Minds and Machines
https://doi.org/10.1007/s11023-022-09592-z

ORIGINAL ARTICLE

Simple Models in Complex Worlds: Occam's Razor and Statistical Learning Theory

Falco J. Bargagli Stoffi[1] · **Gustavo Cevolani**[2] · **Giorgio Gnecco**[2]

Received: 26 March 2021 / Accepted: 2 February 2022

Abstract

The idea that "simplicity is a sign of truth", and the related "Occam's razor" principle, stating that, all other things being equal, simpler models should be preferred to more complex ones, have been long discussed in philosophy and science. We explore these ideas in the context of supervised machine learning, namely the branch of artificial intelligence that studies algorithms which balance simplicity and accuracy in order to effectively learn about the features of the underlying domain. Focusing on statistical learning theory, we show that situations exist for which a preference for simpler models (as modeled through the addition of a regularization term in the learning problem) provably slows down, instead of favoring, the supervised learning process. Our results shed new light on the relations between simplicity and truth approximation, which are briefly discussed in the context of both machine learning and the philosophy of science.

Keywords Simplicity · Complexity · Occam's razor · Machine learning · Statistical learning theory · Vapnik–Chervonenkis dimension · Truth approximation · Sample size

✉ Gustavo Cevolani
gustavo.cevolani@imtlucca.it

Falco J. Bargagli Stoffi
fbargaglistoffi@hsph.harvard.edu

Giorgio Gnecco
giorgio.gnecco@imtlucca.it

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health - Harvard University, Boston, USA

[2] IMT School for Advanced Studies Lucca, Piazza San Francesco 19, 55100 Lucca, Italy

Published online: 08 March 2022

1 Introduction

In many areas of science, a preference for simplicity is often defended as an important methodological principle. Simpler models and theories are not only more manageable from a cognitive and computational point of view, but are also perceived as more likely candidates for true or credible explanations in the relevant domains. This latter idea—that "simplicity is a sign of truth" (*simplex sigillum veri*)—has a venerable history in both science and philosophy. It is often connected with another principle, usually known as Occam's razor: that, all other things being equal, simpler theories, models, and explanations should be preferred over more complex ones. Clarifying these two intuitions, and their relationship, has however proven surprisingly difficult (Baker, 2016; Sober, 2015; Fitzpatrick, 2013).

Following Baker (2016), one can point to three main critical issues. First, an adequate definition of simplicity is needed, distinguishing, for instance, between "simple" as "elegant" (i.e., mathematically or syntactically simple), and "simple" as "parsimonious" (with reference to the number and complexity of entities and commitments assumed on the ontological or metaphysical level). Second, one needs to investigate how the defined notion of simplicity is effectively used by physicists, statisticians, economists, philosophers, and other scholars in their daily work. Third, one should provide a rational justification of the relevant simplicity principles, showing that they are defensible on conceptual grounds and effective in their intended purpose. In short, understanding whether, and why, simplicity is actually evidence of the truth would require a treatment of many different issues on which the philosophical debate is not settled at all.

In this paper, we don't aim at providing a full "theory of simplicity". Instead, we take a more direct approach with a more modest aim, focusing on statistical learning theory, a formal framework in the field of machine learning which allows for a mathematically rigorous treatment of the principles of *simplex sigillum veri* and of Occam's razor. The main aim of statistical learning theory is to investigate properties of learning algorithms according to a statistical framework, deriving bounds on their performance (Bousquet et al., 2004). As we show, statistical learning theory provides interesting insights on the question whether simplicity is a road to the truth, and whether simpler models should be preferred to more complex ones in general. Perhaps not surprisingly, this question has not a simple answer in turn: it subtly depends on a couple of factors, among which the number of observations (training examples) used to find a relevant model in a given family of models—after a suitable training process (which allows, e.g., to properly determine values for the weights of an artificial neural network) – in the first place turns out to be particularly important in our analysis. Another important factor is a measure of complexity (known as the Vapnik–Chervonenkis dimension), which plays a crucial role in the analysis for evaluating the performance of families of "simple" and "complex" models.

We proceed as follows. In Sect. 2, we introduce the basic ideas of statistical learning theory and ask whether a preference for simpler models is always helpful

for approaching the truth in this framework. After answering this question in informal terms, in Sect. 3 we offer a detailed theoretical analysis of the problem and present the main results.[1] In short, we show that, depending on the nature of the data-generating process (either "simple" or "complex"), one of the main tools used by statistical learning theory—namely, regularization—can either reduce or increase the minimal number of training examples necessary to guarantee that a model in the "correct" family of models—i.e., the one including the data-generating process as one of its elements—is selected with probability above any threshold smaller than 1, with respect to the case in which no regularization is used. A discussion of the results obtained in the article appears in Sect. 4, whereas some concluding remarks are reported in Sect. 5; a technical appendix includes the proofs of the main results and discusses some further issues and possible extensions.

2 Simplicity and Truth Approximation in Statistical Learning Theory

Supervised machine learning (SML henceforth) is the sub-field of artificial intelligence studying algorithms that can automatically "learn from experience" using, as available information, labeled data sets. According to the seminal definition by Tom Mitchell, a "computer program is said to learn from experience $\mathcal{E}$ with respect to some class of tasks $\mathcal{T}$ and performance measure $\mathcal{P}$ if its performance at tasks in $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$" (Mitchell, 1997, p. 2). Experience $\mathcal{E}$ is usually constituted by a set of input/output data, colloquially a data set containing the observed realizations of an underlying, unobserved model of data generation. The class of tasks $\mathcal{T}$ relates to the goal of the specific application of SML: e.g., to correctly classify or predict a labeled outcome. Finally, the performance measure $\mathcal{P}$ accounts for the ability of SML to accurately learn the underlying model and varies based on the task being tackled (i.e., predictive error minimization in regression, minimal misclassification rate, and so on).

More formally, given some sample set of data (the "training set"), an SML algorithm *learns* a model—e.g., a function, or a determination of a vector of parameters over a given fixed structure – that should effectively apply also to new data sets, not used during the training (what is usually referred to as the ability of the learned model to *generalize* its predictions to new sets of data). In the case of regression tasks, the training set is a set of points associated with real-valued labels (perhaps representing some observations or measurements of some physical quantity: e.g., pressure and temperature of a gas); in this case, the SML algorithm "learns" a real-valued function modeling the relationship between such points and their labels, and it is able (with high probability) to generalize this relationship to newly available data points outside the training set (the "testing set"). Another typical application of SML is classification, where an SML algorithm learns how to classify different

[1] Part of the theoretical analysis reported in Sect. 3 has been presented without proofs in 2020 at the Sixth International Conference on Machine Learning, Optimization, and Data Science (LOD 2020).

kinds of inputs (e.g., pictures of animals) in different categories (e.g., genera or species). SML has been successfully applied to a number of problems, such as computer vision, speech recognition, e-mail filtering, predictive analytics in economics and business, precision medicine, and others.

In the last few decades, there has been a wide sophistication in approaching simplicity through the lens of statistical theory (Fitzpatrick, 2013). The connections between the SML approach and more traditional discussions of inductive reasoning and human concept learning in philosophy and cognitive science have often been noted and partly investigated.[2] In this connection, statistical learning theory (Vapnik, 2000), to which we now turn, provides an elegant mathematical framework for reasoning about important philosophical and methodological issues, including inductive reasoning, simplicity, and falsificationism (Harman and Kulkarni, 2007, 2011; Corfield et al., 2009; Steel, 2009; Seldin and Schölkopf, 2013).

2.1 Simplicity in Statistical Learning Theory

A basic problem in SML is specifying how to assess different models from a given family W (or different such families) and select the one which performs best given the intended purpose.[3] In statistical learning theory (SLT henceforth), this issue is formulated in terms of a suitable optimization problem. Leaving the technical details for later (see Sect. 3), the basic ideas are the following.

One starts from a set of E examples used to train the family of models, in order to find the "best" such model according to a suitable optimality criterion, based on the available empirical data and hence directly computable.[4] To this purpose, one assumes that a loss function L is defined, measuring the "distance" (read the error) between the prediction of the model and the actual value of each point. Then, the empirical risk $R^{emp,E}(\mathbf{w})$ of each model $\mathbf{w}$ in W is defined as the arithmetic average of the losses (as measured by L) over all E training examples, when the model $\mathbf{w}$ is used.

[2] For philosophical discussions of machine learning see for instance Thagard (1990); Korb (2004); Williamson (2004); Niiniluoto (2005); Williamson (2009); Schubbach (2019, forthcoming); Watson and Floridi (2020, forthcoming); on philosophical issues emerging from the machine learning literature see Corfield et al. (2009); Corfield (2010); Balduzzi (2013); Landgrebe and Smith (2019, forthcoming); Lauc (2020); López-Rubio (2020); finally, we refer the reader to Harman and Kulkarni (2007,, 2011) for introductions at the interface between the two fields.

[3] The choice of the specific family (or families) of models is often guided by some prior knowledge (when available) about the learning problem at hand. For instance, if one wants to learn from supervised examples a function that represents a mass density, then one can choose a family of non-negative functions; if the function to be learned is known to be smooth (based, e.g., on physical considerations), then one can choose a family of smooth functions. When no (or little) prior knowledge is available, one can choose and compare various "general-purpose" families of models: i.e., coming from larger families of functions satisfying properties such as the so-called "universal approximation property" (Cybenko, 1989), which refers to the capability of a family to approximate any continuous function defined over a compact set of an Euclidean space with an arbitrarily small error in the maximum norm.

[4] Here we make the distinction (which is common in SLT) between the training examples, which are used to choose the parameters of the model and the family of models itself, and the test examples, which are used to assess the generalization capability of the resulting "trained" model.

Note that $R^{emp,E}(\mathbf{w})$ can be always calculated for any model $\mathbf{w}$ and set of examples. On the contrary, the so-called *expected risk of the model* can not be in general calculated, even if it can be easily defined as follows. For any model $\mathbf{w}$ in W, its expected risk, denoted $R^{exp}(\mathbf{w})$, quantifies how well the model behaves on average when its predictions are compared to "new" sample data: i.e., new test examples not already employed in the training phase. Accordingly, $R^{exp}(\mathbf{w})$ is defined as the expected loss of the model $\mathbf{w}$—i.e., as the expected value of L for $\mathbf{w}$—with respect to the probability distribution P of the test examples, which in several SML problems (like the ones considered here) is assumed to be the same as the one used to generate the training examples.

The final goal of SML is minimizing the expected risk of models, i.e., to find the model which performs best in its ability to generalize from training to test data. This problem would be immediately solved if the distribution P were known, since then the expected risk of any model $\mathbf{w}$ could be calculated directly. However, in most problems of interest (and most real-world applications of SML) such distribution is typically unknown, or can not be explicitly described. As a consequence, the expected risk is also unknown and must be replaced by a suitable estimate. To this purpose, two strategies can be used.

The first, and the simplest one, is using the empirical risk of the model as a proxy of its expected risk. The Empirical Risk Minimization (ERM henceforth) principle (Vapnik, 2000, Sect. 1.5) amounts to choosing the model $\mathbf{w}$ with the lowest value of $R^{emp,E}(\mathbf{w})$. For a final validation, in order to estimate the expected risk of the specific selected model $\mathbf{w}$, one could then compute its empirical risk for a new set of test examples, different from the training set. One problem with the above ERM strategy, which is crucial in most applications of statistics and SML in general, is "overfitting". In a nutshell, the selected model is "overfitted" to some (training) data when its performance (measured, e.g., by its empirical risk) on those data is very good, but the performance on new data samples is poor. This means that the model is too sensitive to the specific data set used for the training, and it is unable to generalize well to "unknown" data (for instance future data). Using the ERM principle alone (i.e., without any control on the complexity of the family W of models considered), the risk of overfitting can be high, since the best performing model chosen by the algorithm on the training sample may be too sensitive to specific features of this sample which are however not too relevant for the intended learning problem. In this case, given a simplistic view of training data points as a mixture of "signal" and "noise", the algorithm is learning too little signal and too much of the noise in the training data and is, in turn, reproducing that noise on unknown data, hence leading to scarce predictive performance. Overfitting is also closely related to model complexity as the larger is the level of complexity, the higher is the risk that the model will overfit the data (Hastie et al., 2009). A typical example is provided by ERM applied to a family of polynomials in one variable having degree smaller than or equal to a given positive integer D, and the square loss function is used to express the expected and empirical risks. Indeed, if their total number of parameters (i.e., the maximum degree D plus 1, which refers to the constant term) is larger than or equal to E and there are no training examples having the same input vector, then any

model in that family which is learned via the ERM principle has minimum empirical risk, equal to 0 (since any such model is able to interpolate the training data, by an application of Lagrange interpolation theorem, see Barbeau, 2004). However, this holds independently of the generalization capability of the learned model, which is instead expressed by its expected risk.

The second strategy, commonly used to address the overfitting issue and to provide better estimates of $R^{exp}(\mathbf{w})$ than those given by plain ERM, goes under the name of "structural risk minimization" (SRM henceforth). The idea is that the system should minimize not just the empirical risk of the model $\mathbf{w}$, but instead the sum of two terms, the first being $R^{emp,E}(\mathbf{w})$ itself, and the second a "regularization" term (which depends on one or more hyperparameters). Intuitively, "regularizing" a learning problem means making it easier, thus reducing its complexity and potentially mitigating the risk of overfitting. Note that there can be various reasons for introducing a regularization term, for instance higher generalization ability (Hastie et al., 2009, Chapter 7), easier implementation (López-Rubio, 2020), and even improved falsifiability in the sense of Popper (Corfield et al., 2009). More formally, the regularization term employed in SRM is a function of both the size E of the training set and of the so-called Vapnik–Chervonenkis (VC) dimension h of the family of models (or, more precisely, of the associated loss functions, as detailed in Sect. 3). The reader unfamiliar with SLT is referred to Vapnik (2000, Chapter 3) for a precise definition of VC dimension for a family of real-valued functions. Loosely speaking, it is the maximum number h of points that can be classified in all possible 2^h ways by using binary classifiers obtained by applying a threshold to the functions belonging to the family (each classifier corresponds to a specific choice for the function and the threshold). If this maximum does not exist, the VC dimension of the family is $+\infty$. Interestingly, the VC dimension can be interpreted as a measure of the complexity of each family of models (Vapnik, 2000, Chapter 3).[5] In other words, using the VC dimension allows one to keep the absolute value of the difference between the expected and empirical risks "uniformly"—i.e., over the whole family of models – under control with high probability, as highlighted by the regularization term appearing in formulas (4) and (5) in Sect. 3 below.[6] Since the regularization term depends on the family, SRM, as opposed to ERM, aims at minimizing a trade-off between the average loss on the training set

[5] It is worth mentioning that, in SLT, VC dimension is not the only measure of complexity of a family of models: another common such measure is the so-called Rademacher's complexity (Shawe-Taylor and Cristianini, 2004) which, loosely speaking, can be interpreted as the capacity of the family to fit white noise (the smaller that capacity, the less complex the family).

[6] It is worth commenting briefly on the reason why the VC theory is based on a measure of complexity associated with each family of models, instead of a measure of complexity associated with each single model (the reader is referred, e.g., to Mendelson (2003) for more technical details about the following issues). Indeed, it is well-known that "individual" measures of complexity—i.e, one for each function, without reference to the whole family to which it belongs—cannot be easily applied simultaneously to all the elements of a family of models (as this is needed if one wants to replace, e.g., the minimization of the expected risk over that family with the one of the empirical risk). This occurs, e.g., when one tries to apply the so-called Hoeffding's bound (which refers to one function) simultaneously to all the elements of a family of functions with infinite cardinality, by exploiting the so-called union bound technique. In this case, indeed, an infinite—hence, useless—uniform upper bound on the distance between the expected and empirical risk is typically obtained. In the case of families of functions with finite cardinality, instead, such an extension typically provides too loose bounds.

achieved by a model and the complexity of the family to which it belongs (or, which is the same, at maximizing a trade-off between accuracy and simplicity). In this way, several families of models can be easily compared, each one with its own regularization term. It is worth noticing that SRM effectively reduces to ERM for what concerns the choice of a model with the smallest average loss inside each family, since the regularization term is constant over each single family. However, distinct families have typically different regularization terms. Finally, it is the combination of the two terms (the average loss and the complexity of the family of models) that allows one to keep under control the generalization capability of the model finally selected by SRM.

As we shall see in the following, the SLT framework in general, and the principle of structural risk minimization in particular, allow us to discuss interesting issues surrounding both the use of Occam's razor and the idea of simplicity as evidence of truth.

2.2 Simplicity and Truth Approximation in Statistical Learning Theory

In general, one can say that structural risk minimization, as opposed to plain empirical risk minimization, is a mathematical expression of a preference for simpler over more complex models and, in turn, of the principle underlying Occam's razor (Duda et al., 2000; Domingos, 1999). In fact, if two models have the same empirical risk—i.e., they perform equally well with respect to the training set—the one with the smaller regularization term (or equivalently, the one with the smaller VC dimension of the associated family)—i.e., the simpler one – will be chosen under SRM. Thus, "all other things being equal", simpler models are preferred to more complex ones. In this connection, three points are worth noting here.

First, one should note that Occam's razor is often interpreted, following its traditional formulations, as a principle of *ontological* parsimony, according to which "entities should not be multiplied beyond necessity" (Sober, 2015, Chapter 5). In the present, SRM-based framework, this idea can be recovered by noting that a simple (vs. complex) family of models may be interpreted as a family having a small (vs. large) number of parameters (although this is not always the case, as the framework investigated here is more general). In this sense, in several common cases, the VC dimension of a family can be construed as a measure of complexity in terms of the number of parameters involved in its models, where each parameter is associated with a specific input variable (and each of the latter variables may be interpreted as an "entity"): for instance, it is well-known that the VC dimension of a family of linear models is simply equal to the number of its parameters (Vapnik, 2000), hence, to the number of its input variables. It is worth remarking that the case of linear models is quite general, since several families of nonlinear models (e.g., families of polynomial models) can be reduced to families of linear ones, by including additional input variables, derived nonlinearly from the original input variables. For instance, the polynomial model $a + bx + cx^2$ can be also written as the linear model $ax_0 + bx_1 + cx_2$, where $x_0 := 1$, $x_1 := x$, and $x_2 := x^2$.

Second, the *ceteris paribus* clause of the SRM version of Occam's razor is hardly met in practice, since it is virtually impossible that two particular models selected for the comparison (e.g., the two models which minimize the empirical risk on

two different families of models) have exactly the same empirical risk. Thus, SRM embodies a more liberal Occam's razor, where the regularization term quantifies to which extent the simpler model has still to be preferred over the complex model, even when "not all" other things are equal—i.e., when their empirical risks are (either much or even slightly) different. This situation is commonly encountered in practice. In this sense, SRM provides a precise and flexible reconstruction of Occam's razor, allowing to balance in a suitable way the simplicity and the predictive accuracy of a model against each other. Thanks to its solid statistical foundation represented by the VC theory, and in particular, by the so-called VC bound (Vapnik, 2000), SRM provides a principled way for the researcher to weigh predictive accuracy and simplicity (see Sect. 3 for technical details), at the cost of leaving to the researcher's choice only the values to be assigned to a (small) number of tunable (hyper)parameters. Although the choice of the values for these parameters is clearly subject to pragmatic considerations, the parameters themselves (e.g., the confidence parameter δ) have, in any case, a quite precise meaning, which is not left to the discretion of the researcher.

Third, and more interesting for our purposes, one may ask whether SRM favors simplicity and *truth approximation* at the same time, in the sense of selecting models which are both simpler and closer to the "true nature" of the generating process (assuming there is one) underlying the data distribution. If this were the case in general, the *simplex sigillum veri* intuition would be vindicated. As we show, however, the situation is both more interesting and intricate than this. Before proceeding, however, we first need to clarify how SLT in its various forms (i.e., under the ERM or SRM principles) addresses the problem of truth approximation; in other words, what is meant by "the true model" in typical applications and what approximating it amounts to in SLT. In this paper, the "true model" will be defined in a probabilistic sense, i.e., as the model that minimizes the expected risk within the union of the families of models considered (assuming existence and uniqueness of such minimizer, which are both mild conditions according to the results of the analysis reported in the technical Appendix 1). In other words, the true model is construed as the best model that could be obtained in principle (i.e., if the data-generating probability distribution were known) from all the given families of models, using the available input variables. In this connection, two points are worth mentioning. First, the minimum expected risk does not need to be equal to zero in general. A zero expected risk may be obtained by including in the set of input variables all the variables on which the output variable actually depends. In practice, not all these inputs are always available. So, they are usually treated as noise in the statistics and machine learning literature—i.e., linear regression via ordinary least squares. This justifies defining the true model as the best model that can be obtained from the union of the families of models considered, based only on the available input variables, and treating any other unmeasured variable as noise.[7] Second, in case of

[7] In a sense, one can think as if the data-generation process produces a "signal", to which "noise" is overlapped, either linearly or nonlinearly. Such noise can be thought as associated to variables that have not been included in the model of the data-generation process, so they are not available to the learning machine. These variables could be either possibly relevant variables that have not been measured individually, or "disturbances". At this point, two alternative approaches are possible. First, one can simply

multiple models minimizing the expected risk, the true model would be undetermined. However, the minimum expected risk would be still unique and the analysis still focused on finding the minimum expected risk (exactly or approximately).

Accordingly, truth approximation in SLT is construed here as getting as close as possible to the true model, i.e., to get as close as possible to the minimum expected risk in generalizing from training to test data. Since, as said above, the expected risk cannot be typically computed exactly given a finite amount of data, the truth approximation problem translates into one of minimizing a probabilistic upper bound on the expected risk (a bound which is family-dependent in the case of SRM, as we shall see). This means that, even if it will be typically extremely hard to find exactly the true model, one can still come "close to the truth" in probabilistic terms. More precisely, as we show below, SLT allows one both: 1) to find a model with an expected risk not so distant from the one of the true model; 2) to precisely quantify such distance from the truth, which is something that really differentiates SLT from other approaches used in statistics. How closely one can approach the truth via SLT depends on the sample size and on suitable measures of complexity of the family (or families) of models considered (we come back to this issue in Sect. 4).

With the above understanding of what the true model and truth approximation amount to in SLT, we can now turn to our central question—"is simplicity a road to the truth?".

2.3 Simple Models in Complex Worlds

To rigorously investigate the above issue, we frame it, in simplified terms, as follows. First, we assume that the relevant distribution of data may be generated by two different kinds of process, "simple" and "complex". To fix ideas, one may think of a linear vs. a non-linear function generating points on a diagram: the former would count as a simple process, the latter as a complex one. Second, we consider two families of models, the "simple" and the "complex" ones. Again, one may think in terms of linear and non-linear functions, now representing the models to be fitted to the distributions generated by the relevant processes.[8]

Footnote 7 (continued)
define the true model as the signal only. Second, one could define the true model as the union of the signal and the noise, thus providing a complete description of the whole process, which includes every relevant (either omitted or non-omitted) variable. In this interpretation, the signal would represent "one part of the truth", and the noise the "other part of the truth", which cannot be modeled in detail by the machine since it has no individual access to the variables associated to it. In this paper, we follow the first approach, which strikes us as more natural, because in any practical application, this is the most interesting/useful part of the true model according to the second alternative definition; however, the second, alternative approach raises some interesting philosophical issues, whose exploration we leave for future research.

[8] The distinction made by SLT is however subtler than this, since "simple" and "complex" models could both be linear models, but with different dimensions of the basis of the vector space; moreover, one family of non-linear models could be less complex than a second family of linear models with large dimension of such basis. In all these cases, the measure of complexity of each family of models is provided, e.g., by its associated VC dimension.

Finally, we compare two possible situations, where, intuitively, simple and complex models are fitted to both simple and complex "worlds". The two scenarios are as follows:

A. In the "simple world" scenario, the real-valued labels of the sets of training examples are generated by a simple data generating process,[9] to which a small amount of noise is added, to avoid the possibility that the "correct" family of models (that from which the label-generating model comes) is trivially, and immediately, identified by the algorithm. Then, two different families of models are fitted to the training data, using principles from SLT. The first one includes simple models, whereas the other one includes complex models. Finally, the best of such models (again, according to SLT) is selected automatically.
B. In the "complex world" scenario, the same happens but the labels of the sets of training and test examples are generated by a complex process in the first place.

Ideally, the learning algorithm should select a model corresponding to the "true nature" of the data generating process in both worlds: i.e., a simple model in case A, and a complex one in case B. However, we know that, if the selection is guided by SRM, the algorithm will favor simpler models over more complex ones, due to the introduction of the regularization term, depending in turn on the VC dimensions of the respective families.

We can now state our main question as follows: does regularization always favor learning; or: is a preference for simpler over complex models, as embodied in SRM, functional for approximating the truth about the world? Interestingly, the answer depends on both the kind of scenario we deal with (A vs. B) and, crucially, on the number E of training examples used to assess the models. To anticipate the results from the theoretical analysis in the next Sect. 3, we obtain the following:

1. If the number of training examples is sufficiently large, in both cases A and B a model in the "correct" family (i.e., the one from which the data set is generated) is selected with probability above any threshold smaller than 1, independently of whether regularization is used or not (i.e., both under ERM and SRM).
2. In case A (simple world scenario), regularization improves learning even if the number of training examples is relatively small; in other words, regularization

[9] Although test examples are not used in the present analysis, what is stated above holds in principle also for the set of such examples, in the common case in which they are assumed to be generated by the same distribution as the training examples. This depends on the fact that, in machine learning, test examples are used to produce an empirical estimate of the expected risk of the learned model, which is less prone to overfitting than the one obtained from the training set, since the test examples have not been used to select that specific model.

reduces the minimal number of training examples necessary to guarantee that a model in the correct (i.e., simple) family is selected with probability above any threshold smaller than 1, with respect to the case in which no regularization is used (i.e., when only empirical risk is minimized).

3. In case B (complex world scenario), regularization hampers learning if the number of training examples is relatively small: in this case, models in the simple (i.e., incorrect) family tend to be selected even if the world is complex.

The above results are in line with our intuitions concerning the role of simplicity in approaching a "truth" that can be simple or complex in the sense defined. More interesting questions arise if we move from a qualitative to a quantitative analysis. In particular, we may ask: Assuming that the truth is not simple, how many more training examples are needed to identify, with a given probabilistic confidence, a model in the correct family (if not the "true" model itself, which is clearly a more difficult learning task), when one moves from ERM to SRM? Conversely, we also ask: Assuming that the truth is instead simple, how many more training examples are needed to identify, with a given probabilistic confidence, a model in the correct family, when one uses ERM instead of SRM? To the best of our knowledge, the above questions have not been addressed before. Here, we answer both of them, showing how one can precisely quantify the loss in performance (translating here into a larger upper bound on the expected risk) one incurs when SRM is applied under the assumption that the truth is not simple.

In sum, we conclude that the principle of Occam's razor, at least as expressed by the introduction of regularization in SRM, can both favor and hamper learning and hence convergence towards the truth. On the one hand, if "the world" is simple, the SRM "regularization razor" helps in reducing the amount of information (as expressed by the number of training examples) needed to guarantee that a model is chosen (with probability larger than the given threshold), that correctly represents the underlying generating process. On the other hand, if "the world" is complex, a large amount of information is needed to guarantee the above convergence; otherwise, if the size of the set of observations (training examples) is "small", the razor can favor the selection of simple, but incorrect, models. In this sense, simplicity is not necessarily "a road to the truth", even if one could still prefer simpler models for many reasons, like easier implementation, higher computational scalability and so on.

In addition, the theoretical analysis in the following Sect. 3 has a couple of interesting implications for the analysis of simplicity and truth approximation within SML and statistics in general. First, we show how one can rigorously talk about "the true model" and "approximation to the truth" in such contexts, where these notions can be usefully defined. Second, we quantify, under suitable assumptions, the minimal number of training examples needed to achieve a given confidence on the probability of finding the correct family of models by using ERM and SRM, and the behavior of this truth approximation strategy in the two cases in which the truth is, respectively, "simple" or "complex". Third, we show how SRM is more appropriate than other statistical methods used to prevent overfitting—such as the Akaike Information Criterion (see Burnham and Anderson, 2002)—at least as far as one central

issue is concerned: i.e., quantifying the minimal sample size required to meet certain requirements on the performance of the learned model. In Sect. 4, we shall further discuss these and related issues.

3 Theoretical Analysis

The analysis is based on the well-known Vapnik–Chervonenkis (VC) two-sided upper bound (also called VC bound) on the difference between expected risk and empirical risk (the latter based on a training set of size E), see Vapnik (2000, Sect. 3.7). Two families of models, S ("simple") and C ("complex") are considered, each parametrized by a vector ($\mathbf{w}_S$ for S, $\mathbf{w}_C$ for C, which vary respectively in two sets W_S and W_C), with possibly different dimension for each family. The expected and empirical risks of the various models are denoted, respectively, by $R_S^{exp}(\mathbf{w}_S)$ and $R_S^{emp,E}(\mathbf{w}_S)$ for the models in S, and by $R_C^{exp}(\mathbf{w}_C)$ and $R_C^{emp,E}(\mathbf{w}_C)$ for the models in C. All these risks are computed using, respectively, bounded loss functions $L_S(\cdot, \mathbf{w}_S)$ and $L_C(\cdot, \mathbf{w}_C)$, which have the same interval $[A, B]$ (where $A, B \in \mathbb{R}$, with $A < B$) as codomain. Such functions are parametrized, respectively, by the vectors $\mathbf{w}_S$ and $\mathbf{w}_C$. In more details, for a given $\mathbf{w}_S \in W_S$, the expected risk of the loss function $L_S(\cdot, \mathbf{w}_S)$ is its expectation when its first argument (say, $\mathbf{z} = (\mathbf{x}, y) \in Z = X \times Y$) is modeled as a random vector with probability distribution P:

$$R_S^{exp}(\mathbf{w}_S) := \int_Z L_S(\mathbf{z}, \mathbf{w}_S)\, dP(\mathbf{z})\,. \tag{1}$$

In the above, $\mathbf{x} \in X$ represents an input vector, whereas $y \in Y$ is a scalar output, which one would like to model approximately as a function of $\mathbf{x}$ (using a suitably-selected model from one of two families). In the SLT framework, the probability distribution P of $\mathbf{z}$ is typically modeled as unknown. In such situation, the expected risk cannot be computed. However, it can be approximated by its empirical risk (which, instead, can be evaluated). Given a finite number E of so-called training examples $\mathbf{z}_e \in Z$ (for $e = 1, \ldots, E$), the empirical risk of $L_S(\cdot, \mathbf{w}_S)$ is the arithmetic average of the losses $L_S(\mathbf{z}_e, \mathbf{w}_S)$ incurred over such training examples, i.e., one has

$$R_S^{emp,E}(\mathbf{w}_S) := \frac{1}{E} \sum_{e=1}^{E} L_S(\mathbf{z}_e, \mathbf{w}_S)\,. \tag{2}$$

Similar definitions hold for the case of $\mathbf{w}_C \in W_C$. In this work, the same (unknown) probability distribution P for $\mathbf{z}$ is assumed for the expected risks of all the models. To enforce the boundedness of the loss functions $L_S(\cdot, \mathbf{w}_S)$ and $L_C(\cdot, \mathbf{w}_C)$ for cases of practical interest (e.g., the square loss, see later), the only assumption made on the probability distribution P is that it has a given compact support (which excludes from the analysis, e.g., the Gaussian probability distribution, but does not exclude a "truncated" Gaussian probability distribution). Moreover, for a fair comparison, the two loss functions $L_S(\cdot, \mathbf{w}_S)$ and $L_C(\cdot, \mathbf{w}_C)$ are assumed to have the same functional form (e.g., they are both square losses).

Once a model in one of the two families has been selected (according to a suitable criterion, as discussed later), its expected risk can be approximated by its empirical risk computed on a different set of test examples. For fairness purposes, this test set has to be different from the training set (and in principle, its number of elements could be even larger than the number E used to evaluate the empirical risk on the training set, in order to have a better approximation of the expected risk).

In the following, we consider the case in which the loss functions $L_S(\mathbf{z}, \mathbf{w}_S)$ and $L_C(\mathbf{z}, \mathbf{w}_C)$ represent regression. Focusing, e.g., on the case of the models in S, this means that each such model (hence, each choice for $\mathbf{w}_S$) is also associated with an input-output relationship $f_S(\mathbf{x}, \mathbf{w}_S)$ from X to Y, and the loss function $L_S(\mathbf{z}, \mathbf{w}_S)$ has actually the form $L_S(y, f_S(\mathbf{x}, \mathbf{w}_S))$. Examples are provided by the quadratic loss $L_S(y, f_S(\mathbf{x}, \mathbf{w}_S)) = \left(y - f_S(\mathbf{x}, \mathbf{w}_S)\right)^2$ and the absolute loss $L_S(y, f_S(\mathbf{x}, \mathbf{w}_S)) = \left|y - f_S(\mathbf{x}, \mathbf{w}_S)\right|$. To avoid burdening the notation, from now on the function $f_S(\mathbf{x}, \mathbf{w}_S)$ is assumed to be embedded into the expression $L_S(\mathbf{z}, \mathbf{w}_S)$. A similar remark holds for the case of the models in C.

Let both the training and test output data be generated according to a particular model in S or in C,[10] with the output perturbed by 0-mean independent additive noise γ,[11] having small variance σ^2. For instance, in case the particular model is in the family S and is characterized by a specific value $\mathbf{w}_S$ for the parameter vector, this means that

(a) $\mathbf{x}$ is generated according to the associated marginal distribution of P;
(b) $y = f(\mathbf{x}, \mathbf{w}_S) + \gamma$, where $f(\mathbf{x}, \mathbf{w}_S)$ is, again, the input-output relationship modeled by $\mathbf{w}_S$.

Let also

$$\Delta := \min_{\mathbf{w}_C \in W_C} R_C^{exp}(\mathbf{w}_C) - \min_{\mathbf{w}_S \in W_S} R_S^{exp}(\mathbf{w}_S)\,, \tag{3}$$

supposing, without significant loss of generality, that such minima exist and are uniquely achieved (see the technical Appendix 1 for a discussion about this issue). Of course, Δ is in general unknown, but in the following analysis we suppose that its modulus $|\Delta|$ (but not its sign) is provided to the learning machine, e.g., by an oracle. The availability of an oracle is often assumed in the literature on theoretical machine learning (see, e.g., Shalev-Shwartz and Ben-David, 2014; Shi and Iyengarand, 2020 for some examples), in order to prove bounds on the performance of

[10] This assumption is made to simplify the analysis, since it allows to focus the comparison between only two families of models, characterized by two different VC dimensions. However, the analysis of this section is expected to extend easily to the more realistic case in which the data-generating model belongs to one among a larger number of families with different VC dimensions, and, among other assumptions, one does not know a-priori the specific family to which it belongs.

[11] Without this additive noise, one would have always a 0 minimum empirical risk in the correct family of models, which would make its detection trivial, in case the minimum empirical risk on the other family were larger than 0. One can also observe that this assumption translates into partial information about the probability distribution P of $\mathbf{z}$, however, without completely specifying it.

learning machines. Even though such an assumption may be unrealistic in some cases, the bounds obtained are often useful to understand the theoretical limitations/capabilities of learning. The reader is referred to the end of this section for a discussion about which parts of the following analysis do not depend on the availability of an oracle to the learning machine, and how that assumption can be relaxed.

Moreover, we assume that the two families of models are non-nested. This means essentially that S is not strictly contained in C, since the opposite case, where C is strictly contained in S, cannot hold given that S and C refer, respectively, to the "simple" and "complex" family of models. Finally, we assume that $\Delta \neq 0$ holds; this is a technical assumption needed in order to guarantee that the inequalities (14), (15) and (16) reported later in this section hold. Note that, if $\Delta = 0$, the minimizers of the expected risk over the two families achieve the same expected risk and, due to the non-nestedness assumption, such minimizers are guaranteed to be different. If instead $\Delta \neq 0$, as we assume here, one can have either that the best model in S is better than the best model in C (if $\Delta > 0$), or that the best model in C is better than the best model in S (if $\Delta < 0$).[12]

The following theorem follows directly from the classical VC bound from the VC theory (Vapnik, 2000, Sect. 3.7), applied separately to the two families of functions. For a proof of the VC bound, the reader is referred to Vapnik (1998). The theorem is formulated here in terms of the Vapnik–Chervonenkis (VC) dimensions of the two sets of loss functions $\{L_S(\cdot, \mathbf{w}_S), \mathbf{w}_S \in W_S\}$ and $\{L_C(\cdot, \mathbf{w}_C), \mathbf{w}_C \in W_C\}$. It is worth noting that, for most regression problems of practical interest, these are approximately equal to the VC dimensions of the respective families of functions $\{f_S(\cdot, \mathbf{w}_S), \mathbf{w}_S \in W_S\}$ and $\{f_C(\cdot, \mathbf{w}_C), \mathbf{w}_C \in W_C\}$,[13] see Cherkassky and Mulier (2007, Sect. 4.2.1).[14]

Theorem 3.1 *Let h_S and h_C be the VC dimensions, respectively, of the two sets of loss functions $\{L_S(\cdot, \mathbf{w}_S), \mathbf{w}_S \in W_S\}$ and $\{L_C(\cdot, \mathbf{w}_C), \mathbf{w}_C \in W_C\}$, with $h_S < h_C$ (being, indeed, the models in S simpler than those in C), and let the size of the training set be $E > h_C > h_S$. Finally, let the confidence parameter $\delta \in (0, 1)$ be given. Then, the two following bounds hold with corresponding probabilities $p_S \geq 1 - \frac{\delta}{2}$ and $p_C \geq 1 - \frac{\delta}{2}$ with respect to the generation of a training set whose examples are drawn independently from the same probability distribution (i.e., they are independent and identically distributed)*:

[12] Since, as better detailed later, the goal of the successive analysis is to investigate the probability that a model belonging to the family associated with the smallest of the two respective minimum expected risks is selected by applying, respectively, the ERM/SRM principle, the non-nestedness assumption is essential for that analysis. Indeed, without such an assumption, one could simply restrict the attention to models belonging to the "complex" family C. Moreover, the case $\Delta > 0$ considered in the successive analysis can occur only if the two families are non nested.

[13] Examples of computations of VC dimensions of families of (either loss or non-loss) functions are provided in Vapnik (2000).

[14] For this reason, in some parts of the article the term "VC dimension of a family of functions" is used for simplicity as a shortcut for "VC dimension of the family of loss functions associated (via function composition with a given loss) to another family of functions."

$$\sup_{\mathbf{w}_S \in W_S} |R_S^{exp}(\mathbf{w}_S) - R_S^{emp,E}(\mathbf{w}_S)| \leq \varepsilon(h_S, E, \delta)\,, \tag{4}$$

$$\sup_{\mathbf{w}_C \in W_C} |R_C^{exp}(\mathbf{w}_C) - R_C^{emp,E}(\mathbf{w}_C)| \leq \varepsilon(h_C, E, \delta)\,, \tag{5}$$

where $\varepsilon(h, E, \delta) := (B - A)\sqrt{\frac{h \ln \frac{2eE}{h} - \ln \frac{\delta}{8}}{E}}$ *is the regularization term, and* h *can be either* h_S *or* h_C.

By Theorem 3.1, it follows that, with probability $p_S \geq 1 - \delta_S$, the infimum of the empirical risk $R_S^{emp,E}(\mathbf{w}_S)$ over $\mathbf{w}_S \in W_S$ differs at most by $\varepsilon(h_C, E, \delta)$ from the infimum of the expected risk $R_S^{exp}(\mathbf{w}_S)$ over $\mathbf{w}_S \in W_S$. Similarly, with probability $p_C \geq 1 - \delta_C$, the infimum of the empirical risk $R_C^{emp,E}(\mathbf{w}_C)$ over $\mathbf{w}_C \in W_C$ differs at most by $\varepsilon(h_C, E, \delta)$ from the infimum of the expected risk $R_C^{exp}(\mathbf{w}_C)$ over $\mathbf{w}_C \in W_C$. Finally, it follows from Theorem 3.1 that, under its assumptions, both bounds (4) and (5) hold simultaneously with probability $p \geq 1 - \delta$. Indeed, by applying the so-called union bound technique (Mendelson, 2003), the probability that none of them holds is smaller than $\frac{\delta}{2} + \frac{\delta}{2} = \delta$.

The following corollary to Theorem 3.1 is obtained by reversing the roles of the confidence parameter and of the regularization term (this is another standard way of expressing SLT bounds in the literature).

Corollary 3.1 *Let* h_S *and* h_C *be the VC dimensions, respectively, of the two sets of loss functions* $\{L_S(\cdot, \mathbf{w}_S), \mathbf{w}_S \in W_S\}$ *and* $\{L_C(\cdot, \mathbf{w}_C), \mathbf{w}_C \in W_C\}$, *with* $h_S < h_C$, *and let the size of the training set be* $E > h_C > h_S$. *Finally, let the regularization term* $\varepsilon > 0$ *be given. Then, the two following bounds hold with corresponding probabilities* $p_S \geq 1 - \frac{\delta(h_S, E, \varepsilon)}{2}$ *and* $p_C \geq 1 - \frac{\delta(h_C, E, \varepsilon)}{2}$ *with respect to the i.i.d. generation of the training set*:

$$\sup_{\mathbf{w}_S \in W_S} |R_S^{exp}(\mathbf{w}_S) - R_S^{emp,E}(\mathbf{w}_S)| \leq \varepsilon\,, \tag{6}$$

$$\sup_{\mathbf{w}_C \in W_C} |R_C^{exp}(\mathbf{w}_C) - R_C^{emp,E}(\mathbf{w}_C)| \leq \varepsilon\,. \tag{7}$$

where $\delta(h, E, \varepsilon) := \min\left(8 \exp\left(-\frac{E}{(B-A)^2}\varepsilon^2 + h \ln \frac{2eE}{h}\right), 1\right)$ *is the confidence parameter.*

Corollary 3.1 implies that, for each family, the infimum of the empirical risk over the family converges in probability to the infimum of the expected risk over the same family. Indeed, for each $\varepsilon > 0$, both $\delta(h_C, E, \varepsilon)$ and $\delta(h_S, E, \varepsilon)$ tend to 0 as the sample size E tends to $+\infty$. This is essentially due to the functional form of the regularization term in Theorem 3.1.

According to the Empirical Risk Minimization (ERM) principle (Vapnik, 2000, Sect. 1.5), one selects, for each family, the model that minimizes the empirical risk on that family, i.e., the one associated, respectively, with the parameter choice

$$\hat{\mathbf{w}}_S := \text{argmin}_{\mathbf{w}_S \in W_S} R_S^{emp,E}(\mathbf{w}_S) \tag{8}$$

and

$$\hat{\mathbf{w}}_C := \text{argmin}_{\mathbf{w}_C \in W_C} R_C^{emp,E}(\mathbf{w}_C)\,. \tag{9}$$

Again, without significant loss of generality, we assume that these minimizers exist and are unique (see the technical Appendix 1 for a discussion about this issue; relationships among minimizers of the empirical risks and minimizers of the expected risks are discussed in the technical Appendix 2). Finally, of these two parameters $\hat{\mathbf{w}}_S$ and $\hat{\mathbf{w}}_C$, the one achieving the smallest between the empirical risks

$$R_S^{emp,E}(\hat{\mathbf{w}}_S) \tag{10}$$

and

$$R_C^{emp,E}(\hat{\mathbf{w}}_C) \tag{11}$$

is chosen. If the Structural Risk Minimization (SRM) principle (Vapnik, 2000, Sect. 4.1) is chosen, instead, then, of the two parameters, the one associated with the smallest between the regularized empirical risks

$$R_S^{emp,E}(\hat{\mathbf{w}}_S) + \varepsilon(h_S, E, \delta) \tag{12}$$

and

$$R_C^{emp,E}(\hat{\mathbf{w}}_C) + \varepsilon(h_C, E, \delta) \tag{13}$$

is chosen. For simplicity, the case of ties is not considered in the following analysis. It is worth examining how SRM relates to Occam's razor. In the case of identical empirical risks for the simple and complex models (i.e., when $R_S^{emp,E}(\hat{\mathbf{w}}_S) = R_C^{emp,E}(\hat{\mathbf{w}}_C)$), according to SRM, the simplest one is preferred, because its regularization term $\varepsilon(h_S, E, \delta)$ is smaller than the one $\varepsilon(h_C, E, \delta)$ for the more complex model.

This corresponds to the classical Occam's razor version according to which other things being equal, simpler models are better than more complex ones. However, having two models with exactly the same empirical risk is quite unlikely to occur in practice. When the two empirical risks are different, the regularization term quantifies to which extent the simpler model has still to be preferred over the complex model, i.e., the maximum difference in the empirical risks for which this preference can be expressed. Thus, SRM embodies an interesting, more sophisticated reconstruction of intuition underlying the classical version of Occam's razor.

Using (4) and (5), one obtains the following results.

Theorem 3.2 *Let the assumptions of Theorem 3.1 hold. If the ERM principle is applied, then a model in the "correct" family – i.e., one coming from the same family from which the training/test output data are generated, even though it may not coincide with the best such model in terms of the expected risk – is selected with probability $p \geq 1 - \delta$ with respect to the i.i.d. generation of the training set if*[15]

$$\varepsilon(h_S, E, \delta) + \varepsilon(h_C, E, \delta) < |\Delta| \,. \tag{14}$$

Theorem 3.3 *Let the assumptions of Theorem 3.1 hold. If the SRM principle is applied, then one can distinguish two cases*:

1. *Let $\Delta > 0$. In this case, a model in the correct family (S) is selected with probability $p \geq 1 - \delta$ with respect to the i.i.d. generation of the training set if*

$$\varepsilon(h_S, E, \delta) + \varepsilon(h_C, E, \delta) + \varepsilon(h_S, E, \delta) - \varepsilon(h_C, E, \delta) = 2\varepsilon(h_S, E, \delta) < |\Delta| \,. \tag{15}$$

2. *Let $\Delta < 0$. In this case, if the SRM principle is applied, then a model in the correct family (C) is selected with probability $p \geq 1 - \delta$ with respect to the i.i.d. generation of the training set if*

$$\varepsilon(h_S, E, \delta) + \varepsilon(h_C, E, \delta) + \varepsilon(h_C, E, \delta) - \varepsilon(h_S, E, \delta) = 2\varepsilon(h_C, E, \delta) < |\Delta| \,. \tag{16}$$

The proofs of Theorems 3.2 and 3.3 simply require the derivation of the conditions (14), (15), and (16) above, which is reported in the technical Appendix 1.3. It is worth noting that, since the condition $\Delta \neq 0$ has been assumed, all the bounds (14), (15), and (16) are guaranteed to hold if E is large enough. Since, for $E > h_C > h_S$, one has

$$2\varepsilon(h_S, E, \delta) < \varepsilon(h_S, E, \delta) + \varepsilon(h_C, E, \delta) < 2\varepsilon(h_C, E, \delta) \,, \tag{17}$$

one can also conclude the following regarding our two main scenarios.

A. In the "simple world" scenario, the regularization term $\varepsilon(h, E, \delta)$ (for $h = h_S, h_C$) is beneficial for learning. Indeed, the minimal size of the training set for which condition (15) holds is smaller than or equal to the minimal size of the training set for which condition (14) holds. It is worth recalling here that condition (15) is associated to the selection by the SRM principle of a model in the correct "simple" family, whereas condition (14) is associated to the selection by the ERM principle of a model in same correct family.
B. In the "complex world" scenario, no regularization has a better performance guarantee, in the sense that the minimal size of the training set for which condi-

[15] Here and for (15) and (16), the weak inequality $p \geq 1 - \delta$ can be replaced by the strict inequality $p > 1 - \delta$. The former inequality has been preferred to keep the notation uniform in the paper. Instead, the strict inequality in conditions (14), (15), and (16) is needed to avoid ties, thus guaranteeing the selection of the correct family of models (see the technical Appendix 1.3 for a derivation of such conditions).

tion (16) - which is associated to the selection by the SRM principle of a model in the correct "complex" family - holds is larger than or equal to the minimal size of the training set for which condition (14) holds.

We conclude this section by discussing the assumption made previously about the knowledge of $|\Delta|$ by the learning machine, e.g., via an oracle. This assumption has been included in the analysis because it allows the machine to decide autonomously "when to stop collecting training examples" (i.e., the machine can choose autonomously the minimal sample size under which the probability of finding the correct family is above a given threshold). In practice, $|\Delta|$ can be hardly known exactly, and this limits the applicability of the results of the previous analysis. Nevertheless, the assumption itself can be relaxed, without changing significantly the results obtained: e.g., one can replace $|\Delta|$ in (14), (15), (16) with a lower bound on it (also in this case, the learning machine would be able to decide autonomously when to stop collecting examples in order to achieve a similar desired probabilistic guarantee on the selection of the correct family, expressed in terms of that bound). The availability of such a lower bound to the learning machine is surely a milder assumption than the exact knowledge of $|\Delta|$ (although one may wonder in this case, too, under which circumstances this assumption is satisfied in a specific practical application). Nevertheless, it is worth mentioning that, even in the case in which neither $|\Delta|$ nor a lower bound on it were known to the learning machine, the left-hand sides of the inequalities (14), (15), (16) are known to it.[16] Hence, for every possible "guess" of $|\Delta| \neq 0$ by the learning machine, the latter is able to compute the minimal sample size under which each of these inequalities hold, and to compare the resulting minimal sample sizes for the various cases. Interestingly, the conclusions of the analysis of the two main scenarios, which have been reported above, do not depend on the specific value of the guess of $|\Delta|$: e.g., in the "simple world" scenario, the minimal size of the training set for which condition (15) holds is smaller than or equal to the minimal size of the training set for which condition (14) holds, independently on the specific value of the guess of $|\Delta|$.

4 Discussion

Occam's razor expresses the idea that, in the study of natural and social phenomena, simpler theories, models, and explanations should be preferred over more complex ones, other things being equal. The intuition behind this principle is sometimes justified in terms of truth approximation: simpler theories are more likely true than more complex competitors. An analysis of these ideas raises notoriously intricate issues, which have been traditionally discussed in the philosophical literature (Baker, 2016; Sober, 2015; Swinburne, 1997). A more liberal version of Occam's razor is also

[16] This holds when the VC dimensions of the two families are known/easy to compute. Otherwise, again, suitable bounds could be used to replace the VC dimensions, without changing significantly the following conclusions.

needed when "not all" other things are equal, and one has still to perform model selection.

Interestingly, many of these issues critically resurface in many fields, including statistics and artificial intelligence, especially in the context of model selection, where the problems of overfitting and generalizability are critical. Here, we focused on statistical learning theory, a mathematical framework which studies the optimality of model selection for various problems in machine learning, including the field of supervised machine learning. Recent contributions highlighted interesting connections between SLT and earlier proposals in the philosophy of science, including Popper's characterization of simplicity in terms of falsifiability (Corfield et al., 2009) and theories of inductive reasoning (Harman and Kulkarni, 2007, 2011).

This paper contributes to this line of research by focusing on the following question: under which conditions does Occam's razor favor truth approximation construed as the selection of a model from the "right" family—i.e., the one corresponding to the "true" data generating process? Or, stated a bit more technically, under which conditions does a preference for simpler models as construed in SRM by adding a regularization term to the optimization problem favor truth approximation construed as the selection of a model from the family corresponding to the "true" data generating process? We provided an answer to this question with a theoretical analysis of how (families of) "simple" and "complex" models perform when learning is formalized based either on the ERM or on the SRM principle.

The main upshot is that, while the preference for simplicity may indeed favor truth approximation, in some cases it may also slow down the learning process, in the sense that, as compared to ERM, SRM may increase the minimal sample size needed to find a model in the correct class with a desired probabilistic guarantee (although no computational complexity analysis is performed here; see Sect. 4 for a discussion). This happens, roughly, when the training set is relatively small and then "simple" models can be selected even if the process generating the data is "complex" (in a suitably defined sense). In other words, the SML model has too little information on the realized outcomes (colloquially, too little experience of the world) to correctly learn the true underlying model that generated the data. In those cases, plain ERM performs better than SRM. It should be noted, however, that, even in those cases, the relative advantage of complex models over simpler ones would be typically counterbalanced by several disadvantages, such as a more difficult implementation. For instance, training a model characterized by a large number of parameters—i.e., finding the optimal values of its parameters, according to a suitable optimality index—could be subject to the curse of dimensionality (Bellman, 1957). Thus, the practicing scientists may well still prefer a simpler model, because, e.g., of its easier implementation.

Another result of our analysis has to do with the characterization of the idea of truth approximation in a SML context. Such idea is often left implicit and not adequately analyzed; the same holds for its connections with the Occam's razor principle. In this paper, truth approximation is explicitly interpreted as the process of increasingly approaching, as the size of the training set increases, the minimum expected risk over a given family (or families) of models. This is achieved via the minimization of an upper bound on it, which holds with high probability

"simultaneously" for all the models considered. Note that defining the true model as the one minimizing the expected risk over the union of families of models considered is a very reasonable assumption, commonly made in the statistics and machine learning literature. For instance, it is well-known (see, e.g., Cucker and Smale, 2001) that the regression function—defined as the conditional expectation of the output variable given the input variables, hence a good candidate for being considered the true model, according to the discussion presented in Sect. 2.2—coincides with the minimizer of the expected square loss over a family of models (or over the union of the families of models), when it is an element of that family (or union of families). However, it cannot be practically computed if the data generating distribution is unknown.

In view of the above definitions, our results in Sect. 3 can be more precisely stated as follows. First, the probability of finding via SLT, if not the true model itself, at least a sufficiently good approximation of the true model increases with the sample size, where "sufficiently" can be specified, e.g., by making a suitable bound hold. In other words, one can find with increasing probability a model having an expected risk sufficiently close to the minimum one.[17] Moreover, a similar result holds also for finding "exactly" the true model if one is prepared to accept further assumptions. The first assumption (which holds, e.g., in the case of strictly convex learning problems (Shalev-Shwartz and Ben-David, 2014, Ch. 12)) is that the true model is the unique minimizer of the expected risk. The second (which is always satisfied when the uniqueness assumption holds and the number of admissible models is finite) is that its minimum expected risk is "well-separated" from the expected risks of all the other admissible models. This means that the difference between the expected risk of any admissible model which is different from the minimizer and the minimum expected risk is larger than some real number $\alpha > 0$. If both assumptions hold, then also the probability of finding such true model "exactly" increases with the sample size. Finally, when the sample size is "too small" in a suitable specified sense,[18] then SLT cannot typically find nor approach the truth. Of course, this is certainly not a drawback of SLT, because the same holds for every other possible approach of truth-finding or truth-approximating.

[17] It is also worth mentioning that, under additional mild conditions such as smoothness (e.g., Lipschitz continuity) of the loss functions and of the models (also called "hypothesis functions" in this setting), an upper bound on the expected risk translates into an upper bound on the pointwise risk (Zoppoli et al., 2020). In this way, an upper bound on the distance between the pointwise prediction of the learned model and the one coming from the true model can also be obtained.

[18] E.g., less than a constant times the largest of the VC dimensions of the family of models involved. The constant itself could be chosen by taking into account the maximum range of variation $B - A$ of the loss function, and the confidence parameter δ in the VC bound.

5 Concluding Remarks

We conclude by highlighting some pros and cons of our analysis, some connections with other proposals in the literature, and some interesting issues which are left for future research.

First, it is important to remark that algorithms in machine learning typically learn functions, where scientific models are usually thought of as more general mathematical structures. As a consequence, the framework investigated in this work is not able to model all possible facets of Occam's razor (which is also applied to scientific models).

Second, it is worth noting that the bounds computed in this paper only provide sufficient (not necessary) training sample sizes associated with the desired guarantee of finding the correct family of models. Nevertheless, we believe they can be still quite useful in practice, since they belong to the class of so-called "distribution-independent bounds" in SLT: i.e., they hold for any probability distribution that satisfies the mild assumptions of the article: e.g., in the case of the quadratic loss, the fact that the distribution has a given compact support.[19]

Another issue has to do with the assumption we make that the minimum expected risk of a model in the "true family" is strictly smaller than the minimum expected risk in the "wrong family", i.e., that $|\Delta| \neq 0$ holds. This guarantees that the probability of finding the correct family tends to 1 as the sample size E tends to $+\infty$, when either the ERM or SRM principle is applied.[20] One can observe that not knowing a-priori which is the correct family excludes the possibility of comparing nested families of models in our analysis. For example, in an application of this analysis, instances of "simple" and "complex" families of models cannot be simply taken as families of "low-degree" and "high-degree polynomials", respectively, because such families do not satisfy the non-nestedness assumption. Instead, admissible instances are any two families of models, respectively "with small VC dimension" and "with large VC dimension", provided they are non-nested. As already discussed, the non-nestedness assumption is essential for the present analysis. Hence, it may be removed only by considering a different formulation of the comparison between

[19] Even though in principle there could exist "malign" probability distributions for which, under a complex truth, the necessary "distribution-dependent" training sample size for SRM is actually lower than that for ERM, one has to recall that, in the present setting, the actual probability distribution is typically unknown. This implies that, in order to obtain a "distribution-independent" guarantee, one should take the supremum of the necessary "distribution-dependent" guarantee on all admissible probability distributions. When doing this, such "malign" distributions are not expected to make the resulting necessary "distribution-independent" training sample size for SRM be actually lower than that for ERM (of course, further research is needed to confirm this expectation). This is also justified by the fact that such scenarios are expected to be the exception, not the rule. We believe that this is an interesting direction of further research.

[20] This is obtained in the following way, as a by-product of the analysis reported in Sect. 3. For each sample size E, one looks for the minimal value $\delta(E)$ of $\delta > 0$ for which each of the inequalities (14), (15), and (16) holds. Since the choice $\delta = \frac{1}{E}$ satisfies all the three inequalities for E large enough (due to the functional form of $\epsilon(h, E, \delta)$), one gets $\delta(E) \leq \frac{1}{E}$ for large E, then $\lim_{E \to +\infty}(1 - \delta(E)) = 1$.

"simple" and "complex" families of models. For this reason, the possibility of adapting our analysis to the case of nested families is left to further research.

Despite the above limitations, our analysis has a couple of important advantages over other proposals in the literature. In particular, the VC bound on which we base our analysis is a "nonasymptotic bound": i.e., it holds for a finite—possibly even "small"—number of training set data, and depends on such a number. On the contrary, other commonly-used statistical bounds are "asymptotic bounds": they hold (approximately) when the sample size is "sufficiently large", but the threshold on the sample size above which this occurs is typically unspecified. This significantly distinguishes SLT from other related tools for model selection, such as the Akaike Information Criterion (AIC) and the similar Bayesian Information Criterion (BIC). For instance, the AIC, which is also based on principled statistical arguments, can be interpreted as an estimate of the expected relative distance between a fitted model and the unknown true mechanism that generates the data (see Cavanaugh and Neath, 2019, for a short derivation). However, this estimate is typically obtained by making a "large sample approximation" (without quantifying how large the sample should be), and the variance of such an estimate is not addressed (at least in common usage of that criterion). A "corrected AIC" is also used in the statistical literature to take into account the possibly finite sample size, but this is essentially a correction of the bias of the AIC estimate, not of its variance. Hence, the bounds provided by SLT (which is more recent) appear to be, in this regard, better than the ones provided either by the AIC or by the corrected AIC. A similar argument holds for the BIC, whose estimate is analogous to the one provided by the AIC.

In this connection, our analysis can be related to other ideas in machine learning, and extended in different directions. For instance, our results are in line with the so-called "no free lunch theorems" in machine learning (Wolpert, 1996), according to which all training algorithms have the same expected performance, when a suitable average over all possible supervised machine learning problems is taken. These theorems are another important formalization of philosophical principles (Schurz, 2017; Lauc, 2020) and would deserve further investigation in connection with our analysis. Moreover, the connections with philosophical results on inductive learning and truth approximation could be explored from the point of view of machine learning and SLT. As Niiniluoto (2005) notes, both the Carnapian research program on inductive logic and the Popperian one on *verisimilitude* or *truthlikeness* delivered a wealth of results that may be translated and studied in the machine learning framework, in order to shed light on crucial philosophical and methodological problems, including that of cognitive and scientific progress (Niiniluoto, 2019; Cevolani and Tambolo, 2013). In this regard, the frameworks of Probably Approximately Correct (PAC) learning, and Agnostic PAC learning (both excellently presented in Shalev-Shwartz and Ben-David, 2014) seem to be particularly relevant (see, e.g., Herrmann, 2020).

It is worth noting that the analysis proposed in the present article could be extended to the more realistic case in which the data-generation model belongs to one among more than two families of models, since knowing ex-ante that it belongs to one of only two families of models may appear as a strong assumption. Another interesting future research direction concerns extending the comparison made in this

article by taking into account also the computational complexity of learning (Shalev-Shwartz and Ben-David, 2014, Chapter 8), e.g., the computational time needed to solve each of the two optimization problems (8) and (9) as a function of the sample size and of the complexity of the family. For instance, for a given "budget" constraint on the total computational time, one could consider in the comparison the possibility of having a different number of training examples for the two families, taking into account the possibly different computational complexities of the two optimization problems (8) and (9).[21]

Finally, the theoretical analysis proposed here may be supplemented, and validated, by numerical experiments, by simulating "simple" and "complex" learning environments and studying the performance of different families of models with training sets of varying size. In the technical Appendix 1.3, we outline a procedure to implement such numerical experiments; for the time being, however, this is ongoing work for the future.

Appendix 1: Technical appendix

In this appendix, we prove the main results of our analysis in Sect. 3 and discuss some technical details related to it. First, in Appendix 1.1 we discuss minor changes in the theoretical analysis needed to deal with possible nonexistence/nonuniqueness of the minimizers. Second, in Appendix 1.2 we discuss relationships among minimizers of the empirical risks and minimizers of the expected risks, also introducing some notation used later in Appendix 1.3. Third, in Appendix 1.2 we derive conditions (14), (15), and (16) presented in Sect. 3. Finally, in Appendix 1.2 we discuss some ways in which the results of our theoretical analysis could be validated by numerical experiments.

Appendix 1.1: Existence and Uniqueness of Minimizers of the Expected and Empirical Risks

In the work, existence and uniqueness of minimizers of the expected and empirical risks over the two families of models have been assumed to simplify the presentation of the main ideas. In the following, we report the reasons why the analysis is not significantly affected by possible non-existence/non-uniqueness of such minimizers.

With reference to the expected risk:

- If the minima over the two families of models do not exist, one can replace them in the analysis with infima, and each minimizer by an ϵ-minimizer (i.e., for a sufficiently small $\epsilon > 0$, a vector whose expected risk is smaller than or equal to

[21] A similar idea was recently used in different contexts in Gnecco and Nutarelli (2019); Gnecco et al. (2020,, 2021), where the optimal trade-off between the number of training examples and their precision of supervision was investigated for several machine-learning problems, under a given budget constraint on the total cost of their acquisition.

the infimum expected risk plus ϵ). The only change required in the analysis is the replacement of Δ with either $\Delta - 2\epsilon$ (if $\Delta > 0$) or $\Delta + 2\epsilon$ (if $\Delta < 0$). Both have the same sign as Δ if ϵ is sufficiently small.
- Multiple minimizers in each family simply provide the same value of the minimum expected risk in that family. They do not alter the value of Δ. As the goal of the analysis is to find the correct family of models (not to find one specific minimizer, among multiple ones), no changes in the analysis are required in this case.

A similar remark holds for the empirical risk (of course, several combinations of cases could be considered for the empirical/expected risks, for what concerns non-existence/non-uniqueness of minimizers).

Appendix 1.2: Relationship Between Minimizers of the Expected and Empirical Risks

It is worth remarking that both ERM and SRM work by considering the minimizers of the empirical risk over the two families of models. Such minimizers have been denoted in the paper, respectively, as $\hat{\mathbf{w}}_S$ and $\hat{\mathbf{w}}_C$. It is worth recalling that, once the training set has been observed, the empirical risk of each model can be computed exactly, simply by evaluating the summation in the definition of empirical risk. As mentioned in the text, instead, the expected risks cannot be typically computed exactly, which prevents one from finding their minimizers over the respective families of models. In the following, let us denote such minimizers by

$$\mathbf{w}_S^\circ := \text{argmin}_{\mathbf{w}_S \in W_S} R_S^{exp}(\mathbf{w}_S) \tag{18}$$

and

$$\mathbf{w}_C^\circ := \text{argmin}_{\mathbf{w}_C \in W_C} R_C^{exp}(\mathbf{w}_C)\,, \tag{19}$$

respectively. Nevertheless, one can actually find relationships between the expected risks of $\mathbf{w}_M^\circ$ and $\hat{\mathbf{w}}_M$ with $M \in \{C, S\}$ and also between their empirical risks, by exploiting their definitions as minimizers of the respective expected/empirical risks. This is a standard "trick" used in SLT (see, e.g., Zoppoli et al. (2020, Chapter 4)), which is reported here to make the work self-contained. It is based on the application of the triangle inequality and on the fact that, according to Theorem 3.1, with probability $p_S \geq 1 - \frac{\delta}{2}$ the expected/empirical risks are related by (4) for the "simple" family and, still with probability $p_C \geq 1 - \frac{\delta}{2}$, they are related by (5) for the "complex" family (Vapnik, 2000, Sect. 3.7).

Theorem 4.1 *Let the assumptions of Theorem 3.1 hold, and $M \in \{C, S\}$. With probability $p_M \geq 1 - \frac{\delta}{2}$ with respect to the i.i.d. generation of the training set, both the following inequalities hold*:

$$|R^{exp}(\hat{\mathbf{w}}_M) - R^{exp}(\mathbf{w}_M^\circ)| \leq 2\varepsilon\,, \tag{20}$$

$$|R^{emp,E}(\hat{\mathbf{w}}_M) - R^{emp,E}(\mathbf{w}_M^\circ)| \leq 2\varepsilon\,. \tag{21}$$

The proof of Theorem 4.1 is divided into the following steps.

- By Theorem 3.1, with probability $p_M \geq 1 - \frac{\delta}{2}$, one has

$$|R^{exp}(\mathbf{w}_M) - R^{emp,E}(\mathbf{w}_M)| \leq \varepsilon\,, \tag{22}$$

for all $\mathbf{w}_M \in W_M$ (here, ε refers to the right-hand side of either the VC bound (4) or the one (5)).
- Moreover, by their definitions, $\mathbf{w}_M^\circ$ minimizes $R^{exp}(\mathbf{w}_M)$ over W_M, and $\hat{\mathbf{w}}_M$ minimizes $R^{emp,E}(\mathbf{w}_M)$ over W_M.
- Hence, one gets (all the following bounds, from (23) to (28), holding simultaneously with probability $p_M \geq 1 - \frac{\delta}{2}$):

$$R^{exp}(\mathbf{w}_M^\circ) \leq R^{exp}(\hat{\mathbf{w}}_M) \tag{23}$$

(since $\mathbf{w}_M^\circ$ minimizes the expected risk over W_M),

$$R^{emp,E}(\hat{\mathbf{w}}_M) \leq R^{emp,E}(\mathbf{w}_M^\circ) \tag{24}$$

(since $\hat{\mathbf{w}}_M$ minimizes the empirical risk over W_M), and

$$\begin{aligned}
&R^{exp}(\mathbf{w}_M^\circ)\\
&\quad\leq R^{exp}(\hat{\mathbf{w}}_M)\\
&\quad\leq R^{exp}(\hat{\mathbf{w}}_M) - R^{emp,E}(\hat{\mathbf{w}}_M) + R^{emp,E}(\hat{\mathbf{w}}_M)\\
&\quad\leq R^{exp}(\hat{\mathbf{w}}_M) - R^{emp,E}(\hat{\mathbf{w}}_M) + R^{emp,E}(\mathbf{w}_M^\circ)\\
&\quad= R^{exp}(\hat{\mathbf{w}}_M) - R^{emp,E}(\hat{\mathbf{w}}_M) + R^{emp,E}(\mathbf{w}_M^\circ) - R^{exp}(\mathbf{w}_M^\circ) + R^{exp}(\mathbf{w}_M^\circ)\\
&\quad\leq |R^{exp}(\hat{\mathbf{w}}_M) - R^{emp,E}(\hat{\mathbf{w}}_M)| + |R^{emp,E}(\mathbf{w}_M^\circ) - R^{exp}(\mathbf{w}_M^\circ)| + R^{exp}(\mathbf{w}_M^\circ)\\
&\quad\leq 2\varepsilon + R^{exp}(\mathbf{w}_M^\circ)\,.
\end{aligned} \tag{25}$$

Therefore, subtracting $R^{exp}(\mathbf{w}_M^\circ)$ from the above, one gets

$$-2\varepsilon \leq 0 \leq R^{exp}(\hat{\mathbf{w}}_M) - R^{exp}(\mathbf{w}_M^\circ) \leq 2\varepsilon\,, \tag{26}$$

and finally,

$$|R^{exp}(\hat{\mathbf{w}}_M) - R^{exp}(\mathbf{w}_M^\circ)| \leq 2\varepsilon. \tag{27}$$

Similarly, one gets

$$|R^{emp,E}(\hat{\mathbf{w}}_M) - R^{emp,E}(\mathbf{w}_M^\circ)| \leq 2\varepsilon\,. \tag{28}$$

So, also in the typical case for which $\mathbf{w}_M^\circ$ is unknown, it is still possible to relate (with probability $p_M \geq 1 - \frac{\delta}{2}$) its empirical/expected risk to the one of $\hat{\mathbf{w}}_M$. Finally, following the same argument provided in the main text after the statement of Theorem 3.1, the bounds (27) and (28) hold simultaneously for both $M = C, S$ with probability $p \geq 1 - \delta$.

Appendix 1.3: Derivation of Conditions (14), (15), and (16) in Theorems 3.2 and 3.3

Let the assumptions of Theorem 3.1 hold. By the discussion following that theorem, with probability $p \geq 1 - \delta$, the bounds (4) and (5) hold simultaneously for all the models in the two respective families. Then, one can conclude the following (all the following bounds hold simultaneously with probability $p \geq 1 - \delta$):

- the empirical risk of the (unknown) model (associated with $\mathbf{w}_S^\circ$) that minimizes the expected risk over the "simple" family of models is at most

$$\varepsilon(h_S, E, \delta) \tag{29}$$

  away from that expected risk.
- the empirical risk of the (unknown) model (associated with $\mathbf{w}_C^\circ$) that minimizes the expected risk over the "complex" family of models is at most

$$\varepsilon(h_C, E, \delta) \tag{30}$$

  away from that expected risk.
- The following bound—which is derived from (5)—relates the empirical risk of the model associated with $\hat{\mathbf{w}}_C$ and the expected risk of the model associated with $\mathbf{w}_C^\circ$:

$$\begin{aligned} & R_C^{emp,E}(\hat{\mathbf{w}}_C) \\ & \quad \geq R_C^{exp}(\hat{\mathbf{w}}_C) - \varepsilon(h_C, E, \delta) \\ & \quad \geq \min_{\mathbf{w}_C \in W_C} R_C^{exp}(\mathbf{w}_C) - \varepsilon(h_C, E, \delta) \\ & \quad = R_C^{exp}(\mathbf{w}_C^\circ) - \varepsilon(h_C, E, \delta)\,. \end{aligned} \tag{31}$$

  This can be expressed equivalently as

$$-R_C^{emp,E}(\hat{\mathbf{w}}_C) \leq -R_C^{exp}(\mathbf{w}_C^\circ) + \varepsilon(h_C, E, \delta)\,. \tag{32}$$

  Similarly, one gets

$$-R_S^{emp,E}(\hat{\mathbf{w}}_S) \leq -R_S^{exp}(\mathbf{w}_S^\circ) + \varepsilon(h_S, E, \delta)\,. \tag{33}$$

- Since the two models above are "candidate" empirical risk minimizers in the two families, and the difference of their expected risks is Δ, condition (14) guarantees the choice of the correct family, with probability $p \geq 1 - \delta$. Indeed, if $\Delta > 0$, with probability $p \geq 1 - \delta$, under this condition, one gets (using (32) in the fourth inequality of the next formula (34))

$$
\begin{aligned}
&\min_{\mathbf{w}_S \in W_S} R_S^{emp,E}(\mathbf{w}_S) \\
&\quad \leq R_S^{emp,E}(\mathbf{w}_S^\circ) \\
&\quad \leq R_S^{exp}(\mathbf{w}_S^\circ) + \varepsilon(h_S, E, \delta) \\
&\quad \leq R_S^{exp}(\mathbf{w}_S^\circ) - R_C^{emp,E}(\hat{\mathbf{w}}_C) + R_C^{emp,E}(\hat{\mathbf{w}}_C) + \varepsilon(h_S, E, \delta) \\
&\quad \leq R_S^{exp}(\mathbf{w}_S^\circ) - R_C^{exp}(\mathbf{w}_C^\circ) + \varepsilon(h_C, E, \delta) + R_C^{emp,E}(\hat{\mathbf{w}}_C) + \varepsilon(h_S, E, \delta) \\
&\quad < R_S^{exp}(\mathbf{w}_S^\circ) - R_C^{exp}(\mathbf{w}_C^\circ) + R_C^{emp,E}(\hat{\mathbf{w}}_C) + \Delta \\
&\quad = -\Delta + R_C^{emp,E}(\hat{\mathbf{w}}_C) + \Delta \\
&\quad = \min_{\mathbf{w}_C \in W_C} R_C^{emp,E}(\mathbf{w}_C)\,,
\end{aligned}
\tag{34}
$$

hence

$$
\min_{\mathbf{w}_S \in W_S} R_S^{emp,E}(\mathbf{w}_S) < \min_{\mathbf{w}_C \in W_C} R_C^{emp,E}(\mathbf{w}_C)\,, \tag{35}
$$

and the correct family of models is selected. The case $\Delta < 0$ is proved similarly.

- Conditions (15) and (16) are obtained in an analogous way, by taking into account that SRM selects a model that minimizes the regularized empirical risk, and that the regularization terms

$$
\varepsilon(h_S, E, \delta) \tag{36}
$$

and

$$
\varepsilon(h_C, E, \delta) \tag{37}
$$

are different for models belonging to different families, but they are constant within each of the two families.

Appendix 1.4: Possible Design of Numerical Experiments

The following is a possible procedure that could be used to corroborate the theoretical results of the paper with results of numerical experiments:

- First, one could generate the output data according to a model belonging to one of the two families (of course, without letting the machine know the underlying data generating process).
- Then, one would select one of the two families of models by applying ERM and SRM, respectively, for various choices of the training set size, and compare the results obtained in the two cases (e.g., to find situations where SRM chooses the wrong family of models, and ERM chooses the right one, or SRM starts choosing the correct family only for a larger training set size with respect to ERM).

A possible situation for which the comparison above is expected to be in favor of SRM is when all the following holds:

(a) $\Delta < 0$ (so the correct family of models is the "complex" one), but its absolute value $|\Delta|$ is "small".
(b) The training data are generated according to a probability distribution for which the VC bounds (4) and (5) are loose (indeed, since VC bounds are distribution-free, it is well-known from the literature on SLT that they tend to get loose for certain probability distributions; see, e.g., Herbrich and Williamson, 2002). In this case, one would expect, even for a modest training set size E, to have (loosely speaking), with large probability, both

$$R_S^{emp,E}(\mathbf{w}_S) \simeq R_S^{exp}(\mathbf{w}_S) \tag{38}$$

for all the models in the "simple" family, and

$$R_C^{emp,E}(\mathbf{w}_C) \simeq R_C^{exp}(\mathbf{w}_C) \tag{39}$$

for all the models in the "complex" family.
(c) The condition

$$E \simeq h_C \gg h_S \tag{40}$$

holds, so the regularization term for "complex" models is much larger than the one for "simple" models, making SRM choose the "simple" family.

Acknowledgments We wish to thank two anonymous reviewers of this journal for their very detailed and constructive comments on the manuscript. Falco J. Bargagli-Stoffi acknowledges funding from the Alfred P. Sloan Foundation Grant for the development of "Causal Inference with Complex Treatment Regimes: Design, Identification, Estimation, and Heterogeneity" and funding from the 2021 Harvard Data Science Initiative Postdoctoral Research Fund Award. Gustavo Cevolani acknowledges funding from the PRIN 2017 grant "From models to decisions" of the Italian Ministry of University and Research (grant n. 201743F9YE). Giorgio Gnecco acknowledges funding from the 2020 Italian project "Trade-off between Number of Examples and Precision in Variations of the Fixed-Effects Panel Data Model" funded by INdAM-GNAMPA (Istituto Nazionale di Alta Matematica - Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni).

References

Baker, A. (2016). Simplicity. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Balduzzi, D. (2013). Falsification and future performance. In David L. Dowe (Ed.), *Algorithmic probability and friends: Bayesian prediction and artificial intelligence*, volume 7070 of *Lecture notes in computer science* (pp. 65–78). Springer.

Barbeau, E. J. (2004). *Polynomials*. Springer.

Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press.

Bousquet, O., Boucheron, S., & Gábor, L. (2004). Introduction to statistical learning theory. Lecture notes in computer science. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced lectures on machine learning* (Vol. 3176, pp. 169–207). Springer.

Burnham, Kenneth P., & Anderson, David R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer.

Cavanaugh, J. E., & Neath, A. A. (2019). The akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. *WIREs computational statistics*, 11. article number e1460.

Cevolani, G., & Tambolo, L. (2013). Progress as approximation to the truth: A defence of the verisimilitudinarian approach. *Erkenntnis, 78*(4), 921–935.

Cherkassky, V., & Mulier, F. (2007). *Learning from data: Concepts, theory, and methods*. Wiley.

Corfield, D. (2010). Varieties of justification in machine learning. *Minds and Machines, 20,* 291–301.

Corfield, D., Schölkopf, B., & Vapnik, V. N. (2009). Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science, 4,* 51–58.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society, 39,* 1–49.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems, 2,* 303–314.

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

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. Wiley.

Fitzpatrick, S. (2013). Simplicity in the philosophy of science. *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/simplici/.

Gnecco, G., & Nutarelli, F. (2019). On the trade-off between number of examples and precision of supervision in machine learning problems. *Optimization Letters, 15,* 1711–1733.

Gnecco, G., Nutarelli, F., & Selvi, D. (2020). Optimal trade-off between sample size, precision of supervision, and selection probabilities for the unbalanced fixed effects panel data model. *Soft Computing, 24,* 15937–15949.

Gnecco, G., Nutarelli, F., & Selvi, D. (2021). Optimal trade-off between sample size and precision for the fixed effects generalized least squares panel data model. *Machine Learning, 110,* 1549–1584.

Harman, G., & Kulkarni, S. (2007). *Reliable reasoning: Induction and statistical learning theory*. MIT Press.

Harman, G., & Kulkarni, S. (2011). Statistical learning theory as a framework for the philosophy of induction. In Bandyopadhyay, P S., & Forster, M. R. (Eds.), *Philosophy of statistics*, volume 7 of *Handbook of the philosophy of science* (pp. 833–847). North-Holland.

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

Herbrich, R., & Williamson, R. C. (2002). Algorithmic luckiness. *Journal of Machine Learning Research, 3,* 172–212.

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

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

Landgrebe, J., & Smith, B. (2019). Making AI meaningful again. *Synthese*. https://doi.org/10.1007/s11229-019-02192-y.

Lauc, D. (2020). Machine learning and the philosophical problems of induction. In S. Skansi (Ed.), *Guide to deep learning basics* (pp. 93–106). Springer.

López-Rubio, E. (2020). The big data razor. *European Journal of Philosophy of Science, 10,* 1–20.

Mendelson, S. (2003). A few notes on statistical learning theory. In *Advanced lectures on machine learning*, volume 2600 of *Lecture notes in computer science* (pp. 1–40). Springer.

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

Niiniluoto, I. (2019). Scientific progress. In Zalta, E. N. (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition.
Niiniluoto, I. (2005). Inductive logic, verisimilitude, and machine learning. In P. Hàjek, L. Valdés-Villanueva, & D. Westerståhl (Eds.), *Logic, methodology and philosophy of science* (pp. 295–314). College Publications.
Schubbach, A. (2019). Judging machines: Philosophical aspects of deep learning. *Synthese*. https://doi.org/10.1007/s11229-019-02167-z.
Schurz, G. (2017). No free lunch theorem, inductive skepticism, and the optimality of meta-induction. *Philosophy of Science, 84,* 825–839.
Seldin, Y., & Schölkopf, B. (2013). On the relations and differences between popper dimension, exclusion dimension and VC-dimension. In B. Schölkopf, Z. Luo, & V. Vovk (Eds.), *Empirical inference* (pp. 53–57). Springer.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.
Shi, B., & Iyengarand, S. S. (2020). *Mathematical theories of machine learning: Theory and applications*. Springer.
Sober, E. (2015). *Ockham's razors*. Cambridge University Press.
Steel, D. (2009). Testability and Ockham's razor: How formal and statistical learning theory converge in the new riddle of induction. *Journal of Philosophical Logic, 38,* 471–489.
Swinburne, R. (1997). *Simplicity as evidence of truth*. Milwaukee: Marquette University Press.
Thagard, P. (1990). Philosophy and machine learning. *Canadian Journal of Philosophy, 20,* 261–276.
Vapnik, V. N. (1998). *Statistical learning theory*. Wiley-Interscience.
Vapnik, V. N. (2000). *The nature of statistical learning theory*. Springer.
Watson, D. S., & Floridi, L. (2020). The explanation game: A formal framework for interpretable machine learning. *Synthese*. https://doi.org/10.1007/s11229-020-02629-9.
Williamson, J. (2004). A dynamic interaction between machine learning and the philosophy of science. *Minds and Machines, 14,* 539–549.
Williamson, J. (2009). The philosophy of science and its relation to machine learning. In M. M. Gaber (Ed.), *Scientific data mining and knowledge discovery* (pp. 77–90). Springer.
Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation, 8,* 1341–1390.
Zoppoli, R., Sanguineti, M., Gnecco, G., & Parisini, T. (2020). *Neural approximations for optimal control and decision*. Springer.