ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and
Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

Informed Machine Learning: Excess Risk and Generalization

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

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

Abstract. Machine Learning (ML) based predictive models are impacting research, industry, and society at large thanks to their ability to model or surrogate real systems. Two of the main current limitations of ML are the need for large amounts of high quality data and low performance far away from the observed data. For this reason, in certain applications where prior knowledge is available, researchers have developed Informed ML (IML) to decrease ML high quality data voracity and increase ML extrapolation abilities. In this work we study the differences between ML and IML excess risk and generalization using also some examples to elucidate the theoretical discussions. Our findings shed some light on the mechanisms and the conditions under which IML outperforms ML.

1 Introduction

Recent advancements in Machine Learning (ML) impacted both technical principles (e.g., data format, architectures, and training procedures), capabilities (e.g., language, vision, and multimodal), applications (e.g., industry, energy, and healthcare), and society (e.g., alignment, trustworthiness, and sustainability) [1]. ML-based predictive models are nowadays used mainly two scenarios: modeling [2], namely to learn an input-output relation directly from real data to be able to make predictions, and surrogation [3], namely to learn an input-output relation from data partially or totally generated by an already available predictive model or simulator to speed up predictions. Despite the large success of ML-based predictive models there are still many different weak points that need to be addressed like the data voracity [4], poor performance in extrapolation [5], trustworthiness [6], and sustainability [7].

In certain applications, like physics [8] or medicine [9], where prior knowledge is available, a possible mitigation strategy to some of the problems mentioned above, e.g., data voracity [4] and poor performance in extrapolation [5], can be provided by the inclusion of this prior knowledge into ML using the so-called Informed ML (IML) [10]. The infusion of prior information into ML can be implemented at various stages of the ML pipeline, primarily categorized into pre-, in-, and post-processing methods [10]. One can state that pre-processing lays the foundations acting on the data, in-processing embeds the knowledge by modifying the learning mechanisms, and post-processing by aligning the outputs with domain expectations of the machine learning models. Apart from improving IML itself, what remains open, like for many other phenomena of modern

[*] This work is partially supported by (i) EU - NGEU National Sustainable Mobility Center (CN00000023) Italian Ministry of University and Research Decree n. 1033—17/06/2022 (Spoke 10), (ii) project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU, and (iii) project FAIR (PE00000013) under the NRRP MUR program funded by the EU - NGEU.

artificial intelligence [11], is a fundamental understanding of when and why IML outperforms ML [10].

With this goal in mind, after some preliminaries in Section 2, we will first study in Section 3 the excess risk of IML against the one of ML using the approximation-estimation decomposition [12]. Then, we will study in Section 4 how it is possible to assess the generalization ability of IML and ML focusing on the effect of the prior knowledge on generalization using statistical learning theory [13, 14]. Subsequently, Section 5 will leverage some examples to elucidate the theoretical discussions. Finally, Section 6 will conclude the paper.

2 Preliminaries

Let us consider the ML-based predictive model setting [2] where the goal is to map from an input $X \in \mathcal{X}$ to an output $Y \in \mathcal{Y}$ trying to approximate the underlying unknown distribution μ over $\mathcal{X} \times \mathcal{Y}$. This is achieved by learning a model h, chosen from a class of possible ones $\mathcal{H}$, based on a sample $\mathcal{D}_n = \{(X_i, Y_i) | i \in \{1, \cdots, n\}\}$ sampled i.i.d. from μ. The quality of h is measure according to a loss function $\ell(h(X), Y)$ which leads to the definition of risk $\mathtt{R}(h) = \mathbb{E}_{X,Y}\{\ell(f(X_i), Y_i)\}$ and empirical risk $\hat{\mathtt{R}}(h) = {}^{1}\!/_{n} \sum_{i=1}^{n} \ell(f(X_1), Y_1)$. Let us define with $\mathcal{F}$ the set of all measurable functions $\mathcal{X} \to \mathcal{Y}$. The Bayes model, i.e., the desired model, is defined as $f^* = \arg\min_{f \in \mathcal{F}} \mathtt{R}(f)$. Instead the strategy of any ML algorithm to approximate f^* can be formulated as follows $h^* = \arg\tilde{\min}_{h \in \mathcal{H}} \hat{\mathtt{R}}(h)$ namely, we select $\mathcal{H} \subset \mathcal{F}$, we use $\hat{\mathtt{R}}(h)$ as empirical estimator of $\mathtt{R}(h)$, and we try to search for the best model in $\mathcal{H}$ which minimizes $\hat{\mathtt{R}}(h)$ using a practical algorithm $\tilde{\min}$. $\mathcal{H}$ is induced by the choice of the functional form of h (e.g., linear, tree, ensembles, convolutions, and attentions) and the implicit or explicit regularizers (e.g., wights norm, dropout, and early stopping) [2, 15]. $\hat{\mathtt{R}}(h)$ is used since $\mathtt{R}(f)$ cannot be computed as μ is unknown [2]. With $\tilde{\min}$ we meant that, given $\mathcal{H}$ and $\hat{\mathtt{R}}(h)$, actually finding the $h \in \mathcal{H}$ that minimized $\hat{\mathtt{R}}(h)$ can be a computationally expensive operation and consequently practical algorithms are deployed (e.g., gradient descent or greedy algorithms) [2, 15].

3 Excess Risk

The excess risk of the ML algorithm can be decomposed as approximation $\mathtt{E}_{\text{app}}$ (due to the choice of $\mathcal{H}$), estimation $\hat{\mathtt{E}}_{\text{est}}$ (due to the use of $\mathcal{D}_n$), and optimization $\tilde{\mathtt{E}}_{\text{opt}}$ (due to the choice of $\tilde{\min}$) errors as follows [16]

$$\mathtt{R}(h^*) - \mathtt{R}(f^*) = \mathtt{E}_{\text{app}} + \hat{\mathtt{E}}_{\text{est}} + \tilde{\mathtt{E}}_{\text{opt}}, \tag{1}$$

where $\mathtt{E}_{\text{app}} = \mathtt{R}(h^a) - \mathtt{R}(f^*)$ with $h^a = \arg\min_{h \in \mathcal{H}} \mathtt{R}(h)$, $\hat{\mathtt{E}}_{\text{est}} = \mathtt{R}(h^e) - \mathtt{R}(h^a)$ with $h^e = \arg\min_{h \in \mathcal{H}} \hat{\mathtt{R}}(h)$, and $\tilde{\mathtt{E}}_{\text{opt}} = \mathtt{R}(h^*) - \mathtt{R}(h^e)$. These errors can be redistributed into the bias-variance decomposition [17] but this decomposition does not allow to split the difference source of error properly [12].

By carefully optimizing the choice of $\mathcal{H}$ and $\tilde{\min}$ we can reduce this excess risk (e.g., larger $\mathcal{H}$ may reduce $\mathtt{E}_{\text{app}}$ but enlarges $\mathtt{E}_{\text{est}}$, more complex $\mathcal{H}$ may improve $\mathtt{E}_{\text{est}}$ but enlarges $\mathtt{E}_{\text{opt}}$). $\mathtt{E}_{\text{app}}$ cannot be meaningfully bounded without knowing μ [2] while $\mathtt{E}_{\text{est}}$ and $\mathtt{E}_{\text{opt}}$ can be bounded via statistical learning theory (e.g., Rademacher Complexity - RC [13] or Algorithmic Stability - AS [14]) plus a term that depends on the quality ρ of the optimizer such that $\hat{\mathtt{R}}(h^*) - \hat{\mathtt{R}}(h^e) \leq \rho$ [2, 16].

When we deal with IML, with pre-, in-, or post-processing, we are basically modifying the learning algorithm by modifying $\mathcal{H}$. In fact the introduction of the prior knowledge ι into ML may impact, e.g., by input manipulation designing new or different features, by introducing a new regularizer, or by introducing output constraints [10], and, in all cases, these manipulations can be seen always as a modification of $\mathcal{H}$ into a new space $\mathcal{H}_\iota \subset \mathcal{H}$. As a consequence, the larger is ι the smaller is $\mathcal{H}_\iota$. Of course, by modifying $\mathcal{H}$ we may also impact $\widetilde{\min}$ as in Shallow ML where the introduction of, e.g., a differentiable but non-convex regularizer, may lead to larger $\mathtt{E}_{\text{opt}}$ but this additional error is usually manageable and less impacting [10]. What, instead, can be deeply impacted is the computational requirements of $\widetilde{\min}$, or equivalently the time T needed to find h^* [10]. Then, for IML, the best model h^* is found by solving the problem $\widetilde{\min}_{h\in\mathcal{H}_\iota}\hat{\mathtt{R}}(h)$ and, as a consequence, the decomposition of Eq. (1) need to be reformulated as follows

$$\mathtt{R}(h^*) - \mathtt{R}(f^*) = \mathtt{E}_{\text{app}} + \mathtt{E}_{\text{inf}} + \hat{\mathtt{E}}_{\text{est}} + \tilde{\mathtt{E}}_{\text{opt}}, \tag{2}$$

where $\mathtt{E}_{\text{inf}} = \mathtt{R}(h^i) - \mathtt{R}(h^a)$ with $h^i = \arg\min_{h\in\mathcal{H}_\iota}\mathtt{R}(h)$, $\hat{\mathtt{E}}_{\text{est}} = \mathtt{R}(h^e) - \mathtt{R}(h^i)$ with $h^e = \arg\min_{h\in\mathcal{H}_\iota}\hat{\mathtt{R}}(h)$, and $\tilde{\mathtt{E}}_{\text{opt}} = \mathtt{R}(h^*) - \mathtt{R}(h^e)$ which is equivalent to Eq. (1) plus the addition of the error due to the introduction of the prior knowledge $\mathtt{E}_{\text{inf}}$.

	$\mathcal{H}$	n	ρ	ι
$\mathtt{E}_{\text{app}}$	↓			
$\mathtt{E}_{\text{inf}}$				↓
$\hat{\mathtt{E}}_{\text{est}}$	↑	↓		
$\tilde{\mathtt{E}}_{\text{opt}}$			↑	
T	↑	↑	↓	↑

Table 1: Typical variations of $\mathtt{E}_{\text{app}}$, $\mathtt{E}_{\text{inf}}$, $\hat{\mathtt{E}}_{\text{est}}$, $\tilde{\mathtt{E}}_{\text{opt}}$, and T increasing $\mathcal{H}$, n, ρ, and ι.

Note that, in IML, by carefully optimizing the way in which the priory knowledge informs ML we can reduce the excess risk with respect to classical ML [10]. Table 1 reports the typical variations of $\mathtt{E}_{\text{app}}$, $\mathtt{E}_{\text{inf}}$, $\hat{\mathtt{E}}_{\text{est}}$, $\tilde{\mathtt{E}}_{\text{opt}}$, and T increasing $\mathcal{H}$, n, ρ, and ι.

4 Generalization

While the excess risk analysis discussed in Section 3 is an important step to understand the differences between ML and IML, in practical situations what is more important is how to estimate and tune [18] the generalization performance of an IML-based model bounding the following quantity $\mathtt{R}(h^*) - \hat{\mathtt{R}}(h^*)$, namely the distance between what the error observed on the data and the one on the population. In practice this can be done as follows

$$\mathbb{P}\left\{\mathtt{R}(h^*) \leq \hat{\mathtt{R}}(h^*) + \hat{\mathtt{B}}(h^*) + \Delta(n,\delta)\right\} \geq 1-\delta, \tag{3}$$

where $\hat{\mathtt{B}}(h^*)$ is a bias term (related to $\mathtt{E}_{\text{app}}$) and $\Delta(n,\delta)$ is a confidence term (related to $\hat{\mathtt{E}}_{\text{est}}$). To address this issue in a general scenario we have two main options: complexity based method, being the RC the most effective approach [13] and AS [14]. RC has the advantage of being easily bounded if the space of function is explicitly defined while it is hard to compute in practice [19]. Vice versa, AS, and in particular the Hypothesis AS, is very hard to bound but is more easy to compute in practice [19].

In fact, if we consider a quite general shallow IML algorithm (e.g., ensemble, linear, kernel, random projection) we can formulate it [2, 19] by writing $h(X) = \sum_{i=1}^p \alpha_i\phi_i(X)$ with $p\in\mathbb{N}^+$, $\phi_i(X):\mathcal{X}\to\mathbb{R}$, and $\alpha_i\in\mathbb{R}$ with $i\in\{1,\cdots,p\}$ and

$$\hat{\boldsymbol{\alpha}}^* : \arg\min_{\alpha\in\mathbb{R}^p}\hat{\mathtt{R}}(\alpha) + \lambda_1\left(\lambda_2\hat{\mathtt{H}}_{\text{ML}}(\alpha) + (1-\lambda_2)\hat{\mathtt{H}}_{\text{IML}}(\alpha)\right), \tag{4}$$

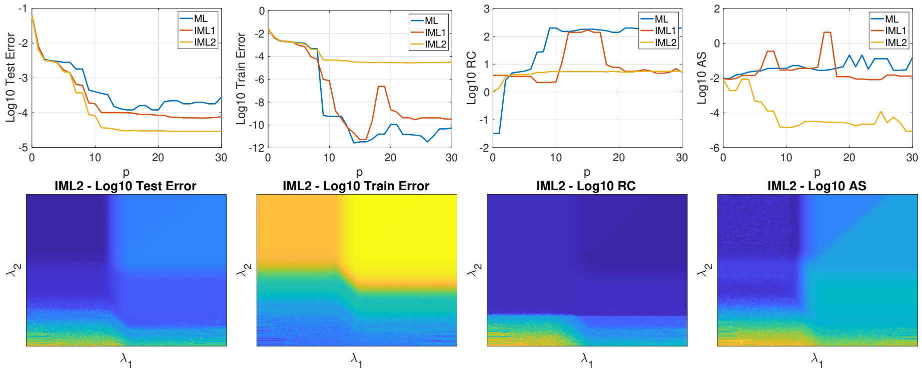

Fig. 2: Ex1 - (Top) For λ_1^* and λ_2^* we report varying p the test error, the train error, the RC, and the AS for ML, IML1, and IML models - (Bottom) For IM2 and p^* test error, the train error, the RC, and the AS varying λ_1 and λ_2.

where, with a little abuse of notation, $\hat{\mathtt{R}}(\alpha)$ is the empirical error of the $h(X)$, $\hat{\mathtt{H}}_{\mathrm{ML}}(\alpha)$ is the usual ML regularizer [2] and the regularizer $\hat{\mathtt{H}}_{\mathrm{IML}}(\alpha)$ and the projection $\boldsymbol{\phi}(X)$ can encapsulate the prior knowledge ι [10]. Note that $\hat{\mathtt{H}}_{\mathrm{IML}}(\alpha)$ represents the adherence of the model h^* (α^*) to ι (e.g., some physical law) that in theory should be matched perfectly but in practice (due to, e.g., noise in the data or approximation of ι in the modeling scenario or to necessity of computational simplicity of h^* in the surrogation scenario) there is a compromise between $\hat{\mathtt{H}}_{\mathrm{IML}}(\alpha)$ and $\hat{\mathtt{H}}_{\mathrm{ML}}(\alpha)$ to find, regulated by λ_2 [10].

Under some mild conditions we can prove that when using the RC $\hat{\mathtt{B}}(h^*) \propto \lambda_2\hat{\mathtt{H}}_{\mathrm{ML}}(\alpha)+(1-\lambda_2)\hat{\mathtt{H}}_{\mathrm{IML}}(\alpha)$ [13, 19]. If we use some implicit regularizers or a deep IML algorithm we cannot rely on RC and we need to use the Hypothesis AS β which can be easily estimated from the data and, in this case, $\hat{\mathtt{B}}(h^*) \propto \beta$ [14, 19].

These tools are extremely important in practical situation since they allow us to understand, for IML, the impact of ι (e.g., ϕ_i with $i \in \{1, \cdots, p\}$ and the shape of $\hat{\mathtt{H}}_{\mathrm{IML}}(\alpha)$) and the classical ML hyperparameters (e.g., the shape of $\hat{\mathtt{H}}_{\mathrm{ML}}(\alpha)$, λ_1, λ_2, and p) on the generalization performance of h^* (α^*) as compromise between $\hat{\mathtt{R}}(h^*)$ and $\hat{\mathtt{B}}(h^*)$ [19].

5 Examples

In this section we will consider two simple examples to test the concepts presented in the previous sections. In the first example (Ex1) we will investigate the effect of IML in an interpolation scenario, namely the quality of the model close to the data samples. The second one (Ex2) will investigate the extrapolation scenario, namely the quality of the model far away from the data samples.

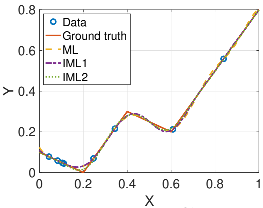

Fig. 1: Ex1 - Ground truth, $\mathcal{D}_n$, and ML, IML1, and IML2 with p^*, λ_1^*, and λ_2^*.

Ex1 is based on the work of [19] where $Y{=}||X-0.4|-0.2| + 0.5X-0.1$, $h(X){=}\sum_{i=0}^{p} \alpha_i X^i$, and $\ell(f(X), Y){=}(f(X){-}Y)^2$ inducing $\hat{\mathtt{R}}(\alpha)$ in Eq. (4). For what concerns the ML model we have to set $\lambda_2 = 1$ and $\hat{\mathtt{H}}_{\mathrm{ML}}(\alpha){=}\int_0^1 [f''(X)]^2 dX{=}\|M\boldsymbol{\alpha}\|_2^2$ in Eq. (4) for a computable M. For what concerns the IML model we have inserted the following knowledge about Y:

$f'(X){=}{-}0.5$ if $X{\in}[0,0.2]$, $f'(X){=}1.5$ if $X{\in}[0.2,0.4]$, $f'(X){=}{-}0.5$ if $X{\in}[0.4,0.6]$, and $f'(X){=}1.5$ if $X{\in}[0.6,0.1]$. We can insert this knowledge only on some data points obtaining our IML1, i.e., the ones in $\mathcal{D}_n$, or on all the space $X{\in}[0,1]$ obtaining our IML2. In both cases, $\hat{\mathtt{H}}_{\text{IML}^*}(\alpha){=}\|M\boldsymbol{\alpha}-\boldsymbol{v}\|_2^2$ in Eq. (4) for computable M and $\boldsymbol{v}$. Figure 1 reports the ground truth, $\mathcal{D}_n$, and ML, IML1, and IML2 models for the best hyperparameters p^*, λ_1^*, and λ_2^*. From Figure 1 we already observe the supremacy of IML over ML. This result can be better observed in Figure 2 where, in the top line for λ_1^* and λ_2^* we report varying p the test error, the train error, the RC [19], and the AS [19] for ML, IML1, and IML2 models. For these results it is possible to observe one expected phenomenon from Section 3, i.e., the expected risk decreases as the amount of knowledge injected into the ML is increased. Moreover, we can see that this model works well in the overparameterization setting, i.e., the more parameters the better the performance, and the effect is more evident for IML showing that combining ML and IML strategies to reduce the excess risk may actually positively resonate. In other words using just $\hat{\mathtt{H}}_{\text{ML}}$ or just $\hat{\mathtt{H}}_{\text{IML}}$, even in this simple case, is worse than a combination of them. Still from the top line of Figure 2 it is possible to observe how the train error coupled with both RC (which is computable just for these explicit regularizers) but much more with AS actually tells us (as expected from Section 4) what hyperparameters and model is the best one to increase our generalization performance. This is even more evident in the bottom line of Figure 2 where for IML2 and the best value of p, namely p^*, we report the test error, the train error, the RC [19], and the AS [19] varying λ_1 and λ_2 (dark blue small value and yellow large values).

Also from the bottom of Figure 2 it is possible to see the effect on the excess risk on the ML and IML regularizers on the performance, i.e., we need a good balance between ML and IML regularizers, and the effectiveness of the training error and AS to actually estimate the generalization performance of the IML2.

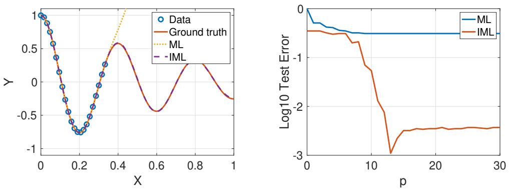

Fig. 3: Ex2 - (Left) ground truth, $\mathcal{D}_n$, ML, and IML models for p^*, λ_1^*, and λ_2^* - (Right) test error and train error for λ_1^* and λ_2^* varying p for ML and IML models.

Ex2 focus on the damped harmonic oscillator[1], using same ML approach of Ex1 while for the IML we took $\hat{\mathtt{H}}_{\text{IML}}(\alpha){=}\int_0^1 (c_1 f''(X){+}c_2 f'(X){+}c_3 f(X))^2\, dX{=}\|M\boldsymbol{\alpha}\|_2^2$, being c_1, c_2, and c_2 parameter of the Ex1, in Eq. (4) for a computable M. Figure 3 on the left reports the ground truth, $\mathcal{D}_n$, and ML and IML models for p^*, λ_1^*, and λ_2^* while on the right, for λ_1^* and λ_2^*, we report varying p the test error and the train error for ML and IML models. From Figure 3 it is possible to see the extreme improvement in extrapolation abilities of the IML model over the ML one and the fact that, in this case, overparameterized models are not the optimal choice. Moreover, Figure 4 reports for Ex2 the equivalent of the bottom line of Figure 2 for IML. Also in this case we can derive the same observation an confirmation derived for Ex1, of the effect of ML and IML regularizers on the excess risk and

[1] `https://github.com/benmoseley/harmonic-oscillator-pinn`

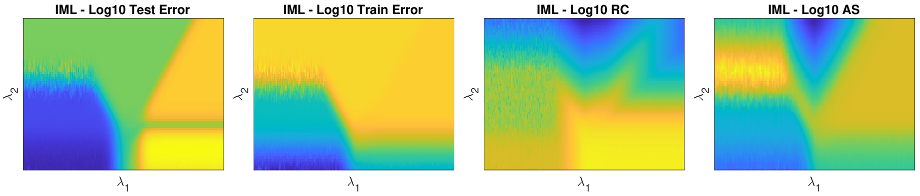

Fig. 4: Ex2 - Equivalent of bottom line of Figure 2 for IML.

generalization ability studied in Sections 3 and 4.

6 Conclusions

In this paper we elaborated on the Informed Machine Learning, namely the ability to include prior knowledge into Machine Learning to decrease its high quality data voracity and increase its extrapolation abilities. In particular we study the differences between Machine Learning and its Informed version via excess risk and generalization error analysis. We elaborated it both from a theoretical and practical perspective using some examples to elucidate the theoretical discussions. Our findings shed some light on the mechanisms and the conditions under which Informed Machine Learning outperforms plain Machine Learning.

References

[1] R. Bommasani, D. A. A. Hudson, E. Adeli, R. Altman, and Others. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258v3*, 2022.

[2] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[3] A. Forrester, A. Sobester, and A. Keane. *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons, 2008.

[4] H. Moseley. In the ai science boom, beware: your results are only as good as your data. *Nature*, 2024.

[5] X. Cao and R. Yousefzadeh. Extrapolation and ai transparency: ... *Big Data & Society*, 10(1):20539517231169731, 2023.

[6] B. Li, P. Qi, B. Liu, S. Di, and Others. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 55(9):1–46, 2023.

[7] A. Van Wynsberghe. Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218, 2021.

[8] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[9] F. Leiser, S. Rank, M. Schmidt-Kraepelin, S. Thiebes, and A. Sunyaev. Medical informed machine learning ... *Artificial Intelligence in Medicine*, 145:102676, 2023.

[10] L. Von Rueden, S. Mayer, K. Beckh, B. Georgiev, and Others. Informed machine learning... *IEEE Transactions on Knowledge and Data Engineering*, 35(1):614–633, 2021.

[11] R. Geirhos, J. H. Jacobsen, C. Michaelis, R. Zemel, and Others. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[12] G. Brown and R. Ali. Bias/variance is not the same as approximation/estimation. *Transactions on Machine Learning Research*, 2024.

[13] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[14] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[15] C. C. Aggarwal. *Neural networks and deep learning*. Springer, 2023.

[16] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. *Neural information processing systems*, 2007.

[17] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

[18] L. Oneto. *Model selection and error estimation in a nutshell*. Springer, 2020.

[19] L. Oneto, S. Ridella, and D. Anguita. Do we really need a new theory to understand over-parameterization? *Neurocomputing*, 543:126227, 2023.