\mathscr{D}$, and can be improved both in the constants and both in the rate of convergence [34,20,21].

Computing $\mathbb{E}_{\mathscr{S}}$ can be computational expensive but we can resort to an estimation [19], via Monte Carlo methods, or we can formulate a bound where only one realization of the sigmas need to be employed [19,35]. Nevertheless, in order to obtain reliable and sharp bounds (avoiding unlucky realization of the sigmas) it is common to resort to a Monte Carlo estimation [19,24].

The RC of Eq. (35) and its associated bound of Eq. (17) are called Global RC (GRC) and GRC based bound respectively since, as the GVCE, all the functions in $\mathscr{H}$ contribute to $\hat{\mathsf{C}}_\ell(\mathscr{H})$ even the ones that will be never chosen by the algorithm, namely the one characterized by high error.

It is then possible to define a localized version of the GRC, namely the Local RC (LRC) [22,24], for a general $[0,1]$-bounded loss controlled by a parameter $r \in [0,1] \subset \mathbb{R}$

$$
\hat{\mathsf{R}}_\ell(\mathscr{H},r) = \mathbb{E}_{\mathscr{S}} \sup_{\alpha\in[0,1], h\in\left\{h:h\in\mathscr{H},\frac{1}{n}\sum_{i=1}^{n}\alpha^2\ell^2(h,Z_i)\leqslant r\right\}} \frac{2}{n}\sum_{i=1}^{n}\sigma_i\alpha\ell(h,Z_i). \tag{38}
$$

$\hat{\mathsf{R}}_\ell(\mathscr{H},r)$ is monotonically increasing in r, namely if $0 \leqslant r_1 \leqslant r_2 \leqslant 1$ then $\hat{\mathsf{R}}_\ell(\mathscr{H},r_1) \leqslant \hat{\mathsf{R}}_\ell(\mathscr{H},r_2)$.

Note that in the same setting of the VCE, namely using the Hard Loss function $\ell(h,Z) = [Yh(X) \leqslant 0]$, we can reformulate Eq. (38) as [22,19]

$$
\begin{aligned}
\hat{\mathsf{R}}_\ell(\mathscr{H},r) &= \mathbb{E}_{\mathscr{S}} \sup_{\alpha\in[0,1]} \sup_{h\in\left\{h:h\in\mathscr{H},\hat{\mathsf{L}}^Y_\ell(h)\leqslant\frac{r}{\alpha^2}\right\}} \frac{2}{n}\sum_{i=1}^{n}\sigma_i\alpha\ell(h,Z_i) \\
&= \mathbb{E}_{\mathscr{S}} \sup_{\alpha\in[0,1]} \sup_{[\ell_1,\cdots,\ell_n]\in\mathscr{L}_{\ell,\mathscr{D},\frac{r}{\alpha^2}}} \frac{2}{n}\sum_{i=1}^{n}\sigma_i\alpha\ell_i.
\end{aligned}
\tag{39}
$$

In this case there is a strong connection between the LRC and the LVCE since LRC is the ability of the vectors of the possible configuration of the $\{0,1\}$-errors distinguishable within $\mathscr{H}$ with respect to the dataset $\mathscr{D}$ with small error (the LVCE) to be aligned with the 2^n possible configurations of the $\mathscr{S}$.

Note also that LRC does not degenerate in the GRC. In fact for $r=1$, based on Eq. (38), we have that

$$\begin{aligned}\hat{\mathsf{R}}_\ell(\mathscr{H},1) &= \mathbb{E}_{\mathscr{S}} \sup_{\alpha\in[0,1],h\in\left\{h:h\in\mathscr{H},\frac{1}{n}\sum_{i=1}^n\alpha^2\ell^2(h,Z_i)\leqslant 1\right\}} \frac{2}{n}\sum_{i=1}^n\sigma_i\alpha\ell(h,Z_i)\\ &= \mathbb{E}_{\mathscr{S}} \sup_{\alpha\in[0,1],h\in\mathscr{H}} \alpha\frac{2}{n}\sum_{i=1}^n\sigma_i\ell(h,Z_i)\\ &\geqslant \mathbb{E}_{\mathscr{S}}\sup_{h\in\mathscr{H}}\frac{2}{n}\sum_{i=1}^n\sigma_i\ell(h,Z_i)\\ &= \hat{\mathsf{R}}_\ell(\mathscr{H}),\end{aligned} \tag{40}$$

since $\sup_{h\in\mathscr{H}}\frac{2}{n}\sum_{i=1}^n\sigma_i\ell(h,Z_i)$ can be negative.

In the general setting of a $[0,1]$-bounded loss it is possible to prove that [24]

$$\begin{aligned}&\mathbb{P}\left\{\mathsf{L}_\ell^Y\left(\hat{h}_\ell\right)\leqslant\min_{K\in(1,\infty)}\frac{K}{K-1}\hat{\mathsf{L}}_\ell^Y\left(\hat{h}_\ell\right)+Kr+2\sqrt{\frac{\ln\left(\frac{3}{\delta}\right)}{2n}}\right\}\geqslant 1-\delta,\\ &\text{s.t.}\begin{cases}r=\hat{\mathsf{R}}_\ell\left(\mathscr{H},3r+\sqrt{\frac{\ln\left(\frac{3}{\delta}\right)}{2n}}\right)+\sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{n}}\\ r>0\end{cases}.\end{aligned} \tag{41}$$

This bound, the LRC based bound, is then able to discard functions with high error thanks to the fact that the functions with high error are not contemplated in the complexity term.

The same comments made for the GRC based bound of Eq. (17) holds also for the LRC based bound just stated in Eq. (41) (namely its localized version). In fact LRC based bound is fully empirical, can be improved in both constants and rate of convergence, and actually holds for any hypothesis chosen in $\mathscr{H}$ according to $\mathscr{D}$ if $\mathscr{H}$ is chosen before observing $\mathscr{D}$ [22,24]. Moreover Eq. (41) can be simplified in the form of the bound of Eq. (21) by setting $c_1=\frac{K^*}{K^*-1}$, $\hat{\mathsf{C}}_\ell(\mathscr{H},c_2)=K^*\hat{\mathsf{R}}_\ell(\mathscr{H},c_2)$, $c_2=3r^*+\sqrt{\frac{\ln\left(\frac{3}{\delta}\right)}{2n}}$, and $\phi(\delta)=\sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{n}}+2\sqrt{\frac{\ln\left(\frac{3}{\delta}\right)}{2n}}$ where K^* and r^* are the values of K and r that solves the problem of the bound of Eq. (41).

3.2.2. Adversarial Setting

Let us consider now the Adversarial Setting in which one has learned $\hat{h}_{\tilde{\ell}_{\mathscr{B}}}$ and has to bound its performance, namely estimate value of $\mathsf{L}_{\tilde{\ell}_{\mathscr{B}}}^Y\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ just based on empirical quantities. Our adversary on $h\in\mathscr{H}$ is modeled according to Eq. (8) using the loss function.

The extension of the notion of GRC to the Adversarial Setting, namely the Adversarial GRC (AGRC), can be easily performed, analogously to what has been done for the GVCE and AVCE, by nothing that $\tilde{\ell}_{\mathscr{B}}(h,Z)\in[0,1]$ and so the bounds presented in the previous section simply holds also in this case with some simple redefinitions and substitutions. What changes is the definition and the properties of the AGRC.

In fact the definition of the empirical AGRC is the following one [17]

$$\begin{aligned}\hat{\mathsf{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) &= \mathbb{E}_{\mathscr{S}}\sup_{h\in\mathscr{H}}\frac{2}{n}\sum_{i=1}^n\sigma_i\tilde{\ell}_{\mathscr{B}}(h,Z_i)\\ &= \mathbb{E}_{\mathscr{S}}\sup_{h\in\mathscr{H}}\frac{2}{n}\sum_{i=1}^n\sigma_i\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},Y_i\right)\right).\end{aligned} \tag{42}$$

In the same setting of the AGVCE, namely $\ell(h,Z)=[Yh(X)\leqslant 0]$, we can reformulate Eq. (42) as

$$\begin{aligned}\hat{\mathsf{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) &= \mathbb{E}_{\mathscr{S}}\sup_{h\in\mathscr{H}}\frac{2}{n}\sum_{i=1}^n\sigma_i\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},Y_i\right)\right)\\ &= \mathbb{E}_{\mathscr{S}}\sup_{[\ell_1,\cdots,\ell_n]\in\mathscr{L}_{\tilde{\ell}_{\mathscr{B}},\mathscr{S}}}\frac{2}{n}\sum_{i=1}^n\sigma_i\ell_i,\end{aligned} \tag{43}$$

where one can immediately see that there is a strong connection between the AGRC and the AVCE.

Assuming, instead, again, as in the previous section, ℓ to be symmetric, the AGRC can be reformulated as follows

$$\begin{aligned}\hat{\mathsf{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) &= \mathbb{E}_{\mathscr{S}}\sup_{h\in\mathscr{H}}\frac{2}{n}\sum_{i=1}^n\sigma_i\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},Y_i\right)\right)\\ &= 1+\mathbb{E}_{\mathscr{S}}\sup_{f\in\mathscr{F}}\frac{2}{n}\left[\sum_{i\in\mathscr{S}^+(\mathscr{S})}\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},Y_i\right)\right)\right.\\ &\qquad\left.-\sum_{i\in\mathscr{S}^-(\mathscr{S})}\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},Y_i\right)\right)-\sum_{i\in\mathscr{S}^+(\mathscr{S})}1\right]\\ &= 1+\mathbb{E}_{\mathscr{S}}\sup_{h\in\mathscr{H}}\frac{2}{n}\left[\sum_{i\in\mathscr{S}^+(\mathscr{S})}\sup_{\widetilde{X}\in\mathscr{B}(X_i)}-\ell\left(h,\left(\widetilde{X},-Y_i\right)\right)\right.\\ &\qquad\left.-\sum_{i\in\mathscr{S}^-(\mathscr{S})}\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},Y_i\right)\right)\right]\\ &= 1-2\mathbb{E}_{\mathscr{S}}\inf_{h\in\mathscr{H}}\frac{1}{n}\sum_{i=1}^n\sigma_i\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\sigma_i\ell\left(h,\left(\widetilde{X},\sigma_iY_i\right)\right)\\ &= 1-2\mathbb{E}_{\mathscr{S}}\inf_{h\in\mathscr{H}}\frac{1}{n}\left[\sum_{i\in\mathscr{S}^+(\mathscr{S})}\sup_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},Y_i\right)\right)\right.\\ &\qquad\left.+\sum_{i\in\mathscr{S}^-(\mathscr{S})}\inf_{\widetilde{X}\in\mathscr{B}(X_i)}\ell\left(h,\left(\widetilde{X},-Y_i\right)\right)\right]\\ &= 1-2\mathbb{E}_{\mathscr{S}}\inf_{h\in\mathscr{H}}\frac{1}{n}\sum_{i=1}^n\tilde{\tilde{\ell}}_{\mathscr{B},\sigma_i}(h(X_i),Y_i)\\ &= 1-2\mathbb{E}_{\mathscr{S}}\inf_{h\in\mathscr{H}}\hat{\mathsf{L}}_{\tilde{\ell}_{\mathscr{B}},\mathscr{S}}(h)\in[0,1],\end{aligned} \tag{44}$$

which allows us to state that the AGRC is the average maximum accuracy of on random labels with a surrogate loss defined as

$$\tilde{\tilde{\ell}}_{\mathscr{B},\sigma}(h(X),Y)=\begin{cases}\sup_{\widetilde{X}\in\mathscr{B}(X)}\ell\left(h\left(\widetilde{X}\right),Y\right) & \text{if}\sigma=+1\\ \inf_{\widetilde{X}\in\mathscr{B}(X)}\ell\left(h\left(\widetilde{X}\right),-Y\right) & \text{if}\sigma=-1\end{cases}. \tag{45}$$

Thanks to this definition, in the general Adversarial Setting, we can state the counterpart of the bound of Eq. (27) based on the AGRC by setting $\hat{\mathsf{C}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H})=\hat{\mathsf{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H})$ and $\phi(\delta)=3\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}$ [30].

As we did for the AGRC, we can define the LRC in the Adversarial Setting, namely the Adversarial LRC (ALRC), as

$$\hat{\mathsf{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H},r)=\mathbb{E}_{\mathscr{S}}\sup_{\alpha\in[0,1],h\in\left\{h:h\in\mathscr{H},\frac{1}{n}\sum_{i=1}^n\alpha^2\tilde{\ell}^2_{\mathscr{B}}(h,Z_i)\leqslant r\right\}}\frac{2}{n}\sum_{i=1}^n\sigma_i\alpha\tilde{\ell}_{\mathscr{B}}(h,Z_i). \tag{46}$$

Analogously to the LRC, the ALRC is monotonically increasing in r, namely if $0 \leqslant r_1 \leqslant r_2 \leqslant 1$ then $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, r_1) \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, r_2)$ and, in the same setting of the ALVCE, namely $\ell(h, Z) = [Yh(X) \leqslant 0]$, we can reformulate Eq. (46) as

$$\begin{aligned} \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, r) &= \mathbb{E}_{\mathscr{S}} \sup_{\alpha \in [0,1]} \sup_{h \in \left\{ h : h \in \mathscr{H}, \hat{\mathrm{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}(h) \leqslant \frac{r}{2^2} \right\}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \alpha \tilde{\ell}_{\mathscr{B}}(h, Z_i) \\ &= \mathbb{E}_{\mathscr{S}} \sup_{\alpha \in [0,1]} \sup_{[\ell_1, \cdots, \ell_n] \in \mathscr{L}_{\tilde{\ell}_{\mathscr{B}}, \mathscr{D}, \frac{r}{2^2}}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \alpha \ell_i, \end{aligned} \tag{47}$$

where one can immediately see that there is a strong connection between the ALRC and the ALVCE but, and as for the LRC and GRC, the ALRC does not degenerate in the AGRC (we can easily prove it using the same argument of Eq. (40)).

Thanks to this definition we can state the counterpart of the LRC bound of Eq. (41) for the Adversarial Setting substituting $\mathrm{L}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\mathrm{L}^{Y}_{\ell}\left(\hat{h}_{\ell}\right), \hat{\mathrm{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}}}\left(\hat{h}_{\tilde{\ell}_{\mathscr{B}}}\right)$ to $\hat{\mathrm{L}}^{Y}_{\ell}\left(\hat{h}_{\ell}\right)$, and $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, \cdot)$ to $\hat{\mathrm{R}}_{\ell}(\mathscr{H}, \cdot)$. Consequently we can state the counterpart of the bound of Eq. (30) for the Adversarial Setting based on the ALRC by setting $c_1 = \frac{K^*}{K^*-1}, \hat{\mathrm{C}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, c_2) = K^* \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, c_2), c_2 = 3r^* + \sqrt{\frac{\ln\left(\frac{3}{\delta}\right)}{2n}}$, and $\phi(\delta) = \sqrt{\frac{2\ln\left(\frac{3}{\delta}\right)}{n}} + 2\sqrt{\frac{\ln\left(\frac{3}{\delta}\right)}{2n}}$ where K^* and r^* are the values of K and r that solves the problem of the ALRC counterpart of the bound of Eq. (41).

3.2.3. Non-adversarial and adversarial settings: a comparison

Let us consider the same properties of $r, r_1, r_2, \mathscr{I}, \mathscr{B}, \mathscr{B}_1, \mathscr{B}_2, \mathscr{X}$ defined in Section 3.1.3. Moreover, analogously to what has been done for the (A) GVCE and (A) LVCE, let us now consider the two settings described in Sections 3.2.1 and 3.2.2 and let us observe the (A) GRC based bounds of Eqns. (17) and (27) and the (A) LRC based bounds of Eqns. (21) and (30). By considering these bounds and observing, by definition, a series of properties.

For what concerns the generalization and the empirical errors the properties of Eq. (31) holds true also for a general $[0,1]$-bounded loss

$$\begin{aligned} &\mathrm{L}^{Y}_{\ell}(h) = \mathrm{L}^{Y}_{\tilde{\ell}_{\mathscr{I}}}(h) \leqslant \mathrm{L}^{Y}_{\tilde{\ell}_{\mathscr{B}_1}}(h) \leqslant \mathrm{L}^{Y}_{\tilde{\ell}_{\mathscr{B}_2}}(h) \leqslant \mathrm{L}^{Y}_{\tilde{\ell}_{\mathscr{X}}}(h) = 1, \\ &\hat{\mathrm{L}}^{Y}_{\ell}(h) = \hat{\mathrm{L}}^{Y}_{\tilde{\ell}_{\mathscr{I}}}(h) \leqslant \hat{\mathrm{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}_1}}(h) \leqslant \hat{\mathrm{L}}^{Y}_{\tilde{\ell}_{\mathscr{B}_2}}(h) \leqslant \hat{\mathrm{L}}^{Y}_{\tilde{\ell}_{\mathscr{X}}}(h) = 1, \quad \forall h \in \mathscr{H}. \end{aligned} \tag{48}$$

For what concerns the complexity terms, also in this case there are many properties that we can state that follows directly from their definitions. In particular, we can state the counterpart of the properties of Eq. (32) for the (A) GRC and the (A) LRC

$$\begin{aligned} &0 = \hat{\mathrm{R}}_{\ell}(\mathscr{H}, 0) \leqslant \hat{\mathrm{R}}_{\ell}(\mathscr{H}, r_1) \leqslant \hat{\mathrm{R}}_{\ell}(\mathscr{H}, r_2) \leqslant \hat{\mathrm{R}}_{\ell}(\mathscr{H}, 1) \leqslant 1, \\ &0 = \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, 0) \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, r_1) \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, r_2) \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, 1) \leqslant 1, \end{aligned} \tag{49}$$

remembering though that

$$\begin{aligned} &0 \leqslant \hat{\mathrm{R}}_{\ell}(\mathscr{H}) \leqslant \hat{\mathrm{R}}_{\ell}(\mathscr{H}, 1) \leqslant 1, \\ &0 \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, 1) \leqslant 1, \end{aligned} \tag{50}$$

since the (A) LRC does not degenerate in the (A) GRC.

For what concerns instead the counterpart of the properties of Eq. (33) for RC, the problem is a bit more tricky. In fact we can surely say that

$$\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H}) \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H}) = \hat{\mathrm{R}}_{\ell}(\mathscr{H}), \quad \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H}, r) \leqslant \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H}, r) = \hat{\mathrm{R}}_{\ell}(\mathscr{H}, r), \tag{51}$$

which follows directly from the properties of Eq. (33). But it is more complex to derive these to limits

$$\lim_{\mathscr{B} \to \mathscr{X}} \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}), \quad \lim_{\mathscr{B} \to \mathscr{X}} \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, \cdot), \tag{52}$$

which are not easy to bound if we do not put some further hypothesis. For example if we use a hard loss function $\ell(h, Z) = [Yh(X) \leqslant 0]$ and then $\tilde{\ell}_{\mathscr{B}}(h, Z) = \sup_{\widetilde{X} \in \mathscr{B}(X)} \left[Yh\left(\widetilde{X}\right) \leqslant 0 \right]$. In this case, if $\mathscr{B} \to \mathscr{X}$, we have that

$$\mathscr{L}_{\tilde{\ell}_{\mathscr{B} \to \mathscr{X}}, \mathscr{D}} = \left\{ \begin{array}{l} [b_1, \cdots, b_n : b_i = \frac{1}{2}(Y_i + 1)], \\ [b_1, \cdots, b_n : b_i = \frac{1}{2}(1 - Y_1)], \\ [b_1, \cdots, b_n : b_i = 1] \end{array} \right\}, \tag{53}$$

since, in this case, as we observed in the case of the AGVCE and the ALVCE, either we correctly label all the $Y_i = +1$ with $h(X) = +1$ or we correctly label all the $Y_i = -1$ with $h(X) = -1$ or we make a mistake on all the points with any other $h \in \mathscr{H}$. Consequently

$$\begin{aligned} \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B} \to \mathscr{X}}}(\mathscr{H}) &= \mathbb{E}_{\mathscr{S}} \sup_{\{\ell_1, \cdots, \ell_n\} \in \mathscr{L}_{\tilde{\ell}_{\mathscr{B} \to \mathscr{X}}, \mathscr{D}}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \ell_i, \\ \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B} \to \mathscr{X}}}(\mathscr{H}, r) &= \mathbb{E}_{\mathscr{S}} \sup_{\alpha \in [0,1]} \sup_{[\ell_1, \cdots, \ell_n] \in \mathscr{L}_{\tilde{\ell}_{\mathscr{B} \to \mathscr{X}}, \mathscr{D}, \frac{r}{2^2}}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \alpha \ell_i, \end{aligned} \tag{54}$$

which are simple and computable quantities bounded by $c\sqrt{2n}$ where c is a universal constant [20]. In Fig. 3 we computed the quantities reported in Eq. (54) for different values of r, n, and $|\{Y : Y \in \mathscr{D}, Y > 0\}| = 1 - |\{Y : Y \in \mathscr{D}, Y < 0\}|$ to support our statements.

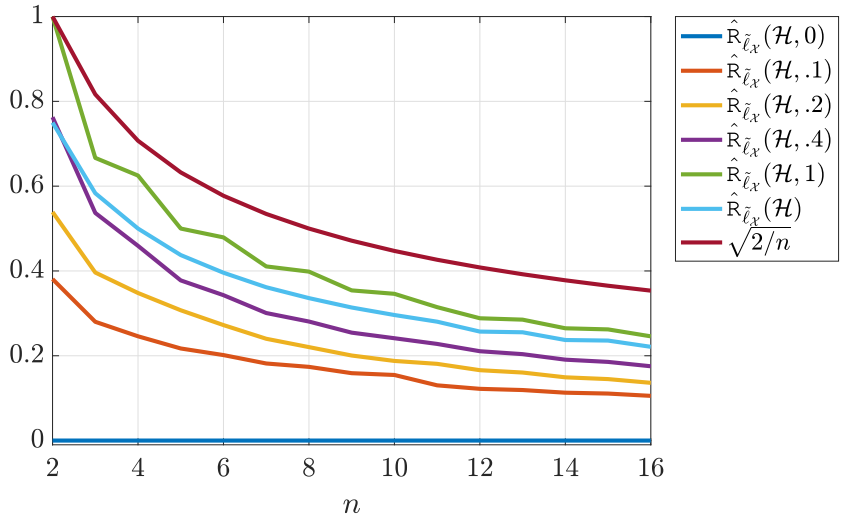
Analogously to the (A) GVCE and (A) LVCE, for the (A) GRC and (A) LRC it is not easy to prove what operator $\circ \in \{\leqslant, \geqslant, =\}$ to insert following relations

$$\begin{aligned} &\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H}) \circ \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}_2}}(\mathscr{H}) \circ \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}_1}}(\mathscr{H}) \circ \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H}), \\ &\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{X}}}(\mathscr{H}, r) \circ \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}_2}}(\mathscr{H}, r) \circ \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}_1}}(\mathscr{H}, r) \circ \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H}, r), \end{aligned} \tag{55}$$

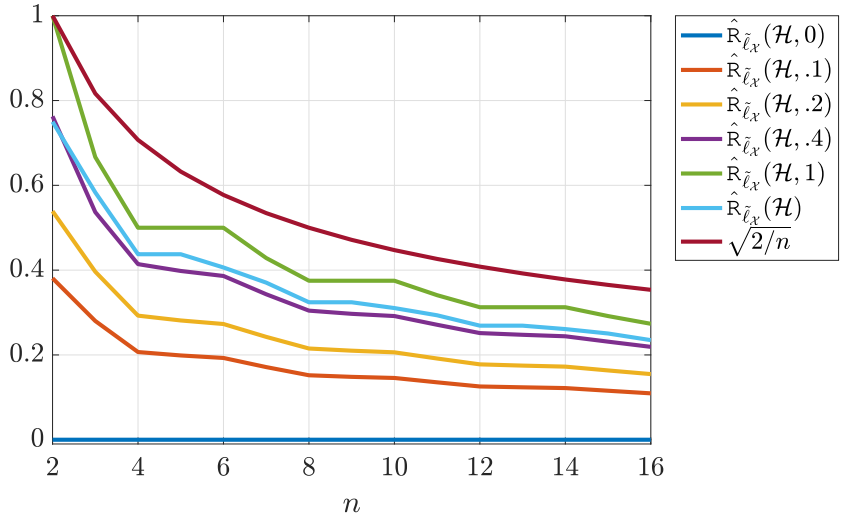
and the answer is the same stated for the (A) GVCE and (A) LVCE: none of them. The (A) GVCE and (A) LVCE can increase for small $\mathscr{B}$ and then can decrease for large $\mathscr{B}$.

In order to support our statement, let us consider the same toy example reported in Section 3.1.3. Let us also use the same loss exploited in Section 3.1.3 $\ell(h, Z) = [Yh(X) \leqslant 0]$ and then $\tilde{\ell}_{\mathscr{B}}(h, Z) = \sup_{\widetilde{X} \in \mathscr{B}(X)} \left[Yh\left(\widetilde{X}\right) \leqslant 0 \right]$. In this setting in Table 1 we have retrieved $\mathscr{L}_{\tilde{\ell}_{\mathscr{B}}, \mathscr{D}, r}$ $\forall \mathscr{B}, r$ of the toy example. So we can also easily compute the (A) GRC using Eq. (43) and the (A) LRC using Eq. (47) remembering that $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H})$ is the GRC, $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H})$ is the AGRC, $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{I}}}(\mathscr{H}, r)$ is the LRC, and $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}, r)$ is the ALRC. The results of this computation are reported in Fig. 4. From Fig. 4 it is possible to observe the same behavior studied, observed, and discussed for (A) GVCE and the (A) LVCE also for (A) GRC and the (A) LRC: there may exist cases in which a perturbation $\mathscr{B}$ can be large enough to not increase the empirical error while decreasing the complexity resulting in sharper bound on generalization error of models learned in the Adversarial Setting.

Finally note that, in some sense, the fact that for small $\mathscr{B}$ the (A) GRC and the (A) LRC can increase should not surprise us. In fact, let us suppose that



(a) $|\{Y : Y \in \mathcal{D}, Y > 0\}| = \lceil n/10 \rceil$



(b) $|\{Y : Y \in \mathcal{D}, Y > 0\}| = \lceil n/2 \rceil$

Fig. 3. Representation of the quantities reported in (Eq. 54) for different values of r, n, and $|\{Y : Y \in \mathscr{D}, Y > 0\}| = 1 - |\{Y : Y \in \mathscr{D}, Y < 0\}|$.

$$\mathscr{X} = \mathbb{R}^d \tag{56}$$

$$\begin{aligned} &h(X) = W \cdot X, \quad \mathscr{H} = \{W : \ W \in \mathbb{R}^d, \|W\|_2 = H\}, \quad H > 0 \\ &\ell(h,Z) = \max\left[0, \min\left[1, \tfrac{1 - YW \cdot X}{2}\right]\right] \\ &\mathscr{B}(X) = \left\{\widetilde{X} : \widetilde{X} \in \mathscr{X}, \|\widetilde{X} - X\|_\infty \leqslant B\right\}, \quad B \in [0,1]. \end{aligned} \tag{57}$$

where $\ell(h,Z)$ is the truncated Hinge (or Ramp or Soft) loss. Then, by exploiting the result of [17] we can state that

$$\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) \leqslant \hat{\mathrm{R}}_{\ell}(\mathscr{H}) + BH\sqrt{\frac{d}{n}}, \tag{58}$$

which is obviously a loose upper bound. In fact, in this case

$$\lim_{B \to \infty} \sup_{X \in \mathscr{B}(X)} \ell(h,Z) = 1, \forall h \in \mathscr{H}, \tag{59}$$

and then, thanks to Eq. (42)

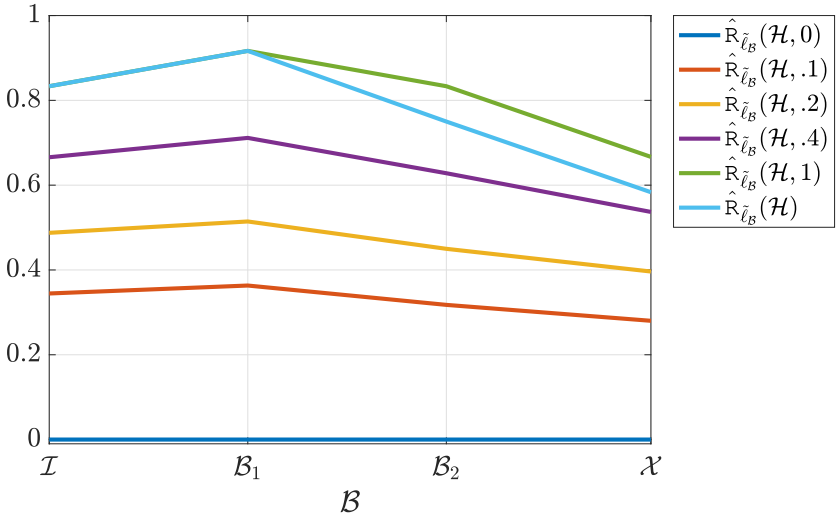
$$\begin{aligned} \lim_{B \to \infty} \hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H}) &= \lim_{B \to \infty} \mathbb{E}_{\mathscr{S}} \sup_{h \in \mathscr{H}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \sup_{\widetilde{X} \in \mathscr{B}(X_i)} \ell\left(h, \left(\widetilde{X}, Y_i\right)\right) \\ &= \lim_{B \to \infty} \mathbb{E}_{\mathscr{S}} \frac{2}{n} \sum_{i=1}^{n} \sigma_i \\ &= 0. \end{aligned} \tag{60}$$

Nevertheless, for small B the upper bound of Eq. (58) [17] can be tight (given what we observed in the toy example for AGRC and ALRC), while for large B it starts to be loose and not useful.

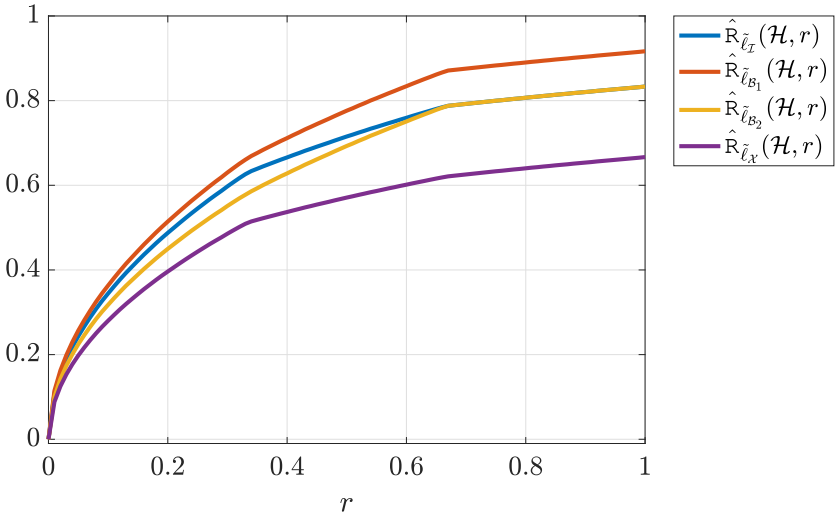
4. Practical analysis of generalization

In this section we will perform a practical analysis to compare the Non-Adversarial and the Adversarial Settings in terms of estimating the generalization abilities of the empirical risk minimizer using the bounds presented in the previous section but, instead of using the toy sample of Sections 3.1.3 and 3.2.3, we will exploit real world data.

Let us consider the case when $\mathscr{X} = \mathbb{R}^d, h(X) = W \cdot X$ where $W \in \mathbb{R}^d$ and the size of $\mathscr{H}$ is regulated by the p-norm of the model weights $\|W\|_p \leqslant H$ where p regulates the shape (e.g, the sparsity or the density) of the solution [36]. Let us also consider the case where $\mathscr{B}(X) \subseteq \mathbb{R}^d$ such that $\mathscr{B}(X) = \left\{\widetilde{X} : \|\widetilde{X} - X\|_q \leqslant B\right\}$ (note that for $B = 0$ we have $\mathscr{B} = \mathscr{I}$). In this case the value of q regulates the shape of the perturbation/attack [37]. Note also that there is a relation between sparsity of the regularizer and robustness to



(a) $\mathcal{B}$ in the x-axis, $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathcal{H})$ and $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathcal{H},r)$ on the y-axis and a curve for different values of r.



(b) r in the x-axis, $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathcal{H},r)$ on the y-axis and a curve for different values of $\mathcal{B}$.

Fig. 4. Quantitative analysis of the properties of the (A) GRC and the (A) LRC for the toy example reported in Section 3.1.3. In particular we reported $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H})$ and $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathscr{B}}}(\mathscr{H},r)$ for $\mathscr{B}\in\{\mathscr{I},\mathscr{B}_1,\mathscr{B}_2,\mathscr{X}\}$ and $r\in[0,1]$.

attacks [38]. For simplicity, in this phase, we will set $p=1$ and $q=\infty$ in the attack. The truncated Hinge (or Ramp or Soft) loss function [32] $\ell(h,Z)=\frac{1}{2}\min[2,\max[0,1-Yh(X)]]$will be exploited (remember that this loss is symmetric).

In this setting, in order to find $\hat{h}_{\tilde{\ell}_{\mathscr{B}}}$ and the $\hat{h}_{\ell}=\hat{h}_{\tilde{\ell}_{\mathscr{I}}}$, namely to find the empirical risk minimizer in both the Non-Adversarial and Adversarial Settings (see Eq. (11)) we have to solve the following problem

$$\inf_{h\in\mathscr{H}}\hat{\mathrm{L}}_{\tilde{\ell}_{\mathscr{B}}}(h)=\inf_{h\in\mathscr{H}}\frac{1}{n}\sum_{i=1}^{n}\sup_{\widetilde{X}\in\mathbb{B}(X_i)}\ell\Big(h,\Big(\widetilde{X},Y\Big)\Big),\tag{61}$$

which, in our setting, can be formulated as

$$\min_{W:W\in\mathbb{R}^d,\|W\|_1\leqslant H}\sum_{i=1}^{n}\max_{\widetilde{X}:\widetilde{X}\in\mathbb{R}^d,\|\widetilde{X}-X_i\|_\infty\leqslant B}\min\Big[2,\max\Big[0,1-Y_iW\cdot\widetilde{X}\Big]\Big].\tag{62}$$

Thanks to the linearity of the Problem (62), it can be simplified as follows [17]

$$\min_{W:W\in\mathbb{R}^d,\|W\|_1\leqslant H}\sum_{i=1}^{n}\min[2,\max[0,1-Y_iW\cdot X_i+\|W\|_1B]].\tag{63}$$

Unfortunately Problem (63) is non-convex but we can approximate its solution by solving its convex relaxation

$$\min_{W:W\in\mathbb{R}^d,\|W\|_1\leqslant H}\sum_{i=1}^{n}\max[0,1-Y_iW\cdot X_i+\|W\|_1B],\tag{64}$$

which a Linear Programming problem, in fact Problem (64) can be reformulated as

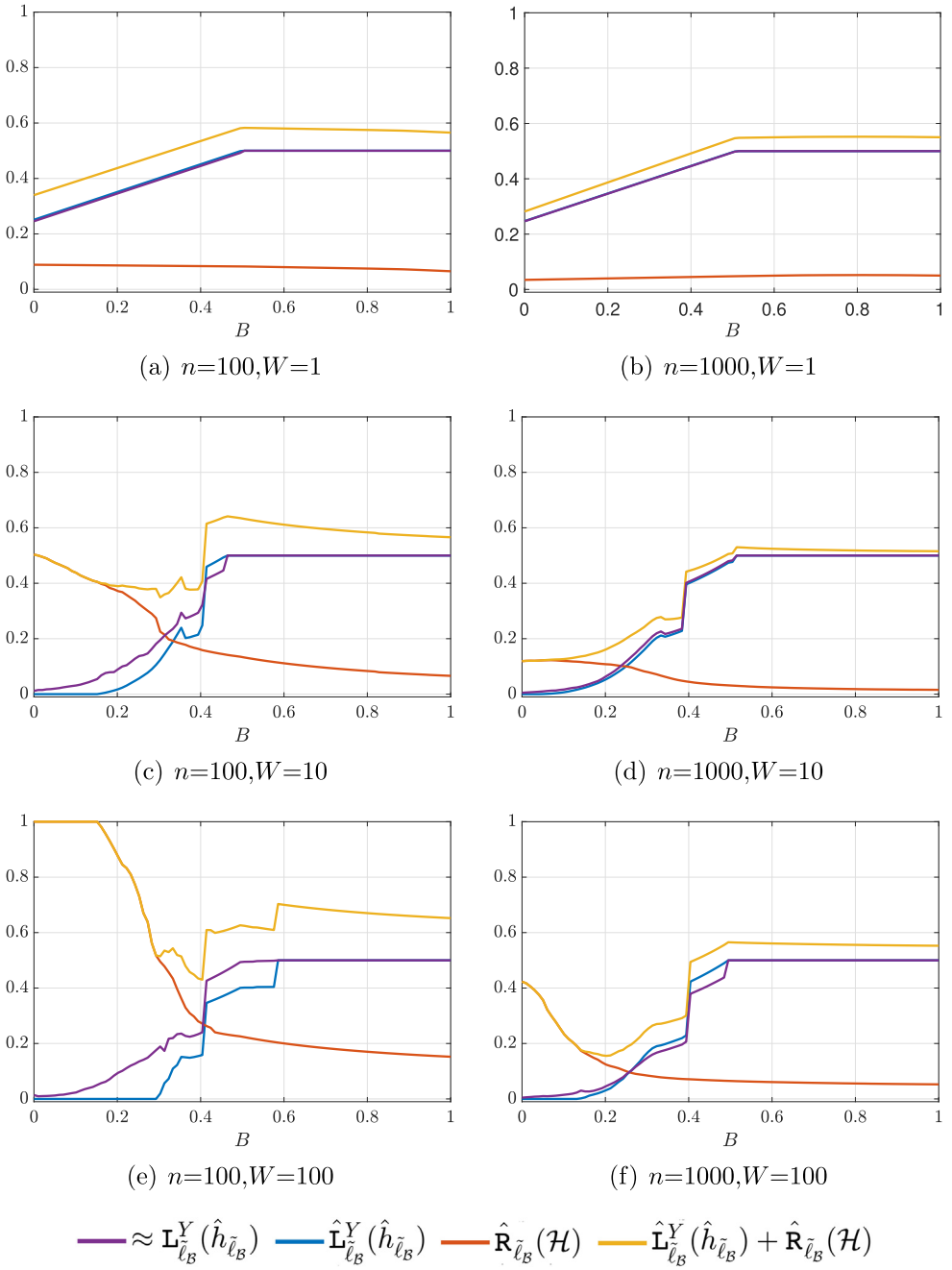


Fig. 5. MNIST-0vs1: we reported the generalization error $\mathtt{L}^Y_{\tilde{\ell}_B}(\hat{h}_{\tilde{\ell}_B})$, approximated with the error on the test set, and the empirical error $\hat{\mathtt{L}}^Y_{\tilde{\ell}_B}(\hat{h}_{\tilde{\ell}_B})$ of the empirical risk minimizer, the (A) GRC $\hat{\mathtt{R}}_{\tilde{\ell}_B}(\mathcal{H})$ and the Empirical Error plus the (A) GRC, namely the bound on the generalization error.

$$\min_{W^+,W^-\in\mathbb{R}^d,\xi\in\mathbb{R}^n} \sum_{i=1}^{n}\xi_i,$$
$$\text{s.t.} \begin{cases} Y_i(W^+ - W^-)\cdot X_i - B\sum_{i=1}^{d}(W_i^+ + W_i^-) \geqslant 1-\xi_i, \\ \qquad\qquad\qquad\qquad\qquad\qquad \forall i \in \{1,\cdots,n\} \\ \sum_{i=1}^{d}(W_i^+ + W_i^-) \leqslant H \\ W_i^+, W_i^- \geqslant 0, \forall i \in \{1,\cdots,d\} \\ \xi_i \geqslant 0, \forall i \in \{1,\cdots,n\} \end{cases}, \tag{65}$$

where $W = W^+ - W^-$. The Linear Programming problems have been solved using the Simplex algorithm [39,40] using the CPLEX.[6] library

Instead, in order to find the complexity term, we will consider in this part only the (A) GRC and the (A) LRC since the (A) VCE and the (A) LVCE cannot be used with the loss function exploited in this section (since it is a $[0,1]$-bounded loss function not a $\{0,1\}$-valued loss function). We could not use the Hard Loss function since optimizing it is computationally prohibitive (NP-Hard problem) and its convex relaxation would result in something similar to the framework we already depicted here. Moreover, remember that, the (A) GRC and (A) LRC are tightly connected to the (A) GVCE and (A) LVCE respectively in the case of the Hard Loss function (see Section 3.2.1 and [33]).

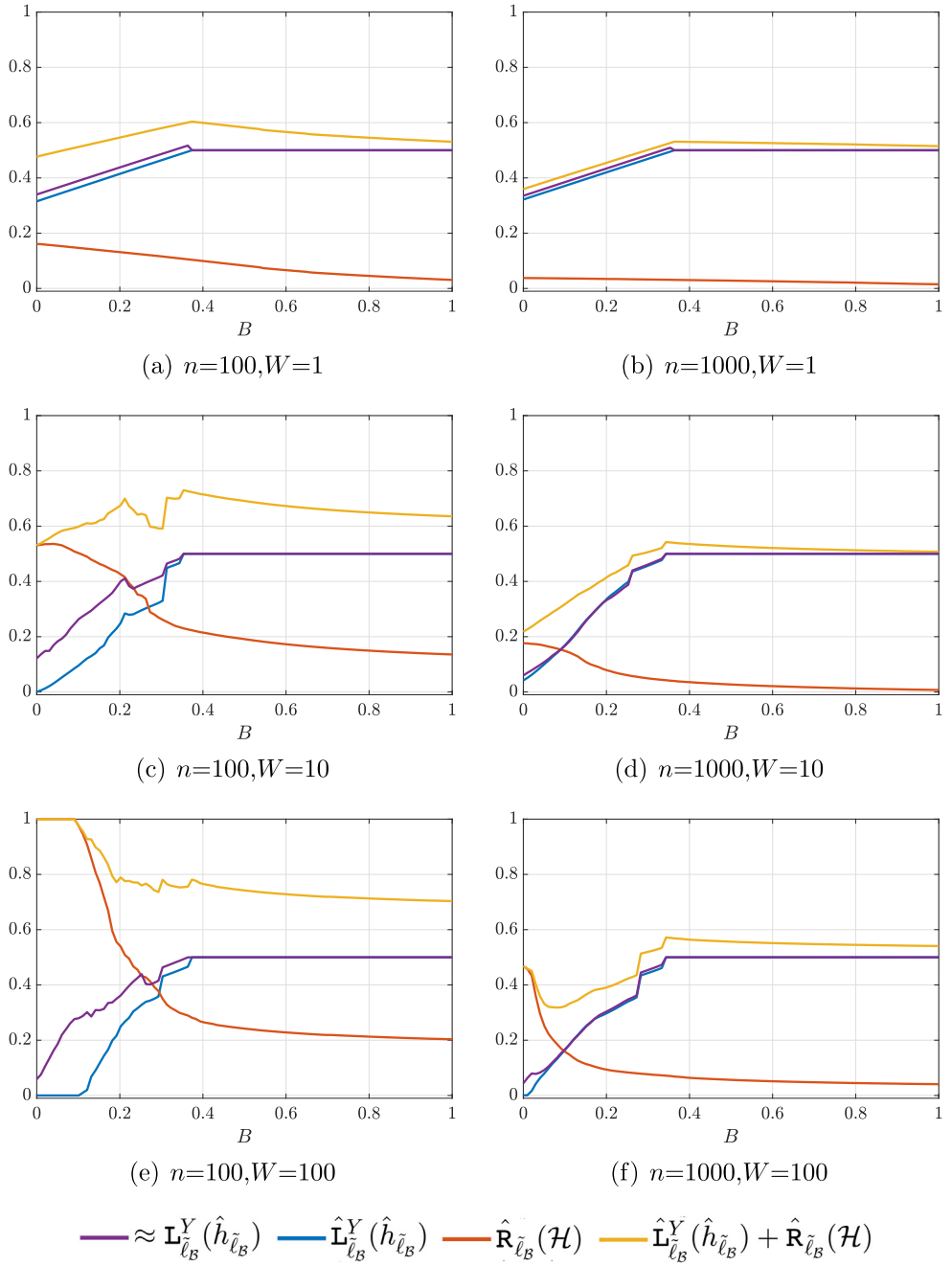


Fig. 6. MNIST-5vs6: we reported the generalization error $\mathtt{L}^{Y}_{\tilde{\ell}_{\mathcal{B}}}(\hat{h}_{\tilde{\ell}_{\mathcal{B}}})$, approximated with the error on the test set, and the empirical error $\hat{\mathtt{L}}^{Y}_{\tilde{\ell}_{\mathcal{B}}}(\hat{h}_{\tilde{\ell}_{\mathcal{B}}})$ of the empirical risk minimizer, the (A) GRC $\hat{\mathtt{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathcal{H})$ and the Empirical Error plus the (A) GRC, namely the bound on the generalization error.

Let us consider then, since the loss is symmetric, the GRC defined in Eq. (36) and the AGRC defined in Eq. (44) are computed remembering that $\hat{\mathtt{R}}_{\ell}(\mathcal{H}) = \hat{\mathtt{R}}_{\tilde{\ell}_{\mathcal{S}}}(\mathcal{H})$. As a consequence, in order to compute the (A) GRC we have to solve the following problem

$$\inf_{h\in\mathcal{H}} \frac{1}{n}\left[\sum_{i\in\mathcal{S}^{+}(\mathcal{S})} \sup_{\tilde{X}\in\mathcal{B}(X_i)} \ell\Big(h,\Big(\tilde{X},Y_i\Big)\Big) + \sum_{i\in\mathcal{S}^{-}(\mathcal{S})} \inf_{\tilde{X}\in\mathcal{B}(X_i)} \ell\Big(h,\Big(\tilde{X},-Y_i\Big)\Big)\right]. \tag{66}$$

Note that for $\mathcal{S} = \{1,\cdots,1\}$ Problem (66) is equivalent to Problem (61). Problem (66) can be formulated as

$$\min_{W:W\in\mathbb{R}^d,\|W\|_1\leq H} \left[\sum_{i\in\mathcal{S}^{+}} \max_{\tilde{X}:\tilde{X}\in\mathbb{R}^d,\|\tilde{X}-X_i\|_\infty\leq B} \min\Big[2,\max\Big[0,1-Y_iW\cdot\tilde{X}\Big]\Big] + \sum_{i\in\mathcal{S}^{-}} \min_{\tilde{X}:\tilde{X}\in\mathbb{R}^d,\|\tilde{X}-X_i\|_\infty\leq B} \min\Big[2,\max\Big[0,1+Y_iW\cdot\tilde{X}\Big]\Big]\right], \tag{67}$$

which, thanks again to the linearity of the problem, can be simplified as follows [17]

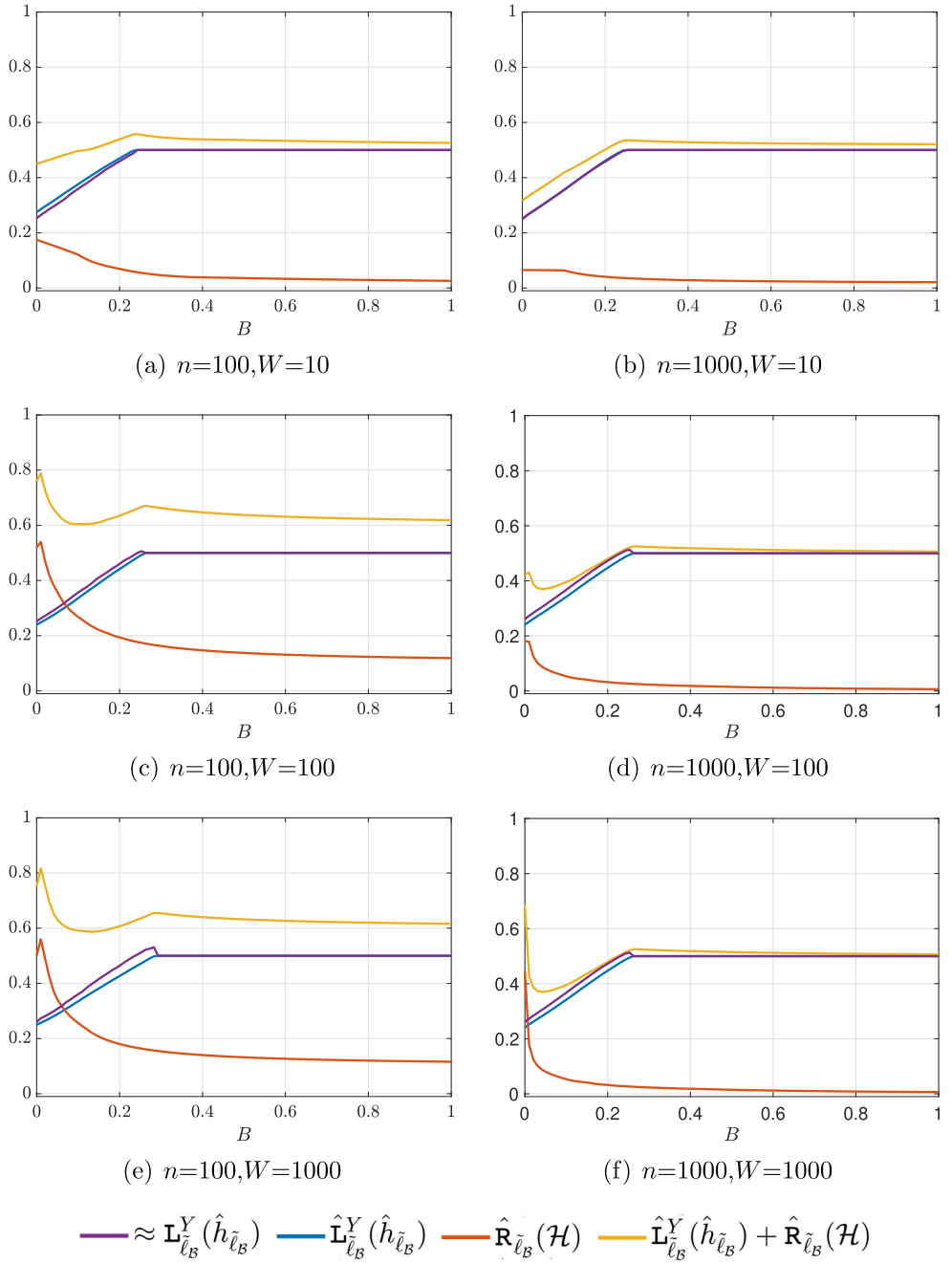


Fig. 7. SVHN-0vs1: we reported the generalization error $\mathtt{L}^Y_{\tilde{\ell}_B}(\hat{h}_{\tilde{\ell}_B})$, approximated with the error on the test set, and the empirical error $\hat{\mathtt{L}}^Y_{\tilde{\ell}_B}(\hat{h}_{\tilde{\ell}_B})$ of the empirical risk minimizer, the (A) GRC $\hat{\mathtt{R}}_{\tilde{\ell}_B}(\mathscr{H})$ and the Empirical Error plus the (A) GRC, namely the bound on the generalization error.

$$\min_{W:W\in\mathbb{R}^d,\|W\|_1\leqslant H} \left[\sum_{i\in\mathscr{S}^+} \min[2,\max[0,1-Y_iW\cdot X_i+\|W\|_1B]] + \sum_{i\in\mathscr{S}^-} \min[2,\max[0,1+Y_iW\cdot X_i-\|W\|_1B]]\right]. \tag{68}$$

$$\min_{W:W\in\mathbb{R}^d,\|W\|_1\leqslant H} \left[\sum_{i\in\mathscr{S}^+} \max[0,1-Y_iW\cdot X_i+\|W\|_1B] + \sum_{i\in\mathscr{S}^-} \max[0,1+Y_iW\cdot X_i-\|W\|_1B]\right], \tag{69}$$

Unfortunately, also Problem (68) is non-convex but we can approximate its solution by easily solving its convex relaxation

which can be rewritten as

$$\min_{W:W\in\mathbb{R}^d,\|W\|_1\leqslant H} \sum_{i=1}^{n} \max[0,1-\sigma_iY_iW\cdot X_i+\sigma_i\|W\|_1B]. \tag{70}$$

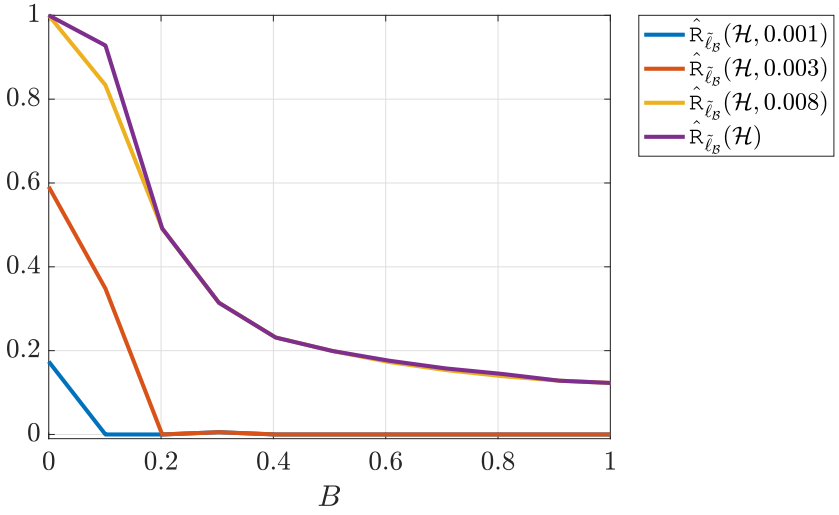
Problem (70) is a Linear Programming problem, in fact Problem (70) can be reformulated as

$$\min_{W^+,W^-\in\mathbb{R}^d,\xi\in\mathbb{R}^n}\sum_{i=1}^{n}\xi_i,$$
$$\text{s.t.}\begin{cases}\sigma_i Y_i(W^+-W^-)\cdot X_i-\sigma_i B\sum_{i=1}^{d}(W_i^++W_i^-)\geqslant 1-\xi_i, & \forall i\in\{1,\cdots,n\}\\ \sum_{i=1}^{d}(W_i^++W_i^-)\leqslant H & \\ W_i^+,W_i^-\geqslant 0,\forall i\in\{1,\cdots,d\} & \\ \xi_i\geqslant 0,\forall i\in\{1,\cdots,n\} & \end{cases}, \quad (71)$$

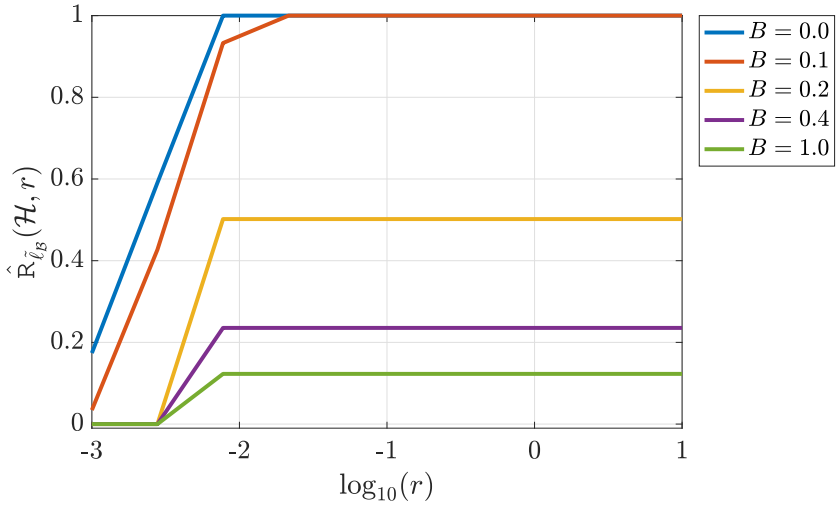
where $W=W^+-W^-$. Note, again, that for $\mathscr{S}=\{1,\cdots,1\}$, namely $\sigma_i=1\ \forall i\in\{1,\cdots,n\}$ Problem (71) is equivalent to Problem (65).

Let us now consider the LRC defined in Eq. (38) and the ALRC defined in Eq. (46). The only differences between the computation of the (A) GRC and the (A) LRC are: (i) we have to find a supremum with respect to $\alpha\in[0,1]$ and (ii) we have to use just functions such that $\frac{1}{n}\sum_{i=1}^{n}\alpha^2\ell^2(h,Z_i)\leqslant r$ remembering that $r\geqslant\min_{h\in\mathscr{H}}\frac{1}{n}\sum_{i=1}^{n}\alpha^2\ell^2(h,Z_i)$. For what concerns (i) we simply perform a brute force search in $\alpha\in 10^{\{-8.00,-7.01,\cdots,0.00\}}$. For what concerns (ii), using the same argument presented for the (A) GRC to obtain Problem (71), we can obtain the following optimization problem that for the (A) LRC

$$\min_{W^+,W^-\in\mathbb{R}^d,\xi,\eta_i\in\mathbb{R}^n}\sum_{i=1}^{n}\xi_i,$$
$$\text{s.t.}\begin{cases}\sigma_i Y_i(W^+-W^-)\cdot X_i-\sigma_i B\sum_{i=1}^{d}(W_i^++W_i^-)\geqslant 1-\xi_i, & \forall i\in\{1,\cdots,n\}\\ Y_i(W^+-W^-)\cdot X_i-B\sum_{i=1}^{d}(W_i^++W_i^-)\geqslant 1-\eta_i, & \forall i\in\{1,\cdots,n\}\\ \sum_{i=1}^{d}(W_i^++W_i^-)\leqslant H & \\ \sum_{i=1}^{d}\eta_i^2\leqslant\frac{nr}{\alpha^2} & \\ W_i^+,W_i^-\geqslant 0,\forall i\in\{1,\cdots,d\} & \\ \xi_i,\eta_i\geqslant 0,\forall i\in\{1,\cdots,n\} & \end{cases}, \quad (72)$$



(a) B in the x-axis, $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathcal{H})$ and $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathcal{H},r)$ on the y-axis and a curve for different values of r.



(b) r in the x-axis, $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathcal{H},r)$ on the y-axis and a curve for different values of B.

Fig. 8. Quantitative analysis of the properties of the (A) GRC and the (A) LRC for the MNIST-0vs1. In particular we reported $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathscr{H})$ and $\hat{\mathrm{R}}_{\tilde{\ell}_{\mathcal{B}}}(\mathscr{H},r)$ for different values of B and r.

which is a convex Quadratically Constrained Linear Programming problem [39,40] that we solved using the CPLEX[6] library.

At this point we are able to compute all the quantities in the GRC based bound of Eq. (17), namely (see Section 3.2.1)

$$\mathsf{L}_{\ell}^{Y}\left(\hat{h}_{\ell}\right) \overset{(1-\delta)}{\leqslant} \hat{\mathsf{L}}_{\ell}^{Y}\left(\hat{h}_{\ell}\right) + \hat{\mathsf{R}}_{\ell}(\mathscr{H}) + 3\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}, \tag{73}$$

and the AGRC based bound of Eq. (27), namely (see Section 3.2.2)

$$\mathsf{L}_{\ell_{\mathscr{B}}}^{Y}\left(\hat{h}_{\ell_{\mathscr{B}}}\right) \overset{(1-\delta)}{\leqslant} \hat{\mathsf{L}}_{\ell_{\mathscr{B}}}^{Y}\left(\hat{h}_{\ell_{\mathscr{B}}}\right) + \hat{\mathsf{R}}_{\ell_{\mathscr{B}}}(\mathscr{H}) + 3\sqrt{\frac{\ln\left(\frac{2}{\delta}\right)}{2n}}. \tag{74}$$

Note that the confidence term is the same and constant in both bounds no matter $\mathscr{B}$ so we can neglect it. In other words we assume that the sample is a good representation of the population so we do not have to pay the associated risk and we have just to pay the risk due to the size of $\mathscr{H}$.

Let us now consider the MNIST dataset [41], a now classical test bench in the adversarial context [6], which consists of 28×28 greyscale (0 white and 1 black) images of numbers from 0 to 9. In particular, we consider the binary classification problems of recognizing 0 against 1 (a simple case named MNIST-0vs1) and 5 against 6 (a more complex one named MNIST-5vs6) exploiting $n = \{100, 1000\}$ samples for train (namely $\{50, 500\}$ from each class) and 10,000 for test (namely 5000 from each class). We will consider also the SVHN dataset [42], which consists of $32 \times 32 \times 3$ colored images of numbers from 0 to 9 taken in natural scene images. In this case, we consider the binary classification problems of recognizing 0 against 1 (SVHN-0vs1) exploiting $n = \{100, 1000\}$ samples for train (namely $\{50, 500\}$ from each class) and 10,000 for test (namely 5000 from each class). Note that SVHN-0vs1 is even more complex than MNIST-5vs6. We exploit 30 random realization of $\mathscr{S}$ to compute both (A) GRC and (A) LRC.

In Figs. 5–7 we reported, for MNIST-0vs1, MNIST-5vs6, and SVHN-0vs1 respectively and for different values of n, W, and B a series of quantities referring to bounds of Eqns. (73) and (74). In particular we reported the generalization error $\mathsf{L}_{\ell_{\mathscr{B}}}^{Y}\left(\hat{h}_{\ell_{\mathscr{B}}}\right)$, approximated with the error on the test set, and the empirical error $\hat{\mathsf{L}}_{\ell_{\mathscr{B}}}^{Y}\left(\hat{h}_{\ell_{\mathscr{B}}}\right)$ of the empirical risk minimizer, the (A) GRC $\hat{\mathsf{R}}_{\ell_{\mathscr{B}}}(\mathscr{H})$ and the Empirical Error plus the (A) GRC, namely the bound on the generalization error. Note that by setting $(B = 0)$ we get the Non-Adversarial Setting while for $(B > 0)$ we get the Adversarial Setting and we tested the trend of these quantities for an increasing size of $\mathscr{B}$.

From the Figs. 5–7 it is possible to see also experimentally the behavior of the empirical error, the complexity, and the generalization bounds the we discussed and expected in the theoretical study performed in Section 3. First let us observe some simple behaviour that we expect knowing the classical theory in both the Non-Adversarial and Adversarial Settings: the larger is n the smaller is the test error, the larger the empirical error, and the smaller the difference between the empirical and test error; the larger is B the larger the empirical and the test error; for W there is an optimal value (not too large not to small) according to the Structural Risk Minimization principle. Then, let us observe the new behaviour. In particular, the complexity can increase with small $\mathscr{B}$ (e.g., Figs. 5(b), 7(c), and 7(e)) while it tends to decrease as $\mathscr{B}$ becomes larger (most of the cases). There is an optimal value of $\mathscr{B}$ (mostly greater than zero) to get the best generalization bound since increasing $\mathscr{B}$ impacts much more the complexity with respect to the empirical error (see Figs. 5(f) and 6(f)). Note also that for this optimal $\mathscr{B}$ the bound on the generalization error is tight, i.e., close to the actual error.

For the sake of completeness we reported in Fig. 8 the counterpart of Fig. 4 for the MNIST-0vs1 dataset. In particular Fig. 8 reports $\hat{\mathsf{R}}_{\ell_{\mathscr{B}}}(\mathscr{H})$ and $\hat{\mathsf{R}}_{\ell_{\mathscr{B}}}(\mathscr{H}, r)$ for different values of B and r. As expected from the theory [24] the (A) LRC is able shrink the (A) GRC.

5. Conclusions

Recent research has shown that models induced by machine learning, and in particular by deep learning, can be easily fooled by an adversary who carefully crafts imperceptible, at least from the human perspective, or physically plausible modifications of the input data. This discovery gave birth to a new field of research, the adversarial machine learning, where new methods of attacks and defense are developed continuously, mimicking what is happening from a long time in cybersecurity.

In this context the scope of the paper was to shift the attention and show that inducing models from data less prone to be fooled by an adversary, while posing many unresolved theoretical (e.g., finding the best perturbation set [31]) and practical (e.g., solving the non-convex optimization problem behind adversarial defense [6]) challenges, actually provides some benefits when it comes to assess their generalization abilities, namely bound their performance on previously unseen samples. For this purpose we first use a theoretical approach, relying on Statistical Learning Theory, exploiting, studying, and extending the (Local) Vapnik–Chervonenkis and (Local) Rademacher Complexity Theories to the Adversarial Setting. We enrich the theoretical discussion with examples and results that focus on giving more insights to the readers and translate the theory into practical concepts. Then we switch from theory to practice with a series of numerical experiments on real data.

More specifically, the proposed generalization bounds for the Adversarial Setting based on the (Local) Vapnik–Chervonenkis and on the Local Rademacher Complexity are novel while the ones based on the Rademacher Complexity have already been studied. Then, we performed a new study on the connection between the (Local) Vapnik–Chervonenkis and on the (Local) Rademacher Complexity in the Adversarial Setting. Finally, theoretical and practical analysis of the behaviour of the (Local) Vapnik–Chervonenkis and the (Local) Rademacher Complexity based bound in the Adversarial Setting when the perturbation domain changes in size has been performed. This study shed new light on a previously unknown phenomenon: increasing the size of the perturbation domain can decrease the complexity of the space of functions and can increase the tightness of the generalization error bounds. In fact, sometimes it exists a perturbation large enough to not increase the empirical error too much while remarkably decreasing the complexity resulting in sharper bound on generalization error of models learned in the Adversarial Setting.

Both theoretical and practical results support the idea that we raise in this paper that dealing with an adversary can actually produce a benefit when it comes to bounding a performance of a model on previously unseen samples.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Cireşan, U. Meier, J. Masci, J. Schmidhuber, A committee of neural networks for traffic sign classification, in: International joint conference on neural networks, 2011.

[2] A. Hekler, J.S. Utikal, A.H. Enk, W. Solass, Others, Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images, Eur. J. Cancer 118 (2019) 91–96.
[3] D. Silver, J. Schrittwieser, K. Simonyan, et al., Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354–359.
[4] J. Jumper, R. Evans, A. Pritzel, T. Green, Others, Highly accurate protein structure prediction with alphafold, Nature 596 (7873) (2021) 583–589.
[5] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, O. Evans, Viewpoint: When will AI exceed human performance? evidence from AI experts, J. Artif. Intell. Res. 62 (2018) 729–754.
[6] B. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Pattern Recogn. 84 (2018) 317–331.
[7] R. Duan, X. Ma, Y. Wang, J. Bailey, A.K. Qin, Y. Yang, Adversarial camouflage: Hiding physical-world attacks with natural styles, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[8] Z. Wu, S.N. Lim, L.S. Davis, T. Goldstein, Making an invisibility cloak: Real world adversarial attacks on object detectors, European Conference on Computer Vision (2020).
[9] S. Komkov, A. Petiushko, Advhat: Real-world adversarial attack on arcface face id system, in: International Conference on Pattern Recognition, 2021.
[10] L. Chen, J. Li, J. Peng, T. Xie, Z. Cao, K. Xu, X. He, Z. Zheng, A survey of adversarial learning on graphs, arXiv preprint arXiv:2003.05730.
[11] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, Adversarial attacks and defences: A survey, arXiv preprint arXiv:1810.00069.
[12] N. Akhtar, A. Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, IEEE Access 6 (2018) 14410–14430.
[13] G.R. Machado, E. Silva, R.R. Goldschmidt, Adversarial machine learning in image classification: A survey toward the defender's perspective, ACM Comput. Surv. 55 (1) (2021) 1–38.
[14] C. Zhang, P. Benz, C. Lin, A. Karjauv, J. Wu, I.S. Kweon, A survey on universal adversarial attack, arXiv preprint arXiv:2103.01498.
[15] J. Khim, P.L. Loh, Adversarial risk bounds via function transformation, arXiv preprint arXiv:1810.09519.
[16] P. Viallard, E.G. Vidot, A. Habrard, E. Morvant, A pac-bayes analysis of adversarial robustness, Neural Information Processing Systems, 2021.
[17] D. Yin, R. Kannan, P. Bartlett, Rademacher complexity for adversarially robust generalization, in: International Conference on Machine Learning, 2019.
[18] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[19] L. Oneto, D. Anguita, S. Ridella, A local vapnik-chervonenkis complexity, Neural Networks 82 (2016) 62–75.
[20] V. Koltchinskii, Rademacher penalties and structural risk minimization, IEEE Trans. Inf. Theory 47 (5) (2001) 1902–1914.
[21] P.L. Bartlett, S. Mendelson, Rademacher and gaussian complexities: Risk bounds and structural results, J. Mach. Learn. Res. 3 (2002) 463–482.
[22] P.L. Bartlett, O. Bousquet, S. Mendelson, Local rademacher complexities, Ann. Stat. 33 (4) (2005) 1497–1537.
[23] V. Koltchinskii, Local rademacher complexities and oracle inequalities in risk minimization, Ann. Stat. 34 (6) (2006) 2593–2656.
[24] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Local rademacher complexity: Sharper risk bounds with and without unlabeled samples, Neural Networks 65 (2015) 115–125.
[25] L. Oneto, Model Selection and Error Estimation in a Nutshell, Springer, 2020.
[26] M. Arjovsky, L. Bottou, I. Gulrajani, D. Lopez-Paz, Invariant risk minimization, arXiv preprint arXiv:1907.02893.
[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: International Conference on Learning Representations, 2018.
[28] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, A. Kurakin, On evaluating adversarial robustness, arXiv preprint arXiv:1902.06705.
[29] A. Shaeiri, R. Nobahari, M.H. Rohban, Towards deep learning models resistant to large perturbations, arXiv preprint arXiv:2003.13370.
[30] A. Ilyas, S. Santurkar, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, Neural information processing systems.
[31] E. Wong, J.Z. Kolter, Learning perturbation sets for robust machine learning, in: International Conference on Learning Representations, 2021.
[32] R. Collobert, F. Sinz, J. Weston, L. Bottou, Trading convexity for scalability, in: International conference on Machine learning, 2006.
[33] D. Anguita, A. Ghio, L. Oneto, S. Ridella, A deep connection between the vapnik-chervonenkis entropy and the rademacher complexity, IEEE Trans. Neural Networks Learn. Syst. 25 (12) (2014) 2202–2211.
[34] L. Oneto, A. Ghio, S. Ridella, D. Anguita, Global rademacher complexity bounds: From slow to fast convergence rates, Neural Process. Lett. 43 (2) (2015) 567–602.
[35] P. Klesk, M. Korzen, Sets of approximating functions with finite vapnik-chervonenkis dimension for nearest-neighbors algorithms, Pattern Recogn. Lett. 32 (14) (2011) 1882–1893.
[36] C. Clason, Regularization of inverse problems, arXiv preprint arXiv:2001.00617.
[37] M. Khoury, D. Hadfield-Menell, On the geometry of adversarial examples, arXiv preprint arXiv:1811.00525.
[38] A. Demontis, P. Russu, B. Biggio, G. Fumera, F. Roli, On security and sparsity of linear classifiers for adversarial settings, in: International Workshops on Statistical Techniques in Pattern Recognition and Structural and Syntactic Pattern Recognition, 2016.
[39] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge University Press, 2004.
[40] R.J. Vanderbei, Linear Programming: Foundations and Extensions, Springer, New York, 2014.
[41] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[42] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A.Y. Ng, Reading digits in natural images with unsupervised feature learning, in: Neural Information Processing Systems – Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

Luca Oneto was born in Rapallo, Italy in 1986. He received his BSc and MSc in Electronic Engineering at the University of Genoa, Italy respectively in 2008 and 2010. In 2014 he received his PhD from the same university in the School of Sciences and Technologies for Knowledge and Information Retrieval with the thesis "Learning Based On Empirical Data". In 2017 he obtained the Italian National Scientific Qualification for the role of Associate Professor in Computer Engineering and in 2018 he obtained the one in Computer Science. He worked as Assistant Professor in Computer Engineering at University of Genoa from 2016 to 2019. In 2018 he was co-funder of the spin-off ZenaByte s.r.l. In 2019 he obtained the Italian National Scientific Qualification for the role of Full Professor in Computer Science and Computer Engineering. In 2019 he became Associate Professor in Computer Science at University of Pisa and currently is Associate Professor in Computer Engineering at University of Genoa. He has been involved in several H2020 projects (S2RJU, ICT, DS) and he has been awarded with the Amazon AWS Machine Learning and Somalvico (best Italian young AI researcher) Awards. His first main topic of research is the Statistical Learning Theory with particular focus on the theoretical aspects of the problems of (Semi) Supervised Model Selection and Error Estimation. His second main topic of research is Data Science with particular reference to the problem of Trustworthy AI and the solution of real world problems by exploiting and improving the most recent Learning Algorithms and Theoretical Results in the fields of Machine Learning and Data Mining.

Sandro Ridella received the 'Laurea' degree in electronic engineering from the University of Genoa, Genoa, Italy, in 1966. Currently, he is a Full Professor at the Department of Biophysical and Electronic Engineering (DIBE, now DITEN Dept.), University of Genoa, where he teaches circuits and algorithms for signal processing. In the last five years, his scientific activity has been mainly focused on the field of neural networks.

Davide Anguita received the 'Laurea' degree in Electronic Engineering and a Ph.D. degree in Computer Science and Electronic Engineering from the University of Genoa, Genoa, Italy, in 1989 and 1993, respectively. After working as a Research Associate at the International Computer Science Institute, Berkeley, CA, on special-purpose processors for neurocomputing, he returned to the University of Genoa. He is currently Associate Professor of Computer Engineering with the Department of Informatics, BioEngineering, Robotics, and Systems Engineering (DIBRIS). His current research focuses on the theory and application of kernel methods and artificial neural networks.