Falsifiable $\implies$ Learnable

David Balduzzi, Victoria University of Wellington

The paper demonstrates that falsifiability is fundamental to learning. We prove the following theorem for statistical learning and sequential prediction: If a theory is falsifiable then it is learnable – i.e. admits a strategy that predicts optimally. An analogous result is shown for universal induction.

> *A theory that explains everything, [predicts] nothing.* – attributed to Karl Popper.

0. INTRODUCTION

To what extent are theory-based predictions justified by prior observations? The question is known as the problem of induction and is fundamental to scientific inference. We address the problem of induction from the perspective of learning theory. That is, we consider which theories, and under what assumptions, can be applied to make optimal predictions.

Our main result is that the more hypotheses a theory falsifies, suitably quantified, the closer the predictive performance of the best strategy (based on the theory) will be to the theory's *post hoc* explanatory performance on observed data.

0.0. Non-technical overview (or, Learning theory for the working scientist)

Learning theorists have characterized the generalization performance of algorithms in a wide range of scenarios. Although none of these scenarios adequately captures the practice of scientific inference, they form a family of minimal models of prediction.

An intuitive understanding of the main results of learning theory therefore belongs in every scientist's conceptual toolkit. Unfortunately, the results are phrased in opaque terminology that depends on specialized concepts such as Rademacher complexity, shattering coefficients and VC-dimensions.

This paper presents basic results from learning theory in terminology that is meaningful to the broader scientific community.

The results cover three scenarios. In each scenario, Forecaster uses a theory (or theories) to predict Nature's next move(s) based on Nature's previous moves.

S2. *Statistical learning* (SLT). Forecaster aims to predict events sampled from an unknown probability distribution based on a finite sample [Vapnik 1995; Boucheron et al. 2000; Bousquet et al. 2004].

S3. *Sequential prediction* (SEQ). Forecaster aims to predict events generated by an adversarial Nature that adapts to Forecaster's previous moves [Cesa-Bianchi and Lugosi 2006; Abernethy et al. 2009; Rakhlin et al. 2014].

S4. *Universal induction* (UNI). Forecaster aims to predict elements drawn from an arbitrarily chosen computable sequence [Solomonoff 1964; Hutter 2011].

The paper develops the following account.

A. *The risk.*
— The **risk of a theory** is how accurately it *explains* a sequence of events.
A theory explains a sequence of events perfectly if it contains a predictor that correctly labels every element. In general, the accuracy of an explanation is the fraction of the sequence that its best predictor explains correctly.
— The **risk of a strategy** is how accurately it *predicts* a sequence of events.
A strategy specifies picks a predictor based on previously observed events, which it then applies to future events. The strategy's predictive accuracy is the fraction of

future events that it labels correctly.

B. *Learnability.*
  — The **predictive risk** (or **regret**) on a sequence is the difference between a strategy's *predictive* accuracy and the theory's *explanatory* accuracy:

$$\{\text{predictive risk}\} = \{\text{how well strategy predicts}\} - \{\text{how well theory explains}\}$$

The predictive risk measures the strategy's effectiveness. It is not an absolute measure. Effectiveness is relative to a baseline – how well the theory explains the sequence in hindsight. Thus, the predictive risk quantifies the cost from not knowing what Nature will do next, independently of the cost of not having a good model of Nature.
  — A **strategy is optimal** if its predictive risk is asymptotically negligible on any sequence:

$$\{\text{strategy optimal}\} \quad \text{if} \quad \left[\lim_{n \to \infty} \{\text{predictive risk}\} = 0 \right]$$

The definition of optimal is subtle. An optimal strategy does not necessarily predict accurately. Rather, it predicts about as accurately as the theory explains.
  — A **theory is learnable** if it admits an optimal strategy:

$$\{\text{theory learnable}\} \quad \text{if} \quad \exists \{\text{optimal strategy}\}$$

In other words, a theory is learnable if it admits a strategy that predicts future events as well as the theory explains them after the fact.

C. *Falsifiability.*
  — The **falsifiability of a theory** is the fraction of effective hypotheses about a sequence that it cannot explain.
Effective hypotheses are hypotheses about finite sequences. The set of effective hypotheses is necessarily finite. We measure falsifiability in two ways, soft and hard:

$$\mathbf{F} := 2 \sum_{\epsilon \in \mathbb{I}} \Big(\text{fraction of effective hypotheses falsified} \Big) \cdot \Big(\text{on fraction } \epsilon \text{ of data} \Big)$$

$$\mathbf{G} := \frac{\log \text{-\# of effective hypotheses that theory falsifies}}{\log \text{\# of effective hypotheses}}$$

The two notions are, respectively, the expectation of a risk-induced distribution on errors and the risk's Bayesian information gain, see section 2.3. They are closely related to the statistical and sequential Rademacher complexities and covering numbers, and Kolmogorov complexity.
  — A **theory is falsifiable** if the fraction of effective hypotheses that it falsifies tends to one asymptotically.

$$\{\text{theory falsifiable}\} \quad \text{if} \quad \left[\lim_{n \to \infty} \{\text{falsifiability}\} = 1 \right]$$

The number of effective hypotheses grows exponentially with sequence length, so the requirement is quite weak. For example, a theory is falsifiable if the number of hypotheses it explains grows polynomially.

D. *Falsifiable $\implies$ Learnable* (SLT, SEQ).
  — **Main theorem (qualitative).** If a theory is falsifiable, then it is learnable:

$$\{\text{falsifiable}\} \implies \{\text{learnable}\}$$

Alternatively, if a theory is falsifiable then it admits a strategy that predicts optimally – that is, a strategy that predicts any sequence as well, asymptotically, as the theory would have explained the sequence in hindsight.
— **Main theorem (quantitative).**

$$\{\text{predictive risk}\} \leq 1 - \{\text{falsifiability}\}$$

The quantitative version of the main theorem provides guarantees – across all sequences of some finite length n – on the expected performance of a theory's best strategy in terms of the falsifiability of the theory. The qualitative version is a corollary of the quantitative.

E. *Falsifiable $\implies$ Learnable* (UNI).
Universal induction differs significantly from the other two scenarios. We reformulate Solomonoff induction to show that Forecaster constructs a nested sequence of theories in response to observations; from which predictors are drawn uniformly at random. Falsifiability is defined as above in this setting, but it admits a different interpretation:

$$\{\text{falsifiability}\} = \{ \log\text{-\# hypotheses Forecaster eliminates whilst adapting theory}\}$$

Importantly, Forecaster eliminates hypotheses prior to – and separately from – making predictions.
— **Main theorem (quantitative).**

$$\{\text{predictive risk}\} \leq \{\text{falsifiability}\}$$

In short, the number of hypotheses eliminated (or falsified) by Forecaster whilst adapting its theory controls its predictive performance.

0.1. Outline of the paper and summary of the main contributions

The paper is organized as follows. Section 1 introduces two basic tools: the induced distribution and the Bayesian information gain. When a function has a finite domain, a natural prior on the domain is the uniform distribution, in which case the induced distribution and information gain can be interpreted as different ways of counting elements in pre-images.

The next three sections consider statistical learning, sequential prediction and universal induction in turn. The sections are variations on a basic template.

The risk is the fundamental object in all three cases, Definition A in sections $x.1$ for $x = 2, 3, 4$. The risk is a function from sequences of events to errors that can be computed with respect to strategies or theories. In the first case, the risk quantifies predictive performance of the strategy; in the second, it quantifies explanatory performance of the theory in hindsight. The predictive risk is the (minimax) difference between predictive and explanatory performance, Definition B in sections $x.2$.

An event is an ordered pair: a process acting on an input. The key step in the paper is to reformulate the risk as a function from hypothetical processes to errors, by fixing the input sequence. The risk is then a function with a finite domain.

We propose two notions of falsifiability,[1] Definition C in sections $x.3$. The first, soft falsifiability is the expected error under the risk-induced distribution on errors. Intuitively, it is a weighted sum of how many potential hypotheses are falsified over different fractions of the data. The second, hard falsifiability, is the risk's Bayesian information gain. Intuitively, it is the "log-fraction" of falsified hypotheses.

---

[1] Only hard falsifiability is relevant to universal prediction.

The main result is that soft and hard falsifiability control the predictive risk in all three scenarios, Theorems D & E in sections $x.4$. Specifically, we show that falsifiability is equivalent to, or upper or lower bounds, the relevant measures of capacity: the statistical and sequential Rademacher complexities and covering numbers, and Kolmogorov complexity. The bounds on predictive risk then follow from standard results in learning theory [Boucheron et al. 2000; Bousquet et al. 2004; Hutter 2011; Rakhlin et al. 2014]. Proofs are collected in sections $x.5$.

The conclusion discusses the results' implications for Popper's account of scientific inference and the problem of induction, section 5.

The main contributions are:

— Relating the formal models of prediction developed by learning theorists to how working scientists think about scientific inference.
— Deriving falsifiability, and so the fundamental measures of capacity and complexity, as natural properties of the optimization problem at hand (the risk, Remark 2).
— Unifying basic notions from information theory, learning theory, and algorithmic complexity under the rubric of falsifiability.

The simplicity of the definitions and resulting theorems – along with the fact that they apply across diverse settings – suggest that falsifiability may be a more natural, flexible concept than capacity.

0.2. Related work

Connections between falsifiability and statistical learning theory were pointed out in [Vapnik 1995; Harman and Kulkarni 2007; Corfield et al. 2009]. However, these works only considered VC dimension, which does not relate to falsifiability as directly as the measures introduced here. Moreover, they only considered the setting of statistical learning.

Preliminary versions of this work were presented in [Balduzzi 2011; 2013].

0.3. Notation

We have endeavored to use similar notation for the three settings. Consequently, we have been forced to overload certain symbols. In particular, superscripts can refer to both Cartesian products, e.g. $X^n = \prod_{t=1}^{n} X$, and disjoint unions, e.g. $Y^\bullet = \bigcup_{n=1}^{\infty} Y^n$.

indicator function	$\mathbf{I}$	unit interval [0,1]	$\mathbb{I}$
0/1 loss	ℓ	set of distributions on X	Δ_X
expectation	$\mathbf{E}$	probability distribution	$\mathbb{P}$ or $\mathbb{Q}$
risk	$\mathbf{R}$	Bayesian information gain	Gain
predictive risk (regret)	$\mathbf{V}$	Rademacher complexity	Radem
soft falsifiability	$\mathbf{F}$	covering number	Cover
hard falsifiability	$\mathbf{G}$	VC-dimension	vc
set of hypotheses	$\mathcal{H}$	Littlestone dimension	ldim
theory	$\mathcal{O}$	Turing machine	$\mathcal{T}$

We restrict to binary classification in this paper.

1. THE BAYESIAN INFORMATION GAIN AND THE INDUCED DISTRIBUTION

This section presents Bayesian information gain and the induced distribution. They will be used to quantify falsifiability in sections $x.3$.

Suppose that X is a finite set, and that we are given a conditional distribution $\mathbb{P}_\mathfrak{m}(y|x)$ and a prior $\mathbb{P}_X$ on X. The conditional distribution models a noisy channel $\mathfrak{m}$ connecting X to Y.

Definition 1 (Bayesian information gain; induced distribution). *The **Bayesian information gain** when $\mathfrak{m}$ outputs y is*

$$\mathsf{Gain}\big(\mathfrak{m}, y, \mathbb{P}_X\big) := \mathbf{D}\Big[\mathbb{P}_{\mathfrak{m}}(X|y)\,\Big\|\,\mathbb{P}_X(X)\Big],$$

where $\mathbf{D}[\mathbb{P}\,\|\,\mathbb{Q}] := \sum_{x \in X} \mathbb{P}(x)\log\frac{\mathbb{P}(x)}{\mathbb{Q}(x)}$ is the Kullback-Leibler divergence. The posterior $\mathbb{P}_{\mathfrak{m}}(x|y)$ is computed via Bayes' rule

$$\mathbb{P}_{\mathfrak{m}}(x|y) = \mathbb{P}_{\mathfrak{m}}(y|x) \cdot \frac{\mathbb{P}_X(x)}{\mathbb{P}_{\mathfrak{m}}(y)},$$

*where $\mathbb{P}_{\mathfrak{m}}(y) = \sum_{x \in X} \mathbb{P}_X(x)\mathbb{P}_{\mathfrak{m}}(y|x)$ is the $\mathfrak{m}$-**induced distribution** on Y.*

The Bayesian information gain quantifies how much observing y reduces uncertainty about X. We remark that

Proposition 1. *The mutual information communicated across $\mathfrak{m}$ is the expected information gain*

$$I_{\mathfrak{m}}(X, Y) = \mathop{\mathbf{E}}_{y \sim \mathbb{P}_{\mathfrak{m}}(Y)} \mathsf{Gain}\big(\mathfrak{m}, y, \mathbb{P}_X\big),$$

where the expectation is with respect to the $\mathfrak{m}$-induced distribution on Y.

Remark 1 (uniform priors on finite sets). *Unless otherwise specified, finite sets are given the uniform prior: $\mathbb{P}_{\text{unif}}(x) = \frac{1}{|X|}$. We write $\mathsf{Gain}(\mathfrak{m}, y)$ as a shorthand for $\mathsf{Gain}(\mathfrak{m}, y, \mathbb{P}_{\text{unif}})$.*

Given a function $f : X \to Y$, define the corresponding conditional distribution

$$\mathbb{P}_f(y|x) = \begin{cases} 1 & \text{if } y = f(x) \\ 0 & \text{else.} \end{cases}$$

Lemma 2. *Given a function $f : X \to Y$, the f-induced distribution on Y is*

$$\mathbb{P}_f(y) = \begin{cases} \frac{|f^{-1}(y)|}{|X|} & \text{if } y \in \text{im}(f) \\ 0 & \text{else.} \end{cases}$$

The Bayesian information gain is

$$\mathsf{Gain}(f, y) = \begin{cases} -\log \mathbb{P}_f(y) & \text{if } y \in \text{im}(f) \\ \text{undefined} & \text{else.} \end{cases}$$

Lemma 3. *The information gain is zero, $\mathsf{Gain}(f, y) = 0$, if and only if $f(x) = y$ for all $x \in X$.*

2. STATISTICAL LEARNING

Statistical learning is concerned with inductive inference under the assumption that observations are drawn independently from an unknown, but fixed, probability distribution.

This section introduces falsifiability in detail. The later sections on sequential prediction and universal induction rely in part on the presentation developed here.

2.0. Setup

Let X be an arbitrary set and $Y = \{0, 1\}$. Let $Z = X \times Y$. A datum $z = (x, y)$ in Z consists of an input x and an outcome or label y. A process is a map $\sigma : X \to Y$ from

inputs to outcomes. The hypothesis space $\mathcal{H} := Y^X = \{\sigma : X \to Y\}$ is the set of all processes. Finally, an *event* (x, σ) is an element of $X \times \mathcal{H}$.

A *theory* is a set of hypotheses, $\mathcal{O} \subset \mathcal{H}$. Elements of the theory are referred to as predictors. Of course, by definition a predictor is also a hypothesis.

Let $\ell : \mathcal{O} \times X \times Y \to \mathbb{I}$ denote the 0/1 loss:

$$\ell(f, x, y) = \mathbf{I}[f(x) \neq y] = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{else.} \end{cases}$$

Predictor f *explains*[2] datum (x, y) if $\ell(f, x, y) = 0$. If not, then (x, y) falsifies f.

2.1. The risk (SLT)

We assume throughout this section that the sample $\vec{x}$ contains n distinct points.

Let $X^\bullet = \bigcup_{t=1}^\infty X^t$ denote the set of finite sequences of elements of X. We typically refer to sequences $\vec{x} = (x_1, \ldots, x_n)$ rather than sets $\{x_1, \ldots, x_n\}$ to keep notation and terminology consistent across sections.

Definition A (risk, SLT). *The **risk** of theory $\mathcal{O}$ on sequences of events is*

$$\mathbf{R}_{\mathcal{O}}^{\mathsf{SLT}} : \mathcal{H} \times X^\bullet \to \mathbb{I} : (\sigma, \vec{x}) \mapsto \inf_{f \in \mathcal{O}} \frac{1}{n} \sum_{t=1}^n \ell\big(f, x_t, \sigma(x_t)\big),$$

where $n = \mathrm{len}(\vec{x})$. The risk on distributions on data is

$$\mathbf{R}_{\mathcal{O}}^{\mathsf{SLT}} : \Delta_Z \to \mathbb{I} : \mathbb{P}_Z \mapsto \inf_{f \in \mathcal{O}} \mathop{\mathbf{E}}_{z \sim \mathbb{P}_Z} \ell\big(f, z\big).$$

The risk quantifies the fraction of events that the best predictor in $\mathcal{O}$ labels incorrectly – that is, the fraction of events that the theory cannot explain:

$$\mathbf{R}_{\mathcal{O}} : \big\{\text{sequence of events}\big\} \mapsto \big\{\text{fraction of sequence that } \mathcal{O} \text{ cannot explain}\big\}.$$

The risk is zero if and only if there is a predictor in $\mathcal{O}$ that explains the entire sequence of events perfectly.

The set of hypotheses is not finite in general. However, since datasets are always finite, it turns out that the *effective* set of hypotheses is finite.

Definition 2 (effective hypotheses). *Given a sequence $\vec{x} = (x_1, \ldots, x_n)$ of inputs, we say that two hypotheses σ_1 and σ_2 in $\mathcal{H}$ are equivalent*

$$\sigma_1 \sim \sigma_2 \text{ if and only if } \sigma_1(x_t) = \sigma_2(x_t) \text{ for all } t \in \{1, \ldots, n\}.$$

*We refer to an equivalence class $[\sigma] = \{\tau \in \mathcal{H} \,|\, \sigma \sim \tau\}$ of hypotheses as an **effective hypothesis** and let $\mathcal{H}_{ef} = \{[\sigma] \,|\, \sigma \in \mathcal{H}\}$ denote the set of effective hypotheses.*

Since $\vec{x}$ contains n elements, it follows that there is a finite number (2^n) of effective hypotheses.

Two hypotheses in the same equivalence class are indistinguishable on the observed data, and thus indistinguishable to the risk. Given a sequence of n inputs $\vec{x}$, the risk can be written as a function taking effective hypotheses about $\vec{x}$ to errors:

$$\mathbf{R}_{\mathcal{O}, \vec{x}}^{\mathsf{SLT}} : \mathcal{H}_{ef} \to \mathbb{I} : [\sigma] \mapsto \inf_{f \in \mathcal{O}} \frac{1}{n} \sum_{t=1}^n \ell\big(f, x_t, \sigma(x_t)\big). \tag{A}$$

Formulated in this way, the risk quantifies how well theory $\mathcal{O}$ explains the action of an hypothetical process σ on input sequence $\vec{x}$. More precisely, the risk $\epsilon = \mathbf{R}_{\mathcal{O}, \vec{x}}(\sigma)$ is

---

[2]Clearly, we are using 'explain' in a very weak, technical sense.

the fraction of the inputs that the best predictor f in $\mathcal{O}$ misclassifies when labels are generated by σ.

2.2. Learnability (SLT)

A theory is learnable if it admits a strategy whose predictions match the theory's best *post hoc* explanation.

A strategy specifies the predictor that Forecaster will deploy in future as a function of previous events. Formally, a *strategy* is a function taking a finite dataset $\vec{z} = (z_1, \ldots, z_n) \in Z^n$ to a predictor in $\mathcal{O}$. Let $\Psi_n = \{Z^n \to \mathcal{O}\}$ denote the set of strategies on datasets of size n.

Example 1 (empirical risk minimization)**.** *A basic strategy is empirical risk minimization (*ERM*), which outputs the predictor that minimizes the training error:*

$$\psi_{\mathsf{ERM}} : Z^n \to \mathcal{O} : (z_1, \ldots, z_n) \mapsto \operatorname*{arginf}_{f \in \mathcal{O}} \frac{1}{n} \sum_{t=1}^n \boldsymbol{\ell}(f, z_t).$$

Following [Abernethy et al. 2009], we formulate learnability via a game played between Forecaster and Nature. Forecaster picks a strategy $\psi \in \Psi^n$. Nature observes Forecaster's strategy and responds by choosing a distribution $\mathbb{P}_Z \in \Delta_Z$ on events.

The value of the game is the generalization error of Forecaster's strategy on Nature's probability distribution: the difference between the predictive errors Forecaster's strategy accumulates and the explanatory errors of the *theory's best predictor, judged after observing the distribution.* Formally, the value of the game is the difference between the risk $\mathbf{R}_{\{\psi(\vec{z})\}}(\mathbb{P}_Z)$ of the strategy $\psi(\vec{z})$ and the risk $\mathbf{R}_{\mathcal{O}}(\mathbb{P}_Z)$ of the entire theory $\mathcal{O}$.

Forecaster aims to minimize the value; Nature aims for the opposite. The minimax value is thus

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}) := \inf_{\psi \in \Psi^n} \sup_{\mathbb{P}_Z \in \Delta_Z} \underbrace{\left[\operatorname*{\mathbf{E}}_{\vec{z} \sim \mathbb{P}_Z} \operatorname*{\mathbf{E}}_{z' \sim \mathbb{P}_Z} \boldsymbol{\ell}\big(\psi(\vec{z}), z'\big) - \inf_{f \in \mathcal{O}} \operatorname*{\mathbf{E}}_{z' \sim \mathbb{P}_Z} \boldsymbol{\ell}(f, z') \right]}_{\text{expected worst-case generalization error of Forecaster's best strategy}}$$

More concisely,

Definition B (predictive risk, learnability; SLT)**.** *The minimax value of the game, or the **predictive risk** of theory $\mathcal{O}$ on datasets of size n is*

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}) = \underbrace{\inf_{\psi \in \Psi^n}}_{\text{Forecaster's best strategy}} \overbrace{\sup_{\mathbb{P}_Z \in \Delta_Z}}^{\text{Nature's worst distribution}} \underbrace{\left[\operatorname*{\mathbf{E}}_{\vec{z} \sim \mathbb{P}_Z} \mathbf{R}_{\psi(\vec{z})}^{\mathsf{SLT}}(\mathbb{P}_Z) - \mathbf{R}_{\mathcal{O}}^{\mathsf{SLT}}(\mathbb{P}_Z) \right]}_{\text{strategy's generalization error on } \mathbb{P}_Z}. \quad \text{(B)}$$

the generalization error of Forecaster's best strategy when exposed to Nature's worst (for Forecaster) sequence of events.

*Theory $\mathcal{O}$ is **learnable** if $\lim_{n \to \infty} \mathbf{V}_n(\mathcal{O}) = 0$.*

The predictive risk is the cost to Forecaster of not knowing what Nature will do next. It is measured against a baseline: Forecaster's best explanation of the entire sequence. The predictive risk thus separates the costs incurred due to predicting from the costs incurred due to having a theory that does not fit reality perfectly.

If theory $\mathcal{O}$ is learnable then, for large n, the cumulative cost to Forecaster of not knowing what Nature will do next is negligible.

Importantly, the predictive risk says nothing about the *absolute* performance of Forecaster's strategy. A theory may have low predictive risk and still predict a particular sequence of events badly since the baseline – the cost of using a theory that does not fit reality – is subtracted.

2.3. Falsifiability (SLT**)**

A theory is falsifiable to the extent that there are hypotheses that it *cannot* explain.
We quantify falsifiability in two ways.

Definition C (falsifiability, SLT)**.** *Let $\mathbb{Q}_{\mathcal{O},\vec{x}}$ denote the $\mathbf{R}_{\mathcal{O},\vec{x}}^{\mathsf{SLT}}$-induced distribution on $\mathbb{I}$.*
*The **soft falsifiability** of $\mathcal{O}$ on $\vec{x}$ is the expected error*

$$\mathbf{F}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) := 2 \mathop{\mathbf{E}}_{\epsilon \sim \mathbb{Q}_{\mathcal{O},\vec{x}}} [\epsilon] \quad and \quad \mathbf{F}_n^{\mathsf{SLT}}(\mathcal{O}) := \inf_{\vec{x} \in X^n} \mathbf{F}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{x}). \tag{C-s}$$

*The **hard falsifiability** of $\mathcal{O}$ on $\vec{x}$ is*

$$\mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) := \frac{1}{n}\mathsf{Gain}\Big(\mathbf{R}_{\mathcal{O},\vec{x}}^{\mathsf{SLT}}, 0\Big) \quad and \quad \mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O}) := \inf_{\vec{x} \in X^n} \mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{x}). \tag{C-h}$$

*A theory is **falsifiable** if $\lim_{n\to\infty} \mathbf{F}_n(\mathcal{O}) = 1$ or $\lim_{n\to\infty} \mathbf{G}_n(\mathcal{O}) = 1$.*

Remark 2 (falsifiability depends on the risk)**.** *Falsifiability is a property of the risk*
$\mathbf{R}_{\mathcal{O},\vec{x}} : \mathcal{H}_{ef} \to \mathbb{I}$. It depends directly on the optimization problem underlying the learning
scenario.

In contrast, capacity measures are typically presented as properties of the theory $\mathcal{O}$
in such a way that their relation to the optimization problem (specifically, finding the
predictor in $\mathcal{O}$ that minimizes the error) is indirect.

Taking the infimum over all possible datasets implies that $\mathbf{F}_n^{\mathsf{SLT}}(\mathcal{O})$ and $\mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O})$
measure worst-case falsifiability: the falsifiability of $\mathcal{O}$ on the least falsifiable input
sequence.

Soft falsifiability is closely related to Rademacher complexity, see Section 2.5. Simi-
larly, hard falsifiability is closely related to the covering number, and so to the shatter-
ing coefficient and VC-dimension.

The coefficients 2 and $\frac{1}{n}$ in Definition C are chosen so that

Lemma 4. *Soft and hard falsifiability take values in the interval $\mathbb{I} = [0, 1]$.*

(1) Theory $\mathcal{O}$ shatters $\{x_1, \ldots, x_n\}$ if and only if $\mathbf{F}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = \mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = 0$.

(2) Theory $\mathcal{O}$ contains a single predictor if and only if $\mathbf{F}_n^{\mathsf{SLT}}(\mathcal{O}) = \mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O}) = 1$ for all n.

Proof. Straightforward. □

To interpret soft falsifiability, recall that the risk, (A), is function that takes an ef-
fective hypothesis σ about $\vec{x}$ to the fraction $\mathbf{V}$ of the sequence that theory $\mathcal{O}$ cannot
explain (i.e. falsifies)

$$\mathbf{R}_{\mathcal{O},\vec{x}}^{\mathsf{SLT}} : \mathcal{H}_{ef} \to \mathbb{I} : \sigma \mapsto \epsilon$$

The pre-image $\mathbf{R}_{\mathcal{O},\vec{x}}^{-1}(\epsilon) \subset \mathcal{H}$ is the subset of hypotheses that, when applied to input
sequence $\vec{x}$, cannot be explain by theory $\mathcal{O}$ on fraction ϵ of $\vec{x}$. Thus, the risk-induced
probability of $\epsilon \in \mathbb{I}$ is the fraction of potential hypotheses that, if true, cause $\mathcal{O}$ to
falsify ϵ of the data:

$$\mathbb{Q}(\epsilon) = \frac{|\mathbf{R}_{\mathcal{O},\vec{x}}^{-1}(\epsilon)|}{|\mathcal{H}_{ef}|}. \tag{1}$$

Finally, soft falsifiability is the weighted sum:

$$\mathbf{F}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = 2 \sum_{\epsilon \in \mathbb{I}} \left(\frac{|\mathbf{R}_{\mathcal{O},\vec{x}}^{-1}(\epsilon)|}{|\mathcal{H}_{ef}|} \cdot \epsilon \right)$$

$$= 2 \sum_{\epsilon \in \mathbb{I}} \left\{ \text{fraction of effective hypotheses falsified} \right\} \cdot \left\{ \text{on fraction } \epsilon \text{ of data} \right\}.$$

To interpret hard falsifiability, apply Lemma 2 to obtain

$$\mathsf{Gain}(\mathbf{R}_{\mathcal{O},\vec{x}},0) = -\log\mathbb{Q}(0) = \overbrace{\log|\mathcal{H}_{ef}|}^{\text{total \# effective hypotheses}} - \overbrace{\log\left|\mathbf{R}_{\mathcal{O},\vec{x}}^{-1}(0)\right|}^{\text{\# hypotheses } \mathcal{O} \text{ explains perfectly}}$$

$$= \Big\{\log\text{-\# of effective hypotheses that } \mathcal{O} \text{ falsifies}\Big\}.$$

If the inputs in $\vec{x}$ are distinct, then the number of effective hypotheses is 2^n, so

$$\mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = \frac{\Big\{\log\text{-\# of effective hypotheses that } \mathcal{O} \text{ falsifies}\Big\}}{\log\Big\{\#\text{ of effective hypotheses}\Big\}}$$

can be interpreted as the "logarithmic fraction" of effective hypotheses that $\mathcal{O}$ falsifies.

2.4. Falsifiable $\implies$ Learnable (SLT)

The main result is that falsifiability controls predictive risk:

Theorem D (main theorem, SLT)**.**

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}) \le 1 - \mathbf{F}_n^{\mathsf{SLT}}(\mathcal{O}) \le d\sqrt{1 - \mathbf{G}_n^{\mathsf{SLT}}(\mathcal{O})}, \tag{D}$$

where $d = \sqrt{8}$.

Surprisingly, the assumption that Nature is *i.i.d.* is not essential to the result – an almost identical theorem holds for sequential prediction, see section 3.

Proof. By Proposition 6, soft falsifiability of a theory is essentially equivalent to its Rademacher complexity

$$\mathbf{F}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = 1 - 2\mathsf{Radem}^{\mathsf{SLT}}\big(\boldsymbol{\ell}(\mathcal{O})|\vec{x}\big).$$

Similarly, by Proposition 7, hard falsifiability recovers the covering number

$$\mathbf{G}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = 1 - \frac{\log\mathsf{Cover}^{\mathsf{SLT}}(\mathcal{O}|\vec{x})}{n}.$$

The result then follows by Theorem 8, which recalls two standard generalization bounds taken from [Rakhlin et al. 2014]. $\qquad\square$

Remark 3 (vacuous bounds)**.** *Two ways in which Theorem D can be vacuous are*

(1) *If a theory is completely unfalsifiable,* $\mathbf{F}_n(\mathcal{O}) = 0$*, then Theorem D provides no guarantees on its predictive performance no matter how* well *it explains empirical data.*
(2) *If a theory is maximally falsifiable,* $\mathbf{F}_n(\mathcal{O}) = 1$*, then it has zero predictive risk, no matter how* badly *it explains empirical data.*

Corollary D' (falsifiability implies learnability, SLT)**.** *A theory is learnable if it is falsifiable:*

$$\lim_{n\to\infty}\mathbf{V}_n(\mathcal{O}) = 0 \text{ if } \lim_{n\to\infty}\mathbf{F}_n(\mathcal{O}) = 1 \text{ or } \lim_{n\to\infty}\mathbf{G}_n(\mathcal{O}) = 1.$$

A much stronger version Theorem D can also be shown.

Theorem D'' (data-dependent bounds, SLT)**.** *Let*

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{z},\mathbb{P}) := \overbrace{\underbrace{\mathbf{R}_{\psi_{\mathsf{ERM}}(\vec{z})}^{\mathsf{SLT}}(\mathbb{P})}_{\textit{expected test error}} - \underbrace{\mathbf{R}_{\psi_{\mathsf{ERM}}(\vec{z})}^{\mathsf{SLT}}(\vec{z})}_{\textit{training error}}}^{\textit{expected generalization error}}$$

be the expected generalization error of a predictor chosen using ERM.

Suppose that $\vec{z}$ is a sequence of n events drawn from probability distribution $\mathbb{P}$ on Z. Let $\vec{x}$ refer to the same sequence, with labels stripped out. Then, for all $\delta > 0$, with probability at least $1 - \delta$,

(1) the expected generalization error is upper bounded by

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{z}, \mathbb{P}) \leq 1 - \mathbf{F}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) + c\sqrt{\frac{1 - \log \delta}{n}} \qquad \text{(D''-s)}$$

where $c = \sqrt{\frac{2}{\log e}}$.

(2) Furthermore,

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{z}, \mathbb{P}) \leq d_1\sqrt{1 - \mathbf{G}^{\mathsf{SLT}}(\mathcal{O}|\vec{x})} + d_2\sqrt{\frac{1 - \log \delta}{n}} \qquad \text{(D''-h)}$$

where $d_1 = \sqrt{\frac{6}{\log e}}$ and $d_2 = \sqrt{\frac{1}{\log e}}$.

Proof. Propositions 6 and 7 connect soft and hard falsifiability to the Rademacher complexity and covering number.

The result then follows from Theorem 9, which collects two theorems from [Boucheron et al. 2000] and [Bousquet et al. 2004]. □

Theorem D'' is a true inductive bound, which requires the *i.i.d.* assumption. It implies that the difference between the observed training error and expected test error depends on how many hypotheses *about the training sequence $\vec{x}$* are falsified by theory $\mathcal{O}$.

In short, if strategy ψ_{ERM} performs well on the training data, and theory $\mathcal{O}$ falsifies many hypotheses about the training data, then the predictor chosen by ψ_{ERM} will perform well in future, with high probability.

2.5. Proofs (SLT)

Our first two results relate soft falsifiability to Rademacher complexity [Koltchinskii 2001].

Definition 3 (Rademacher complexity). *Define a* Rademacher variable ζ *to be a random variable taking values in $\Omega = \{\pm 1\}$ with equal probability.*

*Let $\vec{\zeta} = (\zeta_1, \ldots, \zeta_n)$ be Rademacher variables. The **Rademacher complexity** of theory $\mathcal{O}$ on unlabeled inputs $\vec{x} = (x_1, \ldots, x_n)$ is*

$$\mathsf{Radem}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) := \mathop{\mathbf{E}}_{\vec{\zeta}}\left[\sup_{f \in \mathcal{O}} \frac{1}{n} \sum_{t=1}^{n} \zeta_t \cdot f(x_t) \right].$$

*The **Rademacher complexity of a theory with respect to a loss function** is*

$$\mathsf{Radem}^{\mathsf{SLT}}\big(\ell(\mathcal{O})|\vec{z}\big) := \mathop{\mathbf{E}}_{\vec{\zeta}}\left[\sup_{f \in \mathcal{O}} \frac{1}{n} \sum_{t=1}^{n} \zeta_t \cdot \ell\big(f, (x_t, y_t)\big) \right].$$

Lemma 5.

$$\mathop{\mathbf{E}}_{\vec{\zeta}} \mathbf{R}_{\mathcal{O}}\big(\vec{x}, \zeta \cdot \vec{y}\big) = \frac{1}{2} - \mathsf{Radem}^{\mathsf{SLT}}\big(\ell(\mathcal{O}) \,\big|\, \vec{z}\big) = \frac{1}{2} - \frac{1}{2}\mathsf{Radem}^{\mathsf{SLT}}\big(\mathcal{O} \,\big|\, \vec{z}\big).$$

Proof. For the first equality, observe that

$$\zeta \cdot (1 - 2\ell(f, z)) = \begin{cases} +1 & \text{if } f(x) = \zeta \cdot y \\ -1 & \text{else,} \end{cases}$$

which implies

$$\frac{1}{2} - \zeta \cdot \left(\frac{1}{2} - \ell(f, z) \right) = \ell(f, (x, \zeta \cdot y)).$$

It follows from $\inf_{f \in \mathcal{O}}[-\psi(f)] = -\sup_{f \in \mathcal{O}} \psi(f)$ that

$$\underset{\vec{\zeta}}{\mathbf{E}} \, \mathbf{R}_{\mathcal{O}}(\vec{x}, \vec{\zeta} \cdot \vec{y}) = \underset{\vec{\zeta}}{\mathbf{E}} \inf_{f \in \mathcal{O}} \sum_{t=1}^{n} \ell(f, (\vec{x}, \vec{\zeta} \cdot \vec{y}))$$

$$= \underset{\vec{\zeta}}{\mathbf{E}} \inf_{f \in \mathcal{O}} \sum_{t=1}^{n} \left[\frac{1}{2} - \zeta_t \left(\frac{1}{2} - \ell(f, z_t) \right) \right]$$

$$= \frac{1}{2} - \underset{\vec{\zeta}}{\mathbf{E}} \sup_{f \in \mathcal{O}} \sum_{t=1}^{n} \zeta_t \cdot \ell(f, z_t)$$

$$= \frac{1}{2} - \mathsf{Radem}^{\mathsf{SLT}}(\ell(\mathcal{O}) \,|\, \vec{z}).$$

The second equality follows similarly. □

A corollary of Lemma 5 is that Rademacher complexity is independent of the labels $\vec{y}$. We therefore drop the labels from the notation and write $\mathsf{Radem}^{\mathsf{SLT}}(\mathcal{O}|\vec{x})$ and $\mathsf{Radem}^{\mathsf{SLT}}(\ell(\mathcal{O})|\vec{x})$ below.

Proposition 6 (Rademacher complexity from soft falsifiability, SLT)**.**

$$\frac{1}{2} \mathbf{F}^{\mathsf{SLT}}(\mathcal{O} \,|\, \vec{x}) = \frac{1}{2} - \mathsf{Radem}^{\mathsf{SLT}}(\ell(\mathcal{O}) \,|\, \vec{x}) = \frac{1}{2} - \frac{1}{2}\mathsf{Radem}^{\mathsf{SLT}}(\mathcal{O} \,|\, \vec{x}).$$

Proof. Recall that $\mathbf{F}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) := 2 \, \mathbf{E}_{\epsilon \sim \mathbb{Q}} \left[\, \epsilon \, \right]$ where $\mathbb{Q}$ is the $\mathbf{R}^{\mathsf{SLT}}_{\mathcal{O}, \vec{x}}$-induced distribution on $\mathbb{I}$. The induced distribution is

$$\mathbb{Q}(\epsilon) = \begin{cases} \frac{|\mathbf{R}^{-1}_{\mathcal{O}, \vec{x}}(\epsilon)|}{|\mathcal{H}_{ef}|} & \text{if } \epsilon \in \mathbf{R}_{\mathcal{O}, \vec{x}}(Y^X) \\ 0 & \text{else.} \end{cases}$$

By Lemma 5 it suffices to show that $\mathbf{E}_{\vec{\zeta}} \, \mathbf{R}_{\mathcal{O}}(\vec{x}, \vec{\zeta} \cdot \vec{y}) = \mathbf{E}_{\epsilon \sim \mathbb{Q}} \left[\, \epsilon \, \right]$. Observe that

$$\underset{\vec{\zeta}}{\mathbf{E}} \, \mathbf{R}_{\mathcal{O}}(\vec{x}, \vec{\zeta} \cdot \vec{y}) = \sum_{[\sigma] \in \mathcal{H}_{ef}} \frac{\mathbf{R}_{\mathcal{O}}(\vec{x}, \sigma \circ \vec{x})}{|\mathcal{H}_{ef}|} = \sum_{\epsilon \in \mathrm{im}(\mathbf{R}_{\mathcal{O}, \vec{x}})} \epsilon \cdot \frac{|\mathbf{R}^{-1}_{\mathcal{O}, \vec{x}}(\epsilon)|}{|\mathcal{H}_{ef}|} = \underset{\epsilon \sim \mathbb{Q}}{\mathbf{E}} \left[\, \epsilon \, \right].$$

as required. □

Next, we relate hard falsifiability to the covering number.

Definition 4 (covering number, SLT)**.** *Given unlabeled data $\vec{x} = (x_1, \dots, x_n) \in X^n$ and a theory $\mathcal{O} \subset Y^X$, let q denote the map*

$$q_{\vec{x}} : \mathcal{O} \to \mathbb{R}^n : f \mapsto \big(f(x_1) \dots f(x_n) \big)$$

*taking predictors to labels. The **covering number** of $\mathcal{O}$ on $\vec{x}$ is*

$$\mathsf{Cover}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) := |q_{\vec{x}}(\mathcal{O})|,$$

the number of distinct labellings produced by the predictors in $\mathcal{O}$ applied to $x_1, \ldots, x_n$.

The shattering coefficient and VC-dimension are discussed in Section 3.6, see Definition 8.

The covering number coincides with hard falsifiability:

Proposition 7 (covering number from hard falsifiability, SLT)**.** *The hard falsifiability of theory $\mathcal{O}$ on $\vec{x}$ is*

$$\mathbf{G}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = 1 - \frac{1}{n} \log \mathsf{Cover}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}).$$

Proof. By definition,

$$\mathsf{Gain}(\mathbf{R}_{\mathcal{O},\vec{x}}, 0) = -\log \frac{|\mathbf{R}_{\mathcal{O},\vec{x}}^{-1}(0)|}{|\mathcal{H}_{ef}|}.$$

Since the sample contains n distinct points and $|Y| = 2$, it follows that $\log |\mathcal{H}_{ef}| = n$. It is easy to check that $|q_x(\mathcal{O})| = |\mathbf{R}_{\mathcal{O},\vec{x}}^{-1}(0)|$. $\square$

Theorem 8 (Data-independent bounds in expectation)**.** *Let*

$$\mathsf{Radem}_n^{\mathsf{SLT}}\big(\boldsymbol{\ell}(\mathcal{O})\big) := \sup_{\mathbb{P} \in \Delta_Z} \mathop{\mathbf{E}}_{\vec{z} \sim \mathbb{P}} \mathsf{Radem}^{\mathsf{SLT}}\big(\boldsymbol{\ell}(\mathcal{O}) \,\big|\, \vec{z}\big),$$

where $\mathrm{len}(\vec{z}) = n$*. Then*

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}) \leq 2\mathsf{Radem}_n^{\mathsf{SLT}}\big(\boldsymbol{\ell}(\mathcal{O})\big) \leq 2\sqrt{\frac{2\mathsf{Cover}_n^{\mathsf{SLT}}(\mathcal{O})}{n}}.$$

Proof. [Rakhlin and Sridharan 2014]. $\square$

Theorem 9 (Data-dependent bounds with high probability)**.** *For all $\delta > 0$, the following bounds hold with probability at least $1 - \delta$,*

(1) The predictive risk is upper bounded by

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{z}) \leq 2\mathsf{Radem}^{\mathsf{SLT}}\big(\boldsymbol{\ell}(\mathcal{O})\big|\vec{x}\big) + c\sqrt{\frac{1 - \log \delta}{n}},$$

where $c = \sqrt{\frac{2}{\log e}}$*.*

(2) Furthermore,

$$\mathbf{V}_n^{\mathsf{SLT}}(\mathcal{O}|\vec{z}) \leq d_1 \sqrt{\frac{\mathsf{Cover}^{\mathsf{SLT}}(\mathcal{O}|\vec{x})}{n}} + d_2 \sqrt{\frac{1 - \log \delta}{n}},$$

where $d_1 = \sqrt{\frac{6}{\log e}}$ *and* $d_2 = \sqrt{\frac{1}{\log e}}$*.*

Proof. [Bousquet et al. 2004] and [Boucheron et al. 2000]. $\square$

3. SEQUENTIAL PREDICTION

Sequential prediction is concerned with predicting a finite sequence of binary observations – without any assumptions on how the observations are generated. The *i.i.d.* assumption of statistical learning is replaced by an adversary that observes Forecaster's previous moves and responds maliciously.

We build on the presentation in section 2. The key technical difference between statistical learning and sequential prediction is the introduction of *trees*, which requires us to distinguish between two notions of risk: soft and hard.

Remarkably, the main theorem has an almost identical form in both sequential prediction and statistical learning. However, the stronger data-dependent form, Theorem D", no longer holds, see discussion in section 5.

3.0. Setup

We introduce some useful notation from [Rakhlin et al. 2014].

Definition 5 (trees; paths)**.** *Let* $\Omega = \{-1, +1\}$*. A* Z-***valued tree*** *of depth* n *is an* n*-tuple* $\vec{\mathbf{z}} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ *of functions* $\mathbf{z}_t : \Omega^{t-1} \to Z$*. Trees are denoted with boldface. A* ***path*** *is an element* $\vec{\omega} = (\omega_1, \ldots, \omega_n) \in \Omega^n$*. Combining a path* $\vec{\omega}$ *with a tree* $\vec{\mathbf{z}}$*, obtains a sequence* $\vec{\mathbf{z}}(\vec{\omega}) = (\mathbf{z}_1, \mathbf{z}_2(\omega_1), \ldots, \mathbf{z}_n(\omega_{1:n-1}))$ *of elements in* Z*.*

It will be convenient to use the shorthand $\mathbf{X}^t := X^{\Omega^t} = \{\mathbf{x}_t : \Omega^t \to X\}$. Let $\mathbf{X}^\bullet = \bigcup_{t=1}^\infty \mathbf{X}^t$ denote the set of all X-valued trees.

3.1. The risk (SEQ)

We assume throughout this section that $\vec{\mathbf{x}}$ contains a path with n distinct points.

Definition A (risk, SEQ)**.** *Let* $\mathcal{H} = Y^X = \{\sigma : X \to Y\}$ *denote the set of hypotheses on* X*. The* ***risk*** *for sequential prediction is*

$$\mathbf{R}_{\mathcal{O}}^{\mathsf{SEQ}} : \mathcal{H} \times (\Omega \times \mathbf{X})^\bullet \to \mathbb{I} : (\sigma, \vec{\omega}, \vec{\mathbf{x}}) \mapsto \inf_{f \in \mathcal{O}} \frac{1}{n} \sum_{t=1}^n \ell\Big(f, \mathbf{x}_t(\omega_{1:t-1}), \sigma\big(\mathbf{x}_t(\omega_{1:t-1})\big)\Big),$$

where $n = \mathrm{len}(\vec{\omega}) = \mathrm{len}(\vec{\mathbf{x}})$*.*

The risk for sequential prediction differs from statistical learning in that the inputs are trees, not elements, and the choice of path in Ω^n is an additional degree of freedom. There are two obvious ways to deal with paths:

(1) *Incorporate paths into the input by defining* $\tilde{\mathbf{X}}^n := \Omega^n \times \mathbf{X}^n$*. Given an* X*-valued tree* $\vec{\mathbf{x}} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *and a path* $\vec{\omega} \in \Omega^n$*, we say that two hypotheses* σ *and* τ *in* $\mathcal{H}$ *are equivalent*

$$\sigma \sim \tau \text{ iff } \sigma\big(\mathbf{x}_t(\omega_{1:t-1})\big) = \tau\big(\mathbf{x}_t(\omega_{1:t-1})\big) \quad \forall t \in \{1, \ldots, n\}.$$

Define the **soft risk**,

$$\mathbf{R}_{\mathcal{O}, (\vec{\omega}, \vec{\mathbf{x}})}^{\mathsf{SEQ}} : \mathcal{H}_{ef} \to \mathbb{I} : \sigma \mapsto \inf_{f \in \mathcal{O}} \left[\frac{1}{n} \sum_{t=1}^n \ell\Big(f, \mathbf{x}_t(\omega_{1:t-1}), \sigma\big(\mathbf{x}_t(\omega_{1:t-1})\big)\Big) \right]. \qquad \text{(A-s)}$$

(2) *Incorporate paths into the hypotheses by defining* $\tilde{\mathcal{H}} := \mathcal{H} \times \Omega^n$*. Similarly, two hypotheses* $(\sigma, \vec{\omega})$ *and* $(\tau, \vec{\rho})$ *in* $\tilde{\mathcal{H}} = \mathcal{H} \times \Omega^n$ *are equivalent*

$$(\sigma, \vec{\omega}) \sim (\tau, \vec{\rho}) \text{ iff } \sigma\big(\mathbf{x}_t(\omega_{1:t-1})\big) = \tau\big(\mathbf{x}_t(\rho_{1:t-1})\big) \quad \forall t \in \{1, \ldots, n\}.$$

Let $\tilde{\mathcal{O}} = \mathcal{O} \times \Omega^n$ and define the **hard risk**,

$$\mathbf{R}_{\tilde{\mathcal{O}}, \vec{\mathbf{x}}}^{\mathsf{SEQ}} : \tilde{\mathcal{H}}_{ef} \to \mathbb{I} : (\sigma, \vec{\rho}) \mapsto \inf_{(f, \vec{\omega}) \in \tilde{\mathcal{O}}} \left[\frac{1}{n} \sum_{t=1}^n \ell\Big(f, \mathbf{x}_t(\omega_{1:t-1}), \sigma\big(\mathbf{x}_t(\rho_{1:t-1})\big)\Big) \right]. \qquad \text{(A-h)}$$

3.2. Learnability (SEQ)

Consider the following game played between Forecaster and Nature over n rounds [Abernethy et al. 2009; Rakhlin et al. 2014].

In the first round, Forecaster chooses a probability distribution $\mathbb{P}_1 \in \Delta_{\mathcal{O}}$ on the set of predictors. Nature observes Forecaster's choice, and picks $z_1 \in Z$. A predictor f_1 is then sampled at random from $\mathbb{P}_1$, applied to z_1 and the loss $\ell(f_1, z_1)$ is computed. The game continues for n rounds, where both Forecaster and Nature observe the moves played in previous rounds.

The value of the game is Forecaster's *regret*: the difference between Forecaster's cumulative loss and the loss Forecaster would have accumulated, had it played the best move in hindsight. Forecaster's goal is to minimize its regret; Nature's aims for the opposite:

$$\mathbf{V}_n^{\mathsf{SEQ}}(\mathcal{O}) = \inf_{\mathbb{P}_1 \in \Delta_{\mathcal{O}}} \sup_{z_1 \in Z} \mathop{\mathbf{E}}_{f_1 \sim \mathbb{P}_1} \cdots \inf_{\mathbb{P}_n \in \Delta_{\mathcal{O}}} \sup_{z_n \in Z} \mathop{\mathbf{E}}_{f_n \sim \mathbb{P}_n} \frac{1}{n} \underbrace{\left[\sum_{t=1}^n \ell(f_t, z_t) - \inf_{f \in \mathcal{O}} \sum_{t=1}^n \ell(f, z_t) \right]}_{\text{Forecaster's regret}}$$

Forecaster's move at time t depends on the prior moves by Forecaster and Nature. Forecaster's strategy at time t can be expressed as a function $\psi_t : Z^{t-1} \to \mathcal{O}$. Let $\Psi_t = \{\psi_t : Z^{t-1} \to \mathcal{O}\}$ denote the strategies available to Forecaster at time t, and let $\Psi = \prod_{t=1}^n \Psi_t$ denote the strategies available to Forecaster over an n-round game.

Similarly, Nature's strategy at time t is an element of $\Xi_t = \mathcal{O}^{t-1} \times \Delta_{\mathcal{O}} \to Z$. Let $\Xi = \prod_{t=1}^n \Xi_t$ denote the n-round strategies available to Nature. We can write the minimax value more compactly as

$$\mathbf{V}_n^{\mathsf{SEQ}}(\mathcal{O}) = \inf_{\mathbb{P} \in \Delta_{\Psi}} \sup_{\vec{\xi} \in \Xi} \mathop{\mathbf{E}}_{\vec{\psi} \sim \mathbb{P}} \frac{1}{n} \left[\sum_{t=1}^n \ell\big(\psi_t(\xi_{1:t-1}), \xi_t(\psi_{1:t-1}, \mathbb{P}_t)\big) - \inf_{f \in \mathcal{O}} \sum_{t=1}^n \ell\big(f, \xi_t(\psi_{1:t-1}), \mathbb{P}_t\big) \right],$$

where the sup and inf are understood to unravel recursively as above.

Finally, substituting in the risk obtains

Definition B (predictive risk, SEQ). *The minimax value of an n-round game, or **predictive risk** of theory $\mathcal{O}$, is*

$$\mathbf{V}_n^{\mathsf{SEQ}}(\mathcal{O}) = \inf_{\mathbb{P} \in \Delta_{\Psi}} \sup_{\vec{\xi} \in \Xi} \mathop{\mathbf{E}}_{\vec{\psi} \sim \mathbb{P}} \left[\mathbf{R}_{\vec{\psi}}^{\mathsf{SEQ}}(\vec{\xi}) - \mathbf{R}_{\mathcal{O}}^{\mathsf{SEQ}}(\vec{\xi}) \right]. \tag{B}$$

*Theory $\mathcal{O}$ is **learnable** if $\lim_{n \to \infty} \mathbf{V}_n^{\mathsf{SEQ}}(\mathcal{O}) = 0$.*

The first term, $\mathbf{R}_{\vec{\psi}}(\vec{\xi})$ is the cumulative loss incurred by the best $\mathcal{O}$-based strategy played out on Nature's sequence of moves $\vec{\xi}$. The comparator term, $\mathbf{R}_{\mathcal{O}}(\vec{\xi})$ is the performance of the best predictor in $\mathcal{O}$, taken in hindsight.

3.3. Falsifiability (SEQ)

We use the soft and hard risk to define soft and hard falsifiability:

Definition C (falsifiability, SEQ). *Let $\mathbb{Q}_{\mathcal{O},(\vec{\omega}, \vec{\mathbf{x}})}$ be the $\mathbf{R}_{\mathcal{O},(\vec{\omega}, \vec{\mathbf{x}})}^{\mathsf{SEQ}}$-induced distribution on $\mathbb{I}$. The **soft falsifiability** of theory $\mathcal{O}$ on $\vec{\mathbf{x}}$ is the expected error of the soft risk*

$$\mathbf{F}_n^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}}) := 2 \mathop{\mathbf{E}}_{\vec{\omega} \sim \mathbb{P}_{unif}(\Omega^n)} \mathop{\mathbf{E}}_{\epsilon \sim \mathbb{Q}_{\mathcal{O},(\vec{\omega}, \vec{\mathbf{x}})}} [\epsilon] \quad and \quad \mathbf{F}_n^{\mathsf{SEQ}}(\mathcal{O}) := \inf_{\vec{\mathbf{x}} \in \mathbf{X}} \mathbf{F}_n^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}}). \tag{C-s}$$

*The **hard falsifiability** of theory $\mathcal{O}$ on $\vec{\mathbf{x}}$ is the information gain from the hard risk*

$$\mathbf{G}_n^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}}) := \frac{1}{n}\mathsf{Gain}(\mathbf{R}_{\mathcal{O}\times\Omega^n,\vec{\mathbf{x}}}^{\mathsf{SEQ}},0) \quad and \quad \mathbf{G}_n^{\mathsf{SEQ}}(\mathcal{O}) := \inf_{\vec{\mathbf{x}}\in\mathbf{X}}\mathbf{G}_n^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}}). \qquad \text{(C-h)}$$

*A theory is **falsifiable** if $\lim_{n\to\infty}\mathbf{F}_n(\mathcal{O}) = 1$ or $\lim_{n\to\infty}\mathbf{G}_n(\mathcal{O}) = 1$.*

Hard falsifiability is closely related to the sequential covering number introduced in [Rakhlin et al. 2014]. However, the definition is more intuitive and, importantly, it also leads to combinatorial bounds such as the Littlestone dimension, see Section 3.5 for details.

3.4. Falsifiable $\implies$ Learnable (SEQ)

Finally, we obtain the main theorem for sequential prediction, which is an exact analog of the corresponding theorem for statistical learning:

Theorem D (main theorem, SEQ)**.**

$$\mathbf{V}_n^{\mathsf{SEQ}}(\mathcal{O}) \leq 1 - \mathbf{F}_n^{\mathsf{SEQ}}(\mathcal{O}) \leq c\sqrt{1 - \mathbf{G}_n^{\mathsf{SEQ}}(\mathcal{O})} \qquad \text{(D)}$$

where $c = \sqrt{8}$.

An important point is that hard falsifiability provides a *non-vacuous* upper-bound for the zero-covering number, see Section 3.6.

Proof. By Proposition 10, soft falsifiability is equivalent to the sequential Rademacher complexity

$$\mathbf{F}^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}}) = 1 - 2\mathsf{Radem}^{\mathsf{SEQ}}\big(\ell(\mathcal{O})|\vec{\mathbf{x}}\big).$$

The first inequality then follows from Theorem 11, taken from [Rakhlin et al. 2014].

By Lemma 12 and Proposition 13, hard falsifiability can be used to upper bound the sequential zero-covering number:

$$\frac{\mathsf{Cover}^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}})}{n} \leq 1 - \mathbf{G}^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}}).$$

The second inequality then follows from Theorem 14, also taken from [Rakhlin et al. 2014]. $\qquad\square$

Corollary D' (falsifiability implies learnability, SEQ)**.** *A theory is learnable if it is falsifiable:*

$$\lim_{n\to\infty}\mathbf{V}_n(\mathcal{O}) = 0 \text{ if } \lim_{n\to\infty}\mathbf{F}_n(\mathcal{O}) = 1 \text{ or } \lim_{n\to\infty}\mathbf{G}_n(\mathcal{O}) = 1.$$

3.5. Proofs (SEQ)

This section proves the falsification bounds in Theorem D for sequential prediction.

Definition 6 (Sequential Rademacher complexity)**.**

$$\mathsf{Radem}^{\mathsf{SEQ}}(\mathcal{O}|\vec{\mathbf{x}}) := \mathop{\mathbf{E}}_{\vec{\zeta}}\left[\sup_{f\in\mathcal{O}}\frac{1}{n}\sum_{t=1}^{n}\zeta_t f(\mathbf{x}_t(\zeta_{1:t-1}))\right]$$

Proposition 10 (Rademacher complexity from induced distribution, SEQ)**.** *Let $\mathbb{Q}_{\vec{\omega}} := \mathbb{P}_{\mathbf{R}_{\mathcal{O},(\vec{\omega},\vec{\mathbf{x}})}^{\mathsf{SEQ}}}$ be the distribution on errors in $\mathbb{I}$ induced by the soft risk $\mathbf{R}_{\mathcal{O},(\vec{\omega},\vec{\mathbf{x}})}^{\mathsf{SEQ}} : \mathcal{H} \to \mathbb{I}$. Then,*

$$\mathsf{Radem}^{\mathsf{SEQ}}(\ell(\mathcal{O}),\vec{\mathbf{x}}) = \frac{1}{2} - \mathop{\mathbf{E}}_{\vec{\omega}\sim\mathbb{P}_{\mathrm{unif}}(\Omega^n)}\mathop{\mathbf{E}}_{\epsilon\sim\mathbb{Q}_{\vec{\omega}}}\big[\,\epsilon\,\big].$$

Proof. As for Proposition 6.   □

Theorem 11. *The predictive risk of sequential prediction is bounded by*

$$\mathbf{V}_n^{\mathsf{SEQ}}(\mathcal{O}) \leq 2 \sup_{\vec{\mathbf{x}} \in \mathbf{X}} \mathsf{Radem}^{\mathsf{SEQ}}\big(\boldsymbol{\ell}(\mathcal{O}), \vec{\mathbf{x}}\big),$$

where the sup *is over trees of length* n.

Proof. [Rakhlin et al. 2014].   □

Next, we upper bound the covering number of a tree-process. The following definition is given in [Rakhlin et al. 2014]

Definition 7 (covering number, SEQ). *A **zero-cover** of $\mathcal{O}$ on an X-valued tree $\vec{\mathbf{x}}$ is a set V of Y-valued trees such that*

$$\forall f \in \mathcal{O}, \ \forall (\omega_1, \ldots, \omega_n) \in \Omega^n, \ \exists \mathbf{v} \in V \ \textit{s.t.} \ f(\mathbf{x}_t(\omega_{1:t-1})) = \mathbf{v}_t(\omega_{1:t-1}) \ \forall t \in \{1, \ldots, n\}.$$

*The **covering number** of $\mathcal{O}$ on $\mathbf{x}$ is*

$$\mathsf{Cover}^{\mathsf{SEQ}}(\mathcal{O}, \vec{\mathbf{x}}) = \min\{|V| : V \textit{ is a zero-cover}\}.$$

The sequential covering number is awkward for our purposes since, unlike the statistical covering number in Definition 4, it is not defined as the cardinality of the image of a function. We therefore need the following

Lemma 12 (upper bound for sequential covering number). *Let*

$$q_{\vec{\mathbf{x}}} : \tilde{\mathcal{O}} \to \mathbb{R}^n : (f, \vec{\omega}) \mapsto \Big(f(\mathbf{x}_1), f(\mathbf{x}_2(\omega_1)), \ldots, f(\mathbf{x}_n(\omega_{1:n-1})) \Big)$$

The covering number is upper bounded by

$$\mathsf{Cover}^{\mathsf{SEQ}}(\mathcal{O}, \vec{\mathbf{x}}) \leq |q_{\vec{\mathbf{x}}}(\tilde{\mathcal{O}})|.$$

Proof. We prove the lemma by constructing a zero-cover V_q of $\mathcal{O}$ on $\vec{\mathbf{x}}$ with $|q_{\vec{\mathbf{x}}}(\tilde{\mathcal{O}})|$ elements.

Suppose the image $q_{\vec{\mathbf{x}}}(\mathcal{O} \times \Omega^n)$ has $\mathcal{N}$ elements, $\mathbf{q}^1, \ldots, \mathbf{q}^{\mathcal{N}}$. Define

$$\mathbf{v}^j(\omega_{1:t-1}) := \mathbf{q}_t^j.$$

That is, $\mathbf{v}^j(\omega_{1:t-1})$ is the tth element of $\mathbf{q}^j$ for all paths in Ω^n. Then, by construction $V_q = \{\mathbf{v}^1, \ldots, \mathbf{v}^{\mathcal{N}}\}$ is a zero-cover of $\vec{\mathbf{x}}$ containing $\mathcal{N}$ elements, and we are done.   □

Proposition 13.

$$\mathsf{Gain}(\mathbf{R}_{\tilde{\mathcal{O}}, \vec{\mathbf{x}}}^{\mathsf{SEQ}}, 0) = n - \log|q_{\vec{\mathbf{x}}}(\tilde{\mathcal{O}})|.$$

Proof. As for Proposition 7.   □

Theorem 14. *Let $\vec{\mathbf{x}}$ be an X-valued tree of length n. Then,*

$$\mathsf{Radem}^{\mathsf{SEQ}}(\mathcal{O}, \vec{\mathbf{x}}) \leq \sqrt{\frac{2 \log \mathsf{Cover}^{\mathsf{SEQ}}(\mathcal{O}, \vec{\mathbf{x}})}{n}}$$

Proof. [Rakhlin et al. 2014].   □

It follows from Lemma 12, Proposition 13 and Theorem 14 that hard falsifiability can be used to upper bound the predictive risk for sequential prediction.

3.6. A sequential-to-statistical reduction

Definition 7, of the sequential covering number, is fairly intricate and fragile. For example, slightly changing the definition by reordering the quantifiers gives a quantity that grows much too fast and yields vacuous generalization bounds [Rakhlin and Sridharan 2014].

A natural concern is therefore that the upper bound in Lemma 12 is too loose. In the remainder of this section, we show that $|q_{\vec{x}}(\tilde{\mathcal{O}})|$, and so hard falsifiability, is a *useful*, non-vacuous upper bound.

Definition 8 (shattering, VC and Littlestone dimensions)**.** *We have the following analogous definitions:*

(*1*) Statistical.
    *Theory $\mathcal{O}$ **shatters** input sequence $\vec{x}$ of length n if*

$$\forall \vec{\omega} \in \Omega^n \quad \exists f \in \mathcal{O} \quad \text{s.t.} \quad f(x_t) = \frac{\omega_t + 1}{2} \quad \forall t \in \{1, \dots, n\}.$$

    *Alternatively, $\mathcal{O}$ shatters $\vec{x}$ if $\mathsf{Cover}^{\mathsf{SLT}}(\mathcal{O}|\vec{x}) = 2^n$. The **VC-dimension** is*

$$\mathsf{vc}(\mathcal{O}) := \sup \left\{ n \,\middle|\, \exists \, \textit{input sequence } \vec{x} \textit{ of length } n \textit{ s.t. } \mathcal{O} \textit{ shatters } \vec{x} \right\}$$

(*2*) Sequential.
    *Theory $\mathcal{O}$ SEQ-**shatters** tree $\vec{\mathbf{x}}$ of length n if*

$$\forall \vec{\omega} \in \Omega^n \quad \exists f \in \mathcal{O} \quad \text{s.t.} \quad f\big(\mathbf{x}_t(\omega_{1:t-1})\big) = \frac{\omega_t + 1}{2} \quad \forall t \in \{1, \dots, n\}.$$

    *The **Littlestone dimension** is*

$$\mathsf{ldim}(\mathcal{O}) = \sup_{\vec{\mathbf{x}}} \left\{ n \,\middle|\, \exists X\textit{-valued tree } \vec{\mathbf{x}} \textit{ of length } n \textit{ s.t. } \mathcal{O} \textit{ SEQ-shatters } \vec{\mathbf{x}} \right\}.$$

Let $Y^{\mathbf{X}^\bullet} := \{\sigma : \mathbf{X}^\bullet \to Y\}$ denote the set of hypotheses on the set $\mathbf{X}^\bullet$ of X-valued trees. Given theory $\mathcal{O} \subset Y^X$, define the *new theory*

$$\tilde{\mathcal{O}} := \mathcal{O} \times \Omega^\bullet \subset Y^{\mathbf{X}^\bullet} : (f, \vec{\omega})(\mathbf{x}_t) = f\big(\mathbf{x}_t(\omega_{1:t-1})\big).$$

The lifted theory $\tilde{\mathcal{O}}$ acts on trees, which from our point of view are just another set. The statistical covering number for $\tilde{\mathcal{O}}$ is given, following Definition 4m using the function,

$$q_{\vec{\mathbf{x}}} : \tilde{\mathcal{O}} \to \mathbb{R}^n : (f, \vec{\omega}) \mapsto \Big((f, \vec{\omega})(\mathbf{x}_1), \dots, (f, \vec{\omega})(\mathbf{x}_n) \Big)$$

with $\mathsf{Cover}^{\mathsf{SLT}}(\tilde{\mathcal{O}}|\vec{\mathbf{x}}) = |q_{\vec{\mathbf{x}}}(\tilde{\mathcal{O}})|$. The VC-dimension of $\tilde{\mathcal{O}}$ is then computed straightforwardly.

Proposition 15 (VC-dimension lower bounds Littlestone dimension)**.** *The Littlestone dimension of $\mathcal{O}$ is lower-bounded by the VC-dimension of the lifted theory $\tilde{\mathcal{O}} = \mathcal{O} \times \Omega^\bullet$:*

$$\mathsf{vc}(\tilde{\mathcal{O}}) \leq \mathsf{ldim}(\mathcal{O}).$$

The proposition shows that the Littlestone dimension can be recovered from hard falsifiability. Thus, hard falsifiability can play the same role as the sequential covering number in reducing learning problems into combinatorial problems.

Proof. Suppose there is a tree $\vec{\mathbf{x}}$ of length n shattered by $\tilde{\mathcal{O}}$. We construct a new tree $\vec{\mathbf{z}}$ of length n that is SEQ-shattered by $\mathcal{O}$.

Thus, we assume that

$$\forall (\omega_1, \ldots, \omega_n) \in \Omega^n, \, \exists (f, \vec{b}) \in \tilde{\mathcal{O}} \quad \text{s.t.} \quad f\big(\mathbf{x}_t(b_{1:t-1})\big) = \frac{\omega_t + 1}{2} \quad \forall t \in \{1, \ldots, n\}. \quad (2)$$

Let α denote the function specified by $\alpha(\omega_{1:t-1}) = b_{1:t-1}$, as in (2). Construct the new tree $\vec{z}$ by $\vec{z} = \vec{x} \circ \alpha$. It follows, by the construction of α and by (2), that $\forall (\omega_1, \ldots, \omega_n) \in \Omega^n, \, \exists f \in \mathcal{O}$ such that

$$f\big(\vec{z}(\omega_{1:t-1})\big) = f\big(\mathbf{x}_t \circ \alpha(\omega_{1:t-1})\big) = f\big(\mathbf{x}_t(b_{1:t-1})\big) = \frac{\omega_t + 1}{2} \quad \forall t \in \{1, \ldots, n\}$$

as required. $\square$

The following instructive example, taken from [Rakhlin and Sridharan 2014], was designed to exhibit the intricacy of the sequential covering number's definition. We conclude by computing the statistical covering number of $\tilde{\mathcal{O}}$ on the example, and showing that it yields the correct result.

Example 2. *Consider the function class*

$$\mathcal{O} = \{f_a \mid a \in \mathbb{I}, \, f_a(x) = 0 \, \forall x \neq a, \, f_a(a) = 1\} \subset Y^{\mathbb{I}}.$$

Assuming that the tree $\vec{x}$ takes on 2^{n-1} distinct values (the "worst case"), then for any ordered pair $(f, \vec{\omega})$ we have that

$$q_{\vec{x}}(f_a, \vec{\omega}) = \Big(f_a(\mathbf{x}_1), f_a(\mathbf{x}_2(\omega_1)), \ldots, f_a(\mathbf{x}_n(\omega_{1:n-1})) \Big)$$

is either equal to all zeros, or all zeros with a single coordinate that equals one. The image of $q_{\vec{x}}$ therefore contains at most $n+1$ points and in fact $|q_{\vec{x}}(\tilde{\mathcal{O}})| = n+1$.

4. UNIVERSAL INDUCTION

The third setting is universal induction, which is concerned with predicting computable sequences of binary observations. The setting differs significantly from statistical learning and sequential prediction. For example, universal induction cannot be modeled adversarially since both Nature and Forecaster have too many degrees of freedom.

There are at least two interpretations of universal induction:

U1. *Universal.* Forecaster has a single, universal theory.
U2. *Adaptive.* Forecaster constructs a series of theories in response to successive observations.

The first interpretation is standard. The second, which we advocate here, is new. Both are legitimate.

Under the first interpretation, it does not make sense to evaluate the falsifiability of theories – since there is only one theory and it is universal. The only choice that matters is Nature's choice of sequence $\vec{y}$. It then turns out that the number of hypotheses Nature falsifies (eliminates) whilst choosing $\vec{y}$ controls Forecaster's predictive risk, see section 4.6.

Under the second interpretation, developed in detail below, Forecaster's predictive risk is controlled by the number of hypotheses that Forecaster falsifies whilst adapting its theories.

4.0. Setup

Let $\mathcal{X}$ denote the set of valid programs, where valid programs $\mathcal{X} \subset \bigcup_{t=1}^{\infty} \{0, 1\}^t$ form a prefix-free set. A *prefix-free* universal Turing machine $\mathcal{T}$ takes valid programs to

outputs. Let $Y^\infty = \{0, 1, 00, 01, 10, 11, 000, \ldots\}$ denote the set of all binary sequences, of finite or infinite length. A Turing machine is a function

$$\mathcal{T} : \mathcal{X} \to Y^\infty.$$

Let $\mathcal{Y} = \mathcal{T}(\mathcal{X}) \subset Y^\infty$ denote the set of computable sequences.

Prefix free strings formalize the notion of a computer program. For example, the set of valid C++ programs is a prefix free set since C++'s syntax ensure one program cannot be the prefix of another. The set of valid programs has a complicated structure, since it includes strings of varying length.

It is mathematically convenient to force programs to have a fixed length. First, let

$$\mathcal{X}^n = \{\vec{x} \in \mathcal{X} \mid \operatorname{len}(\vec{x}) = t \text{ for some } t \le n\}.$$

Second, *pad out* short programs: given a program $\vec{x}$ of length $t < n$, construct 2^{n-t} programs of length n by adding arbitrary suffixes to $\vec{x}$. For example, if $\operatorname{len}(\vec{x}) = n - 2$, then the four padded programs are $\{\vec{x}00, \vec{x}01, \vec{x}10, \vec{x}11\}$. The Turing machine ignores the padding. Concretely, a C++ compiler would also ignore the padding, so the padded-out programs are all functionally equivalent.

Let $\mathcal{H}^n$ denote the set of binary strings of length $\le n$ and let $\mathcal{O}^n \subset \mathcal{H}^n$ denote the set of valid, padded programs of length n. Denote the function that strips out the padding by

$$\mathcal{S}^n : \mathcal{H}^n \to \mathcal{X} \cup \{\emptyset\} : \vec{h} \mapsto \begin{cases} \vec{x} & \text{if } \mathcal{O}^n \ni \vec{h} = \vec{x}\vec{s} \text{ for } \vec{x} \text{ a valid program with padding } \vec{s} \\ \emptyset & \text{else.} \end{cases}$$

In other words, if the string contains a valid program as prefix, then $\mathcal{S}^n$ strips out the padding. If the string does not contain a valid program, then $\mathcal{S}^n$ outputs a null character.

The reason for introducing padded strings is that it allows the following simple description of the Solomonoff prior as a limit distribution, induced by the uniform distribution on padded strings:

Definition-Proposition 16 (Solomonoff prior). *Equip $\mathcal{H}^n$ with the uniform distribution for all n. Let $\mathbb{P}_n$ denote the $\mathcal{S}^n$-induced distribution on $\mathcal{X} \cup \{\emptyset\}$. Then*

$$\mathbb{P}_{\mathcal{S}}(\vec{x}) := \lim_{n \to \infty} \mathbb{P}_n(\vec{x}) = 2^{-\operatorname{len}(\vec{x})}$$

for all $\vec{x} \in \mathcal{X}$.

*Let $\mathbb{Q}_n$ denote the $(\mathcal{S}^n \circ \mathcal{T})$-induced distributed on Y^∞. The **Solomonoff prior** is*

$$\mathbb{Q}_{\mathsf{SOL}}(\vec{y}) := \lim_{n \to \infty} \mathbb{Q}_n(\vec{y}) = \sum_{\{\vec{x} \mid \mathcal{T}(\vec{x}) = \vec{y}\bullet\}} 2^{-\operatorname{len}(\vec{x})}.$$

Proof. The standard definition of the Solomonoff prior, and a demonstration that our definition coincides with the standard, are provided in section 4.5. $\square$

Proposition 16 allows us to consider how Solomonoff induction acts on inputs to the Turing machine, instead of its outputs.

4.1. The risk (UNI)

For universal induction, the loss compares the sequences generated by Nature and Forecaster element-wise:

$$\ell : Y \times Y \to \mathbb{R} : (y, y') \mapsto \mathbf{I}[y \ne y'],$$

where as above $Y = \{0, 1\}$.

Definition A (risk, UNI). *The **risk** for universal induction is*

$$\mathbf{R}^n : \mathcal{H}^n \times \mathcal{H}^n \to \mathbb{R}_{\geq 0} : (\vec{x}, \vec{f}) \mapsto \sum_{t=1}^{\infty} \boldsymbol{\ell}\big(\mathcal{T}(\vec{x})_t, \mathcal{T}(\vec{f})_t\big)$$

*The **risk** of theory $\mathcal{O}^n := \mathcal{X}^n$ is*

$$\mathbf{R}^{\mathsf{UNI}}_{\mathcal{O}^n} : \mathcal{H} \to \mathbb{R}_{\geq 0} : \vec{x} \mapsto \inf_{\vec{f} \in \mathcal{O}^n} \sum_{t=1}^{\infty} \boldsymbol{\ell}\big(\mathcal{T}(\vec{f})_t, \mathcal{T}(\vec{x})_t\big).$$

As for statistical learning and sequential prediction, we reinterpret the risk as a function from hypotheses – that is, programs with length at most n – to nonnegative reals

$$\mathbf{R}^n_{\vec{y}} : \mathcal{H}^n \to \mathbb{R}_{\geq 0} : \vec{x} \mapsto \sum_{t=1}^{\infty} \boldsymbol{\ell}\big(\mathcal{T}(\vec{x})_t, y_t\big). \tag{A}$$

In the limit we obtain $\mathbf{R}^{\mathsf{UNI}}_{\vec{y}} := \lim_{n \to \infty} \mathbf{R}^n_{\vec{y}}$ as a function $\mathbf{R}^{\mathsf{UNI}}_{\vec{y}} : \mathcal{H} \to \mathbb{R}_{\geq 0}$.

4.2. Learnability (UNI)

Suppose that Nature chooses a sequence $\vec{y} \in \mathcal{Y}$ and reveals $\vec{y}_{1:t-1} = (y_1, \ldots, y_{t-1})$ at time t. Let $\psi_t = \{\psi_t : Y^{t-1} \to \Delta_Y\}$ denote the set of strategies available to Forecaster in round t, and $\Psi = \prod_{t=1}^{\infty} \psi_t$ the set of all strategies available to Forecaster.

The risk of strategy ψ is

$$\mathbf{R}^{\mathsf{UNI}}_{\psi} : \mathcal{H} \to \mathbb{R}_{\geq 0} : \vec{x} \mapsto \sum_{t=1}^{\infty} \mathbf{E}\,\boldsymbol{\ell}\Big(\psi_t\big(\mathcal{T}(\vec{x})_{1:t-1}\big), \mathcal{T}(\vec{x})_t\Big),$$

where the expectation is over the outputs of the (probabilistic) strategy.

A particularly important strategy is *Solomonoff induction* [Solomonoff 1964]:

Definition-Proposition 17 (Solomonoff induction). *Let*

$$\mathcal{O}^n_t := (\mathbf{R}^n_{y_{1:t-1}})^{-1}(0) = \big\{hypotheses\ of\ length \leq n\ that\ explain\ y_{1:t-1}\big\}.$$

Theory $\mathcal{O}^n_t$ is a finite set; equip it with the uniform distribution. Let $\mathbb{P}_{n,t}(\vec{x})$ denote the $\mathcal{S}^n$-induced distribution on $\mathcal{X}$ and $\mathbb{Q}_{n,t}(\vec{y})$ denote the $(\mathcal{S}^n \circ \mathcal{T})$-induced distribution on Y^{∞}.

***Solomonoff induction** is the strategy:*

$$(\psi_{\mathsf{SOL}})_t : Y^{t-1} \to \Delta_Y : y_{1:t-1} \mapsto \lim_{n \to \infty} \mathbb{Q}_{n,t}(y_t) = \mathbb{Q}_{\mathsf{SOL}}(y_t | y_{1:t-1}).$$

Solomonoff induction depends on the choice of Turing machine, although this dependence is typically not explicit in our notation.

Proof. We show that $\lim_{n \to \infty} \mathbb{Q}_{n,t}(y_t) = \mathbb{Q}_{\mathsf{SOL}}(y_t | y_{1:t-1})$ in section 4.5. $\square$

Solomonoff induction can be interpreted as follows. Forecaster's theory at time step t is $\mathcal{O}_t := \lim_{n \to \infty} \mathcal{O}^n_t$, a limit of finite sets. All hypotheses consistent with the previous observations $y_{1:t-1}$ are weighted equally (recalling that padding entails redundancies). Forecaster predicts the next observation by drawing from $\mathcal{O}_t$ uniformly at random. After observing y_t, and regardless of whether or not Forecaster's prediction at time t was correct, Forecaster constructs new theory $\mathcal{O}_{t+1}$ in the light of y_t.

In short, Solomonoff induction learns by constructing a nested set of progressively smaller theories and predicts by sampling from them uniformly at random.

Definition B (predictive risk, UNI)**.** *The **predictive risk** of strategy ψ and theory $\mathcal{O}^n$ is*

$$\mathbf{V}^{\mathsf{UNI}}(\psi - \mathcal{O}^n | \vec{y}) := \mathbf{R}^{\mathsf{UNI}}_{\psi}(\vec{y}) - \mathbf{R}^{\mathsf{UNI}}_{\mathcal{O}^n}(\vec{y})$$

*The **predictive risk** of strategy ψ is*

$$\mathbf{V}^{\mathsf{UNI}}(\psi | \vec{y}) := \lim_{n \to \infty} \mathbf{V}^{\mathsf{UNI}}_{\psi}(\mathcal{O}^n | \vec{y}). \tag{B}$$

4.3. Falsifiability (UNI)

This subsection and the next relate the error accumulated using Solomonoff induction to the falsifiability of the string chosen by Nature.

Definition C (falsifiability, UNI)**.**

$$\mathbf{G}^{\mathsf{UNI}}_{\mathcal{T}}(\vec{y}) := \lim_{n \to \infty} \mathsf{Gain}(\mathbf{R}^n_{\vec{y}}, 0). \tag{C-h}$$

Remark 4. *The definition for universal induction differs from statistical learning and sequential prediction, in that the coefficient $\frac{1}{n}$ is not present, and so $\mathbf{G}^{\mathsf{UNI}}$ does not necessarily take values in $[0, 1]$.*

To interpret hard falsifiability, first fix an ambient hypothesis space $\mathcal{H}^n$, and consider the hypotheses falsified when observing the substring $y_{1:t}$:

$$\begin{aligned}
\mathbf{G}^n_{\mathcal{T}}(\vec{y}_{1:t}) &= \log 2^n - \log |\mathcal{O}^n_t| \\
&= \big\{ \log\text{-\# strings of length } n \big\} - \big\{ \log\text{-\# strings that output } y_{1:t} \big\} \\
&= \big\{ \log\text{-\# strings of length } n \text{ falsified by } y_{1:t} \big\}.
\end{aligned}$$

Second, consider the hypotheses eliminated when transitioning between theories:

$$\begin{aligned}
\log |\mathcal{O}^n_t| - \log |\mathcal{O}^n_{t-1}| &= \big\{ \log\text{-\# strings outputting } y_{1:t-1} \big\} - \big\{ \log\text{-\# strings outputting } y_{1:t} \big\} \\
&= \big\{ \log\text{-\# strings falsified when modifying } \mathcal{O}_{t-1} \mapsto \mathcal{O}_t \big\}.
\end{aligned}$$

Finally, combining the above obtains

$$\begin{aligned}
\mathbf{G}^{\mathsf{UNI}}_{\mathcal{T}}(\vec{y}) &= \sum_{t=1}^{\infty} \lim_{n \to \infty} \Big(\mathbf{G}^n_{\mathcal{T}}(\vec{y}_{1:t}) - \mathbf{G}^n_{\mathcal{T}}(\vec{y}_{1:t-1}) \Big) \quad \text{where } y_{1:0} := \emptyset \\
&= \sum_{t=1}^{\infty} \lim_{n \to \infty} \Big(\log |\mathcal{O}^n_t| - \log |\mathcal{O}^n_{t-1}| \Big) \\
&= \sum_{t=1}^{\infty} \big\{ \log\text{-\# strings falsified when modifying } \mathcal{O}_{t-1} \mapsto \mathcal{O}_t \big\}.
\end{aligned}$$

Thus, the hard falsifiability of $\vec{y}$ is the number of hypotheses Forecaster eliminates in the process of adapting its theory to the data. Note that theories are falsified *prior* to predicting: at time t, Forecaster first eliminates hypotheses based on $y_{1:t}$ and then uses the new theory $\mathcal{O}_{t+1}$ to predict y_{t+1}.

4.4. Falsifiable $\implies$ Learnable (UNI)

The main theorem for universal induction differs from statistical learning and sequential prediction, in that Forecaster's theory is not fixed. Falsifiability quantifies the hypotheses that Forecaster eliminates whilst adapting its theory. The more Forecaster is required to adapt – *prior* to predicting – the weaker the guarantee on its predictive performance.

Theorem E (main theorem, UNI)**.** *The predictive risk under Solomonoff induction (1) coincides with the expected error and (2) is bounded by the number of hypotheses Nature falsifies when choosing the string $\vec{y}$:*

$$\mathbf{V}(\psi_{\mathsf{SOL}}|\vec{y}) = \mathbf{R}^{\mathsf{UNI}}_{\psi_{\mathsf{SOL}}}(\vec{y}) \le \mathbf{G}^{\mathsf{UNI}}_{\mathcal{T}}(\vec{y}). \tag{E}$$

Proof. By Lemma 19, the predictive risk and risk coincide for universal induction: $\mathbf{V}^{\mathsf{UNI}}(\psi|\vec{y}) = \mathbf{R}^{\mathsf{UNI}}_{\psi}(\vec{y})$.

By Proposition 20, the hard falsifiability of $\vec{y}$ coincides with (the negative logarithm of) the Solomonoff prior

$$\mathbf{G}^{\mathsf{UNI}}_{\mathcal{T}}(\vec{y}) = -\log \mathbb{Q}_{\mathsf{SOL}}(\vec{y}).$$

Finally, the result follows by Solomonoff's Theorem 21.   □

More generally, Theorem E suggests that Bayesian updating is a way of modifying theories, whose cost (measured in errors) can be bounded using falsifiability.

We conclude by relating falsifiability to Kolmogorov complexity. Intuitively, a string is simple if it is the output of a short computer program. More formally,

Definition 9 (Kolmogorov complexity)**.** *The **Kolmogorov complexity** of a string, with respect to Turing machine $\mathcal{T}$, is the length of the shortest program that outputs the string as a prefix [Kolmogorov 1965]:*

$$\mathbf{K}_{\mathcal{T}}(\vec{y}) := \min_{\vec{x} \in \mathcal{X}} \left\{ \mathrm{len}(\vec{x}) \,\middle|\, \mathcal{T}(\vec{x}) = \vec{y} \bullet \right\}$$

The Kolmogorov complexity $\mathbf{K}_{\mathcal{T}}$ depends on the choice of Turing machine up to an additive constant that does not depend on $\vec{y}$ [Li and Vitányi 2008].

Proposition 18 (relation between falsifiability and Kolmogorov complexity)**.** *Falsifiability lower bounds Kolmogorov complexity:*

$$\mathbf{G}^{\mathsf{UNI}}_{\mathcal{T}}(\vec{y}) \le \mathbf{K}_{\mathcal{T}}(\vec{y}).$$

Further, $\mathbf{G}^{\mathsf{UNI}}_{\mathcal{T}}(\vec{y}) = \mathbf{K}_{\mathcal{T}}(\vec{y})$ up to an additive constant that does not depend on $\vec{y}$.

Proof. The inequality follows from the definitions of the Solomonoff prior and Kolmogorov complexity.

By Levin's coding theorem [Li and Vitányi 2008], the Kolmogorov complexity of a string coincides with the negative log probability of the string according to the Solomonoff prior up to an additive constant.   □

4.5. Proofs (UNI)

Equip $\mathcal{H}^n$ with the uniform distribution and let $\mathbb{P}_{\mathcal{S}^n}(\mathcal{X})$ denote the $\mathcal{S}^n$-induced distribution on $\mathcal{X}$. Recall that we defined the Solomonoff prior as the limit of the $\mathcal{T} \circ \mathcal{S}^n$-induced distribution on $\mathcal{Y}$

$$\mathbb{Q}_{\mathsf{SOL}}(\vec{y}) := \lim_{n \to \infty} \mathbb{P}_{\mathcal{T} \circ \mathcal{S}^n}(\vec{y}),$$

where $\mathcal{T} \circ \mathcal{S}^n : \mathcal{H}^n \xrightarrow{\mathcal{S}^n} \mathcal{X} \cup \{\emptyset\} \xrightarrow{\mathcal{T}} \mathcal{Y} \cup \{\emptyset\}$.

Definition-Proposition 16. *The following hold:*

(1) The limit $\mathbb{P}_{\mathcal{S}}(\mathcal{X}) := \lim_{n \to \infty} \mathbb{P}_n(\mathcal{X})$ is well-defined with

$$\mathbb{P}_{\mathcal{S}}(\vec{x}) = 2^{-\mathrm{len}(\vec{x})}.$$

(2) *The limit* $\mathbb{Q}_{\mathsf{SOL}}(\mathcal{Y}) := \mathbb{P}_{\mathcal{T} \circ \mathcal{S}}(\mathcal{Y}) = \lim_{n \to \infty} \mathbb{P}_{\mathcal{T} \circ \mathcal{S}^n}(\mathcal{Y})$ *is well-defined and coincides with the Solomonoff prior. That is,*

$$\mathbb{Q}_{\mathsf{SOL}}(\vec{y}) = \sum_{\{\vec{x} \in \mathcal{X} | \mathcal{T}(\vec{x}) = \vec{y} \bullet\}} 2^{-\operatorname{len}(\vec{x})}.$$

Proof. *Claim* 1. By Lemma 2, the induced probability of a valid program is

$$\mathbb{P}_{\mathcal{S}^n}(\vec{x}) = \begin{cases} \sum_{\vec{x}\vec{s}} \frac{1}{2^n} = \frac{2^{n-\operatorname{len}(\vec{x})}}{2^n} = 2^{-\operatorname{len}(\vec{x})} & \text{if } \operatorname{len}(\vec{x}) \leq n \\ 0 & \text{else.} \end{cases}$$

Thus, $\lim_{n \to \infty} \mathbb{P}_{\mathcal{S}^n}(\vec{x}) = 2^{-\operatorname{len}(\vec{x})}$ for all valid programs.
Claim 2. Also by Lemma 2. $\qquad \square$

Recall that the standard definition of Solomonoff induction is as the strategy:

$$(\psi_{\mathsf{SOL}})_t : Y^{t-1} \to \Delta_Y : y_{1:t-1} \mapsto \mathbb{Q}_{\mathsf{SOL}}(y_t | y_{1:t-1}) := \frac{\mathbb{Q}_{\mathsf{SOL}}(y_{1:t})}{\mathbb{Q}_{\mathsf{SOL}}(y_{1:t-1})}.$$

Definition-Proposition 17. *The two definitions of Solomonoff induction coincide:*

$$\lim_{n \to \infty} \mathbb{Q}_{n,t}(y_t) = \frac{\mathbb{Q}_{\mathsf{SOL}}(y_{1:t})}{\mathbb{Q}_{\mathsf{SOL}}(y_{1:t-1})}.$$

Proof. The theory $\mathcal{O}_t^n$ is the set of all strings of length $\leq n$ consistent with the observation $y_{1:t-1}$. Pushing the uniform distribution on $\mathcal{H}^n$ forward onto Y^∞ yields, asymptotically, the conditional Solomonoff distribution. $\qquad \square$

Lemma 19 (predictive risk reduces to risk)**.** *If $\vec{y}$ is computable then*

$$\mathbf{V}^{\mathsf{UNI}}(\psi | \vec{y}) := \lim_{n \to \infty} \mathbf{V}_\psi^{\mathsf{UNI}}(\mathcal{O}^n | \vec{y}) = \mathbf{R}_\psi^{\mathsf{UNI}}(\vec{y}).$$

Proof. As $n \to \infty$, the theory incorporates all valid programs, and so can match any computable sequence. Thus,

$$\lim_{n \to \infty} \mathbf{R}_{\mathcal{O}^n}^{\mathsf{UNI}}(\vec{y}) = 0$$

and the result follows. $\qquad \square$

Proposition 20 (hard falsifiability and Solomonoff prior)**.** *The hard falsifiability of string $\vec{y}$ for Turing machine $\mathcal{T}$ is*

$$\mathbf{G}_{\mathcal{T}}^{\mathsf{UNI}}(\vec{y}) = -\log \mathbb{Q}_{\mathsf{SOL}}(\vec{y}).$$

Proof. Observe that the risk factorizes as

$$\begin{array}{ccccc} \mathbf{R}_{\vec{y}}^{\mathsf{UNI}} : & \mathcal{X} & \xrightarrow{\mathcal{T}} & \mathcal{Y} & \xrightarrow{\Sigma \boldsymbol{\ell}} & \mathbb{R} \\ & \vec{x} & \mapsto & \mathcal{T}(\vec{x}) & \mapsto & \sum_{t=1}^{\infty} \boldsymbol{\ell}(\mathcal{T}(\vec{x})_t, y_t). \end{array}$$

The proposition follows from the following two claims.

Claim 1. $\mathsf{Gain}(\mathcal{T}, \vec{y}) = -\log \mathbb{Q}_{\mathsf{SOL}}(\vec{y})$ *for all* $\vec{y} \in \mathcal{Y}$.

Consider the function $\mathcal{T} : \mathcal{X} \to Y$, where $\mathcal{X}$ is equipped with the distribution $\mathbb{P}_{\mathcal{S}}(\mathcal{X})$ from Proposition 16. Since Turing machines are deterministic, we have that $\mathbb{P}_{\mathcal{T}}(\vec{y}|\vec{x}) = 1$, and so

$$\mathbb{P}_{\mathcal{T}}(\vec{x}|\vec{y}) = \mathbb{P}_{\mathcal{T}}(\vec{y}|\vec{x}) \cdot \frac{\mathbb{P}_{\mathcal{S}}(\vec{x})}{\mathbb{P}_{\mathcal{T}}(\vec{y})} = \frac{\mathbb{P}_{\mathcal{S}}(\vec{x})}{\mathbb{P}_{\mathcal{T}}(\vec{y})}$$

It follows that

$$\mathsf{Gain}(\mathcal{T}, \vec{y}) = \mathbf{D}\Big[\mathbb{P}_{\mathcal{T}}(\mathcal{X}|\vec{y}) \,\Big\|\, \mathbb{P}_{\mathcal{S}}(\mathcal{X})\Big] = \sum_{\vec{x} \in \mathcal{X}} \mathbb{P}_{\mathcal{T}}(\vec{x}|\vec{y}) \log \frac{\mathbb{P}_{\mathcal{T}}(\vec{x}|\vec{y})}{\mathbb{P}_{\mathcal{S}}(\vec{x})}$$

$$= \sum_{\vec{x} \in \mathcal{X}} \mathbb{P}_{\mathcal{T}}(\vec{x}|\vec{y}) \log \frac{\mathbb{P}_{\mathcal{T}}(\vec{y}|\vec{x}) \cdot \mathbb{P}_{\mathcal{S}}(\vec{x})}{\mathbb{P}_{\mathcal{T}}(\vec{y}) \cdot \mathbb{P}_{\mathcal{S}}(\vec{x})} = \sum_{\vec{x} \in \mathcal{X}} \mathbb{P}_{\mathcal{T}}(\vec{x}|\vec{y}) \log \frac{1}{\mathbb{P}_{\mathcal{T}}(\vec{y})}$$

$$= -\log \mathbb{P}_{\mathcal{T}}(\vec{y})$$

$$= -\log \mathbb{Q}_{\mathsf{SOL}}(\vec{y}).$$

where the last equality follows from Proposition 16.

Claim 2. $\mathbf{G}^{\mathsf{UNI}}(\vec{y}) = \mathsf{Gain}(\mathcal{T}, \vec{y})$.

Follows from $\mathbf{G}^{\mathsf{UNI}}(\vec{y}) = \mathsf{Gain}(\mathbf{R}_{\vec{y}}, 0)$ and $\mathbf{R}_{\vec{y}}^{-1}(0) = \mathcal{T}^{-1}(\vec{y})$.

Concatenating the claims yields the desired result. $\qquad\qquad\square$

Theorem 21 (generalization bound for Solomonoff induction)**.**

$$\sum_{t=1}^{\infty} \mathbf{E}\, \ell\Big(\psi_{\mathsf{SOL}}(y_{1:t-1}), y_t\Big) \le -\log \mathbb{Q}_{\mathsf{SOL}}(\vec{y}).$$

Proof. The following proof is taken from [Hutter 2011]:

$$\sum_{t=1}^{\infty} \mathbf{E}\, \ell\Big(\psi_{\mathsf{SOL}}(y_{1:t-1}), y_t\Big) = \sum_{t=1}^{\infty} \Big|1 - \mathbb{Q}_{\mathsf{SOL}}(y_t|y_{1:t-1})\Big|$$

$$\le -\sum_{t=1}^{\infty} \log \mathbb{Q}_{\mathsf{SOL}}(y_t|y_{1:t-1})$$

$$= -\log \mathbb{Q}_{\mathsf{SOL}}(\vec{y}),$$

where the inequality holds because $1 - x \le -\log x$. $\qquad\qquad\square$

4.6. Interpreting Solomonoff induction as a universal theory

Under the standard interpretation, Forecaster's theory is $\mathcal{O}$ and $\mathbf{G}_{\mathcal{T}}^{\mathsf{UNI}}(\vec{y})$ counts the hypotheses falsified by Nature whilst choosing $\vec{y}$:

$$\mathbf{G}_{\mathcal{T}}^{\mathsf{UNI}}(\vec{y}) = \lim_{n \to \infty}\Big[\log\big\{\# \text{ strings of length } n\big\} - \log\big\{\# \text{ that output } y\big\}\Big]$$

$$= \lim_{n \to \infty}\Big\{\text{log-}\# \text{ strings of length } n \text{ that Nature falsifies}\Big\}.$$

5. DISCUSSION

[A] theory of induction is superfluous. It has no function in a logic of science. The best we can say of a hypothesis[3] is that up to now it has been able to show its worth, and that it has been more successful than other hypotheses although, in principle, it can never be justified, verified,

---

[3]This paper uses 'theory' in the sense that Popper uses 'hypothesis'.

> *or even shown to be probable. This appraisal of the hypothesis relies solely upon deductive conse-*
> *quences (predictions) which may be drawn from the hypothesis: There is no need even to mention*
> *'induction'.*
>
> – from [Popper 1959].

We conclude by discussing the paper's implications for scientific inference, focusing on the ideas of Karl Popper. According to Popper, inductive inference is meaningless. As an alternative, he advocated *hypothetico-deductive inference*, which proceeds as follows [Gelman and Shalizi 2013].

Forecaster makes observations, proposes a theory, and deduces consequences. A theory is scientific if it is *falsifiable*. That is, if it is possible to deduce empirically testable consequences. The scientific method, according to Popper, is: to propose falsifiable theories that are in line with past observations; to subject them to severe empirical tests; and to discard and replace them if and when they are falsified.

Popper's ideas are extremely influential in the scientific community. Indeed, he is essentially the only philosopher that scientists draw on as a resource to evaluate and compare theories. Philosophers, however, consider Popper's approach to be fundamentally flawed [Godfrey-Smith 2011]. The three main problems that have been identified are:

P1. *Infinite alternatives.* The set of imaginable hypotheses is infinite, so that it is trivial to find a collection of specific hypotheses that a specific theory falsifies.
P2. *Stochasticity.* It is unclear how to apply Popper's ideas to stochastic theories, which cannot be definitely falsified.
P3. *No confirmation.* Popper rejected the notion that positive evidence should increase our confidence in a scientific theory. Rejecting confirmation eliminates any rationale, aside from habit, for using a well-tested theory over a brand new theory, assuming both are falsifiable.

Our formulation of falsifiability does not exactly line up with what Popper had in mind. We proceed regardless.

Problem *P1* is solved by restricting attention to the finite set of effective hypotheses. Problem *P2* is also solved as a corollary of our results. Soft and hard falsifiability are defined with respect to *deterministic* hypotheses, whereas the predictive risk allows *probabilistic* hypotheses.

Problem *P3* is more interesting. If Nature is *i.i.d.* then Theorem D" provides a guarantee on a predictor's future accuracy that depend on the theory's falsifiability and the predictor's past performance. Thus, with the addition of the *i.i.d.* assumption, there *is* quantifiable confirmation.

If no assumptions are made about Nature's behavior, then the setting is sequential prediction. The most that can be said is that, if a theory is falsifiable, then its predictive performance can be as good as its explanatory performance in hindsight. Nothing *absolute* can be said about predictive performance *a priori*.

Finally, Solomonoff induction is purported to be a (non-computable) theory that optimally explains and predicts every computable string. However, observe that Theorem E says *nothing* about Solomonoff induction's predictive performance unless $\mathbf{G}_{\mathcal{T}}^{\mathsf{UNI}}(\vec{y})$ or the Kolmogorov complexity $\mathbf{K}_{\mathcal{T}}(\vec{y})$ are known *a priori* – which is never the case. For example, suppose Nature picks a string that contains 10^9 zeros followed by 10^9 coin flips, followed by only zeros. Solomonoff induction's error rate on the first billion instances will not be indicative of its performance on the next billion. Assuming that Nature chooses strings with low Kolmogorov complexity is analogous to, albeit weaker than, assuming Nature is *i.i.d.*

The current state-of-the-art in learning theory therefore supports Popper's intuitions about falsifiability – including his rejection of confirmation. In a more positive vein,

learning theory suggests that inductive inference requires additional assumptions and provides tools for analyzing their implications.

REFERENCES

Jacob Abernethy, Alekh Agarwal, Peter L Bartlett, and Alexander Rakhlin. 2009. A stochastic view of optimal regret through minimax duality. In *COLT*.

David Balduzzi. 2011. Information, learning and falsification, In Philosophy and Machine Learning workshop, Neural Information Processing Systems (NIPS). *arXiv* (2011).

David Balduzzi. 2013. Falsification and Future Performance. In *Algorithmic Probability and Friends: Bayesian Prediction and Artificial Intelligence*, David Dowe (Ed.). LNAI, Vol. 7070. Springer, 65–78.

S Boucheron, G Lugosi, and P Massart. 2000. A Sharp Concentration Inequality with Applications. *Random Structures and Algorithms* 16, 3 (2000), 277–292.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. 2004. Introduction to Statistical Learning Theory. In *Advanced Lectures on Machine Learning*, O Bousquet, U von Luxburg, and G Rätsch (Eds.). Springer, 169–207.

Nicolo Cesa-Bianchi and Gabor Lugosi. 2006. *Prediction, Learning and Games*. Cambridge University Press.

David Corfield, Bernhard Schölkopf, and V Vapnik. 2009. Falsification and Statistical Learning Theory: Comparing the Popper and Vapnik-Chervonenkis Dimensions. *Journal for General Philosophy of Science* 40, 1 (2009), 51–58.

Andrew Gelman and Cosma Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *Brit. J. Math. Statist. Psych.* 66 (2013), 8–38.

Peter Godfrey-Smith. 2011. Popper's Philosophy of Science: Looking Ahead. In *The Cambridge Companion to Popper*, J Shearmur and G Stokes (Eds.). Cambridge University Press.

Gilbert Harman and Sanjeev Kulkarni. 2007. *Reliable Reasoning: Induction and Learning Theory*. MIT Press.

Marcus Hutter. 2011. Universal Learning Theory. In *Encyclopedia of Machine Learning*, Claude Sammut and Geoffrey I Webb (Eds.). Springer.

A N Kolmogorov. 1965. Three approaches to the quantitative definition of information. *Problems Inform. Transmission* 1, 1 (1965), 1–7.

V Koltchinskii. 2001. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory* 47 (2001), 1902–1914.

M Li and P Vitányi. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer.

Karl Popper. 1959. *The Logic of Scientific Discovery*. Hutchinson.

Alexander Rakhlin and Karthik Sridharan. 2014. STAT928: Statistical Learning Theory and Sequential Prediction. Lecture Notes.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. 2014. Online Learning via Sequential Complexities. In *JMLR*.

R J Solomonoff. 1964. A formal theory of inductive inference I, II. *Inform. Control* 7, 1-22, 224-254 (1964).

V Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer.