STATISTICAL LEARNING THEORY AND OCCAM'S RAZOR: REGULARIZATION

TOM F. STERKENBURG

Abstract. The principle of Occam's razor, which instructs us to prefer simplicity in inductive inference, has attracted much scrutiny both in the philosophy of science and in machine learning. In either field, however, a justification for the principle has been elusive. In this paper, building on an earlier "core argument," I spell out a justification from statistical learning theory for the procedure of regularization: for trading off fit for simplicity. The means-ends argument is that in order to profit from theoretical reliability and "what-you-see-is-what-you-get" guarantees, one must implement a certain preference for simplicity over fit. This is a genuine methodological justification, which neither collapses to a purely pragmatic principle that we prefer simplicity for its own sake, nor to an ontological assumption that the truth is simple.

1. Introduction

The methodological role of simplicity in scientific theorizing is a long-standing topic in the philosophy of science (Sober, 2015; Baker, 2022; Dubova et al., 2025). In recent years, following the work of Forster and Sober (1994), a considerable literature has spawned around the lessons we may draw from the role of simplicity in statistics, specifically from methods for model selection (Zellner et al., 2002; Sober, 2015, ch. 2; Baker, 2022, sect. 7; Sprenger and Hartmann, 2019, var. 10).

The contemporary consensus is that these lessons are limited (Kieseppä, 1997; Myrvold and Harper, 2002; Norton, 2021, chs. 6–7). The way, say, the Akaike Information Criterion (AIC) operates will not help us with, say, the rational reconstruction of a preference for the Copernican to the Ptolemaic system. Even so, the methodology of statistics, including the role of simplicity, remains of independent interest to the philosophy of science, because of the importance of statistical inference to many branches of science (Romeijn, 2025). The same holds, increasingly so, for machine learning.

In machine learning, regular reference is made to Occam's razor, understood broadly as the methodological principle to prefer simplicity (Mitchell, 1997; Duda et al., 2001; Shalev-Shwartz and Ben-David, 2014; Goodfellow et al., 2016). These references and associated claims have drawn some critical attention in the philosophy of science (Harman and Kulkarni, 2007; Herrmann, 2020; Bargagli Stoffi et al., 2022). The focus of these works is statistical learning theory (Vapnik, 2000; Shalev-Shwartz and Ben-David, 2014), arguably (still) the main mathematical framework for machine learning; and the question is whether statistical learning theory can support a justification for Occam's razor. What these works illustrate is that to date there has been no successful attempt, neither in the philosophy nor in the machine learning literature, to spell out a justification of this kind.

Date: October 23, 2025. This is a preliminary version. I welcome feedback.

This paper presents such a—*hopefully*, successful—attempt. Building on a previous "core argument" (Sterkenburg, 2025), I give a means-ends argument for the methodological norm to trade off empirical data fit with simplicity (what in machine learning is called *regularization*), in accordance with the learning rule of structural risk minimization (SRM).

The plan is as follows. I start, in section 2, with introducing the main components of statistical learning theory. These are the ingredients for the core argument, which underpins a first methodological simplicity norm ("keep the model simple"). This norm, however, is limited in its applicability, which motivates, as I discuss in section 3, a more general mathematical perspective. This perspective yields the SRM learning rule, which implements explicit regularization and as such a second methodological simplicity norm ("trade fit for simplicity"). In order to spell out the justification for this norm, I take a closer look, in section 4, at the theoretical justification for the SRM rule. I here discuss the *model-relativity* of theoretical justification, which poses a challenge to the possibility of a non-question-begging justification for a simplicity preference. I answer this challenge in section 5, where I give the means-ends argument for a genuine methodological simplicity norm. I conclude in section 6.

In the course of developing my argument, I draw from various earlier insights in the philosophical literature on model selection; and the question may arise what is the point of "supplementing the debate with a separate analysis of a method which only differs in mathematical details" (Kieseppä, 2001, p. 772). My answer is, first, that the framework of statistical learning theory is perhaps not, as Sober (2015, fn. 61) has it, "dramatically" different from that of the methods that took center stage in the earlier debate (essentially, AIC, and BIC—the Bayesian Information Criterion): but it is still different in important ways. What stands out, for instance, are the robust complexity notion of capacity, and the finite-sample bounds justifying the relevant learning rules. Second, the current work makes some progress in connecting related but independent strands in the (philosophy of) statistics and computer science literatures. Third, I do believe that my argument goes beyond what has been proposed in either literature so far.

Kieseppä (1997, p. 41) offers a meta-reflection on the current kind of project. He writes that the aim of statistics is to "provide scientists with (better or worse) methods," whereas "[p]hilosophy of science is concerned with the *justification* of scientific practices." Since we cannot tell "whether the use of some given statistical method is justified in a given situation without describing the method and the background assumptions of the results on which it is based in a rigorous way," Kieseppä concludes that, when dealing with particular statistical methods and results, the philosophy of statistics

> must be mathematically more rigorous than statistics itself. Otherwise it is not clear why such philosophy should be pursued as a field distinct from both statistics and popular science.

I would not say that work in the philosophy of machine learning must strive to be more *mathematically* rigorous than work in machine learning itself; the mathematics, after all, is the job of theoreticians of machine learning. However, I do think there is a job for philosophy in spelling out, with higher standards of rigor, the *arguments* that *use* the maths to justify methods or methodological principles. That is, in any case, the spirit in which I proceed in this paper.

2. The core argument

Statistical learning theory counts as the "received view" in the theoretical analysis of machine learning algorithms (Grote et al., 2024, sec. 2).[1] The most basic type of learning problem, and the exclusive focus of this paper, is binary classification (section 2.1). The crucial element in the theoretical analysis is the property of uniform convergence, which underwrites the learning rule of ERM (section 2.2). The fundamental theorem of statistical learning theory connects uniform convergence and successful learning by ERM with a notion of parsimony or simplicity of sets of classifiers, the VC dimension (section 2.3). The core argument for a simplicity preference is assembled from these components (section 2.4).

2.1. **Binary classification.**

As an "imaginary and somewhat fanciful example" (Duda et al., 2001, pp. 1ff), imagine that, in the context of a wildlife preservation effort, you want to design a system to automatically keep a tally of the number of specimens of European seabass passing by in a certain coastal river. You set up an underwater camera at a particularly narrow strait, which automatically takes pictures of moving objects. Now the problem is to identify the pictures which indeed captured a seabass and not something else.

You decide to use machine learning. You first inspect a number of pictures taken, and select features which are indicative of a seabass present. Say one feature you choose is the relative size of the main object in the picture, and another is the average brightness of the picture (you have software to calculate those). Then you collect a larger sample of pictures generated over several days, and painstakingly annotate those: seabass or no. Finally, you feed this training set of pictures to a learning algorithm, which will automatically infer or learn a general classifier, which will in turn be able to classify new pictures. What is a good learning algorithm?[2]

2.1.1. *The formal framework.*

In statistical learning theory, problems of this type are formalized as follows. We have a domain $\mathcal{X}$ of *instances*, which are usually themselves vectors of real-valued *attributes*. In our example, the instances are the pictures, summarized in two real-valued features: size and brightness. If we assume these values are normalized to the unit interval, then the instances are simply points in the space $[0, 1]^2$.

We further have a *label* set $\mathcal{Y}$; in the example, there are just two labels, seabass or not (1 or 0). This is therefore an example of a *binary classification* problem.

---

[1] The theory sprang out of the pioneering work of Vapnik and Chervonenkis (1971); see (Vapnik, 1998, 2000). I will mainly follow the presentation by Shalev-Shwartz and Ben-David (2014), which is a synthesis of the Vapnik-Chervonenkis theory and Valiant's (1984) model of PAC learning.

[2] In all of the following, the focus will be on the learning or generalization step: the algorithmic inference from training data to learned classifier. The toy example already shows that this focus is inherently limited. What features you pick is obviously important to the quality of the inference; and in reality there will be several iterations of and between stages of problem formulation, algorithmic learning, and evaluation. However, this paper is concerned with the theoretical justification for Occam's razor, and statistical learning theory is a theory of generalization.

A *hypothesis* is a particular classifier, a function $h : \mathcal{X} \to \mathcal{Y}$ from all possible instances to labels.[3] A *learning algorithm* A is a function that receives a training *sample* S, a finite ordered sequence of instance-label pairs, and returns a classifier.[4]

2.1.2. *The goal.* An essential assumption in statistical learning theory is that training instances as well as new data instances are independently and identically distributed (i.i.d.) samples from some true but unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.[5] This distribution thus describes both the selection of instances (new images) and the possibly noisy relationship between features and labels (between object size and average brightness, and depicting a seabass or not).[6]

This assumption allows us to define, for any given classifier h, the probability that it misclassifies a new, randomly generated instance,

$$L_{\mathcal{D}}(h) := \mathbb{P}_{(X,Y)\sim\mathcal{D}}\left[h(X) \neq Y\right]. \tag{1}$$

This is the *true risk* of h, and the goal of learning is to find a classifier which minimizes the true risk.

2.1.3. *Pattern recognition, curve fitting, model selection.* The problem of classification is not quite the same as the problem of curve fitting usually considered in the philosophical literature (Glymour, 1980, ch. VIII; Forster and Sober, 1994). There we assume a set of noisy instances (x_i, y_i) of a functional relationship $f(x) = y$, and the goal is to infer the curve f. This kind of problem is called *regression* in machine learning.

For our purposes, however, more important are the structural similarities between these types of (in early machine learning jargon) "pattern recognition" problems. First, at a formal level, a binary classifier can be seen to select a subset of the instances (those labeled 1); in our example, a subset of the unit square. A type of classifiers are the *separators*: in our example, curves which cut the square into two parts. For instance, a linear separator is a line through the square, with the instances falling on one side of the line classified positively. A curve fitting problem is thus similar to a type of classification problem, where a classifier separates instances under the curve from instances above the curve (cf. Harman and Kulkarni, 2007, p. 65).

Second, the original curve fitting problem, and generally problems in statistical *model selection*, are usually seen as a two-step procedure. First, we select a model or a family of curves (e.g., the quadratic curves); second, we select a particular

[3] A hypothesis is (more) often called a *model* in machine learning, but due to risk of confusion with other uses of the term (inductive model, model selection) I will stick to hypothesis.

[4] Thus a learning algorithm, as defined here, is really a function. I abstract away in this paper from computational considerations.

[5] In the context of the problem of induction, the i.i.d. assumption already functions as a kind of "uniformity of nature" assumption (cf. Forster and Sober, 1994, p. 29)—though, as we will see, the problem of induction still remains (cf. Strevens, 2009, fn. 3; Sterkenburg and Grünwald, 2021).

[6] How well the theory applies to any particular learning problem depends, of course, on how plausible the i.i.d. assumption is. Grünwald (2007, p. 586) writes that "this is one of the few examples of a modeling assumption which may actually be quite realistic in some situations;" Shafer (2009) is more skeptical. Even if the assumption is reasonable for the learning process, the issue of *distribution shift* in domains of application is a threat to the stability of the definition of true risk and as such to the robustness of the learned hypothesis (Grote et al., 2024; Freiesleben and Grote, 2023).

curve from the family (e.g., the least-squared-error quadratic). In the classification problem, we also discern these two steps.

2.1.4. *Hypothesis classes.* The first step is the specification of a family or class $\mathcal{H}$ of hypotheses. For instance, we could choose the $\mathcal{H}_1^{\text{pol}}$ of linear separators, the class $\mathcal{H}_2^{\text{pol}}$ of quadratic separators, or even the class $\mathcal{H}^{\text{pol}}$ of all polynomial separators.

Having chosen a hypothesis class $\mathcal{H}$ ourselves, we now want to delegate the second step to the learning algorithm: to learn, from the training data, a hypothesis with minimal true risk among those in $\mathcal{H}$. As we will see, the analysis of what is a good learning algorithm also has ramifications for how to make our initial choice.

2.2. **Uniform convergence and ERM.** The i.i.d. assumption (sect. 2.1.2) allows us to bring in the law of large numbers, guaranteeing that, for any fixed hypothesis h in our hypothesis class $\mathcal{H}$, the empirical error of h will, as we draw larger and larger samples, converge in probability to its true risk. However, we are not interested in a fixed hypothesis. We are interested in the performance of a learning algorithm, which, depending on the data, can select different hypotheses from $\mathcal{H}$. 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*.

Definition 1 (Uniform convergence)**.** Hypothesis class $\mathcal{H}$ has the uniform convergence property if there exists a sample complexity function $m_{\mathcal{H}}^{\text{uc}} : (0,1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$ and for any $\mathcal{D}$, we have for $m \geq m_{\mathcal{H}}^{\text{uc}}(\epsilon, \delta)$ that

$$\mathbb{P}_{S \sim \mathcal{D}^m} \left[(\forall h \in \mathcal{H}) \left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon \right] \right] \geq 1 - \delta. \tag{2}$$

2.2.1. *Wysiwyg.* The uniform convergence property gives a "what-you-see-is-what-you-get" ("wysiwyg") bound: for large enough sample size, with high probability, the empirical error of each h (which can be ascertained from the training data—what you see) is indicative (up to error term ϵ) of its true risk (what you get):

$$(\forall h \in \mathcal{H}) \left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon \right]. \tag{3}$$

This property motivates and justifies a basic learning rule.

2.2.2. *Empirical risk minimization.* To draw advice from the uniform convergence bound about what hypothesis a good learning algorithm should return for any given data sample S, it is helpful to rewrite it as a bound for any given m. To that end, we define the error function $\epsilon_{\mathcal{H}} : \mathbb{N} \times (0,1) \to (0,1)$ by

$$\epsilon_{\mathcal{H}}(m, \delta) = \min\{\epsilon \in (0,1) : m \geq m_{\mathcal{H}}^{\text{uc}}(\epsilon, \delta)\}, \tag{4}$$

that is, as giving the smallest error ϵ such that an m-length sample will with specified high probability satisfy the ϵ-wysiwyg property (3),

$$(\forall h \in \mathcal{H}) \left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_{\mathcal{H}}(m, \delta) \right]. \tag{5}$$

This bound directly implies that, for all $h \in \mathcal{H}$ simultaneously,

$$L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon_{\mathcal{H}}^{\text{uc}}(m, \delta). \tag{6}$$

Vapnik's first "inductive principle" (2000, p. 20), the learning rule of empirical risk minimization (ERM), is defined by selecting a hypothesis in $\mathcal{H}$ that minimizes

this bound on the true error.[7] Since the error term is identical for all hypotheses in $\mathcal{H}$, this comes down to picking a hypothesis with minimal empirical error.

Definition 2 (ERM)**.** Given hypothesis class $\mathcal{H}$ (with the uniform convergence property). The ERM rule for $\mathcal{H}$ is defined by

$$\mathrm{ERM}_{\mathcal{H}}(S) \in \arg\min_{h \in \mathcal{H}} L_S(h). \tag{7}$$

2.2.3. *ERM's justification (1).* If all hypotheses' training errors are good indication of their true risks, then, in order to obtain an hypothesis with minimal true risk, it makes sense to select an hypothesis with minimal training error.

Formally, by the bound (5) or (6) applied to the particular hypothesis selected by ERM, we first have the wysiwyg bound

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

Next, by the definition of ERM as an empirical error minimizer, and another application of the wysiwyg bound (5),

$$L_S(\mathrm{ERM}_{\mathcal{H}}(S)) + \epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta) \leq \min_{h \in \mathcal{H}} L_S(h) + \epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta) \tag{9}$$

$$\leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta), \tag{10}$$

so that, with (8),

$$L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + 2\epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta). \tag{11}$$

This bound gives a justification for using ERM, as selecting (with high probability, up to a known and constant error term) the best hypothesis in the class.

2.2.4. *Learnability and reliability.* The notion of *PAC* ("probably approximately correct") *learnability* that is central in Shalev-Shwartz and Ben-David (2014)'s presentation follows from retranslating this bound again, fixing error instead of sample size. Namely, hypothesis class $\mathcal{H}$ is PAC learnable by $\mathrm{ERM}_{\mathcal{H}}$ if there is a sample complexity function $m_{\mathcal{H}} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$, every $\mathcal{D}$, and every $h \in \mathcal{H}$, we have for $m \geq m_{\mathcal{H}}(\epsilon, \delta, h)$ that

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon\right] \geq 1 - \delta. \tag{12}$$

We could therefore also refer to bound (11) as a learnability bound. I will rather follow Harman and Kulkarni (2007) in employing the general label of *reliability*.

2.2.5. *ERM's justification (2).* The justification for ERM is thus this reliability bound (11) of probably selecting the approximately best hypothesis in the class. But the wysiwyg bound (8) is in itself also still important, because it means that probably the training error will be an indication of how good approximately this selected hypothesis actually is. That means that if our hypothesis class is not good,

[7] Note that the ERM rule is not quite an algorithm, as defined before, because the definition leaves underdetermined how it breaks ties in case of multiple error-minimizing hypotheses. We can ignore this in the current paper, because the theoretical analysis holds for any specific implementation of the ERM rule. I also ignore the challenge of actually implementing (approximations to) ERM, which is a problem of *optimization* (see, e.g., Hardt and Recht, 2022, ch. 5).

we will probably be able to tell from the training error, and act accordingly (cf. Shalev-Shwartz and Ben-David, 2014, sect. 11.3).[8]

Caution is warranted, though, in interpreting such frequentist guarantees. A pre-sampling wysiwyg guarantee that training error will probably be close to true error does not yet entail that, after sampling and the selection of a particular hypothesis, we can infer from high (low) training error of this hypothesis that its true error is probably high (low), too. (This can be seen most directly from the fact that the probability in the latter assertion must be interpreted epistemically.) Such a conclusion, or a corresponding decision, would need an additional reasoning step, like a Fisherian disjunctive or a Neymanian behaviorial argument.[9] The same holds for the reliability guarantee that ERM probably finds the near-best hypothesis in the class. More could be said about this issue, but in this paper I will simply accept that frequentist guarantees provide a justification.[10]

In sum, we have a two-fold theoretical justification for ERM, a reliability and a wysiwyg bound—provided that the hypothesis class has the uniform convergence property. But what kind of hypothesis classes satisfy this property?

2.3. **Capacity and the fundamental theorem.** The short answer is, *simple* hypothesis classes.

2.3.1. *The VC dimension.* Notions of *capacity* in machine learning are notions of the richness or flexibility of hypothesis classes, in the sense that a higher-capacity hypothesis class can more easily fit a variety of training data. More precisely, in our framework of binary classification, for a given finite set $X = \{x_1, \ldots, x_m\}$ of instances, a hypothesis class $\mathcal{H}$ has maximal such flexibility-of-fit if, for each of the 2^m possible binary labelings of X, there is some h in $\mathcal{H}$ which gives exactly this labeling. In that case we say that $\mathcal{H}$ *shatters* X, and the relevant notion of the capacity of a hypothesis class in our framework is a measure of its ability to shatter sets of instances.

Definition 3. 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.

For example, the hypothesis class $\mathcal{H}_1^{\text{pol}}$ of linear separators has a strictly smaller VC dimension than the class $\mathcal{H}_2^{\text{pol}}$ of quadratic separators. Both are still VC classes, which is no longer so for the class $\mathcal{H}^{\text{pol}}$ of *all* polynomial separators. Another example of a class with infinite VC dimension is the class $\mathcal{H}^{\text{sine}} = \{x \mapsto \sin \alpha x\}_{\alpha \in \mathbb{R}}$ of sine functions. No matter how large a given finite set of labeled instances, the

[8]The wysiwyg justification is not emphasized in (Sterkenburg, 2025), but it will be important for the argument from SRM.

[9]See, e.g., (Sprenger, 2016). Fisher's disjunction, in this context, would be that either training and true error are close, *or* something exceptional has happened; since it is reasonable to rule out the latter, we can conclude the former. The behavioral argument would be that if we always decide to act correspondingly (e.g., discard the selected hypothesis and start over with a different class if training error is high), then in the long run we would only go astray a fraction of the time.

[10]In machine learning practice, one would normally further seek to corroborate that a selected hypothesis is good by assessing it on an independent test set. Since we then have a single fixed hypothesis, we can from the usual law of large numbers (more specifically, Hoeffding's equality) derive a tight wysiwyg bound on the probable difference between the hypothesis' test and true error (Shalev-Shwartz and Ben-David, 2014, thrm. 11.1). Of course, this is still a frequentist bound, so the interpretational issue remains.

corresponding separation in positive and negative classes can be achieved by a sine function with apposite choice of period parameter α (Vapnik, 2000, p. 82).

2.3.2. *Simplicity.* The latter example shows that this notion of capacity does not need to align with traditional notions of simplicity or parsimony in terms of number of parameters. Such an alignment still holds nicely for the polynomials, where $\mathrm{VCdim}(\mathcal{H}_i^{\mathrm{pol}}) < \mathrm{VCdim}(\mathcal{H}_j^{\mathrm{pol}})$ for degrees $i < j$; but the class of sine functions above is specified with only one parameter, yet has maximal capacity.

However, as discussed by Sterkenburg (2025, sect. 3), the capacity notion of VC dimension gives *a* plausible notion of simplicity. Capacity formalizes a conception of complexity as flexibility-of-fit, where a simpler, more parsimonious class covers fewer possible patterns.[11] Moreover, capacity expresses an "inherent" complexity of a class (Romeijn, 2017; Grünwald, 2007), because it is invariant to inessential choices of how to describe the class. This is a kind of robustness which definitions of the complexity of individual hypotheses (and definitions of the complexity of classes derived from those) must lack.[12,13]

More precisely, notions of capacity are invariant under redescriptions which leave the structure of the learning problem untouched—which is an immediate corollary of their role in formal learning guarantees.[14] Indeed, the primary motivation for the definition of VC dimension is not an independent intuitive appeal as a simplicity measure (even if, again, it has such an appeal), but a provable connection, indeed *equivalence*, to uniform convergence.

2.3.3. *The fundamental theorem.* The central result of Vapnik and Chervonenkis (1971), celebrated as the fundamental theorem of statistical learning theory (Shalev-Shwartz and Ben-David, 2014, thrm. 6.7), is that finite VC dimension characterizes uniform convergence.

Theorem 4 (The fundamental theorem)**.** *An hypothesis class has the uniform convergence property if and only if it is a VC class.*

---

[11] Dubova et al. (2025), in a recent overview of Occam's razor in science, call this *parsimony by constraints*.

[12] Definitions that seek to capture how "bumpy" or "wiggly" a single hypothesis looks are inevitably not robust under trivial 1-1 transformations of the coordinate space (Kieseppä, 2001, p. 783). Similarly (Priest, 1976), parameter counting is not robust under trivial redescriptions which introduce fewer or more parameters (Turney, 1990, sect. 7; Kieseppä, 1997, pp. 34f).

[13] Capacity is in a sense independent of the "complexity" of individual hypotheses: a class with a small number of highly "complex" hypotheses (say, high-n-degree polynomials in your favorite coordinate system) still has small capacity (see again Sterkenburg, 2025). The best that can be said is that ways of quantitatively defining the simplicity of individual objects (like in terms of description length) will normally be such that there are *more* complex than simple objects, so that the class with all objects at a certain complexity level will be larger(-capacity) with a higher level. This could go some way towards an explanation why capacity often tracks conceptions of the simplicity of individual hypotheses.

[14] For instance, Steel (2009, sect. 5) writes that while it "cannot be altered by faithful translations of the hypotheses it contains, VC dimension does depend on what counts as an individual data unit," and considers redescribing the learning problem by collapsing every pair of observations into a single one. This makes the VC dimensions smaller; but it also changes the problem, as data sizes are now artificially smaller, too. The formal connection (as given by theorem 4 below) between capacity and data sizes for bounds of uniform convergence is, as it must, preserved.

In particular, $\mathrm{ERM}_{\mathcal{H}}$ satisfies the learnability bound (11) precisely if hypothesis class $\mathcal{H}$ has finite VC dimension.[15] Further, a more fine-grained, quantitative statement of the fundamental theorem tells us that the learnability bound depends on the VC dimension. Namely, we have that, for some constant c,

$$\epsilon^{\mathrm{uc}}_{\mathcal{H}}(m, \delta) = cB, \tag{13}$$

where

$$B = \sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) - \log \delta}{m}}. \tag{14}$$

2.4. **The argument.** The methodological lesson from the fundamental theorem is that *in order to* profit from the reliability and wysiwyg guarantees, one must keep $\mathcal{H}$ simple. This is a means-ends argument for simplicity.

2.4.1. *Means-ends.* I borrow this terminology from the approach in the philosophy of science which goes by the name of *formal learning theory* or also *means-ends epistemology* (Kelly, 1996; Genin, 2018; Schulte, 2017; also see Sterkenburg, 2025, sect. 4.2). The driving idea here is that inductive problems should be analyzed in terms of what types of reliability (ends) are attainable with what assumptions and methods (means); a perspective which aligns well with machine learning theory.[16]

For instance, we can formulate, for a certain type of learning problem, an interesting notion of reliability, and ask for what kind of assumptions this end is achievable. This question would be answered by a characterization theorem, which lays down the necessary and sufficient conditions for the attainability of this end. In Kelly's words, "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" (1996, p. 74).

The fundamental theorem is such a characterization result in the our setting. It makes precise what sort of knowledge (a low-complexity hypothesis class) is necessary for reliable inductive inference (a learnability guarantee for ERM). This gives a means-ends reason to choose a simple hypothesis class.

2.4.2. *The norm.* The means-ends argument yields a methodological simplicity norm.

> **Methodological norm N1 (Occam's razor).** Keep the hypothesis class simple.

For example, (N1) advises against the use of the class $\mathcal{H}^{\mathrm{pol}}$ of all polynomials, because it is too complex—too complex to retain the justification for ERM. Moreover, it advises us to prefer the class $\mathcal{H}^{\mathrm{pol}}_1$ of linear separators to the class $\mathcal{H}^{\mathrm{pol}}_2$ of quadratic separators, because the former comes with a stronger justification—stronger guarantees—for ERM.[17]

---

[15] Learnability is indeed equivalent to learnability by ERM: ERM satisfies learnability if any algorithm does (Shalev-Shwartz and Ben-David, 2014, thrm. 6.7).

[16] The approach indeed grew out of the theoretical computer science branch of *algorithmic learning theory* (Jain et al., 1999).

[17] A slightly different gloss on the theoretical guarantees is that they "tell us *how rich a function space* we can afford to search on the 'budget' given by our sample size, while maintaining the quality of the estimate of the accuracy of the best-fitting function" (Grote et al., 2024, p. 3). But for given "budget" simpler is theoretically still better because the guarantees are stronger.

2.4.3. *Constraints and limitations.* However, if we follow this norm all the way, then we end up choosing a *maximally* simple class, a class of VC dimension 0. That clearly does not make sense in any actual learning problem, since a maximally simple class of VC dimension 0 is a singleton class with only one classifier, meaning there would be no learning problem left.

This is an extreme illustration of the general observation that a methodological norm must be applied in the context of an actual learning problem, in which further epistemic as well as pragmatic factors are at play. For one thing, practical considerations dictate what error and/or confidence bounds would be acceptable, or what training sample sizes are available to us. For another, an essential epistemic factor is what we know about the domain, which informs what assumptions we would be willing to make, and so what hypothesis classes we would still find reasonable.

The latter is especially important because the reliability justification for ERM is *model-relative* (Sterkenburg and Grünwald, 2021): it is about locating the best hypothesis in the *inductive model*, the hypothesis class. The best hypothesis in the class might still not be very good; so it is important to also choose a good class, a class which we expect contains good classifiers. In machine learning terminology, we want the inductive model to have a good *inductive bias*.

These considerations act as a check on the theoretical push towards simplicity that is encoded in (N1). The epistemic norm (N1), purely underwritten by the theoretical justification for a simple inductive model, pushes us in the direction of simplicity; but how far we can go in this direction, to what extent we can reasonably follow this methodological norm, depends on the specifics of the actual learning situation. In particular, our knowledge about the domain informs what hypothesis classes could still be expected to contain good classifiers. This knowledge normally acts as a check on or counterpull to the simplicity norm, because simpler hypothesis classes generally mean stronger assumptions.

There are thus pragmatic and epistemic constraints to the application of simplicity norm (N1). But at this point the worry arises that these factors are normally so constraining that (N1) is hardly applicable at all.

2.4.4. *Beyond the core argument.* In the older days of machine learning, we can see the simplicity norm (N1) routinely evoked and followed.[18] But certainly since the deep learning age the field is rather characterized by a lack of strong modeling assumptions. In the words of Bartlett et al. (2021), "deep learning is a data-driven approach: these are rich but generic models, and the architecture, parametrization and nonlinearities are typically chosen without reference to a specific model for the process generating the data." This does not sit well with the simplicity norm (N1), because simple (small-capacity) hypothesis classes are much more restrictive, and as such do commit one to strong assumptions. The worry is therefore that simplicity norm (N1) is no longer relevant to modern machine learning.[19]

[18]For instance, in their seminal paper introducing convolutional neural nets for handwritten digit recognition, LeCun et al. (1989, p. 541) write that "the basic design principle is to reduce the number of free parameters in the network as much as possible without overly reducing its computational power. Application of this principle increases the probability of correct generalization because it results in a specialized network architecture that has [...] a reduced Vapnik-Chervonenkis dimensionality."

[19]Another worry is that this version of Occam's razor is "a rather different recommendation than the usual exhortation to select the simplest hypothesis compatible with the data" (Grote et al., 2024, fn. 6). The simplicity norm to be discussed in the following, building on the core

Simplicity does still appear to play an important methodological role, however, even in deep learning. It may no longer play a significant role in the initial choice of hypothesis class, but it does pop up again in the subsequent learning process, namely in the standard technique of *regularization*. The theoretical underpinnings for this technique lead us to Vapnik's second "inductive principle," and a methodological simplicity argument which expands on the core argument of this section.

3. Structural risk minimization

3.1. **Generalized uniform convergence and SRM.** The push towards simplicity encoded in (N1) is checked or countered by, in particular, the assumptions we are willing to make. Simpler inductive models generally also encode stronger assumptions, stronger inductive biases: the best hypothesis in a simpler class might be expected to be worse than the best hypothesis in a more complex class. This tension between simpler classes with stronger inductive bias and more complex classes with weaker inductive bias is expressed in a classical trade-off.

3.1.1. *The bias-complexity trade-off.* The true risk of a learned hypothesis $\hat{h}$ can trivially be decomposed as a sum of two errors, the true risk of the best hypothesis in the given hypothesis class $\mathcal{H}$ (the *approximation error*) and the difference between the true risk of $\hat{h}$ and that of the best hypothesis $\mathcal{H}$ (the *estimation error*):

$$L_{\mathcal{D}}(\hat{h}) = \underbrace{\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)}_{\text{appr. error}} + \underbrace{L_{\mathcal{D}}(\hat{h}) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)}_{\text{est. error}}\,. \tag{15}$$

The interplay between these two terms is depicted in figure 1. The simplicity norm (N1) is concerned with minimizing the estimation error, or preventing *overfitting*: the selection of a hypothesis which is significantly worse than the best in the class. It pushes us to the left along the x-axis, to simpler hypothesis classes, with (provably, with high probability) smaller difference between the two curves. However, at a more informal level, less complex classes might be expected to contain fewer good hypotheses, so that the approximation error is higher. To prevent *underfitting*, when even the best hypothesis in the class is not good, we would need to move to the right, towards more complex classes.

There is an asymmetry here, in that the theory, and the resulting epistemic norm (N1), covers only one side, namely the minimization of estimation error. In practice, we would want to find the sweet spot in the figure, where the selected hypothesis does not just have low estimation error, but low total error, low true risk. This asks for a more informal assessment what class, given what we know about the specific learning problem, is likely to have good hypotheses, to have low approximation error. Such an assessment constrains the application of norm (N1).[20]

Purely theoretical analysis cannot tell us how to strike the right balance in the choice of hypothesis class prior to the learning. However, it turns out that, to some extent, the theory can tell us how to strike the balance *during* the learning. The

argument but underwriting regularization in the learning, is a more direct instantiation of the usual Occam norm to prefer simplicity in the inductive inference from data to hypothesis.

[20]The asymmetry in the theoretical focus on estimation error could thus be seen as problematic for the reliability justification for a simple class: the model-relative reliability of ERM counts for little if the approximation error is high. This, however, leaves untouched the wysiwyg justification: it will always be helpful to get an indication of *how* good or bad the (best in the) class is. Also see (Grünwald, 2007, p. 34) on "the inherent difference between under- and overfitting."

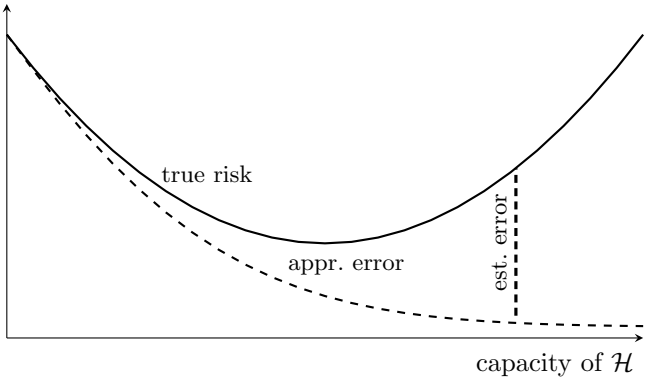

FIGURE 1. The bias-complexity trade-off.

key is a generalization of the uniform convergence theorem, that moves the analysis to the level of *multiple* hypothesis classes.

3.1.2. *Generalized uniform convergence.* Instead of a single hypothesis class somewhere on the complexity-axis of figure 1, imagine we pick a countable *sequence* of different hypothesis classes along the axis. Formally, let $(\mathcal{H}_n)_{n\in\mathbb{N}}$ be such a sequence, where each $\mathcal{H}_n$ is a VC class. I will refer to the union $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$ as the *super*class, and to the individual $\mathcal{H}_n$ as the *sub*classes.

By the fundamental theorem, each subclass $\mathcal{H}_n$ has the uniform convergence property, with a corresponding accuracy function ϵ_n^{uc}. Additionally, we define a *weight function* $w : \mathbb{N} \to [0, 1]$ with $\sum_n w(n) \leq 1$, assigning each subclass $\mathcal{H}_n$ a numerical weight $w(n)$ in the unit interval.

For instance, we could choose the sequence $(\mathcal{H}_n^{\text{pol}})_{n\in\mathbb{N}}$ of all the subclasses of polynomial separators of degree n. Each $\mathcal{H}_n^{\text{pol}}$ is a VC class; and the superclass $\cup_{n\in\mathbb{N}}\mathcal{H}_n^{\text{pol}}$ is the class $\mathcal{H}^{\text{pol}}$ of all polynomials. For the weights, we could, for instance, pick the function $w : n \mapsto 2^{-n}$, which satisfies the property that the sum of all weights does not exceed 1.

Now one can show the following.[21]

Theorem 5 (Generalized uniform convergence)**.** *Given hypothesis subclass sequence $(\mathcal{H}_n)_{n\in\mathbb{N}}$ such that each $\mathcal{H}_n$ has the uniform convergence property with accuracy function ϵ_n^{uc}, and weight function w. Then for all $\delta \in (0, 1)$ and all $\mathcal{D}$ we have with probability at least $1 - \delta$*

$$(\forall n \in \mathbb{N})(\forall h \in \mathcal{H}_n)\left[|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n^{\text{uc}}(m, w(n)\delta)\right], \tag{16}$$

and so in particular, for $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$,

$$(\forall h \in \mathcal{H})\left[L_{\mathcal{D}}(h) \leq L_S(h) + \min_{n:h\in\mathcal{H}_n} \epsilon_n^{\text{uc}}(m, w(n)\delta)\right]. \tag{17}$$

Like in the original uniform convergence result, this wysiwyg bound uniformly holds for all $h \in \mathcal{H}$ simultaneously, but in this case with an accuracy that does depend on the hypothesis subclass(es) that h is in. It depends on h, first, because of the accuracy function ϵ_n^{uc} associated with a subclass $\mathcal{H}_n$ that contains h. Second, it

[21] This is theorem 7.4 in (Shalev-Shwartz and Ben-David, 2014).

depends on h because of the factor $w(n)$ that is applied to the confidence parameter δ in the accuracy function.

3.1.3. *Structural risk minimization.* Analogously to how we defined (sect. 2.2.2) the ERM rule as minimizing the uniform convergence wysiwyg bound (6), we define the SRM rule (Vapnik and Chervonenkis, 1974) as minimizing wysiwyg bound (17).

Definition 6 (SRM)**.** Given hypothesis class sequence $(\mathcal{H}_n)_{n\in\mathbb{N}}$ (such that each $\mathcal{H}_n$ has the uniform convergence property with accuracy function ϵ_n^{uc}) and weight function w. Let $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$. The SRM rule for $(\mathcal{H}_n)_n$ and w is defined by

$$\text{SRM}^{w}_{(\mathcal{H}_n)_n}(S) \in \min_{h\in\mathcal{H}} \left[L_S(h) + \min_{n:h\in\mathcal{H}_n} \epsilon_n^{\text{uc}}(m, w(n)\delta) \right]. \tag{18}$$

3.2. **Regularization.** The SRM rule minimizes, not just the empirical error, but the sum of the empirical error and an additional term. This additional term depends on the hypothesis h, or more precisely, on the earliest hypothesis subclass $\mathcal{H}_n$ in the sequence that the hypothesis is in.

3.2.1. *The regularization term.* SRM can thus be seen to strike a balance between empirical error and an additional *penalty* or (in machine learning terminology) *regularization* term for $\mathcal{H}_n$. This regularization term

$$\min_{n:h\in\mathcal{H}_n} \epsilon_n^{\text{uc}}(m, w(n)\delta) \tag{19}$$

is, by (13) and (14), up to a multiplicative constant equal to[22]

$$\sqrt{\frac{\text{VCdim}(\mathcal{H}_n) - \log w(n) - \log \delta}{m}}. \tag{20}$$

The regularization term is larger with higher VC dimension of $\mathcal{H}_n$ and with lower weight $w(n)$. If we interpret the regularization as a penalty for $\mathcal{H}_n$'s *complexity*, then SRM can be seen to strike a balance between empirical error and simplicity. The rule automatically navigates the bias-complexity trade-off by balancing empirical error (as a proxy for approximation error) with simplicity (as a proxy for estimation error), finding "the best trade-off between the approximation error and a distribution-free upper bound on the estimation error" (Bartlett et al., 2002).

3.2.2. *Model selection.* In the two-step picture of model selection (sect. 2.1.3 above), SRM can thus be seen to automatize both steps: it selects (step 1) a particular "model" (subclass in the sequence) and (step 2) the lowest-empirical-error hypothesis in this model (subclass).[23] This makes SRM structurally similar to methods like AIC and BIC, which likewise select a model by minimizing penalized fit.[24]

[22] In order to actually implement the SRM rule (18), we would of course need to settle on exact values for the penalty term. This is further complicated by the fact that we often only have loose upper bounds for the VC dimension (Hastie et al., 2009, p. 239). Similarly to the case of ERM, I here abstract away from implementation issues.

[23] Of course, it is still up to us to choose the subclass sequence (sect. 4.2 below).

[24] We do not have to automatize model selection to this extent. An alternative is to, first, run $\text{ERM}_{\mathcal{H}_n}$ for each of the subclasses $\mathcal{H}_n$ (realistically, a finite number N) on the same training data set, resulting in a set $\mathcal{H}_{\text{val}}$ of N selected hypotheses. Second, we run $\text{ERM}_{\mathcal{H}_{\text{val}}}$ on a separate *validation* data set to select a final hypothesis. The latter step is to catch overfitting, and so the whole procedure is again meant to manage the bias-complexity trade-off. From this perspective, SRM can be seen to "approximate the validation step [...] automatically" (Hastie et al., 2009, p. 223; also see Shalev-Shwartz and Ben-David, 2014, ch. 11). In practice, especially if data is

3.2.3. *The norm.* If we interpret the regularization penalty as a complexity penalty, then the SRM rule implements a methodological simplicity norm.

> **Methodological norm N2 (Occam's razor).** Trade fit for simplicity.

But what exactly is the justification for this norm? Why indeed is the specific trade-off that SRM implements "the right" (Shalev-Shwartz and Ben-David, 2014) or "the best" (Bartlett et al., 2002), and is the regularization term really best understood as a complexity penalty? In order to address these questions satisfactorily, we have to take a closer look at the theoretical justification for the SRM rule.

4. The justification for SRM

I first discuss the best theoretical justification for SRM (section 4.1). I then highlight an apparent obstacle to a subsequent argument for simplicity norm (N2), namely the model-relativity of this theoretical justification (section 4.2).

4.1. **Theoretical justifications.** Since the worry about simplicity norm (N1) and its theoretical justification was its one-sided concern with estimation error, a good starting point is to see whether in the case of SRM more can be said about the approximation error.

4.1.1. *Universal consistency.* The reliability guarantee of PAC learnability entails that, as we draw more and more training data, the estimation errors of the selected classifiers converge (in probability) to 0. What if we aim for the *true risks* (so estimation *and* approximation errors) to converge (in probability) to 0?

As discussed by von Luxburg and Schölkopf (2011, sect. 7), this can be achieved in two steps.[25] First, we choose a sequence of (nested) subclasses $(\mathcal{H}_n)_{n\in\mathbb{N}}$ such that each possible hypothesis is eventually contained in (or at least arbitrarily closely approximated by) some $\mathcal{H}_n$. Second, we have to devise a learning procedure that only has access to an initial subsequence of subclasses which grows at the right pace with the sample size, namely in such a way that both estimation and approximation error converge to 0. The SRM rule gives such a method that is *Bayes-consistent* (Devroye et al., 1996, thrm. 18.1).

Formally, following again Shalev-Shwartz and Ben-David (2014), we can understand consistency as a learnability notion strictly weaker than PAC learnability, because the sample complexity function depends on more elements. We say that algorithm A is *universally consistent* w.r.t. hypothesis class $\mathcal{H}$ if there is a sample complexity function $m_{\mathcal{H}}^{\text{con}} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$, every $h \in \mathcal{H}$ and every $\mathcal{D}$, we have for $m \geq m_{\mathcal{H}}^{\text{con}}(\epsilon, \delta, h, \mathcal{D})$ the property (12). We further say (now following again von Luxburg and Schölkopf, 2011, p. 660) that A is *universally Bayes-consistent* if $\mathcal{H}$ is the class of *all* hypotheses.

This might at first glance look like an impressive property, but Shalev-Shwartz and Ben-David (2014) disagree. The reason is that we have no control over the speed of convergence. This can be seen from the dependence of the sample complexity function on (aside from a reference h) the true distribution, which we do not know

---

plenty, the more manual validation procedure, or related approaches like cross-validation, might be preferable to implementing regularization in the learning algorithm, like in SRM.

[25] von Luxburg and Schölkopf (2011) do not mention SRM, but as explained this rule is a way of implementing what they describe. Also see (Devroye et al., 1996, ch. 18), which explicitly deals with SRM and which von Luxburg and Schölkopf base part of their discussion on.

in a learning problem (cf. von Luxburg and Schölkopf, 2011, sect. 7.4). Indeed, on a closer look, consistency is so weak that it is satisfied by learning algorithms which do not look reasonable at all. Shalev-Shwartz and Ben-David (2014, pp. 15f) introduce a "bad learner" which simply memorizes the training data and gives constant prediction 0 for unseen instances, but which is universally Bayes-consistent for a countable domain $\mathcal{X}$:[26]

> Intuitively, it is not obvious that the `Memorize` algorithm should be viewed as a *learner*, since it lacks the aspect of generalization [...]. The fact that `Memorize` is a consistent algorithm [...] therefore raises doubt about the usefulness of consistency guarantees. (ibid., p. 67)

But we do not here need to settle the question whether consistency can still count as a minimal kind of justification for learning algorithms,[27] because we can say something stronger about SRM.[28]

4.1.2. *Nonuniform learnability.* Shalev-Shwartz and Ben-David (2014, ch. 7) introduce SRM in the context of a notion of learnability strictly between PAC learnability and consistency. The notion of *nonuniform learnability* involves a sample complexity function which (unlike consistency) does not depend on the unknown true distribution, but which (unlike PAC learnability) does depend on a reference hypothesis h. Thus we say that A nonuniformly learns hypothesis class $\mathcal{H}$ if there is a sample complexity function $m^{\text{nul}}_{\mathcal{H}} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and every $h \in \mathcal{H}$, we have for $m \geq m^{\text{nul}}_{\mathcal{H}}(\epsilon, \delta, h)$ the property (12).[29]

Furthermore, Shalev-Shwartz and Ben-David (2014, thrm. 7.2) state a generalization of the fundamental theorem, which characterizes nonuniform learnability. Namely, a hypothesis class $\mathcal{H}$ is nonuniformly learnable if and only if it is a countable union $\cup_n \mathcal{H}_n$ of VC classes, and indeed if and only if it is nonuniformly learnable by SRM.[30]

However, it is again not clear whether nonuniform learnability is actually a useful notion of learnability. In the words of Shalev-Shwartz and Ben-David themselves,

> When approaching a learning problem, a natural question is how many [data instances] we need to collect in order to learn it. Here, PAC learning gives a crisp answer. However, for both nonuniform

[26]The countability of the domain is important here: von Luxburg and Schölkopf (2011, sect. 4.2) give the same memorize algorithm (note: an instantiation of the ERM rule) for the domain $[0,1]$ as an example of how ERM can *fail* to be consistent.

[27]Consistency could serve as a "sanity check" (cf. Grünwald, 2007, sect. 17.1.1), or a necessary but not sufficient condition for justification. This may be consistent with Shalev-Shwartz and Ben-David's view that "[s]ince it is easy to make any algorithm consistent, it may not be wise to prefer one algorithm over the other just because of consistency considerations" (2014, p. 69).

[28]In discussing the reliability of SRM, Harman and Kulkarni (2007, sects. 3.3–3.4) only mention the property of universal consistency. This is criticized by Kelly and Mayo-Wilson (2008), who also highlight the kind of justification I come to in section 4.1.3 below.

[29]The notion of nonuniform learnability appears to have been first introduced by Benedek and Itai (1988, 1994). Blumer et al. (1989); Linial et al. (1991) discuss the same notion in the PAC setting.

[30]It is relatively straightforward to show that any nonuniformly learnable class can be decomposed into a countable union of individually PAC learnable classes (Shalev-Shwartz and Ben-David, 2014, p. 60; where they refer in the final step to the fundamental theorem, what is more precisely required is their corollary 6.4). The other direction follows by showing that a countable union of VC classes is nonuniformly learnable by SRM (ibid., thrms. 7.3, 7.4).

> learning and consistency, we do not know in advance how many [instances] are required to learn $\mathcal{H}$. In nonuniform learning this number depends on the best hypothesis in $\mathcal{H}$, and in consistency it also depends on the underlying distribution. In this sense, PAC learning is the only useful definition of learnability. (2014, p. 67)

Having such a bound on sample complexity, though, is only one of three possible uses of the theoretical analysis which the authors put forward (ibid., sect. 7.5). The first is to have "an upper bound on the true risk of the learned hypothesis."[31] This points at a kind of justification that is closest to the original derivation of SRM.

4.1.3. *Wysiwyg and oracle bounds.* Here we switch again to the perspective where we draw a training sample of a specific size m, and we are interested in a bound on the true risk of the hypothesis that will be selected. We now reason in a way that is analogous to the justification for ERM (sect. 2.2.3), albeit with some differences.

First, from the generalized uniform convergence bound (16) or (17) for hypothesis class sequence $(\mathcal{H}_n)_n$ and weight function w, applied to the particular hypothesis selected by $\mathrm{SRM}^w_{(\mathcal{H}_n)_n}$, we have (for size-$m$ sample generated from any distribution, with probability at least $1-\delta$)

$$L_{\mathcal{D}}(\mathrm{SRM}^w_{(\mathcal{H}_n)_n}(S)) \leq L_S(\mathrm{SRM}^w_{(\mathcal{H}_n)_n}(S)) + \epsilon^{\mathrm{uc}}_{\hat{n}}(m, w(\hat{n})\delta). \tag{21}$$

Here $\hat{n}$ is the index of the hypothesis class $\mathcal{H}_{\hat{n}}$ that SRM selected the hypothesis from. This is again a wysiwyg bound, with the difference that the error term now depends on the hypothesis (or rather, on the class $\mathcal{H}_{\hat{n}}$ it was selected from).

Furthermore, by the definition of SRM as a minimizer and another application of the generalized uniform convergence bound (16),

$$L_S(\mathrm{SRM}^w_{(\mathcal{H}_n)_n}(S)) + \epsilon^{\mathrm{uc}}_{\hat{n}}(m, w(\hat{n})\delta) \leq \min_{n, h\in\mathcal{H}_n} [L_S(h) + \epsilon^{\mathrm{uc}}_n(m, w(n)\delta)] \tag{22}$$

$$\leq \min_{n, h\in\mathcal{H}_n} [L_{\mathcal{D}}(h) + 2\epsilon^{\mathrm{uc}}_n(m, w(n)\delta)] \tag{23}$$

$$\leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + 2\epsilon^{\mathrm{uc}}_{n^*}(m, w(n^*)\delta). \tag{24}$$

Here $n^*_{\mathcal{D}}$ is the index of the earliest class containing the best (lowest true risk) hypothesis in the superclass, which depends on the true $\mathcal{D}$. In sum, we have

$$L_{\mathcal{D}}(\mathrm{SRM}^w_{(\mathcal{H}_n)_n}(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + 2\epsilon^{\mathrm{uc}}_{n^*_{\mathcal{D}}}(m, w(n^*_{\mathcal{D}})\delta). \tag{25}$$

This is again a reliability bound. It gives a justification for using SRM, as selecting (with high probability, up to an error term) the best in the superclass. However, the difference is that the error term now depends on the subclass which contains the best hypothesis. Such a bound is also called an *oracle bound*, because we do not actually know what this class and hence this error term is.[32]

This reliability bound, not mere consistency, is supposed to show the "real strength" of SRM (Devroye et al., 1996, p. 294), and why the implemented trade-off is the right or "optimal" one (ibid., p. 295). However, this bound less clearly

[31] The third in their list is to have "a crisp way to encode prior knowledge" (Shalev-Shwartz and Ben-David, 2014, p. 68). Here they also discuss SRM for model selection, writing that "the SRM rule enables us to select the right model on the basis of the data itself" (ibid.). But, again, it is not fully clear why SRM does this in the "right" way; and I suggest that the answer lies in the bounds to be discussed next. I will turn to the role of prior knowledge in SRM in section 4.2.

[32] More generally, we can derive such a bound for any reference subclass ("oracle"), without knowing how good this class actually is. See (Shalev-Shwartz and Ben-David, 2014, exc. 7.4).

constitutes an interesting reliability property than the analogous bound for ERM, because of this further "oracle" dependence on what the best class turns out to be. In order to see how exactly this bound can underwrite a methodological justification for SRM, we need to take a step back again, and confront the general issue of the relativity of theoretical justification to the prior choice of inductive model.

4.2. **Model-relativity.** The ERM rule is a generic learning rule, which on each application asks for a particular inductive model, a particular VC class. The fundamental theorem gives a theoretical justification for the ERM rule, but, again (sect. 2.4.3 above), this justification is model-relative. And the inductive model must be restrictive: it must be a hypothesis class that is sufficiently simple.

Something similar holds for SRM, be it for a more general inductive model and a weaker theoretical guarantee. As Shalev-Shwartz and Ben-David (2014, p. 68) write, we now "encode our prior knowledge by specifying weights over (subsets of) hypotheses of $\mathcal{H}$" and then "we again have a generic learning rule – SRM." The theoretical justification for SRM is again model-relative, relative to this more general, but still restrictive type of inductive model.

4.2.1. *No free lunch.* The no-free-lunch theorem of Shalev-Shwartz and Ben-David (2014, thrm. 5.1) shows that (for an infinite domain $\mathcal{X}$) the class of *all* hypotheses is not PAC learnable (ibid., cor. 5.2).[33] This is a consequence of the fundamental theorem, since the class of all hypotheses does not have finite VC dimension.[34] In the case of nonuniform learning, we have an analogous no-free-lunch result (Shalev-Shwartz and Ben-David, 2014, p. 63). For infinite domain, the class of all hypotheses is not nonuniformly learnable, which is a consequence of the fact that the class of all hypotheses is not a countable union of VC classes (ibid., exc. 7.5).

The upshot is that there is no "universal learner" on either of the two definitions of learnability.[35] Both ERM and SRM must be equipped with a restrictive inductive model, which represents a restrictive inductive bias; and the respective learnability guarantees are relative to this choice. In the case of ERM, the inductive model must be a VC class, and learnability gives a model-relative justification of probably finding the near-best in the class. In the case of SRM, both the inductive model and the kind of justification are somewhat more intricate. The inductive model is a (weighted) countable sequence of VC classes, which can be seen to involve both a choice of superclass $\mathcal{H}$ and a choice of how to carve up this class in (and assign weights to) subclasses $\mathcal{H}_n$. The oracle bound guarantee is not just relative to the superclass, but also to this choice of (weighted) subclasses.

Moreover, a choice of (weighted) sequence is automatically a choice of simplicity ordering, because it determines the VC dimensions (and weights) that appear in the regularization term.

---

[33] See (Sterkenburg and Grünwald, 2021) for further discussion of the no-free-lunch theorems.

[34] In the presentation of Shalev-Shwartz and Ben-David (2014), their no-free-lunch theorem forms part of the proof of the fundamental theorem.

[35] There *are* universal learners in the sense of universal consistency: an example is the k-nearest neighbor classifier (von Luxburg and Schölkopf, 2011, sect. 3). However, the rate of convergence can be arbitrarily slow: performance up to any finite sample size can be arbitrarily bad (ibid., p. 695; Devroye et al., 1996, ch. 7). This again illustrates the weakness of consistency guarantees.

4.2.2. *The subclass sequence.* The model-relativity to the choice of hypothesis class sequence is the topic of a recent paper by Bargagli Stoffi et al. (2022). The authors compare, for two disjoint subclasses $\mathcal{H}_S$ and $\mathcal{H}_C$ (one a lower-VC dimension "simple" class, the other a higher-VC dimension "complex" class), the ERM rule on the union $\mathcal{H}_S \cup \mathcal{H}_C$ (which is still a VC class) to the SRM rule on the sequence $(\mathcal{H}_S, \mathcal{H}_C)$ with uniform weights. They then prove upper bounds on the sample size for a probabilistic guarantee of selecting a hypothesis from the "correct" class, where correctness derives from one of two possible scenarios: either the data is actually sampled from an element plus noise in the first class (a "simple world") or from an element plus noise in the second ("complex world"). This bound is sharper for SRM in the simple world than for ERM, but sharper for ERM than for SRM in the complex world.

While it may be an overinterpretation that SRM in the latter case "provably slows down, instead of favoring, the supervised learning process" (Bargagli Stoffi et al., 2022, p. 23),[36] the general lesson is correct that theoretical bounds depend on whether and how we carve up the hypothesis superclass, and this choice may be a better or worse match with the actual learning situation. As a simple illustration inspired by Bargagli Stoffi et al., consider an instance space of two features, and the superclass $\mathcal{H} = \mathcal{H}_2^{\text{pol}}$ of quadratic separators.

This class has VC dimension 5, so the error term for $\text{ERM}_{\mathcal{H}}$ in the bounds (8) and (11) is of the order

$$\sqrt{\frac{5 - \log \delta}{m}}. \tag{26}$$

But we could also choose to use SRM on a uniformly weighted nested decomposition of $\mathcal{H}$ into $\mathcal{H}_1$ of linear separators and $\mathcal{H}_2 = \mathcal{H}$ of quadratic separators. Since $\text{VCdim}(\mathcal{H}_1) = 4$, the error terms in the bounds (21) and (25) for $\text{SRM}^{(.5,.5)}_{(\mathcal{H}_1,\mathcal{H}_2)}$ are now of the order

$$\sqrt{\frac{4 - \log \delta + \log 2}{m}} \quad \text{and} \quad \sqrt{\frac{5 - \log \delta + \log 2}{m}} \tag{27}$$

for $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively.

This means that we have a stronger wysiwyg guarantee for SRM than for ERM for those cases in which the former selects a hypothesis from $\mathcal{H}_1$; but a weaker one for those cases in which SRM selects a hypothesis from $\mathcal{H} \setminus \mathcal{H}_1$. Moreover, we have a stronger reliability guarantee for SRM than the guarantee for ERM in case $\mathcal{H}_1$ contains the best hypothesis in $\mathcal{H}$; but we have a weaker reliability guarantee if it does not. What cases apply, however, is presumed to be unknown in the learning problem: this depends on the true distribution. Thus, depending on the unknown true distribution, it may or may not in fact be optimal (in the sense of having the strongest above bounds) to implement a (simplicity) preference for $\mathcal{H}_1$ via SRM.

4.2.3. *The subfamily problem.* The theory obviously cannot tell us what is the best choice in the previous scenario. Similarly, once we have chosen a non-VC superclass (like the class of all polynomials), the theory cannot tell us what is the best way of

[36] Worst-case upper bounds do not yet determine relative convergence behavior in any particular instance. The derivation of lower bounds would be complicated by the fact that the authors' assumptions of either a "simple" or "complex world" entail corresponding restrictions on the possible true distributions.

carving it up in a sequence of subclasses. In practice, we might have intuitions about the most natural way of doing this (like the carving up of the polynomials by degree), and the theoretical guarantees may even provide the basis for an *explanation* why SRM with this standard choice generally works well. But the theory cannot give a *justification* for this particular choice, because the theoretical guarantees are model-relative; and any other choice of carving up translates to corresponding such guarantees (cf. Sterkenburg, 2025, pp. 11f; Forster, 1995, p. 358).

The analogous point for AIC is made by Kieseppä (1997, p. 40), in the context of a related problem which has received some attention in the philosophical literature. Translated to the current framework, this "subfamily problem," raised by Forster and Sober (1994, sect. 6; also see Kukla, 1995; Forster, 1995), is why the functioning of SRM, interpreted as a methodological advice how to trade off simplicity with fit, is not empty. Namely, *after* we have seen the data S, we could specify any (say, high-degree polynomial) hypothesis $\hat{h}_S$ with perfect fit on this data, and imagine a hypothesis class sequence which includes the singleton class $\mathcal{H}_i = \{\hat{h}_S\}$ with only this hypothesis. Now the SRM rule with this inductive model would have selected this hypothesis, and without any trade-offs. How does this rhyme with the methodological advice to trade off fit with simplicity, let alone with any more specific advice such as to prefer, for equal fit, lower to higher-degree polynomials?

To think there is a problem here reveals a misunderstanding of frequentist guarantees. Or more charitably, it brings out that frequentist guarantees can be quite counter-intuitive, or even of restricted use (cf. Sterkenburg, 2025, p. 14). Frequentist guarantees are *pre-sampling* bounds, relative to a *pre-sampling* choice of inductive model, on the probabilities of outcomes. To try to use such bounds in reasoning about a specific outcome, as in the previous, easily runs into confusions. We could, of course, define as our inductive model a hypothesis class sequence which includes the singleton $\{\hat{h}_S\}$ fitting perfectly the specific data S; and this indeed *would have* led the corresponding SRM procedure to select this hypothesis, in the *hindsight* of this specific outcome S. But *pre-sampling* there is no probabilistic guarantee of this specific outcome (unless we explicitly assume so in the true distribution), so no theoretical guarantee of perfect fit that might be used to justify the selection of this particular hypothesis.

The general point about inferring methodological advice directly from the functioning of SRM is the following. If such advice is inferred from pre-sampling and model-relative guarantees, then it must be pre-sampling and model-relative advice, too: given your choice of inductive model, it is, with high pre-sampling probability, a good idea to trade off fit with a penalty as given by the inductive model. That also means that we cannot infer from these guarantees any "absolute" methodological advice to the effect that, say, one should always for equal fit prefer linear to quadratic hypotheses, because whether SRM implements such a preference depends on the choice of inductive model, and the theory does not tell us what is the right such choice. This holds for the choice of superclass and subclasses, which define the VC dimensions in the penalty terms, but also for the weights of the subclasses.

4.2.4. *The weight function.* Shalev-Shwartz and Ben-David (2014, sect. 7.3) describe a version of the SRM rule where *only* the weights play a role. In their "Minimum Description Length" (MDL) rule, the given countable hypothesis class $\mathcal{H}$ is decomposed in all its singletons, $\mathcal{H} = \cup_n \{h_n\}$. Since each singleton class has

the same VC dimension 1, the VCdim term can simply be crossed out from the regularization (20), and all that matters are the weights.[37]

The weight function now effectively assigns weights to the single hypotheses directly, and the "MDL" label stems from the suggestion (Shalev-Shwartz and Ben-David, 2014, p. 64) to assign hypothesis h a weight that depends on h's *description length* via some description language.[38] They write that "the MDL paradigm gives a formal justification to a philosophical principle of induction called Occam's razor" (ibid., p. 58), because "the more complex a hypothesis h is (in the sense of having a longer description), the larger the sample size it has to fit to guarantee that it has small true risk" (ibid., p. 65).

This reasoning falters on the fact that, due to the dependence on choice of description language, description length is not a robust notion of the simplicity of individual hypotheses;[39] and the authors immediately correct themselves (Shalev-Shwartz and Ben-David, 2014, pp. 65f). In line with the point of the previous section, they rather conclude that (ibid., p. 66)

> there is no inherent generalizability difference between hypotheses. The crucial aspect here is the dependency order between the initial choice of language (or, preference over hypotheses) and the training set. [...] As long as [this choice of inductive model] is done independently of the training sample, our generalization bound holds.

We may add that the bound is model-relative, relative to this initial choice of inductive model. Much like before, the wysiwyg and oracle bounds will be stronger for those hypotheses which have received larger weights, and so will be better in those learning situations in which these preferred hypotheses are in fact good.

Having now discussed the model-relatively in some detail, we must finally confront an obvious question. If the theoretical guarantees are relative to our inductive model, which includes the simplicity ordering, how could we get a non-circular justification for a simplicity preference out of them?

4.2.5. *Circularity?* Although Bargagli Stoffi et al. (2022) shy away from concluding so explicitly,[40] it is easy to read their work as showing that a simplicity preference is good if the truth is simple, and not good if it is not. Similarly, it seems that Shalev-Shwartz and Ben-David's MDL-SRM is a good method if (some) high-weight ("simple") hypotheses indeed have low true risk, but not so good if they do not.

But then it seems that simplicity here ultimately comes down to a particular domain assumption, which may or may not be appropriate. Norton (2021, ch. 6) even argues that talk of simplicity in induction is merely (as Sober, 1988, ch. 2 calls

[37] This rule is only loosely related to (model selection) methods treated in standard references on the MDL principe (Grünwald, 2007; Grünwald and Roos, 2020), and rather an instance of the PAC-Bayes approach as presented in (Grünwald, 2007, sect. 17.10.2; von Luxburg and Schölkopf, 2011, sect. 6.2). In other presentations, the PAC-Bayes label primarily means randomnized prediction via the prior/posterior or weight function, which is updated in accordance with minimizing corresponding bounds (Shalev-Shwartz and Ben-David, 2014, ch. 31; Alquier, 2024).

[38] A weight function corresponds to a prefix-free description language, where higher weights correspond to shorter descriptions (Shalev-Shwartz and Ben-David, 2014, sect. 7.3).

[39] Recall section 2.3.2, and also see (Sterkenburg, 2025, sect. 3.1).

[40] The most they conclude, in answer to their "central question—'is simplicity a road to the truth?'" (ibid., p. 21), is that "the principle of Occam's razor, at least as expressed by introduction of regularization in SRM, can both favor and hamper learning and hence convergence to the truth" (ibid., p. 23).

it) a "surrogate" for specific background facts. According to Norton's material theory (2003; 2021), inductive inferences are warranted solely by background facts; and in his analysis of curve fitting (2021, sect. 6.6), he identifies three such facts.

First, a particular "error model" (ibid., p. 196), which in our framework presumably includes the assumption of a true data-generating distribution. Second, a parameterization or description of the hypotheses. Norton writes that "[d]escriptive complexity can only be a good epistemic guide to the truth [...] if the language of description is chosen so that the truths correspond to simple assertions," which in curve fitting means "a matching with background facts of the parametrization used and the family of its functions from which the curves are drawn" (ibid., p. 199). Third, an "order hierarchy" of families of curves, which "has to be such that curves fitted earlier in the procedure correspond to stronger or more probable processes" (ibid., p. 202). Thus the choice of ordered sequence of subclasses is only good for induction if it matches the actual domain.[41]

If this is so, then the simplicity preference exhibited by SRM reflects an *ontological* commitment, rather than a *methodological* principle (cf. Sober, 1988, ch. 2). It reflects an assumption that "the world" (or, in any particular application, the relevant domain) is simple in the corresponding sense, rather than a principle that a simplicity preference is good even without such an assumption. Consequently, if this is so, any justification for simplicity we may obtain from the theory must be question-begging or circular: we can only show a simplicity preference to be good if we first assume that the domain favors simplicity.

I will now argue that this is not so. Even if the theoretical guarantees are model-relative, we can still obtain a justification for a simplicity preference as a methodological principle.

5. The means-ends justification

The original core argument for a simple hypothesis class is that *in order to* profit from the relevant reliability and wysiwgy justification, we must choose a simple class. The main weakness of this means-ends argument and the corresponding norm (N1) is that it is too restrictive in its applicability.

The new means-ends argument is not so restrictive, because it explicitly deals with a choice of possibly very complex hypothesis class: a countable union of VC classes. It says that *in order to* at least profit from a weaker reliability and wysiwyg justification for such a class, we must carve it up in a sequence of subclasses and implement, as according to norm (N2), a certain preference for simplicity against fit in the learning.

However, in light of the model-relativity discussed previously, more needs to be said about why this would be a *methodological* justification (section 5.2) for a *simplicity* preference (section 5.1). I will start with the latter.

5.1. **Simplicity?**

Recall from section 3.2 that the simplicity interpretation stems from the occurrence of the hypothesis subclasses' capacities in the regularization.

[41] Norton devotes a further chapter to statistical model selection and AIC in particular (2021, ch. 7). His critical discussion of AIC zooms in on the material assumption that "the data are generated by an hypothesis in the model under test" (ibid., p. 205), which does not pertain to statistical learning theory.

5.1.1. *The weight function.* However, there is another variable in the regularization term: the weight of the subclass. In fact, we saw that in the MDL-SRM approach of section 4.2.4, the capacities drop out and *only* the weights matter.

Note, first, that we can also make the weights drop out. Namely, if we have a finite sequence of $(\mathcal{H}_n)_{n<N}$ of hypothesis subclasses, then we can simply use the uniform weight function $w : n \mapsto N^{-1}$. The weights can then still be seen to play a role in the generalized uniform convergence theorem: both the reliability and the wysiwyg guarantee feature the additional $-\log N^{-1} = \log N$ term. But since these terms are the same for all hypothesis subclasses (the penalty term rather expresses the size of the superclass), they do not play a role in the regularization.

In fact, for any finite sample size m, we will effectively always work with a finite sequence of subclasses. In the usual case of a (nested) sequence of subclasses of increasing VC dimension (like the polynomials), we have that subclasses beyond some N will never be considered, whatever the size-m data, because their penalty is larger than good fit could compensate for. So effectively we are working with a finite sequence again, for which we can imagine a uniform weight function, so that the weights again drop out of the regularization. This kind of reasoning might be why in many—most—presentations of SRM the weights do not appear at all.[42]

Of course, it is still always possible to adopt non-uniform weights, effectively introducing modified penalty terms. This allows us to encode as inductive bias a further preference for certain hypothesis subclasses over others. The MDL-SRM rule is an extreme version of doing so, with different penalties for individual hypotheses. By the previous reasoning, for finite data-size m, we are here in a situation where we are effectively working with a *finite* superclass $\mathcal{H}$. Then a choice of uniform weight function comes down to using the ERM rule with the superclass, while a non-uniform weight function expresses a certain preference among hypotheses.

The latter might be a good approach in some scenarios, for instance if we have beliefs about the expected performance of different hypotheses.[43] But special cases of this kind do not show that capacity does not a play a role in regularization by SRM: only that we can engineer it so that all capacities are the same.[44] As soon as we make a (more standard) choice for a nested subsequence of increasing capacity, these capacities explicitly appear in the regularization.

5.1.2. *Something else.* Even in the latter standard case, some authors are wary of talking about simplicity. Harman and Kulkarni (2007, sect. 3) use scare quotes in the title of their section *Induction and "Simplicity"* on SRM, and consistently talk about SRM as trading off fit against "something else." They indeed "prefer to think of VC dimension as providing an *alternative* to simplicity" (2009, p. 53).

Their reservations appear to be those already discussed in section 2.3.2 above, that capacity does not necessarily align with other conceptions of simplicity: the "relevant ordering [...] is not a simplicity ordering, at least if sine curves count as

[42]More generally (not assuming anything about the VC dimensions), for an infinite sequence of subclasses, for any (necessarily non-uniform) weight function over the full sequence, the weight terms $w(n)$ must go to zero as n goes to infinity. That means that the $-\log w(n)$ must go to infinity as well, which again means that subclasses beyond some N would never be considered.

[43]Or by more pragmatic *luckiness* reasoning: see section 5.2.3 and footnote 49.

[44]One could think of wackier choices, like designing, for a finite sequence of classes of increasing capacity, a weight function which exactly counteracts the capacities. Again, this does not show that capacities do not play a role in the regularization: only that one could, with some effort, set up things so as to neutralize this effect!

'simple'" (Harman and Kulkarni, 2007, p. 73). As we discussed, however, capacity gives *a* natural notion of the simplicity of a hypothesis class. Moreover, the notion of capacity possesses a formal robustness that conceptions deriving from the number of parameters or the "bumpiness" of individual hypotheses must lack.

Yet the worry might persist that it is odd to evaluate simplicity at the level of hypothesis classes, rather than at the level of individual hypotheses. What regularization penalty any individual hypothesis is subjected to depends on our initial choice of carving up the superclass in a subclass sequence, and what capacity subclass this hypothesis ends up being in. This, again, need not link to any notion of the simplicity of the individual hypothesis.

This brings us back to the theme of model-relativity. It depends on our initial choice of inductive model how exactly simplicity plays a role in the regularization. However, even if this initial choice of subclasses is ours, the subclasses then come attached with a robust notion of their simplicity qua capacity, which automatically plays a role in the SRM regularization. It would certainly be nice if we had a further formal link to the inherent simplicity of individual hypotheses; but we do not. What we do have, or so I will finally argue shortly, is a methodological justification for making a choice of subclass sequence and then regularize accordingly. The regularization term features the capacities of the subclasses, and so there is a trade-off between fit and capacity. If capacity is a natural notion of simplicity, then it is natural to talk about a trade-off between fit and simplicity, and to phrase norm (N2) in those terms.

5.2. **Methodological?** A simplicity preference would be methodologically justified, as opposed to merely reflecting an (ontological) assumption that the world (the domain) is simple, if this simplicity preference can be shown to be beneficial to successful learning, *even without such a simplicity assumption.*

5.2.1. *Not true: clever.* This is exactly what we have in the case of SRM and the end of (universal) consistency, for uncountable domain (sect. 4.1.1 above). In order to have the guarantee of consistency, it will not do to use ERM: we must employ regularization in accordance with SRM. This is the case even if we do not believe that the domain is simple: the consistency guarantee holds irrespective of whether the Bayes classifier is in a simple class or not. Nor does it matter how we carve up the superclass in a hypothesis class sequence, and so what exact regularization is implemented; but we have to do it somehow, to have the guarantee.

The picture suggested by Norton's material theory, that an inductive method is good if and only if its inductive model matches the material facts (is "true"), is therefore too coarse. As Grünwald (2007, p. 33) writes, a learning method is

> just a *strategy* for inferring models from data ("choose simple models at small sample sizes"), not a statement about how the world works ("simple models are more likely to be true") – indeed, a strategy cannot be true or false, it is "clever" or "stupid." And the strategy of preferring simpler models is clever even if the data-generating process is highly complex [. . .].

5.2.2. *A pragmatic search strategy.* Various authors have observed that it is clever to adopt a "search strategy" of transitioning, as more data comes in, from the simple to the more complex (e.g., Korb, 2004, sect. 3; Harman and Kulkarni, 2007, sect. 3.6). What is also pleasing about this picture is its apparent structural similarity

to the justification in formal learning theory of a simplicity preference as keeping us on "the straightest possible path to the truth" (Kelly, 2007; cf. Steel, 2009).

One might reply, however, that this is more of a *pragmatic* justification for simplicity. The standard pragmatic motivation for simplicity is that we prefer to work with simpler, more convenient theories; the standard example is Mach's view of science as aiming to "compress" our experience. Norton (2021, sec. 6.3) dismisses this "simplicity as mere economy of expression," as one special case where "simplicity is sought merely for pragmatic reasons" (ibid., p. 179). But his other special case is precisely the above "simplicity for economy of search," from Popper's conjectures and refutations[45] to the justification from formal learning theory.

Norton is right that both cases are weaker than the idea that "simplicity functions epistemically as a marker of truth; we are to choose the simpler hypothesis or theory because, we are assured, it is more likely to be true" (2021, p. 178). A defense of a simplicity preference as a clever search strategy does also have a pragmatic flavor.[46] Still, I think it is useful to distinguish the former "merely" pragmatic justification (we like simple hypotheses for reasons of convenience) from a *methodological* justification. I will here understand a methodological justification as one which may contain pragmatic elements, but is still tied to advancing the core epistemic end in statistical learning theory, predictive accuracy (low true risk).[47] Moreover, again, if this is a justification which does not crucially rely on an explicit simplicity assumption, then we can also distinguish it from a "mere" ontological principle.

Nevertheless, the above methodological justification of SRM's simplicity preference as a clever search strategy is not fully satisfying. After all, we discussed how (universal) consistency is arguably too weak to be a useful property, and identified the given-sample-size reliability and wysiwyg guarantees as the strongest theoretical justification (sect. 4.1 above). This redirects us from a picture where we have a "search strategy" for (unboundedly) growing data-sizes, to one where we need to decide on a good learning method for a given-size dataset.[48]

5.2.3. *Luckiness.* But turning to the reliability and wysiwyg guarantees for SRM, it seems more problematic to uphold that a simplicity preference is good regardless of whether the "truth" is simple. After all, in the example inspired by Bargagli

[45]Popper's methodology has also been linked to statistical learning theory (Vapnik, 2000; Corfield et al., 2009; Steel, 2009). I will abstain from discussing the merits of this association.

[46]In the case of formal learning theory, one can say that while finding the truth is an epistemic end, finding it as quickly as possible is already a pragmatic concern (cf. Strevens, 2009, fn. 4).

[47]One might reply that the whole approach and framework of statistical machine learning is already thoroughly pragmatic, because of its instrumentalist concern with predictive accuracy as opposed to truth-finding (see, e.g., Otsuka, 2023, ch. 4). I do not want to deny the difference between truth-finding and predictive accuracy (though see Lin, 2024), but I still think it is sensible to consider predictive accuracy an epistemic end, and talk about methodological principles as being tied to this end.

[48]Both the ontological and the clever search strategy justification for Occam's razor have a long pedigree in the literature. For instance, Duda et al. (2001, sect. 9.2.5) give an evolutionary account of the empirical success of Occam's razor, explaining both that "we are more likely to ignore problems for which Occam's razor does not hold" (restricting ourselves to those where simplicity is a good assumption) and that we are moved towards a pragmatically useful "design methodology" of simple to more complex. The reasoning which I will present now in defense of a methodological justification does not, to the best of my knowledge, have a clear precursor.

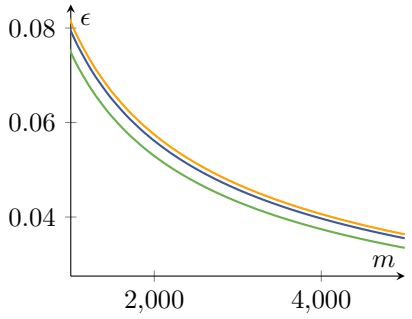

FIGURE 2. Linear v. quadratic. Plotted are the error bound (26) for ERM (blue) on $\mathcal{H} = \mathcal{H}_2$, and the bounds (27) for SRM for $\mathcal{H}_1$ (green) and $\mathcal{H}_2$ (red).

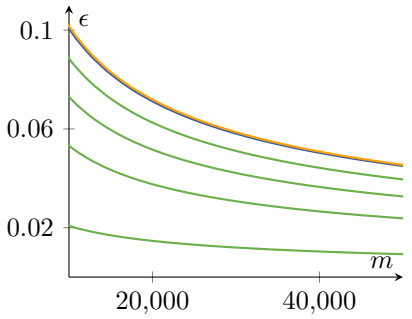

FIGURE 3. Nested sequence $(\mathcal{H}_n)_{n<100}$ with $\text{VCdim}(\mathcal{H}_n) = n$. Plotted are the error bound (26) for ERM on $\mathcal{H} = \mathcal{H}_{100}$ (blue), and the bound (27) for SRM for $\mathcal{H}_{100}$ (red) and for $\mathcal{H}_1, \mathcal{H}_{25}, \mathcal{H}_{50}, \mathcal{H}_{75}$ (green).

Stoffi et al. (2022), we saw that the SRM guarantees are better or worse than those for ERM depending on the "world" (sect. 4.2.2).

Nevertheless, there is a reason why, from the perspective of these guarantees, it is clever to use SRM and regularization, even if we do not want to commit to any simplicity assumption. This reason is related to the principle of *luckiness* in MDL inference (Grünwald, 2007; de Rooij and Grünwald, 2011).[49]

The idea is that by opting for SRM and regularization (rather than for ERM on the superclass), we stand to gain significantly (if we are lucky), at little cost (if we are not). Specifically, if we are lucky and the true distribution is such that a hypothesis class early in the sequence we chose is good (contains hypotheses with low true risk), then we have a significantly stronger reliability guarantee; whereas if we are unlucky and this is not the case, then the reliability guarantee is only little worse. This effect is illustrated in figure 2, which plots (for standard choice of $\delta = 0.05$ and m in the range 1,000–5,000) the reliability error bounds for the example of linear versus quadratic separators.[50]

There are two immediate problems with this idea. First, in our example, the difference between lucky and unlucky does not look *that* significant. But this is in the end not very surprising: that for the (already quite simple) superclass of quadratic separators, a single decomposition gives us a little, but not that much. The effect does get quite significant for larger classes and decompositions. Figure

[49]The original idea actually maps more directly to a choice of non-uniform weight function: by giving a small number of hypothesis classes small weights the reliability bounds for these classes will be significantly better, at the cost of slightly worse bounds for the other classes (see Grünwald, 2007, p. 92).

[50]This observation is similar to but not quite the same as the claim that the "size of the error is about the same as if we had known n beforehand, and minimized the empirical error over $\mathcal{H}_n$" (Devroye et al., 1996, p. 294, my notation; also see Shalev-Shwartz and Ben-David, 2014, p. ?). The authors make a comparison between the ERM and the SRM reliability bounds for subclass $\mathcal{H}_n$, whereas the luckiness reasoning is a practically more relevant comparison between the ERM bound for the superclass $\mathcal{H}$ and the SRM bounds for the subclasses $\mathcal{H}_n$.

3 shows (for $\delta = 0.05$ and m in the range 10,000–50,000) the reliability bounds for a nested sequence $(\mathcal{H}_n)_{n<N}$ for $N = 100$, with $\text{VCdim}(\mathcal{H}_n) = n$.[51] We see that the reliability bounds for SRM are better for by far most subclasses, and many significantly so (e.g., the SRM bound for $\mathcal{H}_{25}$ is about half that of ERM for the superclass).[52]

Still, and this is the second problem: if the hypothesis class that we want to work with is too large to give useful reliability bounds for ERM, so that we are drawn to carving up the class and using SRM, the luckiness principle gives little comfort for the possibility that we are unlucky. It is no good to know that the reliability bound in that unlucky case is not much worse than the reliability bound for ERM, if we already found the ERM bound too weak. To complete the case that SRM is still a clever idea, we need to involve the other component of the theoretical justification, the wysiwyg guarantee.

5.2.4. *What you see.* The wysiwyg guarantees are stronger for the earlier classes in our chosen sequence, whether we are lucky or not. That is, with high probability, for an earlier class in the sequence, the wysiwyg bound (21) is sharp. That means that if the SRM rule selects a hypothesis with low empirical error from an early class, we have some reason to conclude that this hypothesis also has low true error (that we are, in fact, in the lucky case) and/or to act accordingly (to proceed with this hypothesis).[53]

What if the SRM rule selects a hypothesis with *high* empirical error from an early class? We still have that, for earlier classes in the sequence, the wysiwyg bound is sharp. So if the empirical error of the selected hypotheses is high, we have some reason to conclude that the true error of the hypothesis and so indeed the hypothesis class is also bad (that we are, in fact, in the unlucky case) and/or to act accordingly (to go back to the drawing board). In either case, the wysiwyg guarantee tells us that, with high probability, we have an indication whether we are on the right track, which allows us to act accordingly.

This, of course, still leaves the possibility that, despite the larger penalty term, SRM selects a hypothesis far off in the sequence, with weak wysiwyg bounds. But in that case, we in principle have this information: that SRM selected from one of the far-off classes. That in itself strongly indicates that our inductive model was off: that we are not in the lucky case.[54] Hence, again, we have reason to act accordingly: to go back to the drawing board.

---

[51] Here and in the following I assume a uniform weight function $w : n \mapsto 1/N$.

[52] In general, for nested sequences $(\mathcal{H}_n)_{n<N}$, the difference is between a term $\text{VCdim}(n) - \log N^{-1} = n + \log N$ (in the SRM bound for $\mathcal{H}_n$) and $\text{VCdim}(\mathcal{H}_N) = N$ (in the ERM bound). The larger N, the more insignificant the $\log N$ compared to the N term, and by definition all $n \leq N$, so for an increasing fraction of n we have a (significantly) better bound. (Though this effect is tempered by the division by sample size m.)

[53] The earlier proviso regarding frequentist guarantees of course still applies (sect. 2.2.5): strictly speaking, we need some further epistemic or decision-theoretic bridge principle from the guarantee to the particular outcome.

[54] Here the (informal) reasoning is that this would be *very* unlikely to happen if we are, in fact, in the lucky case, because it means that the data is not only such that the good classes do bad, but so bad that the additional penalty for the bad classes is overcome. The reasoning is somewhat informal because by assuming the lucky case we are assuming some subset of possible true distributions, and precise probabilistic statements would depend on what this subset is.

5.2.5. *Pragmatic, ontological, methodological.* There are clearly pragmatic elements to the previous reasoning. The wysiwyg justification is pragmatic rather than epistemic, insofar as it is not a guarantee of accurate prediction, but of having an indication of this accuracy, which derives its importance from the need to make a decision what to do next. The reliability guarantee is a better candidate for a purely epistemic justification, but as we saw it needs the wysiwyg to complete the justification for (N2). In any case, similarly to norm (N1), the actual implementation of the norm would involve various further considerations and decisions. Standing out here are of course the decision that (N1) is infeasible, and the choice of hypothesis class sequence in implementing SRM.

Moreover, an "ontic" element remains in that the reliability guarantees are stronger if the hypothesis sequence is a better match with the "truth." Insofar as this choice of inductive model is a reflection of our prior knowledge, it constitutes a better inductive bias if our prior knowledge is indeed accurate.

Nevertheless, the reasoning still underpins a methodological norm, which neither collapses to a purely pragmatic simplicity preference nor to a purely ontic simplicity assumption. It is not purely pragmatic because it is still tied to the epistemic end of predictive accuracy: the means-ends argument for (N2) says that in order to have reliability (good accuracy) and wysiwyg (good indication of accuracy) guarantees, it is a good idea to regularize. It is not a purely ontic assumption, because regularization is still a good idea even if we are not sure our chosen hypothesis sequence and corresponding regularization terms reflect the domain well.[55]

6. Conclusion

It is important to be clear about the limitations of the argument. Since it is embedded in statistical learning theory, the argument is confined to the framework's idiosyncrasies: its exclusive concern with predictive accuracy, under the assumption of a stable underlying probability distribution. The corresponding frequentist nature of the theoretical justification raises its interpretational questions, and the reasoning essentially relies on randomness in the data; in short, this is certainly not the whole story about simplicity in science (Kelly, 2011, sect. 3). The model-relativity of the theoretical justification, together with a formal notion of simplicity that attaches to classes rather than individual hypotheses, further results in a means-ends reasoning that is less straightforward than earlier attempts to defend Occam's razor in machine learning—but as such also does justice to critiques of these attempts (e.g., Domingos, 1999; also see again Sterkenburg, 2025). Finally, I am not making claims about the extent to which the argument applies to regularization techniques, like drop-out or early stopping in deep learning, which are theoretically less well understood. But I do claim that the argument gives a methodological, non-circular justification for the norm of trading off fit for simplicity—and as such does justice to a central methodological practice in machine learning.

Or have this norm and practice been overtaken by events too? The contemporary debate about "benign interpolation" and "double descent" essentially revolves around the observation that not just norm (N1) but also norm (N2) no longer seem (as) relevant: we achieve good learning with extremely complex classes and no regularization (Belkin, 2021; Bartlett et al., 2021). A philosophical appraisal of the

[55]Indeed, the reasoning shows that it is still a good idea to use regularization if we are given some default ordering of classes, or the ordering is implicit in the adopted regularization procedure.

consequences of this observation, and in particular of the renewed evocation of a principle of Occam's razor at the center of a suitably improved theory of generalization, must be postponed to further work. But an important precondition for such work is clarity on the principle's role in the classical theory, as offered here.

References

P. Alquier. User-friendly introduction to PAC-Bayes bounds. *Foundations and Trends in Machine Learning*, 17(2):174–303, 2024.

A. Baker. Simplicity. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

P. S. Bandyopadhyay and M. R. Forster, editors. *Philosophy of Statistics*, volume 7 of *Handbook of the Philosophy of Science*. Elsevier, 2011.

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

P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

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

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

G. M. Benedek and A. Itai. Nonuniform learnability. In T. Lepistö and A. Salomaa, editors, *Proceedings of the 15th International Colloquium on Automata, Languages and Programming (ICALP 1988)*, volume 317 of *Lecture Notes in Computer Science*, pages 82–92. Springer, 1988.

G. M. Benedek and A. Itai. Nonuniform learnability. *Journal of Computer and System Sciences*, 48(2):311–323, 1994.

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

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

S. de Rooij and P. D. Grünwald. Luckiness and regret in minimum description length inference. In Bandyopadhyay and Forster (2011), pages 865–900.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*. Springer, 1996.

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

M. Dubova, S. Chandramouli, G. Gigerenzer, P. D. Grünwald, W. Holmes, T. Lombrozo, M. Marelli, S. Musslick, B. Nicenboim, L. N. Ross, R. Shiffrin, M. White, E.-J. Wagenmakers, P.-C. Bürkner, and S. J. Sloman. Is Ockham's razor losing its edge? New perspectives on the principle of model parsimony. *Proceedings of the National Academy of Sciences*, 122(5):e2401230121, 2025.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, second edition, 2001.

M. R. Forster. The golfer's dilemma: A reply to Kukla on curve-fitting. *The British Journal for the Philosophy of Science*, 46(3):348–360, 1995.

M. R. Forster and E. Sober. 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, 1994.

T. Freiesleben and T. Grote. Beyond generalization: A theory of robustness in machine learning. *Synthese*, 202:109:1–28, 2023.

K. Genin. *The Topology of Statistical Inquiry*. PhD Dissertation, CMU, Pittsburgh, 2018.

C. Glymour. *Theory and Evidence*. Princeton University Press, 1980.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, 2016.

T. Grote, K. Genin, and E. Sullivan. Reliability in machine learning. *Philosophy Compass*, 19(5):e12974, 2024.

P. D. Grünwald. *The Minimum Description Length Principle*. MIT Series in Adaptive Computation and Machine Learning. MIT Press, 2007.

P. D. Grünwald and T. Roos. Minimum Description Length revisited. *International Journal of Mathematics for Industry*, 11(1), 2020.

M. Hardt and B. Recht. *Patterns, Predictions, and Actions: Foundations of Machine Learning*. Princeton University Press, 2022.

G. Harman and S. Kulkarni. *Reliable Reasoning: Induction and Statistical Learning Theory*. The Jean Nicod Lectures. A Bradford Book. MIT Press, 2007.

G. Harman and S. Kulkarni. Response to Shafer, Thagard, Strevens, and Hanson. *Abstracta*, 4(3):47–56, 2009.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, second edition, 2009.

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

S. Jain, D. N. Osherson, J. S. Royer, and A. Sharma. *Systems That Learn: An Introduction to Learning Theory*. A Bradford Book. MIT Press, 2nd edition, 1999.

K. T. Kelly. *The Logic of Reliable Inquiry*. Logic and Computation in Philosophy. Oxford University Press, 1996.

K. T. Kelly. A new solution to the puzzle of simplicity. *Philosophy of Science*, 74(5): 561–573, 2007.

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

K. T. Kelly and C. Mayo-Wilson. Review of Harman and Kulkarni (2007). *Notre Dame Philosophical Reviews*, 2008.

I. A. Kieseppä. Akaike information criterion, curve-fitting, and the philosophical problem of simplicity. *British Journal for the Philosophy of Science*, 48(1):21–48, 1997.

I. A. Kieseppä. Statistical model selection criteria and the philosophical problem of underdetermination. *The British Journal for the Philosophy of Science*, 52(4):761–794, 2001.

K. B. Korb. Machine learning as philosophy of science. *Minds and Machines*, 14(4): 433–440, 2004.

A. Kukla. Forster and Sober on the curve-fitting problem. *The British Journal for the Philosophy of Science*, 46(2):248–252, 1995.

Y. LeCun, B. E. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. E. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

H. Lin. Unified inductive logic: From formal learning to statistical inference to supervised learning, 2024. arXiv preprint 2412.02969.

N. Linial, Y. Mansour, and R. L. Rivest. Results on learnability and the vapnik-chervonenkis dimension. *Information and Computation*, 90(1):33–49, 1991.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

W. C. Myrvold and W. L. Harper. Model selection, simplicity, and scientific inference. *Philosophy of Science*, 69(S3):S135–S149, 2002.

J. D. Norton. A material theory of induction. *Philosophy of Science*, 70(4):647–670, 2003.

J. D. Norton. *The Material Theory of Induction*, volume 1 of *BSPS open series*. University of Calgary Press, 2021.

J. Otsuka. *Thinking About Statistics: The Philosophical Foundations*. Routledge, 2023.

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

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

J.-W. Romeijn. Philosophy of Statistics. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2025 edition, 2025.

O. Schulte. Formal learning theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2017.

G. Shafer. Comments on (Harman and Kulkarni, 2007). *Abstracta*, 4(3):10–17, 2009.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

E. Sober. *Reconstructing the Past: Parsimony, Evolution, and Inference*. A Bradford Book. The MIT Press, 1988.

E. Sober. *Ockham's Razors: A User's Manual*. Cambridge University Press, 2015.

J. Sprenger. Bayesianism vs. frequentism in statistical inference. In A. Hájek and C. Hitchcock, editors, *The Oxford Handbook of Probability and Philosophy*, pages 382–405. Oxford University Press, 2016.

J. Sprenger and S. Hartmann. *Bayesian Philosophy of Science*. Oxford University Press, 2019.

D. Steel. 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, 2009.

T. F. Sterkenburg. Statistical learning theory and Occam's razor: The core argument. *Minds and Machines*, 35, 3:1–28, 2025.

T. F. Sterkenburg and P. D. Grünwald. The no-free-lunch theorems of supervised learning. *Synthese*, 199(3):9979–10015, 2021.

M. Strevens. Comments on (Harman and Kulkarni, 2007). *Abstracta*, 4(3):27–41, 2009.

P. Turney. The curve fitting problem: A solution. *The British Journal for the Philosophy of Science*, 41(4):509–530, 1990.

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

V. N. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer, 2nd edition, 2000.

V. N. Vapnik and A. J. Chervonenkis. 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.

V. N. Vapnik and A. J. Chervonenkis. *Teoriya Raspoznavaniya Obrazov: Statisticheskie Problemy Obucheniya*. Nauka, Moscow, 1974. Translated as: W. N. Wapnik and A. J. Tscherwonenkis. *Theorie der Zeichenerkennung*. Akademieverlag, Berlin, 1979.

U. von Luxburg and B. Schölkopf. Statistical learning theory: Models, concepts, and results. In Bandyopadhyay and Forster (2011), pages 651–706.

A. Zellner, H. A. Keuzenkamp, and M. McAleer, editors. *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*. Cambridge University Press, 2002.

Munich Center for Mathematical Philosophy (MCMP), LMU Munich
Munich Center for Machine Learning (MCML)
Email address: tom.sterkenburg@lmu.de