On the Foundations of Universal Sequence Prediction

Marcus Hutter

IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland

marcus@idsia.ch http://www.idsia.ch/~marcus

8 February 2006

Abstract

Solomonoff completed the Bayesian framework by providing a rigorous, unique, formal, and universal choice for the model class and the prior. We discuss in breadth how and in which sense universal (non-i.i.d.) sequence prediction solves various (philosophical) problems of traditional Bayesian sequence prediction. We show that Solomonoff's model possesses many desirable properties: Fast convergence and strong bounds, and in contrast to most classical continuous prior densities has no zero p(oste)rior problem, i.e. can confirm universal hypotheses, is reparametrization and regrouping invariant, and avoids the old-evidence and updating problem. It even performs well (actually better) in non-computable environments.

Contents

Keywords

Sequence prediction, Bayes, Solomonoff prior, Kolmogorov complexity, Occam's razor, prediction bounds, model classes, philosophical issues, symmetry principle, confirmation theory, reparametrization invariance, old-evidence/updating problem, (non)computable environments.

1 Introduction

Examples and goal. Given the weather in the past, what is the probability of rain tomorrow? What is the correct answer in an IQ test asking to continue the sequence 1,4,9,16,? Given historic stock-charts, can one predict the quotes of tomorrow? Assuming the sun rose 5000 years every day, how likely is doomsday (that the sun does not rise) tomorrow? These are instances of the important problem of inductive inference or time-series forecasting or sequence prediction. Finding prediction rules for every particular (new) problem is possible but cumbersome and prone to disagreement or contradiction. What we are interested in is a formal general theory for prediction.

Bayesian sequence prediction. The Bayesian framework is the most consistent and successful framework developed thus far [Ear93]. A Bayesian considers a set of environments=hypotheses=models $\mathcal{M}$ which includes the true data generating probability distribution μ. From one's prior belief w_ν in environment $\nu\in\mathcal{M}$ and the observed data sequence $x=x_1...x_n$, Bayes' rule yields one's posterior confidence in ν. In a predictive setting, one directly determines the predictive probability of the next symbol x_{n+1} without the intermediate step of identifying a (true or good or causal or useful) model. Note that classification and regression can be regarded as special sequence prediction problems, where the sequence $x_1y_1...x_ny_nx_{n+1}$ of (x,y)-pairs is given and the class label or function value y_{n+1} shall be predicted.

Universal sequence prediction. The Bayesian framework leaves open how to choose the model class $\mathcal{M}$ and prior w_ν. General guidelines are that $\mathcal{M}$ should be small but large enough to contain the true environment μ, and w_ν should reflect one's prior (subjective) belief in ν or should be non-informative or neutral or objective if no prior knowledge is available. But these are informal and ambiguous considerations outside the formal Bayesian framework. Solomonoff's [Sol64] rigorous, essentially unique, formal, and universal solution to this problem is to consider a single large universal class $\mathcal{M}_U$ suitable for *all* induction problems. The corresponding universal prior w_ν^U is biased towards simple environments in such a way that it dominates=-superior to all other priors. This leads to an a priori probability $M(x)$ which is equivalent to the probability that a universal Turing machine with random input tape outputs x.

History and motivation. Many interesting, important, and deep results have been proven for Solomonoff's universal distribution M [ZL70, Sol78, LV97, Hut04]. The motivation and goal of this paper is to provide a broad discussion of how and in which sense universal sequence prediction solves all kinds of (philosophical) problems of Bayesian sequence prediction, and to present some recent results. Many arguments and ideas could be further developed. I hope that the exposition stimulates such a future, more detailed, investigation.

Contents. In Section 2 we review the excellent predictive performance of Bayesian sequence prediction for generic (non-i.i.d.) countable and continuous model classes.

Section 3 critically reviews the classical principles (indifference, symmetry, minimax) for obtaining objective priors, introduces the universal prior inspired by Occam's razor and quantified in terms of Kolmogorov complexity. In Section 4 (for i.i.d. $\mathcal{M}$) and Section 5 (for universal $\mathcal{M}_U$) we show various desirable properties of the universal prior and class (non-zero p(oste)rior, confirmation of universal hypotheses, reparametrization and regrouping invariance, no old-evidence and updating problem) in contrast to (most) classical continuous prior densities. Finally, we show that the universal mixture performs better than classical continuous mixtures, even in uncomputable environments. Section 6 contains critique and summary.

2 Bayesian Sequence Prediction

Notation. We use letters $t,n\in I\!N$ for natural numbers, and denote the cardinality of a set $\mathcal{S}$ by $\#\mathcal{S}$ or $|\mathcal{S}|$. We write $\mathcal{X}^*$ for the set of finite strings over some alphabet $\mathcal{X}$, and $\mathcal{X}^\infty$ for the set of infinite sequences. For a string $x\in\mathcal{X}^*$ of length $\ell(x)=n$ we write $x_1x_2...x_n$ with $x_t\in\mathcal{X}$, and further abbreviate $x_{t:n}:=x_tx_{t+1}...x_{n-1}x_n$ and $x_{<n}:=x_1...x_{n-1}$. We assume that sequence $\omega=\omega_{1:\infty}\in\mathcal{X}^\infty$ is sampled from the "true" probability measure μ, i.e. $\mu(x_{1:n}):=\mathrm{P}[\omega_{1:n}=x_{1:n}|\mu]$ is the μ-probability that ω starts with $x_{1:n}$. We denote expectations w.r.t. μ by $\mathbf{E}$. In particular for a function $f:\mathcal{X}^n\to I\!R$, we have $\mathbf{E}[f]=\mathbf{E}[f(\omega_{1:n})]=\sum_{x_{1:n}}\mu(x_{1:n})f(x_{1:n})$. If μ is unknown but known to belong to a countable class of environments=models=measures $\mathcal{M}=\{\nu_1,\nu_2,...\}$, and $\{H_\nu:\nu\in\mathcal{M}\}$ forms a mutually exclusive and complete class of hypotheses, and $w_\nu:=\mathrm{P}[H_\nu]$ is our prior belief in H_ν, then $\xi(x_{1:n}):=\mathrm{P}[\omega_{1:n}=x_{1:n}]=\sum_{\nu\in\mathcal{M}}\mathrm{P}[\omega_{1:n}=x_{1:n}|H_\nu]\mathrm{P}[H_\nu]$ must be our (prior) belief in $x_{1:n}$, and $w_\nu(x_{1:n}):=\mathrm{P}[H_\nu|\omega_{1:n}=x_{1:n}]=\frac{\mathrm{P}[\omega_{1:n}=x_{1:n}|H_\nu]\mathrm{P}[H_\nu]}{\mathrm{P}[\omega_{1:n}=x_{1:n}]}$ be our posterior belief in ν by Bayes' rule. For a sequence $a_1,a_2,...$ of random variables, $\sum_{t=1}^\infty\mathbf{E}[a_t^2]\leq c<\infty$ implies $a_t\stackrel{t\to\infty}{\longrightarrow}0$ with μ-probability 1 (w.p.1). Convergence is rapid in the sense that the probability that a_t^2 exceeds $\varepsilon>0$ at more than $\frac{c}{\varepsilon\delta}$ times t is bounded by δ. We sometimes loosely call this the number of errors.

Sequence prediction. Given a sequence $x_1x_2...x_{t-1}$, we want to predict its likely continuation x_t. We assume that the strings which have to be continued are drawn from a "true" probability distribution μ. The maximal prior information a prediction algorithm can possess is the exact knowledge of μ, but often the true distribution is unknown. Instead, prediction is based on a guess ρ of μ. While we require μ to be a measure, we allow ρ to be a semimeasure [LV97, Hut04]:[1] Formally, $\rho:\mathcal{X}^*\to[0,1]$ is a semimeasure if $\rho(x)\geq\sum_{a\in\mathcal{X}}\rho(xa)\,\forall x\in\mathcal{X}^*$, and a (probability) measure if equality holds and $\rho(\epsilon)=1$, where ϵ is the empty string. $\rho(x)$ denotes the ρ-probability that a sequence starts with string x. Further, $\rho(a|x):=\rho(xa)/\rho(x)$ is the "posterior" or "predictive" ρ-probability that the next symbol is $a\in\mathcal{X}$, given sequence $x\in\mathcal{X}^*$.

Bayes mixture. We may know or assume that μ belongs to some countable class $\mathcal{M}:=\{\nu_1,\nu_2,...\}\ni\mu$ of semimeasures. Then we can use the weighted average on $\mathcal{M}$

[1] Readers unfamiliar or uneasy with *semi*measures can without loss ignore this technicality.

(Bayes-mixture, data evidence, marginal)

$$\xi(x) := \sum_{\nu\in\mathcal{M}} w_\nu \cdot \nu(x), \quad \sum_{\nu\in\mathcal{M}} w_\nu \leq 1, \quad w_\nu > 0. \tag{1}$$

for prediction. The most important property of semimeasure ξ is its dominance

$$\xi(x) \;\geq\; w_\nu \nu(x) \quad \forall x \text{ and } \forall \nu\in\mathcal{M}, \quad \text{in particular} \quad \xi(x) \;\geq\; w_\mu \mu(x) \tag{2}$$

which is a strong form of absolute continuity.

Convergence for deterministic environments. In the predictive setting we are not interested in identifying the true environment, but to predict the next symbol well. Let us consider deterministic μ first. An environment is called deterministic if $\mu(\alpha_{1:n})=1\forall n$ for some sequence α, and $\mu=0$ elsewhere (off-sequence). In this case we identify μ with α and the following holds:

$$\textstyle\sum_{t=1}^\infty |1-\xi(\alpha_t|\alpha_{<t})| \;\leq\; \ln w_\alpha^{-1} \qquad \text{and} \qquad \xi(\alpha_{t:n}|\alpha_t) \to 1 \quad \text{for} \quad n \geq t \to \infty \tag{3}$$

where $w_\alpha > 0$ is the weight of $\alpha\widehat{=}\mu\in\mathcal{M}$. This shows that $\xi(\alpha_t|\alpha_{<t})$ rapidly converges to 1 and hence also $\xi(\bar\alpha_t|\alpha_{<t})\to 0$ for $\bar\alpha_t\neq\alpha_t$, and that ξ is also a good multi-step lookahead predictor. Proof: $\xi(\alpha_{1:n})\to c>0$, since $\xi(\alpha_{1:n})$ is monotone decreasing in n and $\xi(\alpha_{1:n})\geq w_\mu\mu(\alpha_{1:n})=w_\mu>0$. Hence $\xi(\alpha_{1:n})/\xi(\alpha_{1:t})\to c/c=1$ for any limit sequence $t,n\to\infty$. The bound follows from $\sum_{t=1}^n 1-\xi(x_t|x_{<t})\leq -\sum_{t=1}^n \ln\xi(x_t|x_{<t})= -\ln\xi(x_{1:n})$ and $\xi(\alpha_{1:n})\geq w_\alpha$.

Convergence in probabilistic environments. In the general probabilistic case we want to know how close $\boldsymbol{\xi}_t := \xi(\,\cdot_t|\omega_{<t}) \in I\!R^{|\mathcal{X}|}$ is to the true probability $\boldsymbol{\mu}_t := \mu(\,\cdot_t|\omega_{<t})$. One can show that

$$\textstyle\sum_{t=1}^n \mathbf{E}[s_t] \;\leq\; D_n(\mu||\xi) := \mathbf{E}[\ln \frac{\mu(\omega_{1:n})}{\xi(\omega_{1:n})}] \;\leq\; \ln w_\mu^{-1}, \tag{4}$$

where $s_t = s_t(\boldsymbol{\mu}_t,\boldsymbol{\xi}_t)$ can be the squared Euclidian or Hellinger or absolute or KL distance between $\boldsymbol{\mu}_t$ and $\boldsymbol{\xi}_t$, or the squared Bayes-regret [Hut04]. The first inequality actually holds for any two (semi)measures, and the last inequality follows from (2). These bounds (with $n=\infty$) imply

$$\xi(x_t|\omega_{<t}) - \mu(x_t|\omega_{<t}) \to 0 \text{ for any } x_t \text{ rapid w.p.1 for } t\to\infty.$$

One can also show multi-step lookahead convergence $\xi(x_{t:n_t}|\omega_{<t})-\mu(x_{t:n_t}|\omega_{<t})\to 0$, (even for unbounded horizon $1\leq n_t-t+1\to\infty$) which is interesting for delayed sequence prediction and in reactive environments [Hut04].

Continuous environmental classes. The bounds above remain approximately valid for most parametric model classes. Let $\mathcal{M} := \{\nu_\theta : \theta\in\Theta\subseteq I\!R^d\}$ be a family of probability distributions parameterized by a d-dimensional continuous parameter θ, and $\mu \equiv \nu_{\theta_0} \in \mathcal{M}$ the true generating distribution. For a continuous weight density[2] $w(\theta) > 0$ the sums (1) are naturally replaced by integrals:

$$\xi(x) \;:=\; \int_\Theta w(\theta)\cdot\nu_\theta(x)\,d\theta, \qquad \int_\Theta w(\theta)\,d\theta \;=\; 1 \tag{5}$$

The most important property of ξ was the dominance (2) achieved by dropping the sum over ν. The analogous construction here is to restrict the integral over θ to a small vicinity of θ_0. Since a continuous parameter can typically be estimated to accuracy $\propto n^{-1/2}$ after n observations, the largest volume in which ν_θ as a function of θ is approximately flat is $\propto (n^{-1/2})^d$, hence $\xi(x_{1:n}) \gtrsim n^{-d/2} w(\theta_0)\mu(x_{1:n})$. Under some weak regularity conditions one can prove [CB90, Hut04]

$$D_n(\mu||\xi) \;:=\; \mathbf{E}\ln\tfrac{\mu(\omega_{1:n})}{\xi(\omega_{1:n})} \;\;\leq\;\; \ln w(\theta_0)^{-1} + \tfrac{d}{2}\ln\tfrac{n}{2\pi} + \tfrac{1}{2}\ln\det\bar{\jmath}_n(\theta_0) + o(1) \tag{6}$$

where $w(\theta_0)$ is the weight density (5) of μ in ξ, and $o(1)$ tends to zero for $n \to \infty$, and the average Fisher information matrix $\bar{\jmath}_n(\theta) = -\frac{1}{n}\mathbf{E}[\nabla_\theta\nabla_\theta^T \ln\nu_\theta(\omega_{1:n})]$ measures the local smoothness of ν_θ and is bounded for many reasonable classes, including all stationary (k^{th}-order) finite-state Markov processes. We see that in the continuous case, D_n is no longer bounded by a constant, but grows very slowly (logarithmically) with n, which still implies that ε-deviations are exponentially seldom. Hence, (6) allows to bound (4) even in case of continuous $\mathcal{M}$.

3 How to Choose the Prior

Classical principles. The probability axioms (implying Bayes' rule) allow to compute posteriors and predictive distributions from prior ones, but are mute about how to choose the prior. Much has been written on the choice of non-informative-=neutral=objective priors (see [KW96] for a survey and references; in Section 6 we briefly discuss how to incorporate subjective prior knowledge). For finite $\mathcal{M}$, Laplace's *symmetry or indifference argument* which sets $w_\nu = \frac{1}{|\mathcal{M}|}\ \forall\nu \in \mathcal{M}$ is a reasonable principle. The analogue uniform density $w(\theta) = [\mathrm{Vol}(\Theta)]^{-1}$ for a compact measurable parameter space Θ is less convincing, since w becomes non-uniform under different parametrization (e.g. $\theta \rightsquigarrow \theta' := \sqrt{\theta}$). Jeffreys' solution is to find a symmetry group of the problem (like permutations for finite $\mathcal{M}$ or translations for $\Theta = I\!\!R$) and require the prior to be *invariant under group transformations*. Another solution is the *minimax approach* by Bernardo [CB90] which minimizes (the quite tight) bound (6) for the worst $\mu \in \mathcal{M}$. Choice $w(\theta) \propto \sqrt{\det\bar{\jmath}_n(\theta)}$ equalizes and hence minimizes (6). Problems are that there may be no obvious symmetry, the resulting prior can be improper, depend on which parameters are treated as nuisance parameters, on the model class, and on n. Other principles are *maximum entropy* and *conjugate priors*. The principles above, although not unproblematic, *can* provide good objective priors in many cases of small discrete or compact spaces, but we will

[2] $w()$ will always denote densities, and $w_{()}$ probabilities.

meet some more problems later. For "large" model classes we are interested in, i.e. countably infinite, non-compact, or non-parametric spaces, the principles typically do not apply or break down.

Occam's razor et al. Machine learning, the computer science branch of statistics, often deals with very large model classes. Naturally, machine learning has (re)discovered and exploited quite different principles for choosing priors, appropriate for this situation. The overarching principles put together by Solomonoff [Sol64] are: Occam's razor (choose the simplest model consistent with the data), Epicurus' principle of multiple explanations (keep all explanations consistent with the data), (Universal) Turing machines (to compute, quantify and assign codes to all quantities of interest), and Kolmogorov complexity (to define what simplicity/complexity means).

We will first "derive" the so called universal prior, and subsequently justify it by presenting various welcome theoretical properties and by examples. The idea is that a priori, i.e. before seeing the data, all models are "consistent," so a-priori Epicurus would regard all models (in $\mathcal{M}$) possible, i.e. choose $w_\nu > 0\ \forall\nu\in\mathcal{M}$. In order to also do (some) justice to Occam's razor we should *prefer* simple hypothesis, i.e. assign high prior/low prior w_ν to simple/complex hypotheses H_ν. Before we can define this prior, we need to quantify the notion of complexity.

Notation. A function $f:\mathcal{S}\to I\!\!R\cup\{\pm\infty\}$ is said to be lower semi-computable (or enumerable) if the set $\{(x,y) : y < f(x), x\in\mathcal{S}, y\in I\!\!\!Q\}$ is recursively enumerable. f is upper semi-computable (or co-enumerable) if $-f$ is enumerable. f is computable (or recursive) if f and $-f$ are enumerable. The set of (co)enumerable functions is recursively enumerable. We write $O(1)$ for a constant of reasonable size: 100 is reasonable, maybe even 2^{30}, but 2^{500} is not. We write $f(x)\stackrel{+}{\leq}g(x)$ for $f(x)\leq g(x)+O(1)$ and $f(x)\stackrel{\times}{\leq}g(x)$ for $f(x)\leq 2^{O(1)}\cdot g(x)$. Corresponding equalities hold if the inequalities hold in both directions.[3] We say that a property $A(n)\in\{true,false\}$ holds for *most* n, if $\#\{t\leq n:A(t)\}/n\stackrel{n\to\infty}{\longrightarrow}1$.

Kolmogorov complexity. We can now quantify the complexity of a string. Intuitively, a string is simple if it can be described in a few words, like "the string of one million ones", and is complex if there is no such short description, like for a random object whose shortest description is specifying it bit by bit. We are interested in effective descriptions, and hence restrict decoders to be Turing machines (TMs). Let us choose some universal (so-called prefix) *Turing machine* U with binary input=program tape, $\mathcal{X}$ary output tape, and bidirectional work tape. We can then define the *prefix Kolmogorov complexity* [LV97] of string x as the length ℓ of the shortest binary program p for which U outputs x:

$$K(x) \;:=\; \min_p\{\ell(p) : U(p) = x\}.$$

For non-string objects o (like numbers and functions) we define $K(o) := K(\langle o\rangle)$, where $\langle o\rangle\in\mathcal{X}^*$ is some standard code for o. In particular, if $(f_i)_{i=1}^\infty$ is an enumeration of all (co)enumerable functions, we define $K(f_i)=K(i)$.

[3] We will ignore this additive/multiplicative fudge in our discussion till Section 6.

An important property of K is that it is nearly independent of the choice of U. More precisely, if we switch from one universal TM to another, $K(x)$ changes at most by an additive constant independent of x. For reasonable universal TMs, the compiler constant is of reasonable size $O(1)$. A defining property of $K:\mathcal{X}^*\to I\!N$ is that it additively dominates all co-enumerable functions $f:\mathcal{X}^*\to I\!N$ that satisfy Kraft's inequality $\sum_x 2^{-f(x)} \leq 1$, i.e. $K(x)\stackrel{+}{\leq} f(x)$ for $K(f)=O(1)$. The universal TM provides a shorter prefix code than any other effective prefix code. K shares many properties with Shannon's entropy (information measure) S, but K is superior to S in many respects. To be brief, K is an excellent universal complexity measure, suitable for quantifying Occam's razor. We need the following properties of K:

$$\begin{array}{rl}
a) & K \text{ is not computable, but only upper semi-computable,} \\
b) & \text{the upper bound } K(n)\stackrel{+}{\leq}\log_2 n+2\log_2\log n, \\
c) & \text{Kraft's inequality } \sum_x 2^{-K(x)}\leq 1, \text{ which implies } 2^{-K(n)}\leq\frac{1}{n} \text{ for most } n, \\
d) & \text{information non-increase } K(f(x))\stackrel{+}{\leq}K(x)+K(f) \text{ for recursive } f:\mathcal{X}^*\to\mathcal{X}^*, \\
e) & K(x)\stackrel{+}{\leq}-\log_2 P(x)+K(P) \text{ if } P:\mathcal{X}^*\to[0,1] \text{ is enumerable and } \sum_x P(x)\leq 1, \\
f) & \sum_{x:f(x)=y}2^{-K(x)}\stackrel{\times}{=}2^{-K(y)} \text{ if } f \text{ is recursive and } K(f)=O(1).
\end{array} \tag{7}$$

Proofs of $(a)-(e)$ can be found in [LV97], and the (easy) proof of (f) in the extended version of this paper.

The universal prior. We can now quantify a prior biased towards simple models. First, we quantify the complexity of an environment ν or hypothesis H_ν by its Kolmogorov complexity $K(\nu)$. The universal prior should be a decreasing function in the model's complexity, and of course sum to (less than) one. Since K satisfies Kraft's inequality ($7c$), this suggests the following choice:

$$w_\nu \;=\; w_\nu^U \;:=\; 2^{-K(\nu)} \tag{8}$$

For this choice, the bound (4) on D_n reads

$$\textstyle\sum_{t=1}^\infty \mathbf{E}[s_t] \;\leq\; D_n \;\leq\; K(\mu)\ln 2 \tag{9}$$

i.e. the number of times, ξ deviates from μ by more than $\varepsilon>0$ is bounded by $O(K(\mu))$, i.e. is proportional to the complexity of the environment. Could other choices for w_ν lead to better bounds? The answer is essentially no [Hut04]: Consider any other reasonable prior w'_ν, where reasonable means (lower semi)computable with a program of size $O(1)$. Then, MDL bound ($7e$) with $P()\leadsto w'_{()}$ and $x\leadsto\langle\mu\rangle$ shows $K(\mu)\stackrel{+}{\leq}-\log_2 w'_\mu+K(w'_{()})$, hence $\ln w'^{-1}_\mu\stackrel{+}{\geq}K(\mu)\ln 2$ leads (within an additive constant) to a weaker bound. A counting argument also shows that $O(K(\mu))$ errors for most μ are unavoidable. So this choice of prior leads to very good prediction.

Even for continuous classes $\mathcal{M}$, we can assign a (proper) universal prior (not density) $w_\theta^U=2^{-K(\theta)}>0$ for computable θ, and 0 for uncomputable ones. This effectively reduces $\mathcal{M}$ to a discrete class $\{\nu_\theta\in\mathcal{M}:w_\theta^U>0\}$ which is typically dense in $\mathcal{M}$. We will see that this prior has many advantages over the classical prior densities.

4 Independent Identically Distributed Data

Laplace's rule for Bernoulli sequences. Let $x=x_1x_2...x_n\in\mathcal{X}^n=\{0,1\}^n$ be generated by a biased coin with head=1 probability $\theta\in[0,1]$, i.e. the likelihood of x under hypothesis H_θ is $\nu_\theta(x)=\mathrm{P}[x|H_\theta]=\theta^{n_1}(1-\theta)^{n_0}$, where $n_1=x_1+...+x_n=n-n_0$. Bayes assumed a uniform prior density $w(\theta)=1$. The evidence is $\xi(x)=\int_0^1\nu_\theta(x)w(\theta)\,d\theta=\frac{n_1!n_0!}{(n+1)!}$ and the posterior probability weight density $w(\theta|x)=\nu_\theta(x)w(\theta)/\xi(x)=\frac{(n+1)!}{n_1!n_0!}\theta^{n_1}(1-\theta)^{n_0}$ of θ after seeing x is strongly peaked around the frequency estimate $\hat{\theta}=\frac{n_1}{n}$ for large n. Laplace asked for the predictive probability $\xi(1|x)$ of observing $x_{n+1}=1$ after having seen $x=x_1...x_n$, which is $\xi(1|x)=\frac{\xi(x1)}{\xi(x)}=\frac{n_1+1}{n+2}$. (Laplace believed that the sun had risen for 5 000 years = 1 826 213 days since creation, so he concluded that the probability of doom, i.e. that the sun won't rise tomorrow is $\frac{1}{1826215}$.) This looks like a reasonable estimate, since it is close to the relative frequency, asymptotically consistent, symmetric, even defined for $n=0$, and not overconfident (never assigns probability 1).

The problem of zero prior. But also Laplace's rule is not without problems. The appropriateness of the uniform prior has been questioned in Section 3 and will be detailed below. Here we discuss a version of the zero prior problem. If the prior is zero, then the posterior is necessarily also zero. The above example seems unproblematic, since the prior and posterior *densities* $w(\theta)$ and $w(\theta|x)$ are non-zero. Nevertheless it is problematic e.g. in the context of scientific confirmation theory [Ear93].

Consider the hypothesis H that all balls in some urn, or all ravens, are black (=1). A natural model is to assume that balls/ravens are drawn randomly from an infinite population with fraction θ of black balls/ravens and to assume a uniform prior over θ, i.e. just the Bayes-Laplace model. Now we draw n objects and observe that they are all black.

We may formalize H as the hypothesis $H':=\{\theta=1\}$. Although the posterior probability of the relaxed hypothesis $H_\varepsilon:=\{\theta\geq 1-\varepsilon\}$, $\mathrm{P}[H_\varepsilon|1^n]=\int_{1-\varepsilon}^1 w(\theta|1^n)\,d\theta=\int_{1-\varepsilon}^1(n+1)\theta^n d\theta=1-(1-\varepsilon)^{n+1}$ tends to 1 for $n\to\infty$ for every fixed $\varepsilon>0$, $\mathrm{P}[H'|1^n]=\mathrm{P}[H_0|1^n]$ remains identically zero, i.e. no amount of evidence can confirm H'. The reason is simply that zero prior $\mathrm{P}[H']=0$ implies zero posterior.

Note that H' refers to the unobservable quantity θ and only demands blackness with probability 1. So maybe a better formalization of H is purely in terms of observational quantities: $H'':=\{\omega_{1:\infty}=1^\infty\}$. Since $\xi(1^n)=\frac{1}{n+1}$, the predictive probability of observing k further black objects is $\xi(1^k|1^n)=\frac{\xi(1^{n+k})}{\xi(1^n)}=\frac{n+1}{n+k+1}$. While for fixed k this tends to 1, $\mathrm{P}[H''|1^n]=\lim_{k\to\infty}\xi(1^k|1^n)\equiv 0\ \forall n$, as for H'.

One may speculate that the crux is the infinite population. But for a finite population of size N and sampling with (similarly without) repetition, $\mathrm{P}[H''|1^n]=\xi(1^{N-n}|1^n)=\frac{n+1}{N+1}$ is close to one only if a large fraction of objects has been observed. This contradicts scientific practice: Although only a tiny fraction of all existing ravens have been observed, we regard this as sufficient evidence for believing strongly in H.

There are two solutions of this problem: We may abandon strict/logical/all-quantified/universal hypotheses altogether in favor of soft hypotheses like H_ε. Although not unreasonable, this approach is unattractive for several reasons. The other solution is to assign a non-zero prior to $\theta=1$. Consider, for instance, the improper density $w(\theta)=\frac{1}{2}[1+\delta(1-\theta)]$, where δ is the Dirac-delta ($\int f(\theta)\delta(\theta-a)\,d\theta=f(a)$), or equivalently $\mathrm{P}[\theta\geq a]=1-\frac{1}{2}a$. We get $\xi(x_{1:n})=\frac{1}{2}[\frac{n_1!n_0!}{(n+1)!}+\delta_{0n_0}]$, where $\delta_{ij}=\{{1 \text{ if } i=j \atop 0 \text{ else}}\}$ is Kronecker's δ. In particular $\xi(1^n)=\frac{1}{2}\frac{n+2}{n+1}$ is much larger than for uniform prior. Since $\xi(1^k|1^n)=\frac{n+k+2}{n+k+1}\cdot\frac{n+1}{n+2}$, we get $\mathrm{P}[H''|1^n]=\lim_{k\to\infty}\xi(1^k|1^n)=\frac{n+1}{n+2}\to 1$, i.e. H'' gets strongly confirmed by observing a reasonable number of black objects. This correct asymptotics also follows from the general result (3). Confirmation of H'' is also reflected in the fact that $\xi(0|1^n)=\frac{1}{(n+2)^2}$ tends much faster to zero than for uniform prior, i.e. the confidence that the next object is black is much higher. The power actually depends on the shape of $w(\theta)$ around $\theta=1$. Similarly H' gets confirmed: $\mathrm{P}[H'|1^n]=\mu_1(1^n)\mathrm{P}[\theta=1]/\xi(1^n)=\frac{n+1}{n+2}\to 1$. On the other hand, if a single (or more) 0 are observed ($n_0>0$), then the predictive distribution $\xi(\cdot|x)$ and posterior $w(\theta|x)$ are the same as for uniform prior.

The findings above remain qualitatively valid for i.i.d. processes over finite non-binary alphabet $|\mathcal{X}|>2$ and for non-uniform prior.

Surely to get a generally working setup, we should also assign a non-zero prior to $\theta=0$ and to all other "special" θ, like $\frac{1}{2}$ and $\frac{1}{6}$, which may naturally appear in a hypothesis, like "is the coin or die fair". The natural continuation of this thought is to assign non-zero prior to all computable θ. This is another motivation for the universal prior $w_\theta^U=2^{-K(\theta)}$ (8) constructed in Section 3. It is difficult but not impossible to operate with such a prior [PH04]. One may want to mix the discrete prior w_ν^U with a continuous (e.g. uniform) prior density, so that the set of non-computable θ keeps a non-zero density. Although possible, we will see that this is actually not necessary.

Reparametrization invariance. Naively, the uniform prior is justified by the indifference principle, but as discussed in Section 3, uniformity is not reparametrization invariant. For instance if in our Bernoulli example we introduce a new parametrization $\theta'=\sqrt{\theta}$, then the θ'-density $w'(\theta')=2\sqrt{\theta}w(\theta)$ is no longer uniform if $w(\theta)=1$ is uniform.

More generally, assume we have some principle which leads to some prior $w(\theta)$. Now we apply the principle to a different parametrization $\theta'\in\Theta'$ and get prior $w'(\theta')$. Assume that θ and θ' are related via bijection $\theta=f(\theta')$. Another way to get a θ'-prior is to transform the θ-prior $w(\theta)\rightsquigarrow\tilde{w}(\theta')$. The reparametrization invariance principle (RIP) states that w' should be equal to $\tilde{w}$.

For discrete Θ, simply $\tilde{w}_{\theta'}=w_{f(\theta')}$, and a uniform prior remains uniform ($w'_{\theta'}=\tilde{w}_{\theta'}=w_\theta=\frac{1}{|\Theta|}$) in any parametrization, i.e. the indifference principle satisfies RIP in finite model classes.

In case of densities, we have $\tilde{w}(\theta')=w(f(\theta'))\frac{df(\theta')}{d\theta'}$, and the indifference principle violates RIP for non-linear transformations f. But Jeffrey's and Bernardo's principle satisfy RIP. For instance, in the Bernoulli case we have $\bar{\jmath}_n(\theta)=\frac{1}{\theta}+\frac{1}{1-\theta}$, hence $w(\theta)=$

$\frac{1}{\pi}[\theta(1-\theta)]^{-1/2}$ and $w'(\theta')=\frac{1}{\pi}[f(\theta')(1-f(\theta'))]^{-1/2}\frac{df(\theta')}{d\theta'}=\tilde w(\theta')$.

Does the universal prior $w_\theta^U=2^{-K(\theta)}$ satisfy RIP? If we apply the "universality principle" to a θ'-parametrization, we get $w'^U_{\theta'}=2^{-K(\theta')}$. On the other hand, w_θ simply transforms to $\tilde w^U_{\theta'}=w^U_{f(\theta')}=2^{-K(f(\theta'))}$ (w_θ is a discrete (non-density) prior, which is non-zero on a discrete subset of $\mathcal{M}$). For computable f we have $K(f(\theta'))\stackrel{+}{\leq}K(\theta')+K(f)$ by (7d), and similarly $K(f^{-1}(\theta))\stackrel{+}{\leq}K(\theta)+K(f)$ if f is invertible. Hence for simple bijections f i.e. for $K(f)=O(1)$, we have $K(f(\theta'))\stackrel{+}{=}K(\theta')$, which implies $w'^U_{\theta'}\stackrel{\times}{=}\tilde w^U_{\theta'}$, i.e. *the universal prior satisfies RIP* w.r.t. simple transformations f (within a multiplicative constant).

Regrouping invariance. There are important transformations f which are *not* bijections, which we consider in the following. A simple non-bijection is $\theta=f(\theta')=\theta'^2$ if we consider $\theta'\in[-1,1]$. More interesting is the following example: Assume we had decided not to record blackness versus non-blackness of objects, but their "color". For simplicity of exposition assume we record only whether an object is black or white or colored, i.e. $\mathcal{X}'=\{B,W,C\}$. In analogy to the binary case we use the indifference principle to assign a uniform prior on $\boldsymbol{\theta'}\in\Theta':=\Delta_3$, where $\Delta_d:=\{\boldsymbol{\theta'}\in[0,1]^d:\sum_{i=1}^d\theta'_i=1\}$, and $\nu_{\theta'}(x'_{1:n})=\prod_i\theta'^{n_i}_i$. All inferences regarding blackness (predictive and posterior) are identical to the binomial model $\nu_\theta(x_{1:n})=\theta^{n_1}(1-\theta)^{n_0}$ with $x'_t=B\rightsquigarrow x_t=1$ and $x'_t=W$ or $C\rightsquigarrow x_t=0$ and $\theta=f(\boldsymbol{\theta'})=\theta'_B$ and $w(\theta)=\int_{\Delta_3}w'(\boldsymbol{\theta'})\delta(\theta'_B-\theta)d\boldsymbol{\theta'}$. Unfortunately, for uniform prior $w'(\theta')\propto 1$, $w(\theta)\propto 1-\theta$ is *not* uniform, i.e. the indifference principle is *not* invariant under splitting/grouping, or general regrouping. Regrouping invariance is regarded as a very important and desirable property [Wal96].

We now consider general i.i.d. processes $\nu_\theta(x)=\prod_{i=1}^d\theta_i^{n_i}$. Dirichlet priors $w(\theta)\propto\prod_{i=1}^d\theta_i^{\alpha_i-1}$ form a natural conjugate class ($w(\theta|x)\propto\prod_{i=1}^d\theta_i^{n_i+\alpha_i-1}$) and are the default priors for multinomial (i.i.d.) processes over finite alphabet $\mathcal{X}$ of size d. Note that $\xi(a|x)=\frac{n_a+\alpha_a}{n+\alpha_1+...+\alpha_d}$ generalizes Laplace's rule and coincides with Carnap's [Ear93] confirmation function. Symmetry demands $\alpha_1=...=\alpha_d$; for instance $\alpha\equiv 1$ for uniform and $\alpha\equiv\frac{1}{2}$ for Bernard-Jeffrey's prior. Grouping two "colors" i and j results in a Dirichlet prior with $\alpha_{i\&j}=\alpha_i+\alpha_j$ for the group. The only way to respect symmetry under all possible groupings is to set $\alpha\equiv 0$. This is Haldane's improper prior, which results in unacceptably overconfident predictions $\xi(1|1^n)=1$. Walley [Wal96] solves the problem that there is no single acceptable prior density by considering sets of priors.

We now show that the universal prior $w_\theta^U=2^{-K(\theta)}$ is invariant under regrouping, and more generally under all simple (computable with complexity O(1)) even non-bijective transformations. Consider prior $w'_{\theta'}$. If $\theta=f(\theta')$ then $w'_{\theta'}$ transforms to $\tilde w_\theta=\sum_{\theta':f(\theta')=\theta}w'_{\theta'}$ (note that for non-bijections there is more than one $w'_{\theta'}$ consistent with $\tilde w_\theta$). In θ'-parametrization, the universal prior reads $w'^U_{\theta'}=2^{-K(\theta')}$. Using (7$f$) with $x=\langle\theta'\rangle$ and $y=\langle\theta\rangle$ we get $\tilde w^U_\theta=\sum_{\theta':f(\theta')=\theta}2^{-K(\theta')}\stackrel{\times}{=}2^{-K(\theta)}=w^U_\theta$, i.e. *the universal prior is general transformation and hence regrouping invariant* (within a multiplicative constant) w.r.t. simple computable transformations f.

Note that reparametrization and regrouping invariance hold for arbitrary classes

$\mathcal{M}$ and are not limited to the i.i.d. case.

5 Universal Sequence Prediction

Universal choice of $\mathcal{M}$. The bounds of Section 2 apply if $\mathcal{M}$ contains the true environment μ. The larger $\mathcal{M}$ the less restrictive is this assumption. The class of all computable distributions, although only countable, is pretty large from a practical point of view, since it includes for instance all of today's valid physics theories. It is the largest class, relevant from a computational point of view. Solomonoff [Sol64, Eq.(13)] defined and studied the mixture over this class.

One problem is that this class is not enumerable, since the class of computable functions $f:\mathcal{X}^*\to I\!R$ is not enumerable (halting problem), nor is it decidable whether a function is a measure. Hence ξ is completely incomputable. Levin [ZL70] had the idea to "slightly" extend the class and include also lower semi-computable semimeasures. One can show that this class $\mathcal{M}_U=\{\nu_1,\nu_2,...\}$ is enumerable, hence

$$\xi_U(x) \;=\; \sum_{\nu\in\mathcal{M}_U} w_\nu^U \nu(x) \tag{10}$$

is itself lower semi-computable, i.e. $\xi_U\in\mathcal{M}_U$, which is a convenient property in itself. Note that since $\frac{1}{n\log^2 n}\stackrel{\times}{\leq} w_{\nu_n}^U\leq\frac{1}{n}$ for most n by (7b) and (7c), most ν_n have prior approximately reciprocal to their index n.

In some sense $\mathcal{M}_U$ is the largest class of environments for which ξ is in some sense computable [Hut04], but see [Sch02] for even larger classes.

The problem of old evidence. An important problem in Bayesian inference in general and (Bayesian) confirmation theory [Ear93] in particular is how to deal with 'old evidence' or equivalently with 'new theories'. How shall a Bayesian treat the case when some evidence $E\widehat{=}x$ (e.g. Mercury's perihelion advance) is known well-before the correct hypothesis/theory/model $H\widehat{=}\mu$ (Einstein's general relativity theory) is found? How shall H be added to the Bayesian machinery a posteriori? What is the prior of H? Should it be the belief in H in a hypothetical counterfactual world in which E is not known? Can old evidence E confirm H? After all, H could simply be constructed/biased/fitted towards "explaining" E.

The universal class $\mathcal{M}_U$ and universal prior w_ν^U formally solve this problem: The universal prior of H is $2^{-K(H)}$. This is independent of $\mathcal{M}$ and of whether E is known or not. If we use E to construct H or fit H to explain E, this will lead to a theory which is more complex ($K(H)\stackrel{+}{\geq}K(E)$) than a theory from scratch ($K(H)=O(1)$), so cheats are automatically penalized. There is no problem of adding hypotheses to $\mathcal{M}$ a posteriori. Priors of old hypotheses are not affected. Finally, $\mathcal{M}_U$ includes *all* hypothesis (including yet unknown or unnamed ones) a priori. So at least theoretically, updating $\mathcal{M}$ is unnecessary.

Other representations of ξ_U. There is a much more elegant representation of ξ_U: Solomonoff [Sol64, Eq.(7)] defined the *universal prior* $M(x)$ as the probability that the output of a universal Turing machine U starts with x when provided with fair coin flips on the input tape. Note that a uniform distribution is also used in the so-called No-Free-Lunch theorems to prove the impossibility of universal learners, but in our case the uniform distribution is piped through a universal Turing machine which defeats these negative implications. Formally, M can be defined as

$$M(x) \;:=\; \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)} \;\stackrel{\times}{=}\; \xi_U(x) \tag{11}$$

where the sum is over all (so-called minimal) programs p for which U outputs a string starting with x. M may be regarded as a $2^{-\ell(p)}$-weighted mixture over all computable deterministic environments ν_p ($\nu_p(x)=1$ if $U(p)=x*$ and 0 else). Now, as a positive surprise, $M(x)$ *coincides with* $\xi_U(x)$ within an irrelevant multiplicative constant. So it is actually sufficient to consider the class of *deterministic* semimeasures. The reason is that the probabilistic semimeasures are in the convex hull of the deterministic ones, and so need not be taken extra into account in the mixture.

Bounds for computable environments. The bound (9) surely is applicable for $\xi=\xi_U$ and now holds for *any* computable measure μ. Within an additive constant the bound is also valid for $M\stackrel{\times}{=}\xi$. That is, ξ_U *and* M *are excellent predictors with the only condition that the sequence is drawn from any computable probability distribution.* Bound (9) shows that the total number of prediction errors is small. Similarly to (3) one can show that $\sum_{t=1}^n|1-M(x_t|x_{<t})|\leq Km(x_{1:n})\ln 2$, where the monotone complexity $Km(x):=\min\{\ell(p):U(p)=x*\}$ is defined as the length of the shortest (nonhalting) program computing a string starting with x [ZL70, LV97, Hut04].

If $x_{1:\infty}$ is a computable sequence, then $Km(x_{1:\infty})$ is finite, which implies $M(x_t|x_{<t})\to 1$ *on every computable sequence.* This means that if the environment is a computable sequence (whichsoever, e.g. 1^∞ or the digits of π or e), after having seen the first few digits, M correctly predicts the next digit with high probability, i.e. it recognizes the structure of the sequence. In particular, observing an increasing number of black balls or black ravens or sunrises, $M(1|1^n)\to 1$ ($Km(1^\infty)=O(1)$) becomes rapidly confident that future balls and ravens are black and that the sun will rise tomorrow.

Universal is better than continuous $\boldsymbol{\mathcal{M}}$. Although we argued that incomputable environments μ can safely be ignored, one may be nevertheless uneasy using Solomonoff's $M\stackrel{\times}{=}\xi_U$ (11) if outperformed by a continuous mixture ξ (5) on such $\mu\in\mathcal{M}\setminus\mathcal{M}_U$, for instance if M would fail to predict a Bernoulli(θ) sequence for incomputable θ. Luckily this is not the case: Although $\nu_\theta()$ and w_θ can be incomputable, the studied classes $\mathcal{M}$ themselves, i.e. the two-argument function $\nu_{()}()$, and the weight function $w_{()}$, and hence $\xi()$, are typically computable (the integral can be approximated to arbitrary precision). Hence $M(x)\stackrel{\times}{=}\xi_U(x)\geq 2^{-K(\xi)}\xi(x)$ by (10) and $K(\xi)$ is often quite small. This implies for *all* μ

$$D_n(\mu||M) \;\equiv\; \mathbf{E}[\ln\tfrac{\mu(\omega_{1:n})}{M(\omega_{1:n})}] \;=\; \mathbf{E}[\ln\tfrac{\mu(\omega_{1:n})}{\xi(\omega_{1:n})}]+\mathbf{E}[\ln\tfrac{\xi(\omega_{1:n})}{M(\omega_{1:n})}] \;\stackrel{+}{\leq}\; D_n(\mu||\xi)+K(\xi)\ln 2$$

So any bound (6) for $D_n(\mu||\xi)$ is directly valid also for $D_n(\mu||M)$, save an additive constant. That is, M is superior (or equal) to all computable mixture predictors ξ based on any (continuous or discrete) model class $\mathcal{M}$ and weight $w(\theta)$, even if environment μ is *not* computable. Furthermore, while for essentially all parametric classes, $D_n(\mu||\xi)\sim\frac{d}{2}\ln n$ grows logarithmically in n for all (incl. computable) $\mu\in\mathcal{M}$, $D_n(\mu||M)\leq K(\mu)\ln 2$ is finite for computable μ. Bernardo's prior even implies a bound for M that is uniform (minimax) in $\theta\in\Theta$. Many other priors based on reasonable principles (see Section 3 and [KW96]) and many other computable probabilistic predictors ρ are argued for. The above actually shows that M is superior to all of them.

6 Discussion

Critique and problems. In practice we often have extra information about the problem at hand, which could and should be used to guide the forecasting. One way is to explicate all our prior knowledge y and place it on an extra input tape of our universal Turing machine U, which leads to the conditional complexity $K(\cdot|y)$. We now assign "subjective" prior $w_{\nu|y}^U=2^{-K(\nu|y)}$ to environment ν, which is large for those ν that are simple (have short description) relative to our background knowledge y. Since $K(\mu|y)\stackrel{+}{\leq}K(\mu)$, extra knowledge never misguides (see (9)). Alternatively we could prefix our observation sequence x by y and use $M(yx)$ for prediction [Hut04].

Another critique concerns the dependence of K and M on U. Predictions for short sequences x (shorter than typical compiler lengths) can be arbitrary. But taking into account our (whole) scientific prior knowledge y, and predicting the now long string yx leads to good (less sensitive to "reasonable" U) predictions [Hut04].

Finally, K and M can serve as "gold standards" which practitioners should aim at, but since they are only semi-computable, they have to be (crudely) approximated in practice. Levin complexity [LV97], Schmidhuber's speed prior, the minimal message and description length principles [Wal05], and off-the-shelf compressors like Lempel-Ziv are such approximations, which have been successfully applied to a plethora of problems [CV05, Sch04].

Summary. We compared traditional Bayesian sequence prediction based on continuous classes and prior densities to Solomonoff's universal predictor M, prior w_ν^U, and class $\mathcal{M}_U$. We discussed: Convergence for generic class and prior, the relative entropy bound for continuous classes, indifference/symmetry principles, the problem of zero p(oste)rior and confirmation of universal hypotheses, reparametrization and regrouping invariance, the problem of old evidence and updating, that M works even in non-computable environments, how to incorporate prior knowledge, the prediction of short sequences, the constant fudges in all results and the U-dependence, M's incomputability and crude but practical approximations. In short, universal prediction solves or avoids or meliorates many foundational and philosophical problems, but has to be compromised in practice.

References

[CB90] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36:453–471, 1990.

[CV05] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression. *IEEE Trans. Information Theory*, 51(4):1523–1545, 2005.

[Ear93] J. Earman. *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* MIT Press, Cambridge, MA, 1993.

[Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer, Berlin, 2004. 300 pages, http://www.idsia.ch/~marcus/ai/uaibook.htm.

[KW96] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[LV97] M. Li and P. M. B. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications.* Springer, Berlin, 2nd edition, 1997.

[PH04] J. Poland and M. Hutter. On the convergence speed of MDL predictions for Bernoulli sequences. In *Proc. 15th International Conf. on Algorithmic Learning Theory (ALT'04)*, volume 3244 of *LNAI*, pages 294–308, Padova, 2004. Springer, Berlin.

[Sch02] J. Schmidhuber. Hierarchies of generalized Kolmogorov complexities and nonenumerable universal measures computable in the limit. *International Journal of Foundations of Computer Science*, 13(4):587–612, 2002.

[Sch04] J. Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54(3):211–254, 2004.

[Sol64] R. J. Solomonoff. A formal theory of inductive inference: Parts 1 and 2. *Information and Control*, 7:1–22 and 224–254, 1964.

[Sol78] R. J. Solomonoff. Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory*, IT-24:422–432, 1978.

[Wal96] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society B*, 58(1):3–57, 1996.

[Wal05] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length.* Springer, Berlin, 2005.

[ZL70] A. K. Zvonkin and L. A. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russian Mathematical Surveys*, 25(6):83–124, 1970.