Minds and Machines (2025) 35:3
https://doi.org/10.1007/s11023-024-09703-y

Statistical Learning Theory and Occam's Razor: The Core Argument

Tom F. Sterkenburg[1]

Received: 15 December 2023 / Accepted: 30 October 2024 / Published online: 27 November 2024

Abstract
Statistical learning theory is often associated with the principle of Occam's razor, which recommends a simplicity preference in inductive inference. This paper distills the core argument for simplicity obtainable from statistical learning theory, built on the theory's central learning guarantee for the method of empirical risk minimization. This core "means-ends" argument is that a simpler hypothesis class or inductive model is better because it has better learning guarantees; however, these guarantees are model-relative and so the theoretical push towards simplicity is checked by our prior knowledge.

1 Introduction

Statistical learning theory is the standard framework for the mathematical analysis of machine learning methods (Shalev-Shwartz & Ben-David, 2014; Vapnik, 2000). The framework offers theoretical learning guarantees for certain learning methods, thus providing a basis for viewing such methods as *good* methods.

An old trope in machine learning, usually evoked under the label of *Occam's razor*, is that a shared trait of good methods is a bias towards *simplicity* (Alpaydin, 2020; Duda et al., 2001; Goodfellow et al., 2016; Mitchell, 1997; Mohri et al., 2018; Shalev-Shwartz & Ben-David, 2014). Occam's razor, understood as the principle that a simplicity preference is integral to good scientific or inductive reasoning, is also a long-standing topic of debate in the philosophy of science (Baker, 2022; Sober, 2015). The central question here is whether we actually have some epistemic *justification* for Occam's razor. That is, we seek a rational reason for holding that a simplicity preference helps us attain desirable epistemic ends, like minimizing error.

A step forwards in the wider debate would be a justification for Occam's razor in machine learning methods, and an obvious place to look for such a justification is statistical learning theory. Indeed, formal results in this framework have been quoted

✉ Tom F. Sterkenburg
tom.sterkenburg@lmu.de

[1] Munich Center for Mathematical Philosophy, LMU Munich, Munich, Germany

in support of such a justification (e.g., Blumer et al,1987, Shalev-Shwartz and Ben-David, 2014). On the other hand, computer scientists and philosophers alike have also relied on formal results to argue *against* such a justification in this framework (e.g., Domingos,1999, Herrmann, 2020).

In this paper, I integrate the intuitions and arguments from both sides into a qualified epistemic justification for Occam's razor from statistical learning theory. Importantly, the notion of simplicity in this justification pertains, *not* to individual classifiers or hypotheses, but to inductive models or *classes* of hypotheses. A further important component of my account is the relativity to such inductive models of the justification obtainable from theoretical learning guarantees, as highlighted by Sterkenburg and Grünwald (2021). This *model-relativity* of learning-theoretic justification aligns well, I argue, with a broad tradition in the philosophy of science which accepts the impossibility of absolute justification, and shifts attention to the project of how to rationally proceed from our current beliefs and assumptions. A final important characteristic of my account is the *means-ends* nature of the justificatory reasoning. In one line, the means-ends justificatory argument says that in order to have better model-relative learning guarantees, we need to codify our assumptions in the form of a simpler inductive model.

This argument is based on the first of the two "inductive principles" (Vapnik, 2000) central to statistical learning theory, namely the method of *empirical risk minimization*. I think this is the "core argument" for simplicity in statistical learning theory, which further underpins the method of *structural risk minimization* and its characteristic simplicity preference. I sketch this at the end of the paper.

The plan of the paper is as follows. In Sect. 2, I present the framework of statistical learning theory and the main technical ingredients for the core argument for Occam's razor. These include the notions of empirical risk minimization, learnability, uniform convergence, and VC dimension, and the fundamental theorem that ties these notions together. In Sect. 3, I argue that VC dimension is a robust notion of the simplicity of a hypothesis class. In Sect. 4, I discuss the theoretical justification for empirical risk minimization and particularly its model-relative nature; and I show how all of the previous comes together into a justificatory argument for simplicity. I conclude in Sect. 5.

1.1 Motivation

Before starting, there is a worry about the paper's general project that I should acknowledge. This worry is that the project engages with a debate of a bygone era. It has been well over a decade since Harman and Kulkarni (2007) initiated a small wave of philosophical interest in statistical learning theory, and Steel 2011, p. 860 concluded that the theory is "worthy of further sustained interest from philosophers of science." This sustained interest has not exactly materialized, while the landscape of machine learning has altered significantly. Especially the advent of deep neural networks (DNN's) has caused a shift in what seem the more pertinent epistemological issues: from the traditional questions around the reliability of inductive inference to questions around interpretability and explainability (Beisbart & Räz, 2022).

Moreover, the advent of these algorithms has problematized the very utility of statistical learning theory. The theory seems simply not equipped to explain the generalization behaviour of learning methods like DNN's (Belkin, 2021; Berner et al., 2022; Hardt & Recht, 2022), prompting a "quest for a new framework for a 'theory of induction'" (Belkin, 2021, p. 217). Putting it bluntly: why a renewed philosophical engagement with the framework of statistical learning theory, if this framework is starting to look like a thing of the past?

One plain answer is that it is still of interest if and how the standard framework already offers justification for Occam's razor. Curiously, statistical learning theory is largely left out the modern shift of the philosophical debate towards various frameworks in mathematical statistics (Baker, 2022, Sect. 5; Sober, 2015, Ch. 2), with Sober (2015, p. 140, fn. 61) perceiving statistical learning theory to be "dramatically" different from the "Bayesian and frequentist ideas" that have informed this debate so far. The current project thus fills a gap in the philosophical literature. Secondly, I may above have put things overly bluntly: it is not at all clear that core components of statistical learning theory will not continue to play an essential role in newer theory (cf. Bartlett et al., 2021). In any case, finally, the current project is a stepping stone towards the philosophical analysis of any new "framework for a theory of induction" in machine learning. Belkin indeed evokes "a very pure form of Occam's razor" as the "guiding principle" in a new framework (2021, p. 218). To assess the role and justification of simplicity in such an emerging new framework, it will at the very least be helpful to actually have clarity on its role in the standard framework. The "generalization puzzle" (Berner et al., 2022, p. 25) that is now hotly debated in the machine learning community is indeed a modern reincarnation of exactly those traditional philosophical questions around the reliability of induction. Work like the current project can, I hope, offer a starting point for philosophers to engage with this exciting but complex debate.

2 The Formal Ingredients

In my presentation of the framework of statistical learning theory, I mainly follow Shalev-Shwartz and Ben-David (2014).[1,2] I restrict attention to the most basic learning paradigm in this framework, the paradigm of binary classification.

[1] Their presentation is essentially a synthesis of Vapnik's (1999; 1998; 2000) "general setting of learning" and Valiant's (1984) model of "probably approximately correct" (PAC) learning (Shalev-Shwartz & Ben-David, 2014, p. 28). The main concern in Vapnik's setting is the statistical analysis of uniform convergence of learning algorithms, and this approach is also simply called *VC theory* after the groundbreaking early work of Vapnik and Chervonenkis (1971). The tradition initiated by Valiant is also called *computational learning theory* [see Anthony & Biggs (1992); Kearns & Vazirani (1994)], and an essential component is the computational efficiency of learning. This computational component is separated from the statistical component in Shalev-Shwartz and Ben-David's presentation, and I will likewise not be concerned with computational considerations in this paper.

[2] A chapter-length introduction to statistical learning theory aimed at philosophers, that I also draw from, is (von Luxburg & Schölkopf, 2011). A more basic philosophical introduction is (Harman & Kulkarni, 2007).

2.1 Binary Classification

In this type of learning problem, we have a domain $\mathcal{X}$ of *instances* (say, images of animals). We seek to assign these instances binary *labels* (say, cat or not cat). More precisely, we seek a general classifier or *hypothesis* $h : \mathcal{X} \rightarrow \mathcal{Y}$ that maps all instances in $\mathcal{X}$ to a label in the binary label set $\mathcal{Y}$.[3]

This is a learning problem because we first draw a finite *training sample* of labeled instances, on the basis of which we then seek to find—to *learn*—a general hypothesis. The assumption in statistical learning theory is that there always is some true but unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, that governs both the sampling of instances and (via the conditional $\mathcal{D}(\mathcal{Y} \mid \mathcal{X})$) the connection between instances and labels. It is assumed we obtain labeled instances by repeatedly drawing from this same distribution: the labeled instances are *independently and identically distributed* (i.i.d.). In this way, we draw a training sample S, that is a finite ordered sequence of input-label pairs. Based on the training sample, we seek to learn a good hypothesis.

To assess hypotheses, we use some error function. The standard choice in binary classification is the 0/1 error function, that returns error 0 (error 1) for a correct (incorrect) classification. Then the *empirical error* of h on a sample S is given by the mean 0/1 error of instances,

$$L_S(h) := \frac{|\{(x, y) \in S : h(x) \neq y\}|}{|S|}. \tag{1}$$

But what we are actually interested in is the quality of a classifier over all possible instances. We express this as the expected error or *risk* of h with respect to the true distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

$$L_{\mathcal{D}}(h) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[L_{(X,Y)}(h)\right]. \tag{2}$$

We thus seek to find a hypothesis, based on a training sample S, with a low risk with respect to the true but unknown distribution $\mathcal{D}$.

2.2 Hypothesis Classes and Learning Methods

In the framework of statistical learning theory, we are fully agnostic about the shape of the distribution $\mathcal{D}$. However (as I discuss in more detail later), we cannot get anywhere unless we impose restrictions elsewhere. The approach in statistical learning theory is to make the analysis relative to some hypothesis class $\mathcal{H}$. We then seek to select a hypothesis h from $\mathcal{H}$ which has *relatively* low risk, among those hypotheses

[3] Hypotheses are often called *models* in the machine learning literature. I will stick here to the terminology of Shalev-Shwartz and Ben-David (2014), also to not risk confusion with the notion of *inductive model* (class of hypotheses) in the model-relative justification I discuss in section 4.

in $\mathcal{H}$, with respect to true but unknown $\mathcal{D}$. That is, we seek to select, based on training data, a hypothesis with risk close to $\min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$.[4]

This is a machine learning problem because we want to specify an automated learning procedure or *learning method* to do this selection—this *learning*—of hypotheses from samples. Formally, we treat a learning method as a function from all possible samples to hypotheses.[5]

A basic such method is the procedure of *empirical risk minimization* (ERM) for given hypothesis class $\mathcal{H}$. This method simply selects for given sample S a hypothesis in $\mathcal{H}$ with minimal error on the sample.

Definition 1 Empirical risk minimization for hypothesis class $\mathcal{H}$, write $\mathrm{ERM}_{\mathcal{H}}$, returns for each $S \in \mathcal{S}$ a hypothesis in $\arg\min_{h \in \mathcal{H}} L_S(h)$.

What makes a learning method like ERM for $\mathcal{H}$ a *good* method? Given the indicated goal of finding a relatively-low-risk hypothesis in $\mathcal{H}$, method $\mathrm{ERM}_{\mathcal{H}}$ can be called good if it has some sufficiently strong guarantee of attaining this goal.

2.3 Learnability

The main formal guarantee of good learning is formulated in terms of the following components.

First, we quantify the "relatively-low-risk" by an accuracy parameter ϵ. This ϵ bounds the difference between the best possible risk $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ and the risk $L_{\mathcal{D}}(A_{\mathcal{H}}(S))$ of a hypothesis selected by method $A_{\mathcal{H}}$ on sample S. This difference is also called the *estimation error*.

Second, because of the randomness in the generation of samples from $\mathcal{D}$, any guarantee can at best be probabilistic. Intuitively, we can only expect a learning method to select a good hypothesis based on samples that are in fact representative of the true distribution $\mathcal{D}$; but we cannot exclude that with small probability we draw a sample that is not representative. Hence we also introduce a confidence parameter δ that quantifies this probability.

Finally, again due to the randomness in drawing samples, the quality of an estimate is inevitably connected to the size of the sample. We will thus formulate our guarantee as a relation between sample size, confidence, and accuracy.

This guarantee is *probably approximately correct* (PAC) *learnability*[6]—or simply, *learnability*. Hypothesis class $\mathcal{H}$ is *learnable* by a learning method $A_{\mathcal{H}}$ if for any

[4] This is what Shalev-Shwartz and Ben-David (2014 p. 23) call *agnostic* learning, as opposed to the more specific paradigm of *realizable* learning, where we make the (very strong) assumption that $\mathcal{H}$ already contains an h^* with zero true risk (ibid., def. 2.3). In computational learning theory this assumption is actually standard.

[5] Again, I abstract away from computational considerations. A restriction of the framework to formal computability [in which learning methods are actual algorithms, i.e., Turing-computable functions; a framework only first studied recently, Agarwal et al. (2020)] does not appear to substantially change the notions and results discussed here (Sterkenburg, 2022).

[6] This notion was formulated (with the additional component of computational complexity) by Valiant (1984), while the term "pac-learning" appears to be due to Angluin and Laird (1988).

given inaccuracy ϵ and confidence $1 - \delta$, there is a large enough sample size m_0 such that for any $m \geq m_0$, we have the following, no matter the true distribution $\mathcal{D}$. With probability at least $1 - \delta$ over the possible size-m samples S^m drawn i.i.d. from $\mathcal{D}$, method $A_{\mathcal{H}}$, on receiving such a sample, returns a hypothesis with estimation error below ϵ. To rephrase,

Definition 2 (Learnability) A hypothesis class $\mathcal{H}$ is learnable if there exists a learning method $A_{\mathcal{H}} : \mathcal{S} \to \mathcal{H}$ and a sample size function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$, for all $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ and any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

$$\mathrm{Prob}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(A_{\mathcal{H}}(S)) \leq \min_{h \in \mathcal{H}}(L_{\mathcal{D}}(h)) + \epsilon\right] \geq 1 - \delta. \tag{3}$$

Note that this guarantee (in particular, the minimum sample size $m_{\mathcal{H}}(\epsilon, \delta)$ for given ϵ and δ) only depends on the hypothesis class $\mathcal{H}$. In line with the agnostic approach of statistical learning theory, it is a *distribution-free* guarantee: the sample size does not depend on the true distribution $\mathcal{D}$.

Learnability is much related to another property of a hypothesis class, namely *uniform convergence*. The former, as we have seen, concerns the estimation error of a learning method; the latter concerns the difference between the empirical errors and the true risks of all hypotheses in the class. This property will allow us to relate the ERM method to learnability.

2.4 Uniform Convergence and Empirical Risk Minimization

The law of large numbers already tells us that, for any fixed hypothesis h, as we draw larger and larger samples S^m i.i.d. from true distribution $\mathcal{D}$, the empirical error of h on S^m will in probability converge to its true risk. However, in our learning problem, we are not interested in fixing a particular hypothesis and estimating its true risk. We are interested in the performance of a learning algorithm, which, depending on the data, can select different hypotheses. For this we need something stronger, namely a "uniform law of large numbers," which bounds the difference between empirical errors and true risks of all hypotheses *uniformly*—simultaneously.

For given hypothesis class $\mathcal{H}$, call a training sample ϵ*-representative* if simultaneously for all hypotheses $h \in \mathcal{H}$ the difference between h's empirical error $L_S(h)$ on S and h's true risk $L_{\mathcal{D}}(h)$ is smaller than ϵ,

$$(\forall h \in \mathcal{H})\left[|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\right]. \tag{4}$$

On such a sample, all empirical errors give good indications of the true errors: "what you see is what you get" ("wysiwyg", terminology from Belkin, 2021). Now a hypothesis class has the *uniform convergence property* if there is a "wysiwyg" guarantee of drawing such representative samples. Precisely,

Definition 3 (Uniform convergence) A hypothesis class $\mathcal{H}$ has the uniform convergence property if there exists a sample size function $m^{\mathrm{uc}}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$, for all $m \geq m^{\mathrm{uc}}_{\mathcal{H}}(\epsilon, \delta)$ and any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ we have

$$\mathrm{Prob}_{S\sim\mathcal{D}^m}\left[(\forall h \in \mathcal{H})\left[|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\right]\right] \geq 1 - \delta. \tag{5}$$

To link this property to the ERM method, we first reformulate it. Namely, we have stated it in terms of the minimum sample size $m^{\mathrm{uc}}_{\mathcal{H}}(\epsilon, \delta)$ we need for given ϵ and δ; but we can also formulate it as the bound on ϵ we get for given δ and sample size m. Precisely, there is an accuracy function $\epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \epsilon)$ such that for given δ and m we have with probability at least $1 - \delta$ that

$$(\forall h \in \mathcal{H})\left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta)\right], \tag{6}$$

which in particular gives a uniform upper bound on true risk in terms of empirical error,

$$(\forall h \in \mathcal{H})\left[L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta)\right]. \tag{7}$$

Now recall that the method $\mathrm{ERM}_{\mathcal{H}}$ for given sample S of length m selects an h that minimizes $L_S(h)$. Since $\epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta)$ is a constant term for fixed m and δ, method $\mathrm{ERM}_{\mathcal{H}}$ can be seen to explicitly minimize this upper bound

$$L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta) \tag{8}$$

on the true risk. Thus, given $\mathcal{H}$ satisfies uniform convergence, $\mathrm{ERM}_{\mathcal{H}}$ selects a hypothesis with the sharpest uniform upper bound on its true risk.

This minimization property, under the assumption of uniform convergence, allows us to derive that $\mathrm{ERM}_{\mathcal{H}}$ *learns* $\mathcal{H}$. Informally,[7] if we have a guarantee that large enough samples are probably representative (uniform convergence), then in particular the lowest-empirical-error hypotheses (selected by $\mathrm{ERM}_{\mathcal{H}}$) probably have approximately lowest true risk, and so small estimation error (learnability).

Uniform convergence thus gives us a sufficient condition for learnability, and learnability by ERM. However, this is still a rather abstract property, that does not give much intuition for what kind of hypothesis classes satisfy it. Fortunately, it turns out that there is a more concrete and intuitive property of hypothesis classes that is equivalent to learnability, and in fact already equivalent to learnability by ERM. This property is a criterion of the *simplicity* of a hypothesis class.

2.5 The VC Dimension

Take a finite set $X = \{x_1, \ldots, x_m\} \subset \mathcal{X}$ of unlabeled instances. There are several different ways in which we can label all instances in X; precisely, for binary labels, there are 2^m possible such labelings. Now take a hypothesis class $\mathcal{H}$. Each hypothesis h in $\mathcal{H}$ gives some such possible labeling of the instances in X. If the hypotheses

[7] See Shalev-Shwartz and Ben-David (2014, Sect. 4.1, specifically lemma 4.2) for the (straightforward) formal derivation.

in $\mathcal{H}$ cover *all* possible labelings, that is, for each possible labeling of X, there is some $h \in \mathcal{H}$ that gives exactly this labeling, then we say that $\mathcal{H}$ *shatters* X.[8]

The crucial notion in demarcating classes that are and are not learnable relies on the ability to shatter sets of instances. Namely, the *Vapnik-Chervonenkis dimension* (VC dimension, after Vapnik and Chervonenkis, 1971) of hypothesis class $\mathcal{H}$ is defined as the largest size of a subset X of instances for which $\mathcal{H}$ can do so.

Definition 4 The *VC dimension* of hypothesis class $\mathcal{H}$ is the maximal size of a set $X \subset \mathcal{X}$ that is shattered by $\mathcal{H}$. If $\mathcal{H}$ shatters sets of arbitarily large size, then the VC dimension of $\mathcal{H}$ is infinite. A *VC class* is a class with finite VC dimension.

In machine learning terminology, VC dimension is a measure of the *capacity* of a hypothesis class.[9] It is a measure of the extent to which a hypothesis class covers—contains hypotheses with good fit on—possible data samples. In that sense VC dimension is a notion of the "richness" or *complexity* of a hypothesis class; and finiteness of VC dimension a criterion of a hypothesis class being sparse or *simple*. I discuss this simplicity interpretation in more detail in Sect. 3 below.

2.6 Bringing it All Together

The central result of statistical learning theory elegantly ties together the main notions of the previous sections.

Theorem 5 (Fundamental theorem of statistical learning theory[10]) *The following are equivalent*:

- $\mathcal{H}$ has the uniform convergence property;
- $\mathcal{H}$ is learnable;
- $\mathcal{H}$ is learnable by $\mathrm{ERM}_{\mathcal{H}}$;

[8] Slightly more formally, define the *restriction of $\mathcal{H}$ to finite set X* as the class $\mathcal{H}_{|X}$ of functions $f : X \to \mathcal{Y}$ such that $f(x) = h(x)$ for some $h \in \mathcal{H}$ and all $x \in \mathcal{X}$. Then $\mathcal{H}$ shatters finite $X \subset \mathcal{X}$ if the restriction of $\mathcal{H}$ to X contains *all* functions $f : X \to \mathcal{Y}$, that is, $|\mathcal{H}_{|X}| = 2^{|X|}$.

[9] There exist several generalizations of VC dimension, like the Natarajan dimension in multiclass categorization (see Shalev-Shwartz and Ben-David 2014, ch. 29), and indeed altogether different capacity notions in different paradigms, like the Littlestone dimension in realizable online learning (see ibid., sect. 21.1), and the parametric complexity in MDL inference (see Grünwald (2007)). An important alternative capacity notion to VC dimension for classification is Rademacher complexity, which can yield stronger data-dependent bounds (see von Luxburg and Schölkopf, von Luxburg and Schölkopf (2011), sect. 5.7 Shalev-Shwartz and Ben-David, 2014, ch. 26). The notion of the capacity of a function class and its relation to generalization appears to have been introduced by Cover (1965).

[10] In their pioneering work, Vapnik and Chervonenkis (1971) established the link between uniform convergence, "consistency" of ERM, and their notion of VC dimension, proving a generalization of the "fundamental theorem of mathematical statistics," the Glivenko-Cantelli theorem of the uniform convergence of the empirical distribution function (see Devroye et al., 1996, ch. 12). The connection between VC theory and computational learning theory (in particular, Valiant's notion of PAC learnability) was first spelled out by Blumer et al. (1986, 1989).

- $\mathcal{H}$ is a VC class.

In particular, if, and only if, hypothesis class $\mathcal{H}$ has finite VC dimension, we have a "wysiwyg" guarantee of a good indication of true risk (uniform convergence), and a method, $\mathrm{ERM}_{\mathcal{H}}$, that is a *good* method, in the sense of satisfying a guarantee of minimizing estimation error (learnability). We therefore have, for VC class $\mathcal{H}$, a certain *justification* for the $\mathrm{ERM}_{\mathcal{H}}$ method.[11]

Actually, we can further fine-grain this picture within the finite VC dimension regime. Namely, the VC dimension of class $\mathcal{H}$ gives a quantitative bound on the sample size for uniform convergence and for learnability.

Theorem 6 (Fundamental theorem, quantitative version[12,13]) *For any VC class $\mathcal{H}$, there are constants C_1, C_2 such that, for any ϵ, δ, we have*

$$C_1 b \leq m^{\mathrm{uc}}_{\mathcal{H}}(\epsilon, \delta), m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 b,$$

where

$$b = \frac{\mathrm{VCdim}(\mathcal{H}) - \log \delta}{\epsilon^2}.$$

At bottom, the fundamental theorem expresses a certain relation between four quantities (VC dimension, sample size, accuracy, and confidence), where, in particular, a lower VC dimension makes room for lower values of the other three quantities (meaning, for stronger bounds). Thus, a lower VC dimension of $\mathcal{H}$ goes with a better guarantee and therefore a *stronger justification* for the $\mathrm{ERM}_{\mathcal{H}}$ method.

I discuss this justification, and how to further turn this into a justificatory argument for simplicity, in more detail in section 4 below. But first I will zoom in on the relevant notion of simplicity, given by the formal notion of VC dimension.

[11] In my presentation, I also follow Shalev-Shwartz and Ben-David (2014) in focusing on the epistemic end of learnability (minimizing estimation error). I only note here that another important epistemic end that is supported by the uniform convergence "wysiwyg" guarantees is *model assessment*, where we use the training error to assess whether the model (the learned hypothesis or indeed the hypothesis class) is actually good. (For instance, in the discussion of VC theory by Hastie et al., 2009, sect. 7.9, the emphasis is rather on this end.)

[12] See Shalev-Shwartz and Ben-David 2014, Thrm. 6.8.

[13] This is a bound on the sample size in terms of given accuracy and confidence parameter (and VC dimension); but we can also infer other bounds by making other choices in what quantities we take as given and what quantity we then solve for. For instance, we can derive that there exists constant C such that for any given m and δ we have an accuracy bound

$$\epsilon^{\mathrm{uc}}(m, \delta) < C\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) - \log \delta}{m}}. \tag{9}$$

3 The Notion of Simplicity

While, I will argue in this section, the notion of VC dimension does not give us a handle on the simplicity of *individual* hypotheses (Sect. 3.1), it does constitute a plausible and robust measure of the simplicity of hypothesis classes (Sect. 3.2).

3.1 Individual Hypotheses and Hypothesis Classes

Capacity notions like VC dimension apply to hypothesis classes, *not* individual hypotheses. Yet in some discussions of simplicity, that also rely on the relation between uniform convergence and small size or capacity of the hypothesis class, the notion of simplicity invoked actually concerns individual hypotheses. These discussions use a notion of the complexity of an hypothesis as its representational complexity in some formal language.

An influential example is the argument of Blumer et al. (1987) that "under very general assumptions, Occam's Razor produces hypotheses that with high probability will be predictive of future observations" (ibid., p. 378).[14] An earlier example still is Pearl (1978), who discusses the connection between simplicity and "credibility" of hypotheses via different notions of capacity and generalization success—including already VC dimension and uniform convergence (ibid., p. 261ff).

Pearl sets the stage as follows. We take some language L with an interpretation function I that maps sentences in the language to hypotheses (ibid., pp. 256f). Then we define some complexity measure on each sentence t, "which may represent either the syntactic aspect of the sentence t, or the work required for the computation of $I(t)$" (ibid., p. 257).[15] Further, "the complexity of a [hypothesis h] with respect to a language L is defined as the complexity of the simplest sentence which represents that [hypothesis]" (ibid.). This allows us to take subsets of sufficiently *simple* hypotheses: a "*complexity bounded sublanguage* of L is a sublanguage $L_c = (T_c, I_c)$ such that $T_c \subseteq T$, $I_c \subseteq I$ and $C(h) \leq c$ for all $h \in I_c$" (ibid., p. 258, slight change in notation).

Now the lower the complexity c, the smaller the size (and in particular, the capacity) of the sublanguage (hypothesis class) L_c.[16] This, via reasoning as in section 2.6 above, leads to a better generalization guarantee or "credibility" of the estimated hypothesis from this class. In this way, Pearl writes in his concluding discussion, "accepted norms of credibility are correlated with [hypotheses'] simplicity" (ibid., p. 263). However, he immediately adds:

[14] The relevant result is derived within Valiant's PAC learning framework, and the requirement of computational efficiency in the definition and generalization guarantee of the relevant "Occam-algorithm" make it a bit more involved than the reasoning I discuss in this paper. For further expositions, see Anthony and Biggs (1992, p. 59ff) Kearns and Vazirani (1994, ch. 2); and for a rebuttal of the argument, see Herrmann (2020).

[15] The standard example is the two-symbol language of bits, where the complexity of a sentence (a bit string) is defined as its length.

[16] For instance, for the language of bits, there can only be 2^{n+1} sentences (bit strings) of complexity (length) up to n, hence at most 2^{n+1} hypotheses of complexity up to n.

> From a philosophical viewpoint it is essential to note that in all cases examined the role of *simplicity* was only incidental to the analysis. We would have gotten identical results if instead of L_c being a complexity bounded sublanguage we were to substitute an arbitrary sublanguage with equal number of [hypotheses]. It is not the nature of the [hypotheses] in I_c but their number $|I_c|$ (more precisely, the number of sample dichotomies induced by the members of I_c) which affects the various plausibility measures considered.

In particular, whereas classes of hypotheses of low representational complexity must be small, the converse does not hold. One can select small classes of (representationally) complex hypotheses, and the same capacity-based reasoning for good generalization still applies (cf. Mitchell, 1997, p. 65; Domingos, 1999, p. 410).[17]

A deeper problem still is that this notion of representational complexity depends on the presupposed formal language and definition of sentence complexity. For any hypothesis that is simple relative to one language, we can design a different language that renders it complex.[18] In that sense representation length does not give us a robust or objective notion of simplicity of individual hypotheses.[19]

Given the bleak prospects for some general mathematical definition to settle what counts as simple for individual hypotheses, one might at this point be inclined to change tack and suggest that in practice, there is often no real problem. For many specific learning problems, we do appear to have clear intuitions about natural representations or parametrizations of hypotheses. In the standard curve-fitting problem (see, e.g., Sober, 2015, pp. 88ff), where we seek to estimate a polynomial function, there exists a natural parametrization by *degree*.[20] The linear functions of degree 1 are simpler than the quadratic functions of degree 2. Moreover, the class of all linear hypotheses is smaller (has lower capacity) than the strict superclass of quadratic hypotheses. Some conception of simplicity of hypotheses is here already taken for granted, which points to a natural ranking of hypothesis classes, and this ranking neatly aligns with their capacity.[21]

[17] This is the main critique of Herrmann (2020) of the argument of Blumer et al. Herrmann derives a parallel result for an "Anti-Occam algorithm" that selects small-cardinality classes of representationally complex hypotheses.

[18] This language-relativity in describing individual hypotheses also clearly arises in some presentations of the *minimum description length* (MDL) approach (e.g., Mitchell, 1997, Sect. 6.6; Shalev-Shwartz & Ben-David, 2014, p. 65f). However, these presentations paint a rather, well, simplistic picture of the approach: in "refined MDL," the focus is on the design of "universal codes," yielding again a robust notion of complexity of *hypothesis classes* (Grünwald, 2007) that plays a role very similar to capacity notions in statistical learning theory (ibid., Sect. 17.10).

[19] It is sometimes held that "idealized MDL" or *Kolmogorov complexity* can offer an objective notion of the representational complexity of individual objects (Li & Vitányi, 2008). See Sterkenburg (2016) for a critique of a suggested justification for Occam's razor via this approach, and Sterkenburg (2018,Sect. 5.2) for a critique of its promise of an objective notion of complexity.

[20] Curve-fitting can be cast as a problem in binary classification by treating the curves as hypotheses separating instances with the one label from instances with the other.

[21] One might seek to base this conception on some formal definition of simplicity in terms of number of adjustable parameters, a line going back at least to Jeffreys (1939). But this still does not suddenly give us a robust and objective notion of simplicity of individual hypotheses: the "grue-like" problems of rep-

If this is a common situation in practice, then, together with the formal connection between low capacity and generalization performance, we may have the basis for an *explanation* of why preferring simple hypotheses generally seems to be a good idea. But even if we accept as given, for many specific learning problems, a standard representation or parametrization of hypotheses, the formal connection between low capacity and generalization performance still falls short of constituting a *justification* for preferring simple hypotheses (for these specific learning problems). The issue remains that the theory does not enforce a connection between simple individual hypotheses (however specified) and classes of low capacity.

3.2 VC Dimension as a Measure of Simplicity

In contrast to definitions of the complexity of an individual hypothesis, definitions of the capacity of a hypothesis class (like VC dimension) do not depend on a specific representation or parametrization, and do therefore possess a certain objectivity or robustness.[22] But does VC dimension also give an objective or robust measure of the *simplicity* of a hypothesis class?

One might deny this on exactly the grounds that VC dimension does not necessarily align with natural parametrizations of individual hypotheses (Domingos, 1999, p. 413). In the case of the usual parametrization of polynomials, the higher the number of free parameters, the higher the capacity of the corresponding hypothesis class; but in other cases the two can come apart. The standard example is the class of sine curves $\{h_\alpha\}_{\alpha\in\mathbb{R}}$ with $h_\alpha(x) = \sin \alpha x$ (ibid.). The elements in this class are given by only one parameter (and in that sense the function class is very simple), yet the class has infinite VC dimension (Vapnik, 2000, p. 78).[23]

Of course, this objection relies on some claim that usual parameterizations do (and exclusively do) track simplicity. But even aside from the ultimate non-robustness of representational notions of complexity, there just exist different and sometimes conflicting intuitions. From one way of looking at it, the class of sine functions *is* maximally complex: exactly because so many possible data configurations can be fit by it (cf. Romeijn, 2017). This is the intuition of richness or complexity (or also *falsifiability*[24]) made precise in a capacity measure. VC dimension is not *the*

Footnote 21 (continued)

resentation invariance remain (Priest, 1976). For a recent discussion and critique of defining simplicity by number of parameters, see Bonk (2023).

[22] More precisely, the capacity of a hypothesis class does not depend on how the individual hypotheses are described: all that matters is their data coverage. Language-relativity only turns up when we start redescribing the instance space (cf. Steel, 2009, p. 482).

[23] Vapnik (1998, p. 698) himself writes that since Occam's razor says that the explanation with "the smallest number of features (free parameters)" is best, and since this is not supported by theoretical results, "Occam's razor principle is misleading and perhaps should be discarded in the statistical theory of inference" (ibid., p. 699). Also see Cherkassky and Mulier (2007, p. 146ff).

[24] Vapnik (2000, p. 42ff) links the capacity of hypothesis classes to Popper's *falsifiability* of theories. Popper famously equated simplicity with falsifiability, and introduced a quantitative notion of falsifiability of theories that he claimed aligned with number of free parameters. Corfield et al. Corfield et al.

measure, but it is *a* plausible and robust measure of the complexity of a hypothesis class. This is enough for my purpose: if a justification is to be had for preferring low capacity, then I think it is reasonable to call this a justification for preferring simplicity—even if there are other reasonable conceptions of simplicity, and even if (to stress again) this notion of simplicity pertains to hypotheses classes and not to individual hypotheses.

4 The Justification for Simplicity

The fundamental theorem of statistical learning theory ties the simplicity—the VC dimension—of a hypothesis class to its learnability, and indeed already to its learnability by the ERM method. This result offers, first of all, a certain justification for the ERM procedure; although this is a justification with several qualifications, chief among them its *model-relativity* (Sect. 4.1). Nevertheless, I will argue that the model-relative justification that learning theory can offer fits right in with a plausible broader epistemological perspective on machine learning methods (Sect. 4.2). Finally, I will assemble from all of the previous elements a qualified justification for a simplicity preference (Sect. 4.3).

4.1 The Justification for Empirical Risk Minimization

The fundamental theorem shows that $\mathrm{ERM}_{\mathcal{H}}$, for VC class $\mathcal{H}$, is a good method. It is good, and good epistemically, because it satisfies a guarantee of attaining a certain epistemic goal. This guarantee therefore constitutes an epistemic *justification* for the method.

An immediate qualification is that this picture of justification or not—learnability or not—is overly black-white. The quantitative version of the fundamental theorem tells us that a smaller VC dimension leads to a stronger guarantee, making ERM an epistemically better method. So we have a graded guarantee that constitutes a graded notion of epistemic justification.

But this step from theoretical guarantees to talk about justification comes with several further qualifications still.

4.1.1 Qualifications

A first elementary point is that the fundamental theorem is a mathematical result. Any epistemic justification derived from it, in the context of a real-world learning problem, needs a story how it maps to this learning problem. Most obviously, for any particular real-world problem, a meaningful application of the fundamental theorem (and justificatory claims derived from it) depends on how well the problem can be modelled in the statistical learning theory framework. This includes the

Footnote 24 (continued)

(2009) and Harman and Kulkarni (2007, p. 50ff) argue that VC dimension is a better measure of falsifiability, though these authors appear to resist linking VC dimension to simplicity.

match of our prior assumptions with the formal assumption of i.i.d. sampling of data from some unknown distribution, but also the match of our goals with the formal choice of the 0/1 error function. The representation of a learning problem in the formal framework of statistical learning theory already forces us to commit to and codify various assumptions (von Luxburg & Schölkopf, 2011, p. 683f), and anything that follows from the mathematics should be appraised with an eye to whether these made sense for the original learning problem.

But even when there are no such modeling concerns, one need not accept that the formal guarantees from the fundamental theorem are sufficiently strong or interesting to warrant talk about justification. There are legitimate reservations one can have about the usefulness of these guarantees.

One possible reservation is that these guarantees are quintessential *frequentist* guarantees. They say something about what we can expect with high probability *before* the sampling and the learning. In that sense we can call methods satisfying these guarantees *reliable* (Harman & Kulkarni, 2007). But these guarantees do not strictly say anything about what we can infer *after* the learning—about the hypothesis that has *actually* been selected, other than that it has been selected by a reliable method (von Luxburg & Schölkopf, 2011, p. 699f). For instance, the "wysiwyg" guarantee of uniform convergence does not strictly say anything about what we have gotten when we see the result. It is easy to misinterpret such guarantees.

A second possible reservation is that the guarantees may be overly weak. In particular, since the guarantees are agnostic about the true distribution, they are worst-case bounds that for many real-world situations—where we feel we can exclude certain classes of ("pathological") distributions—would be overly loose or pessimistic (von Luxburg & Schölkopf, 2011, p. 680, pp. 683f).[25] This motivates, for instance, the study of "fast rates" under further assumptions on the distributions[26] and of guarantees that hold for all distributions but are still distribution-dependent.[27] These studies yield a more complicated picture of the justification for ERM.

Still, the fundamental theorem gives at least *a* plausible theoretical justification for the method of ERM. This does not exclude that one might (also) wish for different kinds of justification, in any specific problem or in general. In any case, my aim here is to flag the above qualifications, to make clear that accepting the justification for ERM, and indeed the justification for simplicity to follow, presupposes accepting those qualifications. One can reject the argument below simply by rejecting the presuppositions of the statistical learning theory framework. For instance, the reasons Kelly (2008, 2011) offers for rejecting arguments for Occam's razor from statistical

[25] That sample sizes can in practice be much smaller was already shown by early experimental results (Cohn & Tesauro, 1992).

[26] Or more precisely, on the relation between the distribution, hypothesis class, and loss function [see T. van Erven et al. (2015)].

[27] This is the idea of the "theory of universal learning" of Bousquet et al. (2021). Their motivation is that the usual "distribution-independent definition of learnability is too pessimistic to explain practical machine learning," showing the need for alternatives that "better capture the practice of machine learning, but still give rise to a canonical mathematical theory of learning rates" (ibid., p. 533).

learning theory are essentially a rejection of the predictive framework.[28] But, I will argue, conditional on those qualifications—including the presuppositions of the framework—we can formulate a justificatory argument.

There is, however, a final aspect to this justification that requires more discussion. Namely, it can be seen to cover only one side of an inevitable trade-off.

4.1.2 The Bias-Complexity Trade-Off

Recall that the notion of learnability of Sect. 2.3 above concerns learning a hypothesis with *relatively* low risk among those hypotheses in given $\mathcal{H}$. Formally, it concerns finding $\hat{h}$ that minimizes the estimation error $L_{\mathcal{D}}(\hat{h}) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$.

Intuitively, this concerns the avoidance of *overfitting*. If a hypothesis class $\mathcal{H}$ has overly high capacity, then for any given data sample, the empirically best hypothesis in $\mathcal{H}$ is likely to overfit to random noise in the sample, in which case it is actually significantly worse than the best—lowest-risk—hypothesis in $\mathcal{H}$. Learnability basically concerns the avoidance of such overfitting, and the fundamental theorem then says that overfitting is avoided if $\mathcal{H}$ is a VC class.

But this leaves out the other direction of error, namely the *underfitting*. Formally, this concerns the *approximation error*, or the (*absolute*) risk $\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ of the best hypothesis in $\mathcal{H}$. The absolute risk of the selected hypothesis can be trivially decomposed as the sum of the two types of errors,

$$L_{\mathcal{D}}(\hat{h}) = \underbrace{\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{\text{approx. error}} + \underbrace{L_{\mathcal{D}}(\hat{h}) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{\text{est. error}} . \tag{10}$$

The opposing pull of these two error terms is also referred to as the *bias-complexity trade-off*. A lower complexity—lower capacity—class excludes more possibilities, and as such embodies, in machine learning terminology, a stronger *inductive bias*.[29] The fundamental theorem yields a guarantee about finding the best in a given class,

[28] Kelly (2008, p. 329) writes that Occam's razor "should help one to select the true theory from among the alternatives," whereas arguments based on risk minimization do not concern "theoretical truth" but "passive prediction" (ibid., 335). "But beliefs are for guiding action and actions can alter the world so that the sampling distribution we drew our conclusions from is altered as well" (ibid.); moreover, "it is clear that the over-fitting story depends, essentially, upon noise in the data [...] One would prefer that the connection between simplicity and theoretical truth not depend essentially upon randomness" (ibid.). These points are all well-taken [in particular the problem of distribution-shift has recently received more attention, Wiles et al. (2022)], but are all already concerns about the scope of statistical learning theory itself.

[29] This is a bit more general than the *bias-variance* trade-off (see Hastie et al., 2009, Sects. 2.6, 2.9). Complexity and (inductive) bias are here in the first instance used as informal terms (even if complexity can be made precise as VC dimension, and some authors refer to the approximation error itself as the bias), while the bias and variance in the latter are well-defined statistical terms in regression with mean squared error.

but this is inevitably a class with some inductive bias. The resulting justification must therefore be relative to this class or inductive bias.[30]

4.1.3 Model-Relative Vs. Absolute Justification

In the terminology of Sterkenburg and Grünwald (2021), this is a *model-relative* justification. It is indeed a justification for a learning method, ERM, that is explicitly *model-dependent*. The method of ERM is a general procedure—a "generic learning rule" (Shalev-Shwartz & Ben-David, 2014, p. 68)—yet one that must, by definition, on each application be supplied with further assumptions. We can view ERM as instantiating a two-place function, that apart from a data sample, also takes a particular hypothesis class. On each specific application, it must be supplied with a hypothesis class or *inductive model* that constitutes further (context-dependent) assumptions, the inductive bias. Correspondingly, the learnability guarantee for ERM is model-relative, because the notion of learnability is relative to a given inductive model or VC hypothesis class.

The analysis of Sterkenburg and Grünwald aims to explain how general learning-theoretic guarantees for generic algorithms are consistent with the skeptical import of the so-called no-free-lunch theorems of supervised learning (going back to Schaffer, 1994; Wolpert, 1992, 1996). Modern versions of these results (Shalev-Shwartz & Ben-David, 2014, p. 36ff; Sterkenburg & Grünwald, 2021, p. 9990f) say that there can exist no *universal* learning algorithm: every particular algorithm is inadequate in some possible learning situations, situations where *another* algorithm *is* adequate. And since there can be no a priori justification for privileging particular learning situations, so the further interpretation goes, there can be no theoretical justification for any particular algorithm.

However, rather than the generic yet model-dependent ERM algorithm, the no-free-lunch statement applies to *any particular instantiation* of ERM with an inductive model $\mathcal{H}$, any particular one-place "data-only" function $\mathrm{ERM}_{\mathcal{H}}$. The no-free-lunch result of Shalev-Shwartz and Ben-David (2014, Thrm. 5.1) essentially states that for any specific inductive model $\mathcal{H}$, the data-only algorithm $\mathrm{ERM}_{\mathcal{H}}$ is inadequate (i.e., with high probability suffers high absolute error) for some situations (i.e., for some true distributions; informally, those that do not match $\mathcal{H}$'s inductive bias) where another data-only algorithm, like $\mathrm{ERM}_{\mathcal{H}}$ for *another* inductive model $\mathcal{H}'$ (that *does* match the situation), is adequate.[31]

[30] Of course, the hypothesis class is not the only—or even the most important (von Luxburg & Schölkopf, 2011, p. 684)—way in which assumptions or biases enter: as discussed in Sect. 4.1 above, important modeling assumptions must already be made in the formalization of the learning problem (including choice of feature space and loss function). But discussions of inductive bias (in particular around the no-free-lunch theorems introduced shortly) usually assume that these elements are already in place, and center on the further inductive assumptions required.

[31] Another way of casting this result is that the class of *all* classifiers is not learnable: since, for any possible distribution, this class has minimum approximation error, its learnability (guarantee of low estimation error) would guarantee low absolute error for any possible distribution. In fact, the proof of Shalev-Shwartz and Ben-David (2014, thrm. 5.1) already shows the failure of learnability of classes with infinite VC dimension, and the no-free-lunch theorem is in their presentation part of the proof of the fundamental theorem (ibid., pp. 45ff).

In other words, while we can have a model-relative justification for model-dependent algorithms (of the rough form, "for any instantiated inductive model of the right form, works well relative to the model"), there is no *absolute* justification ("works well whenever") for any inductive model. The impossibility of such absolute justification is also an important part of the argument of Domingos (1998, 1999) against a possible justification of Occam's razor.

4.1.4 The Failure of an Absolute Justification for Simplicity

Domingos takes issue with what he calls the "second razor": that "given two [hypotheses] with the same [empirical] error, the simpler one should be preferred because it is likely to have lower generalization error" (1999, p. 410). (To be distinguished from the "first razor," that "the simpler [hypothesis] should be preferred because simplicity is desirable in itself," ibid.)[32] He writes that theoretical "zero-sum arguments"—no-free-lunch theorems—"imply that the second razor cannot be true" (1999, p. 413). Namely, "they imply that, for every domain where a simpler [hypothesis] is more accurate than a more complex one, there exists a domain where the reverse is true, and thus no argument which is preferable in general can be made" (ibid.).

This is true, and an expression of the impossibility of absolute justification, applied to the choice of a simple single hypothesis.[33] However, I already observed that the relevant notion of simplicity attaches to hypothesis classes, not single hypotheses. Domingos's critique of the theoretical "PAC-learning argument" for the second razor, which is also a no-free-lunch observation, is more relevant:

> " [uniform convergence results] only say that if we select a sufficiently small set of [hypotheses] prior to looking at the data, and by good fortune one of those [hypotheses] closely agrees with the data, we can be confident that it will also do well on future data. The theoretical results give no guidance as to how to select that [hypothesis class]." (1999, p. 410)

This is again true, and an expression of the impossibility of absolute justification, applied to the choice of (ERM instantiated with) a simple hypothesis class. For a sufficiently small (low-capacity) set of hypotheses, the fundamental theorem gives a guarantee of probably finding the near-best hypothesis in the class. But this hypothesis is only good in an absolute sense (has low risk and so "will also do well on future data") if the class of hypotheses was good to begin with (contains a hypothesis that has low risk and so "closely agrees with the data"). By the no-free-lunch results, we know that for any hypothesis class there are learning situations such that the class is not good. And the theory does not guide us towards a good hypothesis class prior to looking at the data.

[32] We are here, of course, likewise concerned with an *epistemic* justification for Occam's razor, not with claims that simplicity is better because it points to hypotheses or theories that are easier to work with or more aesthetically pleasing (cf. Sober, 2015, pp. 58f).

[33] This point has also been brought out experimentally [e.g., Schaffer (1993); Webb (1996)].

In sum, an absolute justification of (ERM with specific) choice of simple hypothesis class is impossible. Nevertheless, this leaves a model-relative justification of ERM. Such a justification, I will now argue, is still of epistemological interest, and indeed points towards a qualified justification of simplicity.

4.2 Epistemology and Machine Learning Theory

4.2.1 "Model-Relative" Epistemology

Absolute justification is central to a general epistemological project where we are concerned with the foundations of our knowledge. This is a project of turning back, of retracing the justificatory basis for a statement or belief of interest. In the context of machine learning algorithms, we ask: what is the basis, the justification, for trusting what our learning algorithm returns? By the no-free-lunch theorems, we know that our learning algorithm's outputs are grounded in a particular inductive bias. So we are led to ask: what is the foundation, the justification, for this inductive bias? Accepting the Humean argument that neither deductive nor nondeductive justification is forthcoming, we are ultimately led to skepticism (Sterkenburg and Grünwald, 2021, p. 9992ff).

Much of modern philosophy of science implicitly or explicitly views this general epistemological project as a dead end. Reichenbach, in the words of van Fraassen (2000, p. 254), thought that empiricist epistemology must reject "Rationalism's stringent criterion of adequacy: that an epistemology must show how absolutely reliable knowledge is possible." Van Fraassen (1989; 2000; 2004) himself offers an outlook of an empiricism in explicit rejection of "defensive epistemology" which "concentrates on justification, warrant for, and defence of one's belief" (1989, p. 170). Peirce, in the words of Levi (1998, p. 177), "explicitly dismissed doxastic skepticism when he observed that merely writing down a question challenging some current assumption is not sufficient to create the sort of doubt that should occassion an inquiry." Levi (1980; 2004) himself further develops Peirce's doubt-belief model of inquiry in explicit rejection of "pedigree epistemology," under which "one is obliged to justify current beliefs" (2004, p. 11). The broad alternative project that arises is a pragmatist one where we take seriously that one will always already start with a body of beliefs that one at that time does not actively or genuinely doubt. The interesting question is not whether one actually has an ultimate justification for these beliefs. The interesting question is how to proceed from these beliefs: how to *improve* these beliefs. This leads to an epistemological project that investigates how to improve (refine, revise, update) beliefs in the light of new data. And in this project there is still an interesting question of justification: we can theorize about better and worse ways of doing so.[34,35]

The same sentiment can be found in the machine learning literature. Already thirty years ago, Russell (1991, p. 45), after discussing the infeasibility of "tabula rasa" learning, writes that "the picture that is currently fashionable in machine learning is that of an agent that *already knows something* and is trying to learn some more." In similar vein, Domingos (2012, p. 81) himself, after attributing to Hume and Wolpert the insight that "data alone is not enough," writes that "induction (what learners do) is a knowledge lever: it turns a small amount of input knowledge into a large amount of output knowledge." Importantly, these authors, like also Shalev-Shwartz and Ben-David (2014, p. 94), do not just point out the impossibility of inductive inference without assumptions. They presuppose that there is always a starting point of initial knowledge, and put to one side the question of the actual basis for this supposed knowledge. They instead take machine learning to be about how to best proceed from initial knowledge: how to learn more, or turn input knowledge into more output knowledge.

The model-relative guarantees derived within the theory of machine learning serve exactly such a perspective. Model-relative guarantees concern algorithms that presuppose the instantiation, in each application, of an inductive model, that codifies specific prior knowledge. And these guarantees show that such algorithms are good relative to the instantiated inductive model, relative to the prior knowledge. This fits a picture where any real-world learning problem arises in a context where we already take many things for granted, and are willing to accept as prior knowledge. Given our starting point (our particular learning problem and goal, and the way we codify prior knowledge in a formal inductive model), the relevant theoretical model-relative guarantee advises us on how to proceed (what model-dependent algorithm to use). Learning-theoretic results thus provide a normative component to this general epistemological perspective on learning methods.[36]

[34] This broad epistemological approach is certainly not particular to the authors I sampled here. The repudiation of quests for absolute justification is found in the writings of many other prominent philosophers, like Popper ("[t]he piles are driven down from above into the swamp," Popper, 2002/1959, p. 94); the "model-relativity" of all our knowledge is not just central to modern authors on induction like Howson (2000), but already to Carnap and indeed to Kant.

[35] Sober (2015, pp. 86f) also sets apart, with reference to Neurath's boat, a "foundationalist" picture from a "more defensible" alternative picture where we "now have numerous beliefs about the world. Our task is to take new observations into account so as to improve our system of beliefs. We don't start from zero; we start from where we are" (ibid., p. 87). Notably, though, he presents these two pictures as two different versions of *Bayesianism*; there is no reference to these different epistemological views in his discussion of frequentist ideas. Other authors, including Levi and van Fraassen, also assume a Bayesian framework. I do not do that here: I am not after a formal account of the wider epistemic context of agents using machine learning methods to inform their beliefs. I am here interested in the epistemological lessons we may draw from a formal account of the machine learning methods themselves, namely, machine learning theory.

[36] Following up on footnote 29, prior knowledge is not only instantiated in the inductive model qua hypothesis class, but also already in the other formal components of the learning problem. But clearly learning-theoretic guarantees must always be relative to these choices as well.

4.2.2 The Theory and the Practice

In evoking a wider epistemological perspective on machine learning, we have a responsibility to do justice, not only to the normative role of the theory, but also to the actual practice of (modern) machine learning. Here arises a worry that the preceding picture is excessively neat. For one thing, it is not exactly standard machine learning practice to carefully specify an inductive model strictly on the basis of well-formulated domain-specific assumptions; the practice, say in deep learning, is to a significant extent one of trial-and-error of different general architectures and hyperparameters, that codify largely ill-understood inductive biases. For another, the kind of justification that practitioners offer is less based on the (formal) properties of the learning algorithm than on the empirical performance of the output classifier (the trained model), in the first instance on a separate test data set or in a cross-validation procedure.[37]

A full reply to this worry would have to engage more systematically with the relation between the theory and the practice of machine learning, which I do not attempt in this paper. Here I will just observe the following. Whatever the several back-and-forths of design and evaluation in an actual machine learning pipeline, the core remains the training of an algorithm to return a classifier that generalizes well. And if there is one theoretical lesson that will always stand, then it is that the algorithm must possess restrictive inductive biases, and that the algorithm will only do well if the inductive biases are appropriate. It is also nothing short of a practical necessity to constrain, amidst all the further guesswork and trial-and-error, the possible inductive models to at least some extent; and this will still always involve at least some amount of knowledge about what is likely to be appropriate in the current problem. Going further, we can make the minimal normative point that it is indeed *better* to try and implement inductive biases that are aligned with what we believe or are prepared to assume about the relevant domain. With all the qualifications and idealizations involved in linking the theory to the practice (which also ties back to the qualifications listed in Sect. 4.1.1 above), if there is a normative role for the theory to play, then it is in grounding learning procedures that can capitalize on an inductive model encoding inductive assumptions, fitting in a broad "model-relative" epistemological picture on machine learning methods.

4.2.3 "Means-Ends" Epistemology

The idea that theoretical learning guarantees provide the basis for a normative epistemology is also central to the philosophical tradition best known as *formal learning theory* (Genin, 2018; Kelly, 1996, 2016; Schulte, 2017). I will briefly make a connection to this tradition, in order to identify another crucial ingredient for the qualified justification for a simplicity preference that I will propose after.

[37] This turn away from explicit modeling and towards predictive performance is what is indeed often seen to separate machine learning from "traditional" statistical inference (Breiman, 2001).

The core principle of formal learning theory, which also has its roots in machine learning theory,[38] is that inductive problems call for a context-dependent *means-ends* analysis of what epistemic notions of success (ends) are attainable with what assumptions and methods (means). Schulte (1999) therefore also speaks of "means-ends" epistemology.

This means-ends analysis is context-dependent in the sense that given a particular learning problem, which usually comes with restrictive "background assumptions," the analysis does not question these assumptions (Kelly, 1996, p. 11; Schulte, 2017, Sects. 1.1, 2.2). For a particular learning problem and notion of success, the analysis is concerned with showing that certain methods can or cannot "solve" the problem (can or cannot attain the notion of success), given the background assumptions. This again fits a "model-relative" epistemological perspective (cf. Kelly, 2016, p. 713f), where the analysis provides a model-relative justification (with the inductive model constituted by the background assumptions) for methods that solve the problem.

However, this "problem solvability analysis" of whether and which methods can solve a given learning problem is not the only possible direction of theoretical analysis (Kelly, 1996, p. 37f). Indeed, each of the different parameters at play (the learning problem, the background assumptions, the notion of success, the methods) we can either vary or keep fixed (Kelly, 2016, p. 696). In particular, we can fix a learning problem and notion of success, and ask what background assumptions are needed for a method to possibly solve the problem. Here we are after characterization results that give necessary and sufficient conditions for the attainability of (i.e., the existence of a method that attains) the relevant notion of success (Kelly, 1996, p. 74). In the words of Kelly (ibid.),

> a characterization theorem isolates exactly the kind of background knowledge necessary and sufficient for scientific reliability, given [...] the sense of success demanded. To revive Kant's expression, such results may be thought of as *transcendental deductions* for reliable inductive inference, since they show what sort of knowledge is necessary if reliable inductive inference is to be possible.

In the logical framework studied in formal learning theory, there arises a neat hierarchy where different notions of success are characterized by the topological structure of the problem and background assumptions (Kelly, 1996). In the statistical learning theory framework, the fundamental theorem characterizes the main notion of success in terms of the combinatorial structure of the background assumptions: for a method to have the success guarantee of learnability, the hypothesis class must have finite VC dimension. Thus, adopting Kelly's words, the fundamental theorem shows what sort of knowledge (form of hypothesis class) is necessary for reliable inductive inference (a learnability guarantee). This provides a means-ends reason for modelling, if we can, our background assumptions in the form of a class of hypotheses with finite VC dimension, a simple class of hypotheses. From the combination of a

[38] Specifically, the approach of *algorithmic learning theory* going back to Putnam (1965) and Gold (1967); see Jain et al. (1999).

model-relative perspective and a means-ends analysis thus arises a justification for preferring simplicity—albeit with important qualifications.

4.3 A Qualified Justification for Simplicity

We can now assemble from the building blocks of the previous sections a justificatory argument.

4.3.1 The Argument

We face a certain problem of classification, which we are prepared to model as a problem in statistical learning. We enter this problem with further prior knowledge still; and we are interested in a method that is good relative to this prior knowledge. As a formalization of what it means for a method to be good relative to prior knowledge, we adopt the model-relative notion of learnability. Now the fundamental theorem tells us that for there to exist a method with this guarantee of learnability, we need to formulate a hypothesis class, as formalization of our prior knowledge, that is a VC class—that is *simple*. Only when the hypothesis class is simple, does there exist a method with the guarantee of learnability relative to this hypothesis class. This "transcendental deduction" gives us a means-ends justification for modeling, if we can, our prior knowledge in the shape of a simple class of hypotheses.

The previous reasoning was based on the black-and-white picture of learnability or no. The quantitative version of the fundamental theorem offers a more fine-grained version; but the essence of the argument is the same. Accepting the guarantee of learnability, we recognize that stronger bounds give a stronger guarantee, and we take a method with a stronger guarantee to be better. Now the quantitative version of the fundamental theorem tells us that a VC class of lower VC dimension—a *simpler* hypothesis class—gives a stronger guarantee. This gives us a means-ends justification for modeling, to the extent we can, our prior knowledge in the shape of class that is maximally simple (of maximally low VC dimension).

4.3.2 Qualifications

The above argument comes with a series of presuppositions and restrictions in scope, most of which I have discussed earlier. The argument only applies to learning problems that fit the statistical learning theory framework, and it presumes that a learning method is (more) justified if it satisifies (stronger bounds for) the formal criterion of learnability (Sect. 4.1). In particular, it presumes the epistemological value of model-relative justification (as I have argued for in Sect. 4.2). And it presumes that the VC dimension of a hypothesis class is a plausible criterion of simplicity (as I have argued for in Sect. 3.2).

But perhaps most importantly, the final step of the argument still comes with a crucial qualifier. We have a means-ends justification for modeling our prior knowledge in the form of a simpler class, *to the extent we can*. The theory can be seen to push into one direction, towards simplicity. However, the actual context of the learning problem, and in particular our prior knowledge, acts as a check on this push towards simplicity.

The prior knowledge we have in each instance is a crucial constraint in the argument. In line with the epistemological "model-relative" perspective sketched above, we always enter a learning problem with informal prior knowledge (beliefs we have, assumptions we are willing to make), and it is this informal prior knowledge that we need to codify in some formal hypothesis class. There will always be some leeway here, simply because it will be a significant translation step from our informal knowledge to the formal object of a hypothesis class, but it is still a constraint. It will never be reasonable, for instance, to adopt a singleton hypothesis class (a maximally simple class of VC dimension 0), because that would mean there would not be a learning problem to begin with. Without this constraint by informal prior knowledge, the argument would lose the connection to the epistemic context of an actual learning problem, and collapse into the useless advice to always choose a class of VC dimension 0.

The downside is that there will be learning problems where this constraint, this check from complexity, is too strong for the simplicity argument to still be meaningful. There will be situations, in particular, where our prior knowledge is too weak to plausibly translate into a class of VC dimension sufficiently small to still yield useful bounds. If such situations are the rule rather than the exception, then this seems to be a serious restriction in the scope of the argument.

The good news is that we can embed the argument into a more general setting, presupposing a much weaker kind of prior knowledge. Namely, in rough terms, we can simultaneously evaluate a (ranked) sequence of *multiple* VC classes, where the theoretical push towards simplicity manifests itself again in a preference for lower-VC-dimension classes; but one which is now checked by the classes' empirical error. In fact, this gives a trade-off which can be automated into a learning procedure. This is Vapnik's second "inductive principle," or the method of *structural risk minimization*.

4.3.3 Structural Risk Minimization

Indeed, rather than on uniform convergence and ERM, discussions of Occam's razor in statistical learning theory tend to focus on this method (Harman & Kulkarni, 2007, ch. 3; Kelly, 2008, p. 335; Shalev- Shwartz & Ben-David,2014, Sect. 7.3; Bargagli Stoffi et al,2022). Moreover, this method is directly applicable to the problem of model selection, which takes center stage in the modern philosophical debate about Occam's razor (Forster & Sober, 1994; Sober, 2015). Very briefly, the formal route to the method of structural risk minimization (SRM) is the following.

We start with a generalization of uniform convergence (Definition 3 above). This generalization applies to a weighted countable sequence of VC classes, and gives a uniform accuracy bound for each hypothesis which depends on the class that the hypothesis is in (Shalev-Shwartz & Ben-David, 2014, Thrm. 7.4). Next, similar to how the ERM method is defined as minimizing a uniform convergence bound, which leads to a minimization of empirical risk (Sect. 2.4 above), the SRM method is defined by minimizing a generalized uniform convergence bound, which leads to a minimization of a function of empirical risk *and* VC dimension (ibid., p. 62). This, finally, translates into a bound on SRM's performance; and we can also prove a certain weaker guarantee of *nonuniform* learnability (ibid., def. 7.1, thrm. 7.5).

I think that SRM can be understood as implementing a parallel application, over multiple VC classes at the same time, of the core argument for simplicity that I gave above. The theoretical push towards simplicity is again a push towards hypotheses from classes of lower VC dimension. This push is again checked by the classes' adequacy for the learning problem, or the adequacy of the inductive assumptions they codify. However, this is now not done by an informal evaluation of how well the formal inductive assumptions match our background knowledge. Instead, this adequacy is directly estimated by empirical error. This gives rise to a quantitative trade-off between these two elements, which SRM automates. By taking into account empirical errors, SRM can thus be seen to automate the evaluation of the adequacy of the inductive assumptions in individual VC classes; but of course the whole procedure is still relative to an initial choice of a weighted sequence of VC classes, which now constitutes the inductive model.

Clearly, it needs more work to spell out this view in the proper amount of detail, including how it fits with earlier discussions of Occam's razor in SRM, and with the philosophical discussion of Occam's razor in model selection methods. But that work must wait for another occasion.

5 Conclusion

In this paper, I described a means-ends justificatory argument for Occam's razor from statistical learning theory, based on the method of empirical risk minimization (ERM). I think this argument accomodates both intuitions and arguments in the machine learning literature in favor of the possibility of such a justification (by actually providing one) and those against (in its various qualifications). It is an honest epistemic justification, that connects a simplicity preference to guarantees of predictive accuracy; and it does not rely on any extra-theoretical assumptions about the true or best hypotheses being simple. But it does presuppose acceptance of the justificatory force of the statistical learning theory framework, including the model-relative nature of the theoretical guarantees. And both the notion of simplicity (as pertaining to classes rather than individual hypotheses) and the means-ends nature of the argument (pushing for simplicity as a constraint on inductive assumptions, rather than directly on inductive conclusions) is perhaps different and more minimal than what one might have hoped from an argument for Occam's razor.

In its essence, the argument capitalizes on a push from the mathematical theory towards specifying a simple hypothesis class. The theoretical push towards simplicity, however, is checked or indeed opposed by the informal knowledge one brings to the learning problem, which will generally rather pull in the direction of complexity. This points at a certain trade-off, which is in fact automated in the method of structural risk minimization (SRM); and this methods thus appears to directly implement a simplicity preference in inductive inference. It is indeed this method that is usually brought up in discussions of Occam's razor in statistical learning theory, also as representative for a wider family of regularization techniques in machine learning; further, it is closely related to the statistical methods for model selection discussed in the modern philosophical debate. To complete the case that the current account accomodates earlier arguments and intuitions pro and con, it therefore needs to be spelled out how exactly it figures in the method of SRM. This I intend to do in future work; but I think the core argument is already here.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Agarwal, S., Ananthakrishnan, N., Ben-David, S., Lechner, T., & Urner, R. (2020) On learnability wih computable learners. In A. Kontorovich and G. Neu, (eds), *Proceedings of the 31st International Conference on Algorithmic Learning Theory (ALT 2020)*, volume 117 of *Proceedings of Machine Learning Research*, pp. 48–60

Alpaydin, E. (2020) *Introduction to machine learning*. Adaptive computing and machine learning. MIT Press, Fourth Edition.

Angluin, D., & Laird, P. (1988). Learning from noisy examples. *Machine Learning, 2*, 343–370.

Anthony, M., Biggs, N. (1992) Computational learning theory, Volume 30 of *Cambridge Tracts in Theoretical Computer Science*. Cambridge University Press

Baker, A. (2022) Simplicity. In E. N. Zalta (Eds), *The Stanford encyclopedia of philosophy*. Metaphysics research lab, Stanford University, Summer 2022 Edition.

Bandyopadhyay, P. S., Forster, M. R. (2011) *Philosophy of statistics*, Volume 7 of *Handbook of the Philosophy of Science*. Elsevier.

Bargagli Stoffi, F. J., Cevolani, G., & Gnecco, G. (2022). Simple models in complex worlds: Occam's razor and statistical learning theory. *Minds & Machines, 32*(1), 13–42.

Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: A statistical viewpoint. *Acta Numerica, 30*, 87–201.

Beisbart, C., & Räz, T. (2022). Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass, 17*(6), e12830.

Belkin, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica, 30*, 203–248.

Berner, J., Grohs, P., Kutyniok, G., & Petersen, P. (2022). The modern mathematics of deep learning. In P. Grohs & G. Kutyniok (Eds.), *Mathematical aspects of deep learning* (pp. 1–111). Cambridge University Press.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M.K. (1986) Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. In J. Hartmanis (eds), *Proceedings of the 18th Annual ACM Symposium on Theory of Computing (STOC '86)*, pp. 273–282.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam's razor. *Information Processing Letters, 24*, 377–380.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery, 36*(4), 929–965.

Bonk, T. (2023). Functionspaces, simplicity and curve fitting. *Synthese, 201*, 58.

Bousquet, O., Hanneke, S., Moran, S., van Handel, R., Yehudayoff, A. (2021) A theory of universal learning. In S. Khuller, V. V. Williams (Eds), *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21)*, pp. 532–541. ACM.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231.

Cherkassky, V., & Mulier, F. (2007). *Statistical learning from data: Concepts, theory, and methods* (2nd ed.). Wiley.

Cohn, D., & Tesauro, G. (1992). How tight are the Vapnik-Chervonenkis bounds? *Neural Computation, 4*(2), 249–269.

Corfield, D., Schölkopf, B., & Vapnik, V. (2009). Falsificationism and statistical learning theory: Comparing the popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science, 40*(1), 51–58.

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers, 14*(3), 326–334.

Devroye, L., Györfi, L., Lugosi, G. (1996) *A probabilistic theory of pattern recognition*, volume 31 of *Applications of mathematics: stochastic modelling and applied probability*. Springer.

Domingos, P. (1998) Occam's two razors: The sharp and the blunt. In R. Agrawal, P. E. Stolorz, and G. Piatetsky-Shapiro (Eds), *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98)*, pp. 37–43. AAAI Press.

Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery, 3*(4), 409–425.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM, 55*(10), 78–87.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). Wiley.

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*(1), 1–35.

Genin, K. (2018) *The topology of statistical inquiry*. PhD Dissertation, CMU Pittsburgh.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*(5), 447–474.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Adaptive computation and machine learning, MIT Press.

Grünwald, P.D. (2007) *The minimum description length principle*. MIT series in adaptive computation and machine learning. MIT Press.

Hardt, M., Recht, B. (2022) *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press.

Harman, G., & Kulkarni, S. (2007). *Reliable reasoning: Induction and statistical learning theory*. The Jean Nicod lectures. A Bradford book, MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009) *The elements of statistical learning: Data mining, inference, and prediction*. Springer series in statistics. Springer, Second Edition.

Herrmann, D. A. (2020). PAC learning and Occam's razor: Probably approximately incorrect. *Philosophy of Science, 87*(4), 685–703.

Howson, C. (2000) *Hume's problem: Induction and the justification of belief*. Oxford University Press.

Jain, S., Osherson, D. N., Royer, J. S., Sharma, A. (1999) *Systems that learn: An introduction to learning theory*. A Bradford book. MIT Press, 2nd Edition.

Jeffreys, H. (1939). *Theory of probability*. Clarendon Press.

Kearns, M. J., & Vazirani, U. V. (1994). *An introduction to computational learning theory*. MIT Press.

Kelly, K.T. (1996) *The logic of reliable inquiry*. Logic and computation in philosophy. Oxford University Press, 1996

Kelly, K. T. (2008). Ockham's razor, truth, and information. In J. F. van Benthem & P. W. Adriaans (Eds.), *Handbook of the philosophy of information* (Vol. 8, pp. 321–360). Elsevier.

Kelly, K. T. (2011) Simplicity, truth, and probability. In *Bandyopadhyay and Forster* (2011), pp. 983–1024.

Kelly, K.T. (2016) Learning theory and epistemology. In H. Arló-Costa, V. F. Hendricks, and J. F. A. K. van Benthem, (Eds), *Readings in formal epistemology*, volume 1 of *Graduate Texts in Philosophy*, pp. 695–716. Springer, 2016.

Levi, I. (1980). *The enterprise of knowledge: An essay on knowledge, Credal probability, and chance*. MIT Press.

Levi, I. (1998). Pragmatism and change of view. In C. Misak (Ed.), *Pragmatism* (pp. 177–201). Cambridge University Press.

Levi, I. (2004) *Mild contraction: Evaluating loss of information due to loss of belief.* Clarenford Press.

Li, M., & Vitányi, P. M. B. (2008) *An introduction to Kolmogorov complexity and its applications*. Texts in computer science. Springer, 3rd edition.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mohri, M., Rostamizadeh, A., Talwalkar, A. (2018) *Foundations of machine learning*. Adaptive computation and machine learning. MIT Press, Second edition.

Pearl, J. (1978). On the connection between the complexity and credibility of inferred models. *International Journal of General Systems, 4*(4), 255–264.

Popper, K. (2002) *The logic of scientific discovery*. Hutchinson, 2002/1959. Republished, Routledge Classics.

Priest, G. (1976). Gruesome simplicity. *Philosophy of Science, 43*(3), 432–437.

Putnam, H. (1965). Trial and error predicates and the solution to a problem of Mostowski. *Journal of Symbolic Logic, 30*(1), 49–57. https://doi.org/10.2307/2270581

Romeijn, J.-W. (2017). Inherent complexity: A problem for statistical model evaluation. *Philosophy of Science, 84*(5), 797–809.

Russell, S. (1991). Inductive learning by machines. *Philosophical Studies, 64*(1), 37–64.

Schaffer, C. (1993). Overfitting avoidance as bias. *Machine Learning, 10*, 153–178.

Schaffer, C. (1994) A conservation law for generalization performance. In W. W. Cohen and H. Hirsch, (Ed.) *Proceedings of the 11th International Conference on Machine Learning (ICML 1994)*, pp. 259–265, Morgan Kaufmann.

Schulte, O. (1999). Means-ends epistemology. *The British Journal for the Philosophy of Science, 50*(1), 1–31.

Schulte, O. (2017). Formal learning theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Metaphysics Research Lab, Stanford University.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Sober, E. (2015). *Ockham's razors: A user's manual*. Cambridge University Press.

Steel, D. (2009). Testability and Ockham's razor: How formal and statistical learning theory converge in the new riddle of induction. *Journal of Philosophical Logic, 38*(5), 471–489.

Steel, D. (2011). Testability and statistical learning theory. In *Bandyopadhyay and Forster* (2011), pp. 849–861.

Sterkenburg, T. F. (2016). Solomonoff prediction and Occam's razor. *Philosophy of Science, 83*(4), 459–479.

Sterkenburg, T. F. (2018). *Universal prediction: A philosophical investigation*. PhD Dissertation, University of Groningen.

Sterkenburg, T. F. (2022). On characterizations of learnability with computable learners. In P.-L. Loh and M. Raginsky (Eds), *Proceedings of the Thirty-Fifth Conference on Learning Theory (COLT 2022), volume 178 of Proceedings of Machine Learning Research*, pp. 3365–3379. PMLR.

Sterkenburg, T. F., & Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese, 199*, 9979–10015.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the Association for Computing Machinery, 27*(11), 1134–1142.

Van Erven, T., Grünwald, P. D., Mehta, N. A., Reid, M. D., & Williamson, R. C. (2015). Fast rates in statistical and online learning. *Journal of Machine Learning Reseach, 16*(54), 1793–1861.

van Fraassen, B. C. (1989). *Laws and symmetry*. Clarendon Press.

van Fraassen, B. C. (2000). The false hopes of traditional epistemology. *Philosophy and Phenomenological Research, 60*(2), 253–280.

van Fraassen, B. C. (2004). *The empirical stance*. The terry lectures: Yale University Press.

Vapnik, V. N. (1998). *Statistical learning theory*. Wiley.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks, 10*(5), 988–999.

Vapnik, V.N. (2000) *The nature of statistical learning theory*. Statistics for engineering and information science. Springer, 2nd edition.

Vapnik, V.N., Chervonenkis, A.J. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16 (2): 264–280, 1971. Translation of the Russian original in *Teoriya Veroyatnostei i ee Primeneniya*, 16(2): 264–279, 1971.

von Luxburg, U., & Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Inductive logic, volume 10 of handbook of the history of logic* (pp. 651–706). Elsevier.

Webb, G. I. (1996). Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research, 4*, 397–417.

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S., Ktena, I., Dvijotham, K., & Cemgil, A.T. (2022) A fine-grained analysis on distribution shift. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, pp. 1–15.

Wolpert, D. H. (1992). On the connection between in-sample testing and generalization error. *Complex Systems, 6*, 47–94.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341–1390.