The Benefits of Adversarial Defence in Generalisation

Luca Oneto, Sandro Ridella, Davide Anguita

University of Genoa - Via Opera Pia 11a, 16145, Genova, Italy

Abstract. Recent researches have been shown that models induced by machine learning, 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 defence 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 actually provides some benefits when it comes to assess their generalisation abilities.

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 of the vast amount of data collected. In some tasks (e.g., vision and games) recent deep-learning algorithms have shown super-human performance [1, 2]. 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 [3, 4]. 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 [3]. In fact new methods of attacks (i.e., adversarial attacks) and defence (i.e., adversarial defense) are developed continuously, 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 paper 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 generalisation capabilities of a model learned in the classical setting, where no adversary is present, against the ones of a model designed to be less prone to attacks and then less exposed to adversaries. Exploiting the Global Rademacher Complexity (GRC) theory and Local Rademacher Complexity (LRC) theories [5, 6] we will show that the introduction of a mechanism of defence in the learning phase of a model actually improve our ability to accurately estimate its generalisation performance (i.e., the tightness of the generalisation bound). In particular we will recall first some background notions in Section 2, then we will study the problem and derive some results

leveraging both on the Statistical Learning Theory in Section 3 and both on experimental result on real data in Section 4 concluding the paper in Section 5.

2 Background

Let us consider the binary classification problem[1] [7] under evasion attach [3]. Based on a random observation of $\boldsymbol{x} \in \mathcal{X} \in \mathbb{R}^d$ one has to estimate $y \in \mathcal{Y} \subseteq \{\pm 1\}$ by choosing a suitable hypothesis $h : \mathcal{X} \to \hat{\mathcal{Y}}$ in a set of possible ones $\mathcal{H}$. Note that choosing the right $\mathcal{H}$ is the so-called model selection problem [8] 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 $\boldsymbol{x}$ according to a set of possible modifications $\mathcal{B}(\boldsymbol{x})$, namely

$$\tilde{\boldsymbol{x}}^* : \arg\sup_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x})} [h(\tilde{\boldsymbol{x}}) \neq h(\boldsymbol{x})], \tag{1}$$

where the Iverson bracket notation is exploited. A learning algorithm selects $h \in \mathcal{H}$ by exploiting a set of n labelled samples $\mathcal{D}_n^y : \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$. $\mathcal{D}_n^y$ consists of a sequence of independent samples distributed according to μ over $\mathcal{X} \times \mathcal{Y}$. The generalisation error (i.e., the risk) $\mathsf{L}^y(h) = \mathbb{E}_{(\boldsymbol{x},y)} \ell(h(\boldsymbol{x}), y)$ associated to an hypothesis $h \in \mathcal{H}$, is defined through a loss function $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to [0, 1]$. As μ is unknown, $\mathsf{L}^y(h)$ cannot be explicitly computed, but we can compute the empirical error (i.e., the empirical risk) namely the empirical estimator of the generalisation error $\hat{\mathsf{L}}^y(h) = \frac{1}{n} \sum_{(\boldsymbol{x},y) \in \mathcal{D}_n^y} \ell\left(h\left(\boldsymbol{x}\right), y\right)$. The purpose of any learning procedure is to find the minimizer h^* of the generalisation error $\mathsf{L}^y(h)$ ($h^* = \arg\min_{h \in \mathcal{H}} \mathsf{L}^y(h)$) but since $\mathsf{L}^y(h)$ is unknown we have to estimate h^* exploiting an empirical estimator $\hat{h}$ which is the empirical risk minimizer ($\hat{h} = \arg\min_{h \in \mathcal{H}} \hat{\mathsf{L}}^y(h)$). $\hat{h}$ is effective when $\mathcal{H}$ is carefully tuned [8]. 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 Defence [3]. The idea is that the attack of Eq. (1) can be reformulated as $\tilde{\boldsymbol{x}}^* : \arg\sup_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x})} \ell(h(\tilde{\boldsymbol{x}}), y)$, and then we can consider the now-classical problem of Adversarial Defence [3] $\tilde{h}^* : \arg\inf_{h \in \mathcal{F}} \tilde{\mathsf{L}}^y(h)$ where $\tilde{\mathsf{L}}^y(h) = \mathbb{E}_{(\boldsymbol{x},y)} \sup_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x})} \ell(h(\tilde{\boldsymbol{x}}), y)$ and its empirical estimator

$$\hat{\tilde{h}} : \arg\inf_{h \in \mathcal{F}} \hat{\tilde{\mathsf{L}}}^y(h), \tag{2}$$

where $\hat{\tilde{\mathsf{L}}}^y(h) = \frac{1}{n} \sum_{(\boldsymbol{x},y) \in \mathcal{D}_n} \sup_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x})} \ell(h(\tilde{\boldsymbol{x}}), y)$. Note that when $\mathcal{B}(\boldsymbol{x}) = \boldsymbol{x}$ we have that $\tilde{\mathsf{L}}^y(h) = \mathsf{L}^y(h)$, $\hat{\tilde{\mathsf{L}}}^y(h) = \hat{\mathsf{L}}^y(h)$, $\tilde{h}^* = h^*$, and $\hat{\tilde{h}} = \hat{h}$

3 Statistical Learning Theory

The main topic that we want to investigate in this work is how to estimate the risk of $\hat{h}$ and $\hat{\tilde{h}}$ showing that there is a benefit in assessing the generalisation

[1] Everything we will present can be easily generalised to the whole supervised learning framework but, for simplicity and clarity of the notation, we will restrict the presentation to the binary classification framework.

performance of $\hat{\tilde{h}}$ with respect to $\hat{h}$. Let us consider the case when $\hat{\mathcal{Y}} = [-1, 1]$ and a symmetric loss function is exploited (i.e., $\ell(\cdot, -y) = 1 - \ell(\cdot, y)$). In this setting it is possible to prove the GRC-based bound on the generalisation ability of $\hat{h}$ [5, 9]

$$\mathbb{P}\{\mathsf{L}^y(\hat{h}) \leq \hat{\mathsf{L}}^y(\hat{h}) + \mathsf{R}(\mathcal{H}) + \phi_1(\delta)\} \geq 1 - \delta, \tag{3}$$

where [6] $\mathsf{R}(\mathcal{H}) = \sup_{h \in \mathcal{H}} \frac{2}{n} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_n^y} \sigma_i \ell(h(\boldsymbol{x}), y)$ or, equivalently, $\mathsf{R}(\mathcal{H}) = 1 - 2 \inf_{h \in \mathcal{H}} \hat{\mathsf{L}}^\sigma(h)$, is the GRC and where $\mathcal{D}_n^\sigma = \{(\boldsymbol{x}_1, \sigma_1), \cdots, (\boldsymbol{x}_n, \sigma_n)\}$, $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = 1/2$, $\hat{\mathsf{L}}^\sigma(h) = \frac{1}{n} \sum_{(\boldsymbol{x}, \sigma) \in \mathcal{D}_n^\sigma} \ell\left(h\left(\boldsymbol{x}\right), \sigma\right)$, and $\phi_1(\delta)$ is the confidence term [5]. $\mathsf{R}(\mathcal{H})$ can be estimated with a single extraction of $\boldsymbol{\sigma}$, with multiple realisation of $\boldsymbol{\sigma}$, or by computing $\mathbb{E}_{\boldsymbol{\sigma}}$ [6].

Under the same setting it is possible to prove the LRC-based bound on the generalisation ability of $\hat{h}$ [10]

$$\mathbb{P}\left\{\mathsf{L}^y(\hat{h}) \leq \hat{\mathsf{L}}^y(\hat{h}) + \mathsf{R}\left(\left\{h \middle| h \in \mathcal{H}, \hat{\mathsf{L}}^y(h) \leq \hat{\mathsf{L}}^y(\hat{h}) + \phi_2(\delta)\right\}\right) + \phi_3(\delta)\right\} \geq 1 - \delta, \tag{4}$$

where $\phi_2(\delta)$ and $\phi_3(\delta)$ are constant confidence terms that can be computed from the data.

It is then possible, to also bound the generalisation ability of $\hat{\tilde{h}}$ via GRC-based bound [11]

$$\mathbb{P}\left\{\tilde{\mathsf{L}}^y(\hat{\tilde{h}}) \leq \hat{\tilde{\mathsf{L}}}^y(\hat{\tilde{h}}) + \tilde{\mathsf{R}}(\mathcal{H}) + \phi_1(\delta)\right\} \geq 1 - \delta, \tag{5}$$

where $\tilde{\mathsf{R}}(\mathcal{H}) = \sup_{h \in \mathcal{H}} \frac{2}{n} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_n} \sigma_i \sup_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x})} \ell(h(\tilde{\boldsymbol{x}}), y)$ or, equivalently, $\tilde{\mathsf{R}}(\mathcal{H}) = 1 - 2 \inf_{h \in \mathcal{H}} \hat{\tilde{\mathsf{L}}}^\sigma(h)$ where $\hat{\tilde{\mathsf{L}}}^\sigma(h) = \frac{1}{n} \sum_{(\boldsymbol{x}, \sigma) \in \mathcal{D}_n^\sigma} \sup_{\tilde{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{x})} \ell(h(\tilde{\boldsymbol{x}}), \sigma)$. Moreover, with a a slightly more complex and technical proof that we do not report because of space constraints, it is possible to prove the LRC-based bound on the generalisation ability of $\hat{\tilde{h}}$

$$\mathbb{P}\left\{\tilde{\mathsf{L}}^y(\hat{h}) \leq \hat{\tilde{\mathsf{L}}}^y(\hat{h}) + \tilde{\mathsf{R}}\left(\left\{h \middle| h \in \mathcal{H}, \hat{\tilde{\mathsf{L}}}^y(h) \leq \hat{\tilde{\mathsf{L}}}^y(\hat{h}) + \phi_2(\delta)\right\}\right) + \phi_3(\delta)\right\} \geq 1 - \delta. \tag{6}$$

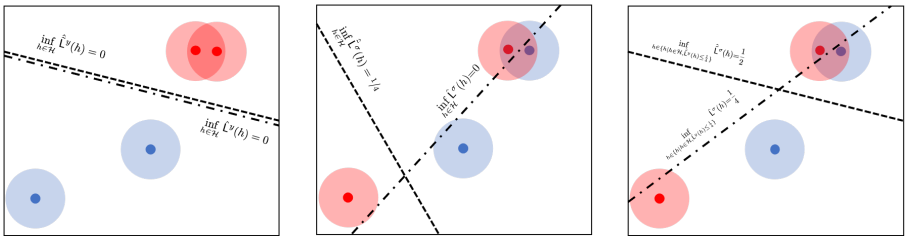
Note that, it is possible to easily observe and prove that [6]

$$\mathsf{R}\Big(\Big\{h \Big| h \in \mathcal{H}, \hat{\mathsf{L}}^y(h) \leq \hat{\mathsf{L}}^y(\hat{h}) + \phi_2(\delta)\Big\}\Big) \leq \mathsf{R}(\mathcal{H}), \quad \tilde{\mathsf{R}}\Big(\Big\{h \Big| h \in \mathcal{H}, \hat{\tilde{\mathsf{L}}}^y(h) \leq \hat{\tilde{\mathsf{L}}}^y(\hat{h}) + \phi_2(\delta)\Big\}\Big) \leq \tilde{\mathsf{R}}(\mathcal{H}), \tag{7}$$

but, with a a slightly more complex and technical proof that we do not report because of space constraints since it follows from [11], also that

$$\tilde{\mathsf{R}}(\mathcal{H}) \leq \mathsf{R}(\mathcal{H}), \quad \tilde{\mathsf{R}}\Big(\Big\{h \Big| h \in \mathcal{H}, \hat{\tilde{\mathsf{L}}}^y(h) \leq \hat{\tilde{\mathsf{L}}}^y(\hat{h}) + \phi_2(\delta)\Big\}\Big) \leq \mathsf{R}\Big(\Big\{h \Big| h \in \mathcal{H}, \hat{\mathsf{L}}^y(h) \leq \hat{\mathsf{L}}^y(\hat{h}) + \phi_2(\delta)\Big\}\Big). \tag{8}$$

The interpretation of Inequalities (7) and (8) are well exemplified in Figure 1 where, for a toy example, it is represented the same problem, with and without an adversary, and it is reported the empirical minimizer, the GRC, and the LRC. Basically when one takes into account the whole space of attach around



(a) True Labels: Empirical Error is zero with or without the adversary.

(b) Random Labels: GRC is 1 without the adversary and $1/2$ with the adversary.

(c) Random Labels: LRC is $1/2$ without the adversary and 0 with the adversary.

Fig. 1: Toy Example: blue and red points are respectively negatively and positively labelled samples, lines are the minimisers in the space of linear models, and circles around samples are the spaces of attach.

the labelled sample, and not just the labelled sample, this reduces the ability of $\mathcal{H}$ to fit random labels (i.e., it reduces the complexity of $\mathcal{H}$ when measured with the GRC or even more when measured with the LRC) much more than the ability to learn good models (i.e., to fit meaningful labels).

What is not easy to derive and is not yet been proved, apart the naive Inequalities (7) and (8) and the ones presented in [11], is how much, quantitatively, can be the benefit in generalisation of the introduction of the adversary. This analysis will be the subject of Section 4.

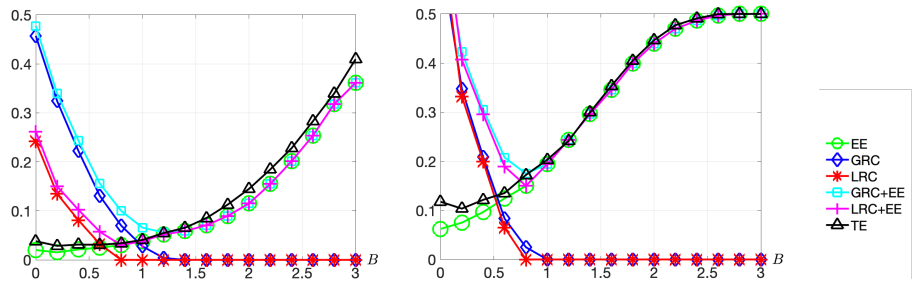
Note, finally, that all the bounds presented above can be generalised as follows, $\forall \dot{h} \in \mathcal{H}$

$$\mathbb{P}\left\{\dot{\mathsf{L}}^{y}(h_1) \leq \hat{\dot{\mathsf{L}}}^{y}(h_1) + \left[1 - 2\inf_{h_2 \in \left\{h_2 \middle| h_2 \in \mathcal{H}, \hat{\dot{\mathsf{L}}}(h_2) \leq \Delta_1(\delta)\right\}} \hat{\dot{\mathsf{L}}}^{\sigma}(h)\right] + \Delta_2(\delta)\right\} \geq 1 - \delta, \quad (9)$$

namely the generalisation error of a function is bounded by the empirical error, plus a complexity term which measures the size of the space (GRC or LRC), plus a confidence term. In fact, if, for example, we set $\dot{\mathsf{L}} = \mathsf{L}$, $\dot{h} = \hat{h}$, $\Delta_1(\delta) = 1$, and $\Delta_2(\delta) = \phi_1(\delta)$ we get the GRC-based bound on the generalisation ability of $\hat{h}$. If, instead, we set $\dot{\mathsf{L}} = \tilde{\mathsf{L}}$, $\dot{h} = \hat{\tilde{h}}$, $\Delta_1(\delta) = \hat{\tilde{\mathsf{L}}}^{y}(\hat{h}) + \phi_2(\delta)$, and $\Delta_2(\delta) = \phi_3(\delta)$ we get the LRC-based bound on the generalisation ability of $\hat{\tilde{h}}$.

4 From Theory to Practice

Let us consider the case when $h(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x}$ where $\boldsymbol{w} \in \mathbb{R}^d$ and the size of $\mathcal{H}$ is regulated by the p-norm of the model weights $\|\boldsymbol{w}\|_p \leq W$ where p regulates the sparsity of the solution [12]. Let us also consider the case where $\mathcal{B}(\boldsymbol{x})$ is the subset of $\mathbb{R}^d$ such that $\mathcal{B}(\boldsymbol{x}) = \{\tilde{\boldsymbol{x}} \mid \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_p \leq B\}$. In this case the value of p regulates the sparsity of the attack [13]. Note also that there is a relation between sparsity of the regularizer and robustness to attacks [14]. For simplicity, in this section, we will set $p = 2$ in both regularisation and attack. Since we are dealing with binary classification problems, the Hinge loss $\ell(h(\boldsymbol{x}), y) = \max[0, 1 - yh(\boldsymbol{x})]$ will be exploited [15]. Unfortunately, the Hinge loss is not bounded and consequently the truncated Hinge loss will be exploited $\ell(h(\boldsymbol{x}), y) = \max[0, \min[1, \frac{1-yh(\boldsymbol{x})}{2}]] \doteq \lceil \lfloor \frac{1-yh(\boldsymbol{x})}{2} \rfloor_0 \rceil^1$.



(a) 0 vs 1 (b) 5 vs 6

Fig. 2: Results on the Mnist dataset varying $B \in [0,3]$ (0 the case when no adversary is present): EE, GRC, LRC, EE plus GRC, EE plus LRC, and TE.

In this setting, in order to find the $\hat{h}$ (or $\hat{\hat{h}}$), and to estimate its generalisation abilities with the GRC or the LRC-based bounds according to Eq. (9) we have to solve the following problem

$$\min_{\substack{\boldsymbol{w}:\ \|\boldsymbol{w}\|\leq W,\\ \sum_{i=1}^{n}\max_{\tilde{\boldsymbol{x}}:\|\tilde{\boldsymbol{x}}-\boldsymbol{x}_i\|<B}\lceil\lfloor\frac{1-y_i\boldsymbol{w}\cdot\tilde{\boldsymbol{x}}}{2}\rfloor_0\rceil^1\leq r}} \sum_{i=1}^{n}\max_{\substack{\tilde{\boldsymbol{x}}:\\ \|\tilde{\boldsymbol{x}}-\boldsymbol{x}_i\|<B}}\lceil\lfloor\tfrac{1-c_i\boldsymbol{w}\cdot\tilde{\boldsymbol{x}}}{2}\rfloor_0\rceil^1, \tag{10}$$

where $c_i = y_i$, $r = \inf$, and $B = 0$ for finding $\hat{h}$ (or $B > 0$ for $\hat{\hat{h}}$), $c_i = \sigma_i$, $r = \inf$, and $B = 0$ for computing the GRC (or $B > 0$ for the GRC with Adversarial Defence), and $c_i = \sigma_i$, $r \leq n$, and $B = 0$ for finding the LRC (or $B > 0$ for the LRC with Adversarial Defence). Problem (10) can be reformulated as follows

$$\begin{aligned}&\min_{\boldsymbol{w}}\max_{\boldsymbol{\eta}_1,\cdots,\boldsymbol{\eta}_n,\boldsymbol{\gamma}_1,\cdots,\boldsymbol{\gamma}_n}\sum_{i=1}^{n}\left\lfloor\left\lceil\tfrac{1-c_i\boldsymbol{w}\cdot(\boldsymbol{x}_i+\boldsymbol{\eta}_i)}{2}\right\rfloor_0\right\rceil^1 \\ &\text{s.t.}\sum_{i=1}^{n}\left\lfloor\left\lceil\tfrac{1-y_i\boldsymbol{w}\cdot(\boldsymbol{x}_i+\boldsymbol{\gamma}_i)}{2}\right\rfloor_0\right\rceil^1\leq r,\ \|\boldsymbol{w}\|\leq W,\ \|\boldsymbol{\eta}_i\|\leq B,\ \|\boldsymbol{\gamma}_i\|\leq B,\ \forall i\in\{1,\cdots,n\}\end{aligned} \tag{11}$$

which can be solved via alternating minimisation but, because the linearity of the model, Problem (11) is equivalent to

$$\min_{\boldsymbol{w}}\sum_{i=1}^{n}\left\lfloor\left\lceil\tfrac{1+\|\boldsymbol{w}\|B-c_i\boldsymbol{w}\cdot\boldsymbol{x}_i}{2}\right\rfloor_0\right\rceil^1,\ \text{s.t.}\sum_{i=1}^{n}\left\lfloor\left\lceil\tfrac{1+\|\boldsymbol{w}\|B-y_i\boldsymbol{w}\cdot\boldsymbol{x}_i}{2}\right\rfloor_0\right\rceil^1\leq r,\ \|\boldsymbol{w}\|\leq W \tag{12}$$

Using instead the kernel trick [16], namely working with linear models in a reproducing kernel Hilbert space, the problem would become much more complex to solve [3], but this is out of the scope of this paper.

Problem (12) allows to understand, in more realistic case, the meaning of the Inequalities (7) and (8) and what has been reported in Figure 1 for a toy sample. For this reason let us consider the Mnist dataset [17], a now classical test bench in the adversarial context [3], 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 recognising 0 against 1 (a simple case) and 5 against 6 (a more complex one) exploiting 100 sample from each class from train and 1000 for test. Each experiment has been repeated 30 times to ensure statistical validity of the results.

In Figure 2 reports, setting $W = 1$ and varying the amplitude of the space of attach B, the Empirical Error (EE), the GRC, the LRC, the EE plus the GRC,

the EE plus the LRC, and Test Error (TE). Note that EE plus the GRC, the EE plus the LRC are basically the GRC- and LRC-based generalisation bounds of Eq. (9) when the confidence terms (constants) are disregarded. Figure 2 shows, analogously to Figure 1, that increasing B makes GRC (and even more LRC) decrease, while less increasing the EE. This means that there is a trade-off, namely an optimal size B, which allows to improve our ability to estimate the final performance of the model (i.e., the tightness of the bound).

5 Conclusions

The scope of this paper was to show that inducing models from data less prone to be fooled by an adversary, while posing many unresolved challenges, actually provides some benefits when it comes to assess their generalisation abilities. In particular, we studied the problem first from a theoretical perspective deriving some results leveraging both on the Statistical Learning Theory and then practically with a series of numerical experiments. The results presented in the paper are surely promising but require deeper theoretical and experimental analysis since they open a quite new perspective in the field of adversarial machine learning that deserve to be further investigated.

References

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

[2] D. Silver, J. Schrittwieser, K. Simonyan, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[3] B. Biggio and F. Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[4] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang. Adversarial camouflage: Hiding physical-world attacks with natural styles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[5] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

[6] L. Oneto, S. Ridella, and D. Anguita. Local rademacher complexity machine. *Neurocomputing*, 342:24–32, 2019.

[7] V. N. Vapnik. *Statistical Learning Theory*. Wiley New York, 1998.

[8] L. Oneto. *Model Selection and Error Estimation in a Nutshell*. Springer, 2020.

[9] L. Oneto, D. Anguita, and S. Ridella. A local vapnik-chervonenkis complexity. *Neural Networks*, 82:62–75, 2016.

[10] L. Oneto, A. Ghio, S. Ridella, and D. Anguita. Local rademacher complexity: Sharper risk bounds with and without unlabeled samples. *Neural Networks*, 65:115–125, 2015.

[11] D. Yin, R. Kannan, and P. Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, 2019.

[12] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*. Springer Science & Business Media, 1996.

[13] M. Khoury and D. Hadfield-Menell. On the geometry of adversarial examples. *arXiv preprint arXiv:1811.00525*, 2018.

[14] A. Demontis, P. Russu, B. Biggio, G. Fumera, and 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.

[15] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural computation*, 16(5):1063–1076, 2004.

[16] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.