On Universal Prediction and Bayesian Confirmation

Marcus Hutter

RSISE @ ANU and SML @ NICTA

Canberra, ACT, 0200, Australia

marcus@hutter1.net www.hutter1.net

May 23, 2007

Abstract

The Bayesian framework is a well-studied and successful framework for inductive reasoning, which includes hypothesis testing and confirmation, parameter estimation, sequence prediction, classification, and regression. But standard statistical guidelines for choosing the model class and prior are not always available or can fail, in particular in complex situations. Solomonoff completed the Bayesian framework by providing a rigorous, unique, formal, and universal choice for the model class and the prior. I discuss in breadth how and in which sense universal (non-i.i.d.) sequence prediction solves various (philosophical) problems of traditional Bayesian sequence prediction. I show that Solomonoff's model possesses many desirable properties: Strong total and future bounds, and weak instantaneous 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, Black raven paradox, reparametrization invariance, old-evidence/updating problem, (non)computable environments.

1 Introduction

"... in spite of it's incomputability, Algorithmic Probability can serve as a kind of 'Gold Standard' for induction systems"

— Ray Solomonoff (1997)

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 induction 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 I am interested in is a formal general theory for prediction.

The Bayesian framework is the most consistent and successful framework developed thus far [Ear93, Jay03]. 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 prequential [Daw84] or transductive [Vap99, Sec.9.1] 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. With the exception of Section 4, this paper concentrates on *prediction* rather than model identification. The ultimate goal is to make "good" predictions in the sense of maximizing one's profit or minimizing one's loss. 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.

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, and the shortest program computing x produces the most likely continuation (prediction) of x.

Many interesting, important, and deep results have been proven for Solomonoff's universal distribution M [ZL70, Sol78, Gác83, LV97, Hut01, 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.

In Section 2, I review the excellent predictive and decision-theoretic performance results 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 $\mathcal{M}_U$) I 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. I also complement the general total bounds of Section 2 with some universal and some i.i.d.-specific instantaneous and future bounds. Finally, I show that the universal mixture performs better than classical continuous mixtures, even in uncomputable environments. Section 6 contains critique, summary, and conclusions.

The reparametrization and regrouping invariance, the (weak) instantaneous bounds, the good performance of M in non-computable environments, and most of the discussion (zero prior and universal hypotheses, old evidence) are new or new in the light of universal sequence prediction. Technical and mathematical non-trivial new results are the Hellinger-like loss bound (8) and the instantaneous bounds (14) and (17).

2 Bayesian Sequence Prediction

I now formally introduce the Bayesian sequence prediction setup and describe the most important results. I consider sequences over a finite alphabet, assume that the true environment is unknown but known to belong to a countable or continuous class of environments (no i.i.d. or Markov or stationarity assumption), and consider general prior. I show that the predictive distribution converges rapidly to the true sampling distribution and that the Bayes-optimal predictor performs excellent for any bounded loss function.

Notation. I use letters $t,n \in \mathbb{N}$ for natural numbers, and denote the cardinality of a set $\mathcal{S}$ by $\#\mathcal{S}$ or $|\mathcal{S}|$. I 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 $\ell(x)=n$ I write $x_1x_2...x_n$ with $x_t \in \mathcal{X}$, and further abbreviate $x_{t:n} := x_t x_{t+1}...x_{n-1}x_n$ and $x_{<n} := x_1...x_{n-1}$.

I 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 $x_{1:n}$. I denote expectations w.r.t. μ by $\mathbf{E}$. In particular for a function $f:\mathcal{X}^n \to \mathbb{R}$, we have $\mathbf{E}[f] = \mathbf{E}[f(\omega_{1:n})] = \sum_{x_{1:n}} \mu(x_{1:n}) f(x_{1:n})$. Note that in Bayesian learning, measures, environments, and models coincide, and are the same objects; let $\mathcal{M} = \{\nu_1,\nu_2,...\}$

denote a countable class of these measures. Assume that (a) μ is unknown but known to be a member of $\mathcal{M}$, (b) $\{H_\nu : \nu \in \mathcal{M}\}$ forms a mutually exclusive and complete class of hypotheses, and (c) $w_\nu := \mathrm{P}[H_\nu]$ is the given 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. I 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. I 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 I require μ to be a measure, I 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 $\mathcal{M}$ (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. One may interpret $w_\nu = \mathrm{P}[H_\nu]$ as prior belief in ν and $\xi(x) = \mathrm{P}[x]$ as the subjective probability of x, and $\mu(x) = \mathrm{P}[x|\mu]$ is the sampling distribution or likelihood. 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:

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

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

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 $\xi(x_t|x_{<t})$ is to the true probability $\mu(x_t|x_{<t})$. One convenient distance measure is the (squared) Hellinger distance

$$h_t(\omega_{<t}) \;:=\; \sum_{a\in\mathcal{X}}(\sqrt{\xi(a|\omega_{<t})}-\sqrt{\mu(a|\omega_{<t})})^2 \tag{4}$$

One can show [Hut03a, Hut04] that

$$\sum_{t=1}^{n}\mathbf{E}\Big[\Big(\sqrt{\tfrac{\xi(\omega_t|\omega_{<t})}{\mu(\omega_t|\omega_{<t})}}-1\Big)^2\Big] \;\leq\; \sum_{t=1}^{n}\mathbf{E}[h_t] \;\leq\; D_n(\mu||\xi):=\mathbf{E}[\ln\tfrac{\mu(\omega_{1:n})}{\xi(\omega_{1:n})}] \;\leq\; \ln w_\mu^{-1} \tag{5}$$

The first two inequalities actually hold for any two (semi)measures, and the last inequality follows from (2). These bounds (with $n=\infty$) imply $s_t\to\infty$ and hence

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

An improved bound $\mathbf{E}[\exp(\frac{1}{2}\sum_t h_t)]\leq w_\mu^{-1/2}$ [HM04] even shows that the probability that $\sum_t h_t$ additively exceeds $\ln w_\mu^{-1}$ by c (e.g. $c>10$) is tiny $\mathrm{e}^{-c/2}$. 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]. Since ξ rapidly converges to μ, one can anticipate that also decisions based on ξ are good.

Bayesian decisions. Let $\ell_{x_t y_t}\in[0,1]$ be the received loss when predicting $y_t\in\mathcal{Y}$, but $x_t\in\mathcal{X}$ turns out to be the true t^{th} symbol of the sequence. The ρ-optimal predictor

$$y_t^{\Lambda_\rho}(\omega_{<t}) \;:=\; \arg\min_{y_t}\sum_{x_t}\rho(x_t|\omega_{<t})\ell_{x_t y_t} \tag{6}$$

minimizes the ρ-expected loss. For instance for $\mathcal{X}=\mathcal{Y}=\{0,1\}$, Λ_ρ is a threshold strategy with $y_t^{\Lambda_\rho}=0/1$ for $\rho(1|\omega_{<t})\gtrless\gamma$, where $\gamma:=\frac{\ell_{01}-\ell_{00}}{\ell_{01}-\ell_{00}+\ell_{10}-\ell_{11}}$. The instantaneous loss at time t and the total μ(=true)-expected loss for the first n symbols are

$$l_t^{\Lambda_\rho}(\omega_{<t}) \;:=\; \mathbf{E}[\ell_{\omega_t y_t^{\Lambda_\rho}}|\omega_{<t}] \qquad\text{and}\qquad L_n^{\Lambda_\rho} \;:=\; \sum_{t=1}^{n}\mathbf{E}[\ell_{\omega_t y_t^{\Lambda_\rho}}] \tag{7}$$

Let Λ be *any* prediction scheme (deterministic or probabilistic) with no constraint at all, taking *any* action $y_t^\Lambda\in\mathcal{Y}$ with total expected loss L_n^Λ. If μ is known, Λ_μ is

obviously the best prediction scheme in the sense of achieving minimal expected loss $L_n^{\Lambda_\mu} \leq L_n^{\Lambda}$ for any Λ. For the predictor Λ_ξ based on the Bayes mixture ξ, one can show (proof in Appendix A; see also [MF98, Hut03a] for related bounds)

$$(\sqrt{L_n^{\Lambda_\xi}} - \sqrt{L_n^{\Lambda_\mu}})^2 \;\leq\; \sum_{t=1}^{n} \mathbf{E}[(\sqrt{l_t^{\Lambda_\xi}} - \sqrt{l_t^{\Lambda_\mu}})^2] \;\leq\; \sum_{t=1}^{n} 2\mathbf{E}[h_t] \tag{8}$$

which actually holds for any two (semi)measures. Chaining with (5) implies, for instance, $l_t^{\Lambda_\xi} \to l_t^{\Lambda_\mu}$ rapid w.p.1, $\sqrt{L_n^{\Lambda_\xi}}$ exceeds $\sqrt{L_n^{\Lambda_\mu}}$ by at most $\sqrt{2\ln w_\mu^{-1}}$, $L_n^{\Lambda_\xi}/L_n^{\Lambda_\mu} \to 1$ for $L_n^{\Lambda_\mu} \to \infty$, or if $L_\infty^{\Lambda_\mu}$ is finite, then also $L_\infty^{\Lambda_\xi}$. This shows that ξ (via Λ_ξ) performs also excellent from a decision-theoretic perspective, i.e. suffers loss only slightly larger than the optimal Λ_μ predictor.

One can also show that Λ_ξ *is Pareto-optimal* (admissible) in the sense that every other predictor with smaller loss than Λ_ξ in some environment $\nu \in \mathcal{M}$ must be worse in another environment [Hut03c].

Continuous environmental classes. I will argue later that countable $\mathcal{M}$ are sufficiently large from a philosophical and computational perspective. On the other hand, countable $\mathcal{M}$ exclude all continuously parameterized families (like the class of all i.i.d. or Markov processes), common in statistical practice. I show that the bounds above remain approximately valid for most parametric model classes. Let

$$\mathcal{M} \;:=\; \{\nu_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$$

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:

$$\xi(x_{1:n}) \;:=\; \int_\Theta w(\theta) \cdot \nu_\theta(x_{1:n})\, d\theta, \qquad \int_\Theta w(\theta)\, d\theta \;=\; 1 \tag{9}$$

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, Hut03c]

$$D_n(\mu||\xi) \;:=\; \mathbf{E} \ln \tfrac{\mu(\omega_{1:n})}{\xi(\omega_{1:n})} \;\leq\; \ln w(\theta_0)^{-1} + \tfrac{d}{2} \ln \tfrac{n}{2\pi} + \tfrac{1}{2} \ln \det \bar{\jmath}_n(\theta_0) + o(1) \tag{10}$$

where $w(\theta_0)$ is the weight density (9) 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. See Section 4 for an application

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

to the i.i.d. ($k=0$) case. 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, (10) allows to bound (5) and (8) even in case of continuous $\mathcal{M}$.

3 How to Choose the Prior

I showed in the last section how to predict if the true environment μ is unknown, but known to belong some class $\mathcal{M}$ of environments. In this section, I assume $\mathcal{M}$ to be given, and discuss how to (universally) choose the prior w_ν. After reviewing various classical principles (indifference, symmetry, minimax) for obtaining objective priors for "small" $\mathcal{M}$, I discuss large $\mathcal{M}$. Occam's razor in conjunction with Epicurus' principle of multiple explanations, quantified by Kolmogorov complexity, leads us to a universal prior, which results in a better predictor than any other prior over countable $\mathcal{M}$.

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 priors (see [KW96] for a survey and references). A main classification is between objective and subjective priors. An *objective prior* w_ν is a prior constructed based on some rational principles, which ideally everyone without (relevant) extra prior knowledge should adopt. In contrast, a *subjective prior* aims at modelling the agents personal (subjective) belief in environment ν prior to observation of x, but based on his past personal experience or knowledge (e.g. of related phenomena). In Section 6, I show that one way to arrive at a subjective prior is to start with an objective prior, make all past personal experience explicit, determine a "posterior" and use it as subjective prior. So I concentrate in the following on the more important objective priors.

Consider a very simple case of two environments, e.g. a biased coin with head *or* tail probability 1/3. In absence of any extra knowledge (which I henceforth assume) there is no reason to prefer head probability $\theta=1/3$ over $\theta=2/3$ and vice versa, leaving $w_{1/3}=w_{2/3}=\frac{1}{2}$ as the only rational choice. More generally, for finite $\mathcal{M}$, the *symmetry or indifference argument* [Lap12] suggests to set $w_\nu=\frac{1}{|\mathcal{M}|}$ $\forall\nu\in\mathcal{M}$. Similarly for a compact measurable parameter space Θ we may choose a uniform density $w(\theta)=[\mathrm{Vol}(\Theta)]^{-1}$. But there is a problem: If we go to a different parametrization (e.g. $\theta\rightsquigarrow\theta':=\sqrt{\theta}$ in the Bernoulli case), the prior $w(\theta)\rightsquigarrow w'(\theta')$ becomes non-uniform. Jeffreys' [Jef46] solution is to find a symmetry group of the problem (like permutations for finite $\mathcal{M}$) and require the prior to be *invariant under group transformations*. For instance, if $\theta\in\mathbb{R}$ is a location parameter (e.g. the mean) it is natural to require a translation-invariant prior. Problems are that there may be no obvious symmetry, the resulting prior may be improper (like for the translation group), and the result can depend on which parameters are treated as nuisance parameters.

The *maximum entropy principle* extends the symmetry principle by allowing certain types of constraints on the parameters. *Conjugate priors* are classes of priors such that the posteriors are themselves again in the class. While this can lead to interesting classes, the principle itself is not selective, since e.g. the class of all priors forms a conjugate class.

Another *minimax approach* by Bernardo [Ber79, CB90] is to consider bound (10), which can actually be improved within $o(1)$ to an equality. Since we want D_n to be small, we minimize the r.h.s. for the worst $\mu \in \mathcal{M}$. Choice $w(\theta) \propto \sqrt{\det \bar{j}_n(\theta)}$ equalizes and hence minimizes (10). The problems are the same as for Jeffrey's prior (actually often both priors coincide), and also the dependence on the model class and potentially on n.

The principles above, although not unproblematic, *can* provide good objective priors in many cases of small discrete or compact spaces, but we will meet some more problems later. For "large" model classes I am 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).

I 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 hypotheses, i.e. assign high prior (low) prior w_ν to simple (complex) hypotheses H_ν. Before I can define this prior, I need to quantify the notion of complexity.

Notation. A function $f : \mathcal{S} \to \mathbb{R} \cup \{\pm\infty\}$ is said to be lower semi-computable (or enumerable) if the set $\{(x,y) : y < f(x), x \in \mathcal{S}, y \in \mathbb{Q}\}$ 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. I write $O(1)$ for a constant of reasonable size: For instance, a sequence of length 100 is reasonable, maybe even 2^{30}, but 2^{500} is not. I 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 $\#\{t \leq n : A(t)\}/n \stackrel{n\to\infty}{\longrightarrow} 1$.

[3] I will ignore these additive and multiplicative fudges in the discussion till Section 6.

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* [Cha75, Gác74, Kol65, Lev74] of string x as the length ℓ of the shortest binary program p for which U outputs x:

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

Simple strings like 000...0 can be generated by short programs, and, hence have low Kolmogorov complexity, but irregular (e.g. random) strings are their own shortest description, and hence have high Kolmogorov complexity. 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 $K(f_i) = K(i)$.

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 natural universal TMs, the compiler constant is of reasonable size $O(1)$. A defining property of $K : \mathcal{X}^* \to \mathbb{N}$ is that it additively dominates all co-enumerable functions $f : \mathcal{X}^* \to \mathbb{N}$ 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:

$$\begin{aligned}
&a)\ K \text{ is not computable, but only upper semi-computable,}\\
&b)\ \text{the upper bound } K(n) \stackrel{+}{\leq} \log_2 n + 2\log_2\log n,\\
&c)\ \text{Kraft's inequality } \textstyle\sum_x 2^{-K(x)} \leq 1, \text{ which implies } 2^{-K(n)} \leq \frac{1}{n} \text{ for most } n,\\
&d)\ \text{information non-increase } K(f(x)) \stackrel{+}{\leq} K(x) + K(f) \text{ for recursive } f : \mathcal{X}^* \to \mathcal{X}^*,\\
&e)\ K(x) \stackrel{+}{\leq} -\log_2 P(x) + K(P) \text{ if } P : \mathcal{X}^* \to [0,1] \text{ is enumerable and } \textstyle\sum_x P(x) \leq 1,\\
&f)\ \textstyle\sum_{x:f(x)=y} 2^{-K(x)} \stackrel{\times}{=} 2^{-K(y)} \text{ if } f \text{ is recursive and } K(f) = O(1).
\end{aligned} \tag{11}$$

The proof of (f) can be found in Appendix A and the proofs of $(a)-(e)$ in [LV97].

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 (11c), this suggests the following choice:

$$w_\nu \;=\; w_\nu^U \;:=\; 2^{-K(\nu)} \tag{12}$$

For this choice, the bound (5) on D_n (which bounds (5) and (8)) reads

$$\sum_{t=1}^{\infty}\mathbf{E}[h_t] \;\leq\; D_\infty \;\leq\; K(\mu)\ln 2 \tag{13}$$

i.e. the number of times, ξ deviates from μ or l^{Λ_ξ} deviates from l^{Λ_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 (11e) 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) $w^U_\theta=2^{-K(\theta)}>0$ for computable θ, and 0 for uncomputable ones. This effectively reduces $\mathcal{M}$ to a discrete class $\{\nu_\theta\in\mathcal{M}:w^U_\theta>0\}$ 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

I now compare the classical continuous prior densities to the universal prior on classes of i.i.d. environments. I present some standard critiques to the former, illustrated on Bayes-Laplace's classical Bernoulli class with uniform prior: the problem of zero p(oste)rior, non-confirmation of universal hypotheses, and reparametrization and regrouping non-invariance. I show that the universal prior does not suffer from these problems. Finally I complement the general total bounds of Section 2 with some i.i.d.-specific instantaneous bounds.

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 [Bay63] 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 [Lap12] 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 I 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 (or ravens) are drawn randomly from an infinite population with fraction θ of black balls (or 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. This quantifies [Mah04, Thm.11] and shows that Maher does *not* solve the problem of confirmation of universal hypotheses.

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 $\delta_{ij} = \{{}^{1 \text{ if } i=j}_{0 \text{ else}}\}$ 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 $w_\theta^U=2^{-K(\theta)}$ (12) constructed in Section 3. It is difficult but not impossible to operate with such a prior [PH04, PH06]. 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 (11d), 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 $w'(\boldsymbol{\theta'}) \propto 1$, $w(\theta) \propto 1-\theta$ 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].

I 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 [Car52] 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.

I now show that the universal prior $w^U_\theta = 2^{-K(\theta)}$ 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 $(11f)$ 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.

Instantaneous bounds. The cumulative bounds (5) and (10) stay valid for i.i.d. processes, but instantaneous bounds are now also possible. For i.i.d. $\mathcal{M}$ with continuous, discrete, and universal prior, respectively, one can show (in preparation; see [Kri98, PH04, PH06] for related bounds)

$$\mathbf{E}[h_n] \;\stackrel{\times}{\leq}\; \tfrac{1}{n}\ln w(\theta_0)^{-1} \quad \text{and} \quad \mathbf{E}[h_n] \;\stackrel{\times}{\leq}\; \tfrac{1}{n}\ln w^{-1}_{\theta_0} \;=\; \tfrac{1}{n}K(\theta_0)\ln 2 \tag{14}$$

Note that, if summed up over n, they lead to weaker cumulative bounds.

5 Universal Sequence Prediction

Section 3 derived the universal prior and Section 4 discussed i.i.d. classes. What remains and will be done in this section is to find a universal class of environments, namely Solomonoff-Levin's class of all (lower semi)computable (semi)measures. The resulting universal mixture is equivalent to the output distribution of a universal Turing machine with uniform input distribution. The universal prior avoids the problem of old evidence and the universal class avoids the necessity of updating $\mathcal{M}$. I discuss the general total bounds of Section 2 for the specific universal mixture, and supplement them with some weak instantaneous bounds. Finally, I show that the universal mixture performs better than classical continuous mixtures, even in uncomputable environments.

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. (Finding a non-computable physical system would overturn the Church-Turing thesis.) 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 $f:\mathcal{X}^*\to\mathbb{R}$ 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

$$\xi_U(x) \;=\; \sum_{\nu\in\mathcal{M}_U} w_\nu^U \nu(x) \tag{15}$$

is itself lower semi-computable, i.e. $\xi_U\in\mathcal{M}_U$, which is a convenient property in itself. Note that since $\frac{1}{n\log^2 n} \overset{\times}{\leq} w_{\nu_n}^U \leq \frac{1}{n}$ for most n by (11b) and (11c), most ν_n have prior approximately reciprocal to their index n, as advocated by Jeffreys [Jef61, p238] and Rissanen [Ris83].

In some sense $\mathcal{M}_U$ is the largest class of environments for which ξ is in some sense computable [Hut03b, Hut06], but see [Sch02a] for even larger classes. Note that including non-semi-computable ν would not affect ξ_U, since $w_\nu^U=0$ on such environments.

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* hypotheses (including yet unknown or unnamed ones) a priori. So at least theoretically, updating $\mathcal{M}$ is unnecessary.

Other representations of $\boldsymbol{\xi_U}$. Definition (15) is somewhat complex, relying on enumeration of semimeasures and Kolmogorov complexity. I now approach ξ_U from a different perspective. Assume that our world is governed by a computable *deterministic* process describable in $\leq l$ bits. Consider a standard (not prefix) Turing machine U' and programs p generating environments starting with x. Let us pad all programs so that they have length exactly l. Among the 2^l programs of length l there are $N_l(x) := \#\{p \in \{0,1\}^l : U'(p) = x*\}$ programs consistent with observation x. If we regard all environmental descriptions $p \in \{0,1\}^l$ a priori as equally likely (Epicurus) we should adopt the relative frequency $N_l(x)/2^l$ as our prior belief in x. Since we do not know l and we can pad every p arbitrarily, we could take the limit $M(x) := \lim_{l\to\infty} N_l(x)/2^l$ (which exists, since $N_l(x)/2^l$ increases). Or equivalently: $M(x)$ is the probability that U' outputs a string starting with x when provided with uniform random noise on the program tape. Note that a uniform distribution is also used in the No Free Lunch theorems [WM97] 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. Yet another representation of M is as follows: For every q printing $x*$ there exists a shortest prefix (called minimal) p of q printing x. p possesses $2^{l-\ell(p)}$ prolongations to length l, all printing $x*$. Hence all prolongations of p together yield a contribution $2^{l-\ell(p)}/2^l = 2^{-\ell(p)}$ to $M(x)$. Let $U(p) = x*$ iff p is a minimal program printing a string starting with x. Then

$$M(x) \;=\; \sum_{p:U(p)=x*} 2^{-\ell(p)} \tag{16}$$

which 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. One can also get an explicit enumeration of all lower semi-computable semimeasures $\mathcal{M}_U = \{\nu_1, \nu_2, ...\}$ by means of $\nu_i(x) := \sum_{p:T_i(p)=x*} 2^{-\ell(p)}$, where $T_i(p) \equiv U(\langle i \rangle p)$, $i=1,2,...$ is an enumeration of all monotone Turing machines.

Bounds for computable environments. The bound (13) 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 (13) 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.

Total bounds (3) and (13) are suitable in an online setting, but *given* a fixed number of n observations, they give no guarantee on the next instance.

Weak instantaneous bounds. In Section 4, I derived good instantaneous bounds for i.i.d. classes. For coin or die flips or balls drawn from an urn this model is appropriate. But ornithologists do not really sample ravens independently at random. Although not strictly valid, the i.i.d. model may in this case still serve as a useful proxy for the true process. But to model the rise of the sun as an i.i.d. process is more than questionable. On the other hand it is plausible that these examples (and other processes like weather or stock-market) are governed by *some* (probabilistic) computable process. So model class $\mathcal{M}_U$ and predictor M seem appropriate. While excellent total bounds (3) and (13) exist, the essentially only *instantaneous bound* I was able to derive (proof in Appendix A) is

$$2^{-K(n)}\stackrel{\times}{\leq}M(\bar{x}_n|x_{<n})\stackrel{\times}{\leq}2^{2Km(x_{1:n})-K(n)} \tag{17}$$

valid for all n and $x_{1:n}$ and $\bar{x}_n\neq x_n$. I discuss the bound for the sequence $x_{1:\infty}=1^\infty$, but most of what I say remains valid for any other computable sequence. Since $Km(1^n)=O(1)$, we get

$$M(0|1^n)\stackrel{\times}{=}2^{-K(n)}$$

Since $2^{-K(n)}\leq\frac{1}{n}$ for most n, this shows that M quickly disbelieves in non-black objects and doomsday, similarly as in the i.i.d. model, but now only for most n.

Magic numbers. This 'most' qualification has interesting consequences: $M(0|1^n)$ spikes up for simple n. So M is cautious at magic instance numbers, e.g. fears doom on day 2^{20} more than on a comparable random day. While this looks odd and pours water on the mills of prophets, it is not completely absurd. For instance,

major software problems have been anticipated for the magic date, 1st of January 2000. There are many other occasions, where something happens at "magic" dates or instances; for instance solar eclipses.

Also, certain processes in nature follow fast growing sequences like those of the powers of two (e.g. the number of cells in an early human embryo) or the Fibonacci numbers (e.g. the number of petals or the arrangement of seeds in some flowers). Finally, that numbers with low (Kolmogorov) complexity cause high probability in real data bases can readily be verified by counting their frequency in the world wide web with Google [CV06].

On the other hand, (returning to sequence prediction) on most simple dates, nothing exceptional happens. Due to the total bound $\sum_{n=0}^{\infty} M(0|1^n) \leq O(1)$, M cannot spike up too much too often. M tells us to be more prepared but not to expect the unexpected on those days. Another issue is that often we do not know the exact start of the sequence. How many ravens exactly have ornithologists observed, and how many days exactly did the sun rise so far? In absence of this knowledge we need to Bayes-average over the sequence length which will wash out the spikes.

Universal is better than continuous $\mathcal{M}$. Although I argued that incomputable environments μ can safely be ignored, one may be nevertheless uneasy using Solomonoff's $M \stackrel{\times}{=} \xi_U$ (16) if outperformed by a continuous mixture ξ (9) 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 (15) 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 (10) 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 are argued for (see Section 3 and [KW96]). The above shows that M is superior to all of them. Actually the existence of *any* computable probabilistic predictor ρ with $D_n(\mu||\rho) = o(n)$ is sufficient for M to predict μ equally well.

Future bounds. Another important question is how many errors are still to come after some grace or learning period. Formally, given $x_{1:n}$, how large is the future expected error $r_n := \sum_{t=n+1}^{\infty} \mathbf{E}[h_t|\omega_{1:n} = x_{1:n}]$? The total bound (5)+(13) only implies that r_n asymptotically tends to zero w.p.1, and the instantaneous bounds (14) and (17) are weak and do not sum up finitely. Since the complexity of μ bounds the total

loss, a natural guess is that something like the conditional complexity of μ given x (on an extra input tape) bounds the future loss. Indeed one can show [Hut04, CH05]

$$\sum_{t=n+1}^{\infty} \mathbf{E}[h_t|\omega_{1:n}] \stackrel{+}{\leq} [K(\mu|\omega_{1:n})+K(n)]\ln 2 \tag{18}$$

i.e. if our past observations $\omega_{1:n}$ contain a lot of information about μ, we make few errors in future. For instance, consider the large space $\mathcal{X}$ of pixel images, and all observations are identical $\mu \widehat{=} \omega = x_1x_1x_1...$, where x_1 is a "typical" image of complexity, say, $K(x_1) \stackrel{+}{=} 10^6 \stackrel{+}{=} Km(\omega)$. Obviously, after seeing a couple of identical images we expect the next one to be the same again. While total bound (13) quite uselessly tells us that M makes less than 10^6 errors, future bound (18) with $n=1$ shows that M makes only $K(\mu|x_1)=O(1)$ errors. The $K(n)$ term can be improved to the complexity of the randomness deficiency of $\omega_{1:n}$ if a more suitable variant of algorithmic complexity that is monotone in its condition is used [CH05, CHS07]. No future bounds analogous to (18) for general prior or class are known.

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 $w^U_{\nu|y} = 2^{-K(\nu|y)}$ 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 (13)). 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]. For an interesting attempt to make M unique see [Mül06].

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], the speed prior [Sch02b], the minimal message and description length principles [Ris89, Wal05], and off-the-shelf compressors like Lempel-Ziv [LZ76] are such approximations, which have been successfully applied to a plethora of problems [CV05, Sch04].

Summary. I compared traditional Bayesian sequence prediction based on continuous classes and prior densities to Solomonoff's universal predictor M, prior w^U_ν, and class $\mathcal{M}_U$. I discussed the following advantages (+) and problems (−) of Solomonoff's approach:

+ general total bounds for generic class, prior, and loss,
+ universal and i.i.d.-specific instantaneous and future bounds,
+ the D_n 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 practical approximations.

In short, universal prediction solves or avoids or meliorates many foundational and philosophical problems, but has to be compromised in practice.

Conclusion. The goal of the paper was to establish a single, universal theory for (sequence) prediction and (hypothesis) confirmation, applicable to all inductive inference problems. I started by showing that Bayesian prediction is consistent for any countable model class, provided it contains the true distribution. The major (agonizing) problem Bayesian statistics leaves open is how to choose the model class and the prior. Solomonoff's theory fills this gap by choosing the class of all computable (stochastic) models, and a universal prior inspired by Ockham and Epicurus, and quantified by Kolmogorov complexity. I discussed in breadth how and in which sense this theory solves the inductive inference problem, by studying a plethora of problems other approaches suffer from. In one line: All you need for universal prediction is Ockham, Epicurus, Bayes, Solomonoff, Kolmogorov, and Turing. By including Bellman, one can extend this theory to universal decisions in reactive environments [Hut04].

Acknowledgements. I would like to thank Frank Stephan for his detailed feedback on earlier drafts.

References

[Bay63] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:376–398, 1763. [Reprinted in *Biometrika*, 45, 296–315, 1958].

[Ber79] J. M. Bernardo. Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society*, B41:113–147, 1979.

[Car52] R. Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.

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

[CH05] A. Chernov and M. Hutter. Monotone conditional complexity bounds on future prediction errors. In *Proc. 16th International Conf. on Algorithmic Learning Theory (ALT'05)*, volume 3734 of *LNAI*, pages 414–428, Singapore, 2005. Springer, Berlin.

[Cha75] G. J. Chaitin. A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340, 1975.

[CHS07] A. Chernov, M. Hutter, and J. Schmidhuber. Algorithmic complexity bounds on future prediction errors. *Information and Computation*, 205(2):242–261, 2007.

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

[CV06] R. Cilibrasi and P. M. B. Vitányi. Similarity of objects and the meaning of words. In *Proc. 3rd Annual Conferene on Theory and Applications of Models of Computation (TAMC'06)*, volume 3959 of *LNCS*, pages 21–45. Springer, 2006.

[Daw84] A. P. Dawid. Statistical theory. The prequential approach. *Journal of the Royal Statistical Society*, Series A 147:278–292, 1984.

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

[Gác74] P. Gács. On the symmetry of algorithmic information. *Soviet Mathematics Doklady*, 15:1477–1480, 1974.

[Gác83] P. Gács. On the relation between descriptional complexity and algorithmic probability. *Theoretical Computer Science*, 22:71–93, 1983.

[HM04] M. Hutter and An. A. Muchnik. Universal convergence of semimeasures on individual random sequences. In *Proc. 15th International Conf. on Algorithmic Learning Theory (ALT'04)*, volume 3244 of *LNAI*, pages 234–248, Padova, 2004. Springer, Berlin.

[Hut01] M. Hutter. New error bounds for Solomonoff prediction. *Journal of Computer and System Sciences*, 62(4):653–667, 2001.

[Hut03a] M. Hutter. Convergence and loss bounds for Bayesian sequence prediction. *IEEE Transactions on Information Theory*, 49(8):2061–2067, 2003.

[Hut03b] M. Hutter. On the existence and convergence of computable universal priors. In *Proc. 14th International Conf. on Algorithmic Learning Theory (ALT'03)*, volume 2842 of *LNAI*, pages 298–312, Sapporo, 2003. Springer, Berlin.

[Hut03c] M. Hutter. Optimality of universal Bayesian prediction for general loss and alphabet. *Journal of Machine Learning Research*, 4:971–1000, 2003.

[Hut04] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2004. 300 pages, http://www.hutter1.net/ai/uaibook.htm.

[Hut06] M. Hutter. On generalized computable universal priors and their convergence. *Theoretical Computer Science*, 364:27–41, 2006.

[Jay03] E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, MA, 2003.

[Jef46] H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proc. Royal Society London*, volume Series A 186, pages 453–461, 1946.

[Jef61] H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 3rd edition, 1961.

[Kol65] A. N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information and Transmission*, 1(1):1–7, 1965.

[Kri98] R. E. Krichevskiy. Laplace's law of succession and universal encoding. *IEEE Transactions on Information Theory*, 44:296–303, 1998.

[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.

[Lap12] P. Laplace. *Théorie analytique des probabilités*. Courcier, Paris, 1812. [English translation by F. W. Truscott and F. L. Emory: *A Philosophical Essay on Probabilities*. Dover, 1952].

[Lev74] L. A. Levin. Laws of information conservation (non-growth) and aspects of the foundation of probability theory. *Problems of Information Transmission*, 10(3):206–210, 1974.

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

[LZ76] A. Lempel and J. Ziv. On the complexity of finite sequences. *IEEE Transactions on Information Theory*, 22:75–81, 1976.

[Mah04] P. Maher. Probability captures the logic of scientific confirmation. In C. Hitchcock, editor, *Contemporary Debates in Philosophy of Science*, chapter 3, pages 69–93. Blackwell Publishing, 2004.

[MF98] N. Merhav and M. Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

[Mül06] Markus Müller. Stationary algorithmic probability. Technical Report http://arXiv.org/abs/cs/0608095, TU Berlin, Berlin, 2006.

[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.

[PH06] J. Poland and M. Hutter. MDL convergence speed for Bernoulli sequences. *Statistics and Computing*, 16(2):161–175, 2006.

[Ris83] J. J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11(2):416–431, 1983.

[Ris89] J. J. Rissanen. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore, 1989.

[Sch02a] 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.

[Sch02b] J. Schmidhuber. The speed prior: A new simplicity measure yielding near-optimal computable predictions. In *Proc. 15th Conf. on Computational Learning Theory (COLT-2002)*, volume 2375 of *LNAI*, pages 216–228, Sydney, 2002. Springer, Berlin.

[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.

[Vap99] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, Berlin, 2nd edition, 1999.

[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.

[WM97] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[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.

A Proofs of (8), (11f), and (17)

Proof of loss bound (8). Let X and Y be real-valued random variables. Taking the square root of the well-known Schwarz inequality $(\mathbf{E}[XY])^2 \leq \mathbf{E}[X^2]\mathbf{E}[Y^2]$ we get

$$\mathbf{E}[(X-Y)^2] - (\sqrt{\mathbf{E}[X^2]} - \sqrt{\mathbf{E}[Y^2]}\,)^2 \;\equiv\; 2\sqrt{\mathbf{E}[X^2]\mathbf{E}[Y^2]} - 2\mathbf{E}[XY] \;\geq\; 0.$$

Substituting $X\leadsto\sqrt{a_i}$, $Y\leadsto\sqrt{b_i}$, $\mathbf{E}[...]\leadsto\frac{1}{v_\Sigma}\sum_i v_i...$ with $v_\Sigma:=\sum_i v_i$, we get, after multiplying with v_Σ, the "Hellinger" bound

$$(\sqrt{\textstyle\sum_i v_i a_i}-\sqrt{\textstyle\sum_i v_i b_i})^2 \;\leq\; \textstyle\sum_i v_i(\sqrt{a_i}-\sqrt{b_i})^2 \tag{19}$$

for real $a_i,b_i,v_i\geq 0$ (also valid for $v_\Sigma=0$). I will use (19) three times in proving (8). With the abbreviations $m=y_t^{\Lambda_\mu}$ and $s=y_t^{\Lambda_\xi}$ and

$$\mathcal{X}=\{1,...,N\},\quad N=|\mathcal{X}|,\quad i=x_t,\quad y_i=\mu(x_t|\omega_{<t}),\quad z_i=\xi(x_t|\omega_{<t})$$

the loss (7) and Hellinger distance (4) can then be expressed by $l_t^{\Lambda_\xi}=\sum_i y_i\ell_{is}$, $l_t^{\Lambda_\mu}=\sum_i y_i\ell_{im}$ and $h_t=\sum_i(\sqrt{z_i}-\sqrt{y_i})^2$. By definition (6) of $y_t^{\Lambda_\mu}$ and $y_t^{\Lambda_\xi}$ we have

$$\sum_i y_i\ell_{im}\leq\sum_i y_i\ell_{ij}\quad\text{and}\quad\sum_i z_i\ell_{is}\leq\sum_i z_i\ell_{ij}\quad\text{for all}\quad j. \tag{20}$$

Actually, I need the first constraint only for $j=s$ and the second for $j=m$. From (20) we get

$$\begin{aligned}&\sqrt{\textstyle\sum_i y_i\ell_{is}}-\sqrt{\textstyle\sum_i y_i\ell_{im}}\;\geq\;0\qquad\text{and}\\ &[\tfrac{\partial}{\partial\ell_{is}}+\tfrac{\partial}{\partial\ell_{im}}](\sqrt{\textstyle\sum_i y_i\ell_{is}}-\sqrt{\textstyle\sum_i y_i\ell_{im}})=\frac{y_i}{2}\Big(\frac{1}{\sqrt{\sum_i y_i\ell_{is}}}-\frac{1}{\sqrt{\sum_i y_i\ell_{im}}}\Big)\leq 0.\end{aligned} \tag{21}$$

That is, if we decrease $\ell_{is}\leadsto\ell'_{is}:=\ell_{is}-\delta_i$ and $\ell_{im}\leadsto\ell'_{im}:=\ell_{im}-\delta_i$ by the same amount δ_i, then (21) increases. The maximal possible $\delta_i:=\min\{\ell_{is},\ell_{im}\}$ makes ℓ'_{is} or ℓ'_{im} zero, hence $0\leq\ell'_{is}+\ell'_{im}\leq 1$. Similarly

$$0\;\leq\;\sqrt{\textstyle\sum_i z_i\ell_{im}}-\sqrt{\textstyle\sum_i z_i\ell_{is}}\;\leq\;\sqrt{\textstyle\sum_i z_i\ell'_{im}}-\sqrt{\textstyle\sum_i z_i\ell'_{is}}$$

This implies

$$\begin{aligned}0\;&\leq\;\sqrt{l_t^{\Lambda_\xi}}-\sqrt{l_t^{\Lambda_\mu}}\;\equiv\;\sqrt{\textstyle\sum_i y_i\ell_{is}}-\sqrt{\textstyle\sum_i y_i\ell_{im}}\\ &\leq\;\sqrt{\textstyle\sum_i y_i\ell'_{is}}-\sqrt{\textstyle\sum_i y_i\ell'_{im}}\;+\;\sqrt{\textstyle\sum_i z_i\ell'_{im}}-\sqrt{\textstyle\sum_i z_i\ell'_{is}}\\ &\leq\;\sqrt{\textstyle\sum_i\ell'_{is}(\sqrt{y_i}-\sqrt{z_i})^2}+\sqrt{\textstyle\sum_i\ell'_{im}(\sqrt{y_i}-\sqrt{z_i})^2}\\ &\leq\;\sqrt{2\textstyle\sum_i(\ell'_{is}+\ell'_{im})(\sqrt{y_i}-\sqrt{z_i})^2}\;\leq\;\sqrt{2\textstyle\sum_i(\sqrt{y_i}-\sqrt{z_i})^2}\;\equiv\;\sqrt{2h_t}\end{aligned}$$

In the third inequality I used the Hellinger bound (19) twice, and in the fourth inequality I used $\sqrt{a}+\sqrt{b}\leq\sqrt{2(a+b)}$. Without the reduction $\ell\leadsto\ell'$ the bound would have been a factor of $\sqrt{2}$ worse. Taking the square, expectation, and sum over t proves the last inequality in (8). The first inequality in (8) is again an instantiation of (19) with $i\leadsto(t,\omega_{<t})$ and $v_i\leadsto\mu(\omega_{<t})$, i.e. $\sum_i v_i...\leadsto\sum_t\mathbf{E}[...]$ and $a_i\leadsto l_t^{\Lambda_\xi}$ and $b_i\leadsto l_t^{\Lambda_\mu}$. ■

Proof of equation (11*f*). Function $P(y):=\sum_{x:f(x)=y}2^{-K(x)}$ is lower semi-computable, since $K(x)$ is upper semi-computable, all $x\in\mathcal{X}^*$ can be enumerated,

and $f(x)=y$ is decidable. Further, $\sum_y P(y)=\sum_x 2^{-K(x)}\leq 1$, hence MDL bound (11$e$) implies $K(y)\stackrel{+}{\leq}-\log_2 P(y)+K(P)$. Let $g(y)=\min\{x:f(y)=x\}$ be the lexicographically first inverse of f. With $K(P)\stackrel{+}{\leq}K(f)=O(1)$, also function g has complexity $O(1)$. Hence

$$2^{-K(y)} \stackrel{\times}{\geq} P(y) \equiv \sum_{x:f(x)=y} 2^{-K(x)} \geq 2^{-K(g(y))} \geq 2^{-K(y)}$$

where I dropped all but the contribution from $g(y)$ in the sum, and used (11d) for g. ■

Proof of bound (17) $M(\bar{x}_n|x_{<n})\stackrel{\times}{\geq}2^{-K(n)}$. For $x=x_{<n}\in\mathcal{X}^{n-1}$ and $a=\bar{x}_n\in\mathcal{X}$ we have

$$M(a|x) \stackrel{(a)}{=} \frac{M(xa)}{M(x)} \stackrel{(b)}{=} \frac{\sum_{p:U(p)=xa*}2^{-\ell(p)}}{\sum_{p:U(p)=x*}2^{-\ell(p)}} \stackrel{(c)}{\geq} \frac{\sum_{p:U(\tilde{p})=xa}2^{-\ell(\tilde{p})}}{\sum_{p:U(p)=x*}2^{-\ell(p)}} \stackrel{(d)}{=} 2^{-\ell(qn^*)} \stackrel{(e)}{\stackrel{\times}{=}} 2^{-K(n)}$$

In (a) and (b) I simply inserted the definition (16) of M. I now (c) restrict the sum over all $p:U(p)=xa*$ in the numerator to programs $\tilde{p}$ of the following form: $\tilde{p}=qn^*p$, where $U(p)=x*$, n^* is the shortest code of n, and q simulates p until $n-1$ symbols are printed, then prints a, and thereafter halts, i.e. $U(\tilde{p})=xa$. The numerator now sums over exactly the same programs p as the denominator. Since $2^{-\ell(\tilde{p})}=2^{-\ell(qn^*)}2^{-\ell(p)}$, and $2^{-\ell(qn^*)}$ is a constant independent of p, numerator and denominator cancel and (d) follows. (e) follows from the definition of n^* and from $\ell(q)=O(1)$. ■

Proof of bound (17) $M(\bar{x}_n|x_{<n})\stackrel{\times}{\leq}2^{2Km(x_{1:n})-K(n)}$. Assume $x_{1:\infty}$ is a computable sequence, $\mathcal{X}$ is binary, and $\bar{x}_n\neq x_n$, and define $P(n):=M(x_{<n}\bar{x}_n)$. Given $x_{1:\infty}$, P can be semi-computed from below, hence $K(P)\stackrel{+}{\leq}Km(x_{1:\infty})$. Also $\sum_n P(n)\leq 1$, since $\{x_{<n}\bar{x}_n:n\in\mathbb{N}\}$ forms a prefix-free set. Hence $K(n)\stackrel{+}{\leq}-\log_2 P(n)+K(P)$ by (11e), which implies $M(x_{<n}\bar{x}_n)\stackrel{\times}{\leq}2^{Km(x_{1:\infty})-K(n)}$. Since $M(x_{<n})\geq 2^{-Km(x_{<n})}\geq 2^{-Km(x_{1:\infty})}$, we get $M(\bar{x}_n|x_{<n})\stackrel{\times}{\leq}2^{2Km(x_{1:\infty})-K(n)}$, which nearly is (17). Since the l.h.s. is independent of $x_{n+1:\infty}$, a bound independent of it should be (and is) possible, as we will now show.

Consider sequence $x_{1:n}$ and shortest program p printing $x_{1:n}*$. Let U_t be U stopped after t time steps and define corresponding M_t. Then $U_t(p)=x_{1:n_t}$ (for some $x_{n+1:n_t}$ if $n_t>n$). I define $P_t(n'):=\sum_{a\neq x_{n'}}M_t(x_{<n'}a)$ for $n'\leq n_t$ and 0 for $n'>n_t$. With n_t also P_t is computable and increasing, hence $P(n'):=\lim_{t\to\infty}P_t(n')=\sup_t P_t(n')$ is lower semi-computable. Clearly $P(n')=\sum_{a\neq x_{n'}}M(x_{<n'}a)$ for $n'\leq n_\infty$ and $P(n')=0$ for $n'>n_\infty$ ($n'_\infty=\lim_t n_t\leq\infty$). Hence $\sum_{n'}P(n')\leq 1$, since $\{x_{<n'}a:a\neq x_{n'},n'\leq n_\infty\}$ is a prefix free set, which implies $K(n)\stackrel{+}{\leq}-\log_2 P(n)+K(P)$ by (11e). Since $n\leq n_\infty$ and $K(P)\stackrel{+}{\leq}\ell(p)=Km(x_{1:n})$, we get $\sum_{a\neq x_n}M(x_{<n}a)\stackrel{\times}{\leq}2^{Km(x_{1:n})-K(n)}$. Using $M(x_{<n})\geq 2^{-Km(x_{<n})}\geq 2^{-Km(x_{1:n})}$, we get the desired bound $M(\bar{x}_n|x_{<n})\leq\sum_{a\neq x_n}M(a|x_{<n})\stackrel{\times}{\leq}2^{2Km(x_{1:n})-K(n)}$. ■