Falsification and future performance

David Balduzzi

MPI for Intelligent Systems, Tuebingen, Germany
david.balduzzi@tuebingen.mpg.de

Abstract. We information-theoretically reformulate two measures of capacity from statistical learning theory: empirical VC-entropy and empirical Rademacher complexity. We show these capacity measures count the number of hypotheses about a dataset that a learning algorithm *falsifies* when it finds the classifier in its repertoire minimizing empirical risk. It then follows from that the future performance of predictors on *unseen* data is controlled in part by how many hypotheses the learner falsifies. As a corollary we show that empirical VC-entropy quantifies the message length of the true hypothesis in the optimal code of a particular probability distribution, the so-called actual repertoire.

1 Introduction

This note relates the number of hypotheses falsified by a learning algorithm to the expected future performance of the predictor it outputs. It does so by reformulating two basic results from statistical learning theory information-theoretically.

Suppose we wish to predict an unknown physical process $\sigma^* : \mathcal{X} \to \mathcal{Y}$ occurring in nature after observing its outputs $(y_1, \ldots, y_l)$ on sample $\mathcal{D} = (x_1, \ldots, x_l)$ of its inputs, where inputs arise according to unknown distribution P. One method is to take a repertoire $\mathcal{F}$ of functions from $\mathcal{X} \to \mathcal{Y}$ and choose the predictor $\hat{f} \in \mathcal{F}$ that best approximates σ^* on the observed data. How confident can we be in $\hat{f}$'s future performance on unseen data?

Statistical learning theory provides bounds on $\hat{f}$'s expected future performance by quantifying a tradeoff implicit in the choice of repertoire $\mathcal{F}$. At first glance, the bigger the repertoire the better since the best approximation to σ^* in $\mathcal{F}$ can only improve as more more functions are added to $\mathcal{F}$. However, increasing $\mathcal{F}$, and improving the approximation on observed data, can *reduce* future performance due to overfitting. As a result, the bounds depend on both the accuracy with which $\hat{f}$ approximates σ^* on the observed data and the capacity of repertoire $\mathcal{F}$, see Theorems 9 and 10.

We wish to connect statistical learning theory with Popper's ideas about falsification. Popper argued that no amount of positive evidence confirms a theory [11]. Rather, theories should be judged on the basis of how many hypotheses they falsify. A theory is *falsifiable* if there are possible hypotheses about the world (i.e. data) that are not consistent with the theory. A bold theory falsifies (disagrees with) many potential hypotheses about observed data. Testing a bold theory, by checking that the hypotheses it disagrees with are in fact false,

provides corroborating evidence. If a theory has been thoroughly tested then (perhaps) we can have confidence in its predictions. Popper's criticism of positive confirmation was devastating. However, and hence the "perhaps", he failed to provide a rationale for trusting the predictions of severely tested theories.

To understand how falsifying hypotheses affects future performance we reformulate learning as a kind of *measurement*. Before doing so, we need to describe precisely what we mean by measurement.

Given physical system X with state space $S(X)$, a classical measurement is a function $f : S(X) \to \mathbb{R}$. For example a thermometer f maps configurations (positions and momenta) of particles in the atmosphere to real numbers. When the thermometer outputs $15^{\circ}C$ it generates information by specifying that atmospheric particles were in a configuration in $f^{-1}(15) \subset S(X)$. The information generated by the thermometer is a brute physical fact depending on how the thermometer is built and its output. We quantify the information, see §2, by comparing the size of the total configuration space $S(X)$ with the size of the pre-image $f^{-1}(15)$. The smaller the pre-image, the more informative the measurement, see §2 for details.

More generally, any (classical) physical process $f : \mathcal{X} \to \mathcal{Y}$ can be thought of as performing measurements by taking inputs in $\mathcal{X}$ to outputs in $\mathcal{Y}$. Section §4 introduces an important example, the *min-risk* $\mathbf{R}_{\mathcal{F},\mathcal{D}} : \Sigma(\mathcal{X}, \mathcal{Y}) \to \mathbb{R}$, which outputs the minimum value of the empirical risk over repertoire $\mathcal{F}$ on a hypothesis space $\Sigma(\mathcal{X}, \mathcal{Y})$. Finding the min-risk is a necessary step in finding the best approximation $\hat{f}$ to σ^* in $\mathcal{F}$. Since computing the min-risk requires actually implementing it as a physical process somehow or other, the measurements it performs and the effective information it generates are brute physical facts, no different in kind than the information generated by a thermometer.

It turns out that the min-risk categorizes hypotheses in Σ according to how well they are approximated by predictors in repertoire $\mathcal{F}$. Proposition 12 shows that the effective information generated by the min-risk is (essentially) the empirical VC-entropy. Moreover, the effective information generated by the min-risk "counts" the number of hypotheses about $\mathcal{D}$ that $\mathcal{F}$ falsifies, see Eq. (13). As a consequence, Corollary 13, we obtain that the future performance of predictor $\hat{f}$ is controlled by (i) how well $\hat{f}$ fits the observed data; (ii) how many hypotheses about the data the min-risk rules out and (iii) a confidence term.

It follows that, assuming the assumptions of the theorems below hold, bounds on future performance are brute physical facts resulting from the act of minimizing empirical risk, and so falsifying potential hypotheses, on observed data.

A consequence of our results, Corollary 15, is that empirical VC-entropy is essentially the minimal length of the true hypothesis under the optimal code for the actual repertoire (a distribution depending on the min-risk). This suggests there may be interesting connections between VC-theory and the minimum message length (MML) approach to induction proposed by Wallace and Boulton [15, 16].

Finally, section §4.2 reformulates empirical Rademacher complexity via falsification. Here we build on Solomonoff's probability distribution introduced in [12]. In short, we take Solomonoff's definition and substitute the *min-risk* in

place of the universal Turing machine, thereby obtaining what we refer to as the Rademacher distribution – a *non-universal* analog of Solomonoff's distribution. Rademacher complexity is then computed using the expectation of the min-risk over the Rademacher distribution, see Proposition 17.

The min-risk thus provides a bridge that not only connects VC-theory to a computable analog of Solomonoff's seminal distribution, but also sheds light on how falsification provides guarantees on future performance.

Related work. The connection between Popper's ideas on falsifiability and statistical learning theory was pointed out in [5,7,14]. However, these works focus on VC-dimension, which does not relate to falsification as directly as VC-entropy and Rademacher complexity which we consider here. Further, VC-entropy is a more fundamental concept in statistical learning theory than VC-dimension since VC-dimension is defined in terms of the limit behavior of the growth function, which is an upper bound on VC-entropy [14]. For more details on the link between MML and algorithmic probability, see [17].

Acknowledgements. I thank David Dowe and Samory Kpotufe for useful comments on an earlier version of this paper.

2 Measurement

We consider a toy universe containing probabilistic mechanisms (input/output devices) of the following form

Definition 1 *Given finite sets $\mathcal{X}$ and $\mathcal{Y}$, a **mechanism** is a Markov matrix $\mathfrak{m}$ defined by conditional probability distribution $p_{\mathfrak{m}}(y|x)$.*

Mechanisms generate information about their inputs by assigning them to outputs [1, 2].

Definition 2 *The **actual repertoire** (or **measurement**) specified by $\mathfrak{m}$ outputting y is the probability distribution*

$$p_{\mathfrak{m}}(x|y) := \frac{p_{\mathfrak{m}}(y|x)}{p(y)} \cdot p_{unif}(x),$$

*where $p_{unif}(x) = \frac{1}{|\mathcal{X}|}$ is the uniform distribution. The **effective information** generated by the measurement is*

$$ei(\mathfrak{m}, y) := H\Big[p_{\mathfrak{m}}(X|y)\Big\|p_{unif}(X)\Big],$$

where $H[p\|q] = \sum_i p_i \log_2 \frac{p_i}{q_i}$ is Kullback-Leibler divergence.

The Kullback-Leibler divergence $H[p\|q]$ can be interpreted informally as the number of Y/N questions needed to get from distribution q to distribution p. However, as pointed out in [6], Kullback-Leibler divergence is invariant with respect to the "framing of the problem" – the ordering and structure of the questions – suggesting it is a suitable measure of information-theoretic "effort".

The definition of measurement is motivated by the special case where $p_{\mathfrak{m}}$ assigns probabilities that are either 0 or 1; in other words, when it corresponds to a set-valued function $f : \mathcal{X} \to \mathcal{Y}$. The measurement performed by f is

$$p_f(x|y) = \begin{cases} \frac{1}{|f^{-1}(y)|} & \text{if } f(x) = y \\ 0 & \text{else,} \end{cases}$$

where $|\cdot|$ denotes cardinality. The support of $p_f(X|y)$ is the preimage $f^{-1}(y) \subset \mathcal{X}$. All elements of the support are assigned equal probability – they are treated as an undifferentiated list. The measurement $p_{\mathfrak{m}}(X|y)$ therefore generalizes the notion of preimage to the probabilistic setting.

The effective information generated by f outputting y is $ei(f, y) = \log_2 \frac{|\mathcal{X}|}{|f^{-1}(y)|}$:

$$\begin{aligned} ei(f, y) = & \quad \log_2 |\mathcal{X}| \quad - \quad \log_2 |f^{-1}(y)| \\ = & \Big(\text{no. potential inputs}\Big) - \Big(\text{no. inputs in pre-image}\Big) \\ = & \Big(\text{no. inputs ruled out}\Big), \end{aligned} \tag{1}$$

where inputs are counted in bits (after logarithming). Effective information is maximal ($\log_2 |\mathcal{X}|$ bits) when a single input leads to y, and is minimal (0 bits) when *all* inputs lead to y. In the first case, observing f output y tells us exactly what the input was, and in the latter case, it tells us nothing at all.

2.1 Semantics

Next we consider two approaches to characterizing the meaning of measurements. The first relates to possible world semantics [9]. Here, the meaning of a sentence is given by the set of possible worlds in which it is true. Meaning is thus determined by considering all counterfactuals. For example, the meaning of "That car is 10 years old" is the set of possible worlds where the speaker is pointing to a car manufactured 10 years previously. Since the set of contains cars of many different colors, we see that color is irrelevant to the meaning of the sentence.

More precisely, the meaning of sentence $\mathcal{S}$ is a map from possible worlds W to truth values $v_{\mathcal{S}} : W \to \{0, 1\}$. Equivalently, the meaning of a sentence is

$$\begin{aligned} W \quad & \supset \quad v_{\mathcal{S}}^{-1}(1) \\ \Big(\text{possible worlds}\Big) & \supset \Big(\text{worlds where } \mathcal{S} \text{ is true}\Big). \end{aligned} \tag{2}$$

Inspired by possible world semantics, we propose

Definition 3 *The **meaning** of output y by mechanism $\mathfrak{m}$ is*

$$\begin{aligned} p_{unif}(X) \quad & \to \quad p_{\mathfrak{m}}(X|y) \\ \Big(\textit{possible inputs}\Big) & \to \Big(\textit{inputs that cause } y\Big). \end{aligned} \tag{3}$$

For a deterministic function this reduces to $\mathcal{X} \supset f^{-1}(y)$.

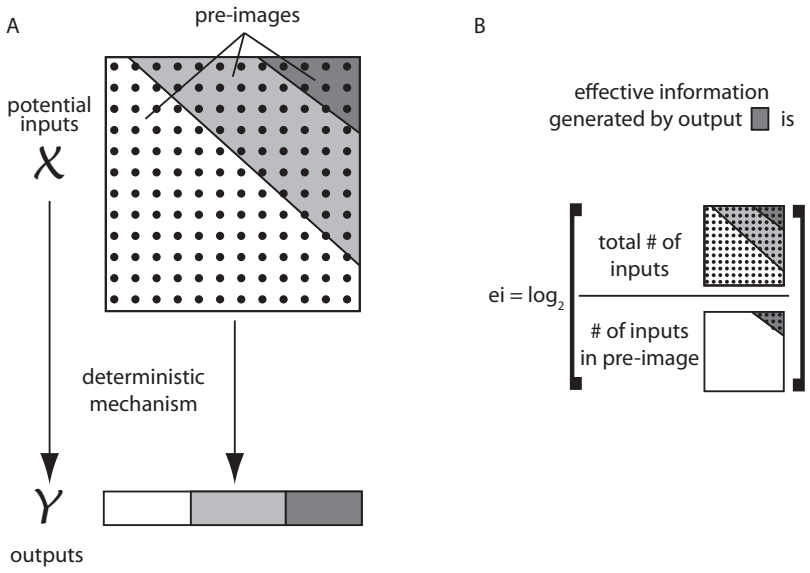

Fig. 1: **The effective information generated by measurements.** (A) A deterministic device can receive 144 inputs and produce 3 outputs. (B): Each input is implicitly assigned to a category (shaded areas). The information generated by the dark gray output is $\log_2 144 - \log_2 9 = 4$ bits.

Grounding meanings in mechanisms yields four advantages over the possible worlds approach. First, it replaces the difficult to define notion of a possible world with the concrete set of inputs the mechanism is physically capable of receiving. Second, in possible world semantics the work of determining whether or not a sentence is true is performed somewhat mysteriously offstage, whereas the meaning of a measurement is determined via Bayes' rule. Third, the approach generalizes to probabilistic mechanisms. Finally, we can compute the effective information generated by a measurement, whereas there is no way to quantify the information content of a sentence in possible world semantics.

2.2 Risk

The second, pragmatic notion of meaning characterizes usefulness. We consider a special case, well studied in statistical learning theory, where usefulness relates to predictions [14].

Let $\Sigma(\mathcal{X}, \mathcal{Y}) = \{\sigma : \mathcal{X} \to \mathcal{Y}\}$ be the set of all functions (deterministic mechanisms) mapping $\mathcal{X}$ to $\mathcal{Y} = \{-1, +1\}$. We will often write Σ for short. Suppose there is a random variable X taking values in $\mathcal{X}$ with unknown distribution P and an unknown mechanism $\sigma^* \in \Sigma$, the *supervisor*, who assigns labels to elements of $\mathcal{X}$.

Definition 4 *The **risk** quantifies how well mechanism f approximates an unknown or partially known mechanism σ^*:*

$$\mathbf{R}(f) = \sum_{x \in \mathcal{X}} \mathbb{I}\big[f(x) \neq \sigma^*(x)\big] \cdot p(x). \tag{4}$$

It is the probability that f and σ^* disagree on elements of $\mathcal{X}$.

Unfortunately, the risk cannot be computed since P and σ^* are unknown.

Definition 5 *Given a finite sample $\mathcal{D} = (x_1, \ldots, x_l) \in \mathcal{X}^l$ with labels $\mathcal{L} = \sigma^*\mathcal{D} = (y_1, \ldots, y_l) \in \mathcal{Y}^l$, the **empirical risk** of $f : \mathcal{X} \to \mathcal{Y}$*

$$\mathbf{R}(f, \mathcal{D}, \mathcal{L}) = \frac{1}{l} \sum_{i=1}^{l} \mathbb{I}\big[f(x_i) \neq y_i\big] \tag{5}$$

is the fraction of the data $\mathcal{D}$ on which f and σ^* disagree.

The empirical risk provides a computable approximation to the (true) risk.

Remark 6 *Note that in this paper, sets $\mathcal{X}$ and $\mathcal{Y}$ are both finite. Similarly, the training data $\mathcal{D} \in \mathcal{X}^l$ and labels $\mathcal{L} \in \mathcal{Y}^l$ also live in finite sets.*

3 Statistical learning theory

Suppose we wish to predict the unknown supervisor σ^* based on its behavior on labeled data $(\mathcal{D}, \mathcal{L})$. A simple way to find a mechanism in repertoire $\mathcal{F} \subset \Sigma(\mathcal{X}, \mathcal{Y})$ that approximates σ^* well is to minimize the empirical risk.

Definition 7 *Given repertoire $\mathcal{F} \subset \Sigma$ and unlabeled data $\mathcal{D} \in \mathcal{X}^l$, define **learning algorithm***

$$\mathcal{A}_{\mathcal{F},\mathcal{D}} : \Sigma \to \mathcal{F} : \sigma \mapsto \arg\min_{f \in \mathcal{F}} \mathbf{R}(f, \mathcal{D}, \sigma\mathcal{D}) \tag{6}$$

which finds the mechanism in $\mathcal{F}$ that minimizes empirical risk.

Learning algorithm $\mathcal{A}_{\mathcal{F},\mathcal{D}}$ finds the mechanism in $\mathcal{F}$ that appears, based on the empirical risk, to best approximate σ^*. Empirical risk stays constant or decreases as $\mathcal{F}$ is enlarged, suggesting that the larger the repertoire the better.

This is not true in general since minimizing risk – and *not* empirical risk – is the goal. There is a tradeoff: increasing the size of $\mathcal{F}$ leads to overfitting the data which can increase risk even as empirical risk is reduced.

The tendency of a repertoire to overfit data depends on its size or capacity. We recall two measures of capacity that are used to bound risk: empirical VC-entropy [13] and empirical Rademacher complexity [8].

Definition 8 *Given unlabeled data* $\mathcal{D} \in \mathcal{X}^l$ *and repertoire* $\mathcal{F} \subset \Sigma$ *let*

$$q_{\mathcal{D}} : \mathcal{F} \to \mathbb{R}^l : f \mapsto \Big(f(x_1), \ldots, f(x_l)\Big). \tag{7}$$

The empirical ***VC-entropy***[1] *of* $\mathcal{F}$ *on* $\mathcal{D}$ *is* $\mathcal{V}(\mathcal{F}, \mathcal{D}) := \log_2 |q_{\mathcal{D}}(\mathcal{F})|$*, where* $|q_{\mathcal{D}}(\mathcal{F})|$ *is the number of distinct points in the image of* $q_{\mathcal{D}}$*.*

The empirical ***Rademacher complexity*** *of* $\mathcal{F}$ *on* $\mathcal{D}$ *is*

$$\mathcal{R}(\mathcal{F}, \mathcal{D}) = \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{l} \sum_{i=1}^{l} \sigma(x_i) \cdot f(x_i) \right]. \tag{8}$$

VC-entropy "counts" how many labelings of $\mathcal{D}$ the classifiers in $\mathcal{F}$ fit perfectly. Rademacher complexity is a weighted count of how many labelings of $\mathcal{D}$ functions in $\mathcal{F}$ fit well.

The following theorems are shown in [3] and [4] respectively:

Theorem 9 (empirical VC-entropy bound)
With probability $1 - \delta$*, the expected risk is bounded by*

$$\mathbf{R}(f) \leq \mathbf{R}(f, \mathcal{D}, \mathcal{L}) + c_1 \sqrt{\frac{\mathcal{V}(\mathcal{F}, \mathcal{D})}{l}} + c_2 \sqrt{\frac{1 - \log_2 \delta}{l}} \tag{9}$$

for all $f \in \mathcal{F}$*, where the constants are* $c_1 = \sqrt{\frac{6}{\log_2 e}}$ *and* $c_2 = \sqrt{\frac{1}{\log_2 e}}$*.*

Theorem 10 (empirical Rademacher bound)
For all $\delta > 0$*, with probability at least* $1 - \delta$*,*

$$\mathbf{R}(f) \leq \mathbf{R}(f, \mathcal{D}, \mathcal{L}) + \mathcal{R}(\mathcal{F}, \mathcal{D}) + c_3 \sqrt{\frac{1 - \log_2 \delta}{l}}, \tag{10}$$

for all $f \in \mathcal{F}$*, where* $c_3 = \sqrt{\frac{2}{\log_2 e}}$*.*

The tradeoff between empirical risk and capacity is visible in the first two terms on the right-hand sides of the bounds.

The left-hand sides of Eqs (9) and (10) cannot be computed since P and σ^* are unknown. Remarkably, the right-hand sides depend only on mechanism f chosen from repertoire $\mathcal{F}$, labeled data $(\mathcal{D}, \mathcal{L})$ and desired confidence δ. The theorems assume data is drawn *i.i.d.* according to P and labeled according to σ^*; it make no assumptions about the distribution P on $\mathcal{X}$ or supervisor σ^*, except that they are *fixed*.

[1] VC-entropy is the *expectation* of empirical VC-entropy [14]. Also, note the standard definition of VC-entropy uses $\log_e$ rather than $\log_2$.

4 Falsification

This section reformulates the results from statistical learning theory to show how the past falsifications performed by a learning algorithm control future performance. We show that the empirical VC-entropies and Rademacher complexities admit interpretations as "counting" (in senses made precise below) the number of hypotheses falsified by a particular measurement performed when learning.

We start by introducing a special mechanism, the min-risk, which is used implicitly in learning algorithm $\mathcal{A}_{\mathcal{F},\mathcal{D}}$. As we will see, the structure of the measurements performed by the min-risk determine the capacity of the learning algorithm.

Definition 11 *Given repertoire $\mathcal{F} \subset \Sigma$ and unlabeled data $\mathcal{D} \in \mathcal{X}^l$, define the* ***min-risk*** *as the minimum of the empirical risk on $\mathcal{F}$:*

$$\mathbf{R}_{\mathcal{F},\mathcal{D}} : \Sigma \to \mathbb{R} : \sigma \mapsto \min_{f \in \mathcal{F}} \mathbf{R}(f, \mathcal{D}, \sigma\mathcal{D}). \tag{11}$$

The min-risk is a mechanism mapping supervisors σ in Σ to the empirical risk of their best approximations $\mathcal{A}_{\mathcal{F},\mathcal{D}}(\sigma)$ in $\mathcal{F}$, see Fig. 2. Note that inputs to the min-risk are themselves mechanisms.

We suggestively interpret the setup as follows. Suppose a scientist studies a universe where inputs in $\mathcal{X}$ appear according to distribution P, and are assigned labels in $\mathcal{Y}$ by unknown physical process σ^*. The *hypothesis space* is $\Sigma(\mathcal{X}, \mathcal{Y})$, the set of all possible (deterministic) physical processes that take $\mathcal{X}$ to $\mathcal{Y}$.

The scientist's goal is to learn to predict physical process σ^*, on the basis of a small sample of labeled data $(\mathcal{D}, \mathcal{L})$. She has a *theory*, repertoire $\mathcal{F}$, and a method, $\mathcal{A}_{\mathcal{F},\mathcal{D}}$, which she uses to fit some particular $\hat{f} \in \mathcal{F}$ given $\mathcal{L}$.

The most important question for the scientist is: How reliable are predictions made by $\hat{f}$ on *new* data? We will show that $\hat{f}$'s reliability depends on the measurements performed by the min-risk – i.e. on the work done by the scientist when she applies method $\mathcal{A}_{\mathcal{F},\mathcal{D}}$ to find $\hat{f}$.

4.1 Empirical VC entropy

Empirical VC-entropy is, essentially, the effective information generated by the min-risk when it outputs a perfect fit:

Proposition 12 (VC-entropy via effective information)
Empirical VC entropy is

$$\mathcal{V}(\mathcal{F}, \mathcal{D}) = l - ei\left(\mathbf{R}_{\mathcal{F},\mathcal{D}}, 0\right). \tag{12}$$

Proof: Let $\mathcal{X} = \mathcal{D} \cup \mathcal{D}^c$ and $|\mathcal{X}| = m$. Then $\Sigma = \{\sigma : \mathcal{D} \to \mathcal{Y}\} \times \{\sigma : \mathcal{D}^c \to \mathcal{Y}\}$. By definition

$$ei\left(\mathbf{R}_{\mathcal{F},\mathcal{D}}, 0\right) = \log_2 |\Sigma| - \log_2 |\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(0)|,$$

with $\log_2 |\Sigma| = m$. It remains to show that $|\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(0)| = 2^{m-l} \cdot |q_{\mathcal{D}}(\mathcal{F})|$. Points in the image of $q_{\mathcal{D}}$ correspond to labelings σ of the data by functions in $\mathcal{F}$. Thus,

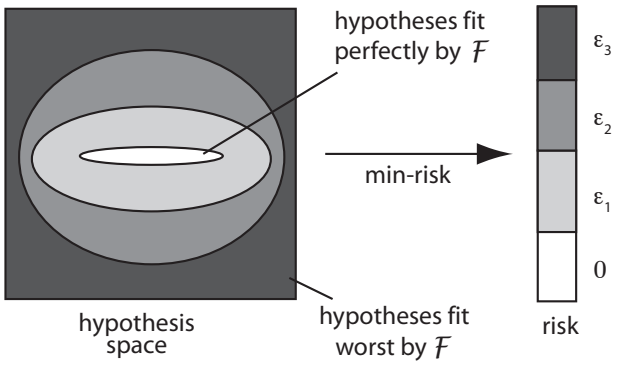

Fig. 2: **The structure of the measurement performed by the min-risk.** The min-risk categorizes potential hypothesis in Σ according to how well they are fit by mechanisms in theory $\mathcal{F}$.

$|q_{\mathcal{D}}(\mathcal{F})|$ counts distinct labelings of $\mathcal{D}$ that $\mathcal{F}$ fits perfectly. These occur with multiplicity 2^{m-l} in the pre-image by the product decomposition of Σ above. ■

We interpret the result as follows. Suppose the scientist applies theory $\mathcal{F}$ to explain her labeled data and perfectly fits function $\hat{f} = \mathcal{A}_{\mathcal{F},\mathcal{D}}(\sigma^*)$ with risk $\epsilon = 0$.

By Definition 3, the meaning of her work is $\Sigma \supset \mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(0)$: the set of mechanisms that her theory $\mathcal{F}$ fits perfectly. The effective information generated by her work is

$$\begin{aligned} ei(\mathbf{R}_{\mathcal{F},\mathcal{D}}, 0) &= \log_2|\Sigma| - \log_2|\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(0)| \\ &= \Big(\text{total no. of hypotheses}\Big) - \Big(\text{no. that theory fits}\Big) \\ &= \Big(\text{no. of hypotheses falsified}\Big), \end{aligned} \tag{13}$$

where hypotheses are counted in bits (after logarithming). A theory is informative if it rules out many potential hypotheses [11].

The number of hypotheses the scientist falsifies when using theory $\mathcal{F}$ to fit $\hat{f}$ has implications for its future performance:

Corollary 13 (information-theoretic empirical VC bound) *With probability $1-\delta$, the risk of predictor $\hat{f} = \mathcal{A}_{\mathcal{F},\mathcal{D}}(\sigma^*)$ outputted by learning algorithm $\mathcal{A}_{\mathcal{F}}$ is bounded by*

$$\mathbf{R}(f) \leq \mathbf{R}(f, \mathcal{D}, \mathcal{L}) + c_1\sqrt{1 - \frac{ei(\mathbf{R}_{\mathcal{F},\mathcal{D}}, 0)}{l}} + c_2\sqrt{\frac{1 - \log_2\delta}{l}}. \tag{14}$$

Proof: By Theorem 9 and Proposition 12. ■

The corollary states that minimizing empirical risk embeds expectations about the future into predictors. So long as the corollary's assumptions hold, future performance by $\hat{f}$ is controlled by: (i) the output of the min-risk, i.e. the fraction ϵ of the data that $\hat{f}$ fits; (ii) the effective information generated by the min-risk, i.e. the number (in bits) of hypotheses the learning algorithm falsifies if it fits perfectly; and (iii) a confidence term. The only assumption made by the corollary is that P and σ^* are *fixed*.

Remark 14 *The theorem provides no guarantees on the future performance of a theory that "explains everything", i.e. $\mathcal{F} = \Sigma$, no matter how well it fits the data. This follows since effective information is zero when $\mathcal{F} = \Sigma$, and so the second term on the right-hand side of Eq.* (14) *is $c_1 \approx 2$.*

Reformulating the above result in terms of code lengths suggests a connection between VC-theory and minimum message length (MML), see [16] and §6.6 of [6]. Recall that, given probability distribution $p(X)$, the message length of event x in an optimal binary code is $\text{len}(x) := -\log_2 p(x)$.

Corollary 15 (VC-entropy controls code length of true hypothesis) *Denote the min-risk by $\mathfrak{m} = \mathbf{R}_{\mathcal{F},\mathcal{D}}$. The length of the true hypothesis $\hat{\sigma}$ in the optimal code for the actual repertoire specified by the min-risk, $p_{\mathfrak{m}}(\Sigma|\epsilon = 0)$, is*

$$len(\hat{\sigma}) = \mathcal{V}(\mathcal{F}, \mathcal{D}) + \big(|\mathcal{X}| - |\mathcal{D}|\big).$$

Proof: By Proposition 12 we have $-\log_2 p_{\mathfrak{m}}(\hat{\sigma}|\epsilon = 0) = \log_2 |\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(0)|$. ■

The length of the message describing the true hypothesis in the actual repertoire's optimal code is the empirical VC-entropy plus a term, $(|\mathcal{X}|-|\mathcal{D}|) = (m-l)$, that decreases as the amount of training data increases. The shorter the message, the better the predictor's expected performance (for fixed empirical risk).

4.2 Empirical Rademacher complexity

VC-entropy only considers hypotheses that theory $\mathcal{F}$ fits perfectly. Rademacher complexity is an alternate capacity measure that considers the distribution of risk across the entire hypothesis space. This section explains Rademacher complexity via an analogy with Solomonoff probability [12,17].

We first recall Solomonoff's definition. Given universal Turing machine T, define (unnormalized) **Solomonoff probability**

$$p_T(s) := \sum_{\{i|T(i)=s\bullet\}} 2^{-\text{len}(i)}, \tag{15}$$

where the sum is over strings[2] i that cause T to output s as a prefix, and $\text{len}(i)$ is the length of i. We adapt Eq. (15) by replacing Turing machine T with min-risk $\mathbf{R}_{\mathcal{F},\mathcal{D}} : \Sigma \to \mathbb{R}$.

[2] A technical point is that no proper prefix of i should output s.

Definition 16 *Equipping hypothesis space with the uniform distribution $p_{unif}(\Sigma)$, all hypotheses have length $len(\sigma) = |\mathcal{X}| = \log_2 |\Sigma|$ in the optimal code. Set the* ***Rademacher distribution*** *for the min-risk $\mathfrak{m} = \mathbf{R}_{\mathcal{F},\mathcal{D}}$ as*

$$p_{\mathfrak{m}}(\epsilon) := \sum_{\{\sigma | R_{\mathcal{F},\mathcal{D}}(\sigma)=\epsilon\}} 2^{-len(\sigma)} = \begin{cases} \frac{\left|\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(\epsilon)\right|}{|\Sigma|} & \text{if } \epsilon \in \mathbf{R}_{\mathcal{F},\mathcal{D}}(\Sigma) \\ 0 & \text{else.} \end{cases} \tag{16}$$

The Rademacher distribution is constructed following Solomonoff's approach after substituting the min-risk as a "special-purpose Turing machine" that only accepts hypotheses in finite set Σ as inputs. It tracks the fraction of hypotheses in Σ that yield risk ϵ.

The Rademacher distribution arises naturally as the denominator when using Bayes' rule to compute the actual repertoire $p_{\mathfrak{m}}(\Sigma|\epsilon)$:

$$p_{\mathfrak{m}}(\sigma|\epsilon) = \frac{p_{\mathfrak{m}}(\epsilon|\sigma)}{p_{\mathfrak{m}}(\epsilon)} \cdot p_{unif}(\sigma), \quad \text{where } p_{\mathfrak{m}}(\epsilon|\sigma) = \begin{cases} 1 & \text{if } \mathbf{R}_{\mathcal{F},\mathcal{D}}(\sigma) = \epsilon \\ 0 & \text{else.} \end{cases}$$

Proposition 17 (Rademacher complexity via min-risk)

$$\mathcal{R}(\mathcal{F},\mathcal{D}) = 1 - 2 \cdot \mathbb{E}\big[\epsilon \,\big|\, p_{\mathfrak{m}}(\epsilon)\big]. \tag{17}$$

Proof: We refer to $\mathbb{E}\big[\epsilon \,\big|\, p_{\mathfrak{m}}(\epsilon)\big]$ as the expected min-risk. From Eq. (8),

$$\mathcal{R}(\mathcal{F},\mathcal{D}) = \frac{1}{|\Sigma|} \sum_{\sigma \in \Sigma} \left[\sup_{f \in \mathcal{F}} \frac{1}{l} \sum_{i=1}^{l} \sigma(x_i) \cdot f(x_i) \right].$$

Observe that $\frac{1}{l}\sum_{i=1}^{l} \sigma(x_i) \cdot f(x_i) = 1 - 2\mathbf{R}(f,\mathcal{D},\sigma)$. It follows that $\sup_{f\in\mathcal{F}} \frac{1}{l}\sum_{i=1}^{l} \sigma(x_i) \cdot f(x_i) = 1 - 2\mathbf{R}_{\mathcal{F},\mathcal{D}}(\sigma)$, which implies

$$\mathcal{R}(\mathcal{F},\mathcal{D}) = 1 - 2\sum_{\sigma\in\Sigma} \frac{\mathbf{R}_{\mathcal{F},\mathcal{D}}(\sigma)}{|\Sigma|} = 1 - 2\sum_{\epsilon} \epsilon \cdot \frac{\left|\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(\epsilon)\right|}{|\Sigma|}. \blacksquare$$

Rademacher complexity is low if the expected min-risk is high. The expected min-risk admits an interesting interpretation. For any hypothesis $\sigma \in \mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(\epsilon)$ the classifier $\hat{f}_\sigma := \mathcal{A}_{\mathcal{F},\mathcal{D}}(\sigma) \in \mathcal{F}$ outputted by the learning algorithm yields incorrect answers on fraction $\epsilon = \frac{1}{l}\sum_{i=1}^{l} \mathbb{I}\big[\hat{f}_\sigma(x_i) \neq \sigma(x_i)\big]$ of the data. It follows that

$$\begin{aligned} \textstyle\sum_\epsilon p_{\mathfrak{m}}(\epsilon) \cdot \epsilon &= \textstyle\sum_\epsilon \quad \frac{\left|\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(\epsilon)\right|}{|\Sigma|} \quad \cdot \quad \frac{1}{l}\sum_l \mathbb{I}\big[\hat{f}_\sigma(x_i) \neq \sigma(x_i)\big] \\ &= \textstyle\sum_\epsilon \Big(\text{fraction of hypotheses falsified}\Big) \cdot \Big(\text{on fraction } \epsilon \text{ of the data}\Big). \end{aligned}$$

A bold theory $\mathcal{F}$ is one for which $\mathbb{E}[\epsilon|p_{\mathfrak{m}}(\epsilon)]$ is high, meaning that its predictors (the classifiers it tries to fit to data) are sufficiently narrow that it would falsify most hypotheses on most of the data.

When a bold theory happens to fit labeled data well, it is guaranteed to perform well in future:

Corollary 18 (information-theoretic empirical Rademacher bound) *With probability* $1-\delta$*, the risk of predictor* $\hat{f} = \mathcal{A}_{\mathcal{F}}(\mathcal{D}, \mathcal{L})$ *outputted by learning machine* $\mathcal{A}_{\mathcal{F}}$ *is bounded by*

$$\mathbf{R}(f) \leq \mathbf{R}(f, \mathcal{D}, \mathcal{L}) + \left[1 - 2\sum_{\epsilon} \epsilon \cdot 2^{-ei(\mathbf{R}_{\mathcal{F},\mathcal{D}},\epsilon)}\right] + c_3\sqrt{\frac{1-\log_2 \delta}{l}} \tag{18}$$

Proof: By Proposition 17 and definition of effective information we have

$$\mathcal{R}(\mathcal{F}, \mathcal{D}) = 1 - 2\sum_{\epsilon} \epsilon \cdot \frac{\left|\mathbf{R}^{-1}_{\mathcal{F},\mathcal{D}}(\epsilon)\right|}{|\Sigma|} = 1 - 2\sum_{\epsilon} \frac{\epsilon}{2^{ei(\mathbf{R}_{\mathcal{F},\mathcal{D}},\epsilon)}}.$$

The result follows by Theorem 10. ■

Rademacher complexity is low if the min-risk's sharp measurements (high ei) are accurate (low ϵ), and conversely. Analogously to Corollary 13, the Rademacher bound implies the future performance of a classifier depends on: (i) the fraction ϵ of the data that $\hat{f}$ fits; (ii) the weighted (by the fraction ϵ of data that falsifies them) sum of the fraction of hypotheses falsified; and (iii) a confidence term. Once again, the only assumption is that P and σ^* are *fixed*.

5 Discussion

Learning according to algorithm $\mathcal{A}_{\mathcal{F},\mathcal{D}}$ entails computing the min-risk, which classifies hypotheses about $\mathcal{D}$ according to how well they are approximated by predictors in repertoire $\mathcal{F}$. Repertoires that rule out many hypotheses when they fit labeled data $(\mathcal{D}, \mathcal{L})$ generate more effective information than repertoires that "approximate everything". As a consequence, when and if an informative repertoire fits labeled data well, Corollary 13 implies we can be confident in future predictions on unseen data.

A pleasing consequence of reformulating empirical VC-entropy and empirical Rademacher complexity in terms of falsifying hypotheses is that it directly connects Popper's intuition about falsifiable theories to statistical learning theory, thereby providing a rigorous justification for the former.

Our motivation for reformulating learning theory information-theoretically arises from a desire to better understand the role of information in biology. Although Shannon information has been heavily and successfully applied to biological questions, it has been argued that it does not fully capture what biologists mean by information since it is not semantic. For example, Maynard Smith states that "In biology, the statement that A carries information about B implies that A has the form it does because it carries that information" [10]. Shannon information was invented to study communication across prespecified channels,

and lacks any semantic content. Maynard Smith therefore argues that a different notion of information is needed to understand in what sense evolution and development embed information into an organism.

It may be fruitful to apply statistical learning theory to models of development. One possible approach is to consider analogs of repertoire $\mathcal{F}$. For example, $\mathcal{F}$ may correspond to the repertoire of possible adult forms a zygote could develop into. The particular adult form chosen, $\hat{f} \in \mathcal{F}$, depends on the historical interactions $(\mathcal{D}, \mathcal{L})$ between the organism and its environment, assuming these can be suitably formalized. The information generated by the organism's development would then have implications for its future interactions with its environment. More speculatively, a similar tactic could be applied to quantify the information embedded in populations by inheritance and natural selection.

References

1. Balduzzi, D., Tononi, G.: Integrated Information in Discrete Dynamical Systems: Motivation and Theoretical Framework. PLoS Comput Biol 4(6), e1000091 (2008)
2. Balduzzi, D., Tononi, G.: Qualia: the geometry of integrated information. PLoS Comput Biol 5(8), e1000462 (Aug 2009)
3. Boucheron, S., Lugosi, G., Massart, P.: A Sharp Concentration Inequality with Applications. Random Structures and Algorithms 16(3), 277–292 (2000)
4. Bousquet, O., Boucheron, S., Lugosi, G.: Introduction to Statistical Learning Theory. In: Bousquet, O., von Luxburg, U., Rätsch, G. (eds.) Advanced Lectures on Machine Learning, vol. 3176, pp. 169–207. Springer (2004)
5. Corfield, D., Schölkopf, B., Vapnik, V.: Falsification and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis Dimensions. Journal for General Philosophy of Science 40(1), 51–58 (2009)
6. Dowe, D.L.: Handbook of the Philosophy of Science. Volume 7: Philosophy of Statistics, chap. MML, hybrid Bayesian network graphical models, statistical consistency, invariance and uniqueness, pp. 901–982. Elsevier (2011)
7. Harman, G., Kulkarni, S.: Reliable Reasoning: Induction and Learning Theory. MIT Press (2007)
8. Koltchinskii, V.: Rademacher penalties and structural risk minimization. IEEE Trans. Inf. Theory 47, 1902–1914 (2001)
9. Lewis, D.: On the Plurality of Worlds. Oxford New York: Basil Blackwell (1986)
10. Maynard Smith, J.: The Concept of Information in Biology. Philosophy of Science 67, 177–194 (2000)
11. Popper, K.: The Logic of Scientific Discovery. Hutchinson (1959)
12. Solomonoff, R.J.: A formal theory of inductive inference I, II. Inform. Control 7(1-22, 224-254) (1964)
13. Vapnik, V.: Estimation of Dependencies Based on Empirical Data. Springer (1982)
14. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons (1998)
15. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. Springer (2005)
16. Wallace, C.S., Boulton, D.M.: An information measure for classification. The Computer Journal 11, 185–194 (1968)
17. Wallace, C.S., Dowe, D.L.: Minimum Message Length and Kolmogorov Complexity. The Computer Journal 42(4), 270–283 (1999)