Computable priors sharpened into Occam's razors

David R. Bickel

December 30, 2016

Ottawa Institute of Systems Biology

Department of Biochemistry, Microbiology, and Immunology

Department of Mathematics and Statistics

University of Ottawa

451 Smyth Road

Ottawa, Ontario, K1H 8M5

+01 (613) 562-5800, ext. 8670

dbickel@uottawa.ca

Abstract

The posterior probabilities available under standard Bayesian statistics are computable, apply to small samples, and coherently incorporate previous information. Modifying their priors according to results from algorithmic information theory adds the advantage of implementing Occam's razor, giving simpler distributions of data higher prior probabilities.

Keywords: algorithmic probability; Bayesian inference; entropy rate; Kolmogorov complexity; prior probability; universal prior

Two general formalisms for data analysis based on Bayes's theorem include standard Bayesian statistics and Solomonoff's approach to universal distributions under algorithmic information theory. The posterior probabilities available under the former are computable and explicitly represent previous knowledge in terms of a probability distribution. While lacking those advantages, Solomonoff's use of universal distributions implements Occam's razor by assigning more prior probability to strings of symbols that are simpler in the sense of having lower Kolmogorov complexity.

In this note, the two approaches are combined to make their advantages simultaneously applicable to data analysis, inference, and decision. The complexity of each stochastic process under consideration will determine its prior probability for Bayesian inference and decision making consistent with Occam's razor. That prior probability of a process will be defined as the limit of its probability that an observation coincides with a symbol distributed according to a universal prior under algorithmic information theory.

To make that precise, consider a stationary process $X_{\theta,1}, X_{\theta,2}, \dots$ such that the law of $X_{\theta,t}$ is a probability mass function $f_{\theta,t}$ on a finite alphabet $\mathcal{X}$ for all $t = 1, 2, \dots$ and for each parameter value θ in a set Θ. Finite strings of length τ are written as $X_\theta^\tau \coloneqq (X_{\theta,1}, \dots, X_{\theta,\tau})$. The length of any string x is $\ell(x)$. Throughout, $\log = \log_2$. Convergence in probability and weak convergence as $\tau \to \infty$ are both denoted by $\xrightarrow{\text{weak}}$ (see Billingsley, 1999, p. 27).

Let ϑ denote a random variable with distribution π_0, a probability measure on the measurable space $(\Theta, \mathfrak{H})$, where $\mathfrak{H}$ is a σ-field of subsets of Θ. For any $\mathcal{H} \in \mathfrak{H}$, $\pi_0(\mathcal{H})$ is then $P(\vartheta \in \mathcal{H})$, the probability of the hypothesis that the process $X_{\vartheta,1}, X_{\vartheta,2}, \dots$ is in $\{(X_{\theta,1}, X_{\theta,2}, \dots) : \theta \in \mathcal{H}\}$. Since π_0 enables but does not depend on the complexity considerations of Occam's razor, it is called the *blunted prior distribution*. It incorporates application-specific information into the *universal prior* M, defined such that $M(x)$ is the algorithmic probability that a given universal monotonic Turing machine generates output beginning with x from a random program (Hutter, 2006, §2.4.1). M is a universal lower semicomputable continuous semimeasure rather than a probability measure (Li and Vitányi, 2008, §4.5).

According to the *Solomonoff measure* S, the conditional probability of a symbol followed by τ previous symbols based on M is

$$S(x_{\tau+1}|x^\tau) = \frac{S(x^{\tau+1})}{S(x^\tau)} \propto \frac{M(x^{\tau+1})}{M(x^\tau)}, \tag{1}$$

normalized such that $\sum_{y \in \mathcal{X}} S(y|x^\tau) = 1$, as required for $S(\bullet|x^\tau)$ to be a probability mass function, for all

$x^\tau \in \mathcal{X}^\tau$ (Li and Vitányi, 2008, §4.5.3). While S is incomputable and was originally intended for making predictions without any blunted π_0 (Solomonoff, 1978, 2008) or other explicit reliance on prior knowledge (Rathmanner and Hutter, 2011, §10.1), the leverage of such a π_0 leads to a computable posterior distribution suitable for predictions and other Bayes actions, decisions that minimize posterior expected loss.

Let $U_{\theta,\tau+1}$ be a random variable distributed according to the random probability mass function $S\left(\bullet|X_\theta^\tau\right)$ for each $\theta \in \Theta$. For any $\mathcal{H} \in \mathfrak{H}$ and $t = 1, 2, \ldots$, the τ-*whetted probability* $\pi_\tau\left(\mathcal{H}\right)$ and the *whetted probability* $\pi\left(\mathcal{H}\right)$ that $\vartheta \in \mathcal{H}$ are defined by

$$\pi_\tau\left(\mathcal{H}\right) \coloneqq P\left(\vartheta \in \mathcal{H} | X_{\vartheta,\tau+1} = U_{\vartheta,\tau+1}\right) \tag{2}$$

$$\pi\left(\mathcal{H}\right) \coloneqq \frac{\pi_0\left(\mathcal{H}\right)\int_{\mathcal{H}} d\pi_0\left(\theta|\mathcal{H}\right) 2^{-H(\theta)}}{\int d\pi_0\left(\theta\right) 2^{-H(\theta)}}, \tag{3}$$

where the function H on Θ gives the entropy rate $H\left(\theta\right) \coloneqq \lim_{\tau\to\infty} H\left(X_{\theta,1},\ldots,X_{\theta,\tau}\right)/\tau$ on the basis of the Shannon entropy $H\left(X_{\theta,1},\ldots,X_{\theta,\tau}\right)$ for each $\theta \in \Theta$. Equations (2)-(3) define two prior probability measures on $\left(\Theta, \mathfrak{H}\right)$, the τ-*whetted distribution* π_τ and the *whetted distribution* π. Describing both with the same adjective will be justified by $\pi_\tau \xrightarrow{\text{weak}} \pi$.

Example 1. Suppose Θ is a countable set such as the set of nonnegative integers. For any $\theta \in \Theta$ and $t = 1, 2, \ldots$, the τ-whetted probability $\pi_\tau\left(\theta\right)$ and the whetted probability $\pi\left(\theta\right)$ that the process is $X_{\theta,1}, X_{\theta,2}, \ldots$ are, with $\pi_0\left(\theta\right) \coloneqq \pi_0\left(\{\theta\}\right)$,

$$\pi_\tau\left(\theta\right) \coloneqq \pi_\tau\left(\{\theta\}\right) = P\left(\vartheta = \theta | X_{\vartheta,\tau+1} = U_{\vartheta,\tau+1}\right)$$

$$\pi\left(\theta\right) \coloneqq \pi\left(\{\theta\}\right) = \frac{\pi_0\left(\theta\right) 2^{-H(\theta)}}{\sum_{i\in\Theta} \pi_0\left(i\right) 2^{-H(i)}}. \tag{4}$$

Example 2. In the finite-Θ setting, suppose $X_{\theta,1}, X_{\theta,2}, \ldots$ are mutually independent, $\Theta = \mathcal{X} = \{1, 2, \ldots, m\}$, $\pi_0\left(\theta\right) = 1/m$, and $f_{\theta,t}\left(x\right) = 1/\theta$ if $x \le \theta$ but $f_{\theta,t}\left(x\right) = 0$ if $x > \theta$ for all $\theta, x = 1, 2, \ldots, m$ and $t = 1, 2, \ldots$. The entropy rate $H\left(\theta\right)$ is the entropy $-\sum_{x=1}^m f_{\theta,t}\left(x\right) \log f_{\theta,t}\left(x\right) = \log\theta$ under independence. Thus, according to equation (4), the whetted distribution is given by $\pi\left(\theta\right) \propto m^{-1} 2^{-\log\theta} = m^{-1}\theta^{-1}$.

Under ergodicity, the τ-whetted distribution converges to the whetted distribution, which is computable since π_0 and $f_{\theta,t}$ are computable.

Theorem 1. *If the process $X_{\theta,1}, X_{\theta,2}, \ldots$ is ergodic for every $\theta \in \Theta$, then $\pi_\tau \xrightarrow{weak} \pi$.*

Proof. Let $\mathcal{H}$ denote any continuity set in $\mathfrak{H}$. For any $t = 1, 2, \ldots$,

$$\pi_\tau(\mathcal{H}) = \frac{P(\vartheta \in \mathcal{H}) P(U_{\vartheta,\tau+1} = X_{\vartheta,\tau+1} | \vartheta \in \mathcal{H})}{P(U_{\vartheta,\tau+1} = X_{\vartheta,\tau+1})} = \frac{\pi_0(\mathcal{H}) \int_{\mathcal{H}} d\pi_0(\theta|\mathcal{H}) P(U_{\theta,\tau+1} = X_{\theta,\tau+1})}{P(U_{\vartheta,\tau+1} = X_{\vartheta,\tau+1})}. \tag{5}$$

The key factor is $\int_{\mathcal{H}} d\pi_0(\theta|\mathcal{H}) P(U_{\theta,\tau+1} = X_{\theta,\tau+1})$, in which

$$\begin{aligned}
P(U_{\theta,\tau+1} = X_{\theta,\tau+1}) &= \sum_{x \in \mathcal{X}} P(X_{\theta,\tau+1} = x) P(U_{\theta,\tau+1} = X_{\theta,\tau+1} | X_{\theta,\tau+1} = x) \\
&= \sum_{x \in \mathcal{X}} f_{\theta,\tau+1}(x) P(U_{\theta,\tau+1} = x | X_{\theta,\tau+1} = x) = \sum_{x \in \mathcal{X}} f_{\theta,\tau+1}(x) E(S(x|X_\theta^\tau) | X_{\theta,\tau+1} = x) \\
&= E(S(X_{\theta,\tau+1}|X_\theta^\tau)) = E\left(\frac{S(X_{\theta,\tau+1})}{S(X_\theta^\tau)}\right) \qquad (6)
\end{aligned}$$

for all $\theta \in \mathcal{H}$. Since $X_{\theta,1}, X_{\theta,2}, \ldots$ is an ergodic process, $C(X_\theta^\tau | \tau)/\tau \xrightarrow{\text{weak}} H(\theta)$, where $C(x|\tau)$ is the Kolmogorov complexity of a sequence x conditional on $\ell(x) = \tau$ (Horibe, 2003). That implies $C(X_\theta^\tau)/\tau \xrightarrow{\text{weak}} H(\theta)$ (cf. Li and Vitányi, 2008, p. 605), for there are constants $c_1, c_2 > 0$ that satisfy

$$C(x^\tau | \tau) - c_1 \leq C(x^\tau) \leq C(x^\tau | \tau) + 2\log\tau + c_2 \; \forall x^\tau \in \mathcal{X}^\tau$$

with the lower bound from Li and Vitányi (2008, p. 119) and the upper bound from Cover and Thomas (2006, Theorem 14.2.3). Writing $o_P(U_\tau)$ for any random sequence such that $o_P(U_\tau)/U_\tau \xrightarrow{\text{weak}} 0$ given a random sequence U_τ (Serfling, 1980, §1.2.5),

$$\frac{C(X_\theta^\tau)}{\tau} + o_P(1) = \frac{C(X_\theta^\tau | \tau)}{\tau} + o_P(1) = H(\theta).$$

Relations between C and M (Uspensky, 1992; Uspensky and Shen, 1996) entail that there are a $c_3 > 0$ and a $c_4 > 2$ satisfying

$$C(x^\tau) - \log\tau - c_3 \leq -\log M(x^\tau) \leq C(x^\tau) + c_4 \log\tau \; \forall x^\tau \in \mathcal{X}^\tau,$$

yielding $\left(-\log M\left(X_\theta^\tau\right)-C\left(X_\theta^\tau\right)\right)/\tau \xrightarrow{\text{weak}} 0$ and thus

$$\begin{aligned}-\log M\left(X_\theta^{\tau+1}\right)-\left(-\log M\left(X_\theta^\tau\right)\right) &= (\tau+1)\left(\frac{C\left(X_\theta^{\tau+1}\right)}{\tau+1}+o_P(1)\right)-\tau\left(\frac{C\left(X_\theta^\tau\right)}{\tau}+o_P(1)\right)\\ &= (\tau+1)\left(H(\theta)+o_P(1)\right)-\tau\left(H(\theta)+o_P(1)\right)\xrightarrow{\text{weak}} H(\theta).\end{aligned}$$

The continuous mapping theorem then yields

$$\frac{M\left(X_\theta^{\tau+1}\right)}{M\left(X_\theta^\tau\right)} = 2^{-\left(-\log M\left(X_\theta^{\tau+1}\right)-\left(-\log M\left(X_\theta^\tau\right)\right)\right)} \xrightarrow{\text{weak}} 2^{-H(\theta)}.$$

$$\therefore \lim_{\tau\to\infty} P\left(U_{\theta,\tau+1}=X_{\theta,\tau+1}\right) \propto 2^{-H(\theta)}$$

according to equations (1) and (6) since $\exp\left(-\left(M\left(X_\theta^{\tau+1}\right)-M\left(X_\theta^\tau\right)\right)\right)$ is bounded. From equation (5), $\pi_\tau(\mathcal{H})$ converges to a quantity proportional to $\pi_0(\mathcal{H})\int_{\mathcal{H}} d\pi_0(\theta|\mathcal{H})\exp(-H(\theta))$. Since $\int d\pi_\tau(\theta)=1$ for all $\tau=1,2,\ldots$, it follows that $\pi_\tau(\mathcal{H})\to\pi(\mathcal{H})$ for every continuity set $\mathcal{H}$ in $\mathfrak{H}$. The portmanteau theorem makes that equivalent to $\pi_\tau \xrightarrow{\text{weak}} \pi$. □

Since $\theta\mapsto 2^{-H(\theta)}$ in equation (3) is mathematically equivalent to a likelihood function and yet does not depend on any data, it is an example of a prior likelihood function, the logarithm of which is a "prior support" function as defined by Edwards (1992). As likelihood is defined only up to a multiplicative constant, $L(\bullet)=c2^{-H(\bullet)}$ for any $c>0$ may be called the *whetted likelihood function*. It relates the *whetted* and *blunted probability densities* by $(d\pi/d\eta)(\theta)\propto L(\theta)(d\pi_0/d\eta)(\theta)$ for any $\theta\in\Theta$, where η is any measure that dominates π_0 and π. In the sense that the use of the whetted likelihood justifies the frequent claim that the prior probability density should be higher for simpler hypotheses, it resembles the *simplicity postulate*, which requires laws with fewer free parameters to have higher prior probabilities (Jeffreys, 1948, pp. 100-101).

Example 3. Let $X_{\theta,1},X_{\theta,2},\ldots$ denote independent Bernoulli variates of unknown probability θ of success. By independence, the entropy rate $H(\theta)$ is the entropy, $-\theta\log\theta-(1-\theta)\log(1-\theta)$. The resulting whetted likelihood $L(\theta)=\theta^\theta(1-\theta)^{(1-\theta)}$ is the dashed curve in Figure 1. The other two plotted curves are the products of $L(\theta)$ and the likelihood functions of the binomial distribution for $n=1$ and $n=10$ observations, all successes. (The finite sample size n should not be confused with τ, which goes to infinity.) If π_0 is uniform, the density of π is proportional to $L(\theta)$, differing markedly from Jeffreys's prior density, instead proportional

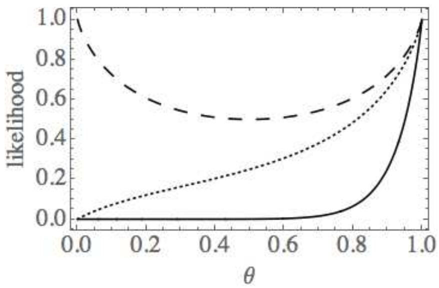

Figure 1: Likelihood functions for the binomial probability as θ. The whetted likelihood function is combined with the likelihood function for $n = 0$ (dashed), $n = 1$ (dotted), and $n = 10$ (solid) consecutive successes.

to $\theta^{-1/2} (1-\theta)^{-1/2}$ (Robert et al., 2012, p. 73), and the posterior density is proportional to the likelihood functions in Figure 1.

Example 4. Consider the two-state Markov chain $X_{\theta,1}, X_{\theta,2}, \ldots$ with probability transition matrix

$$\begin{bmatrix} 1-\theta & \theta \\ \phi & 1-\phi \end{bmatrix},$$

where θ has a uniform blunted prior distribution π_0 and ϕ is known. The entropy rate of the process is

$$H(\theta) = \frac{\phi}{\theta+\phi} \left(-\log \Lambda(\theta)\right) + \frac{\theta}{\theta+\phi} \left(-\log \Lambda(\phi)\right),$$

where $\Lambda(\bullet) = \bullet^{\bullet} (1-\bullet)^{(1-\bullet)}$ (Cover and Thomas, 2006, pp. 73, 78), which is the whetted likelihood function of Example 3. The whetted prior distribution is determined by

$$\pi(\theta) \propto L(\theta) = (\Lambda(\theta))^{\frac{\phi}{\theta+\phi}} (\Lambda(\phi))^{\frac{\theta}{\theta+\phi}}.$$

To extend the whetted distribution to continuous random variables, let $Y_{\theta,1}, Y_{\theta,2}, \ldots$ denote a stationary process such that, for all $\theta \in \Theta$ and $t = 1, 2, \ldots$, the distribution of $Y_{\theta,t}$ is a Riemann-integrable probability density function $g_{\theta,\tau}$ on an interval $\mathcal{Y}$ for all $\tau = 1, 2, \ldots$. The *differential whetted probability* $\widetilde{\pi}(\mathcal{H})$ that $\vartheta \in \mathcal{H}$ is

$$\widetilde{\pi}(\mathcal{H}) := \frac{\widetilde{\pi}_0(\mathcal{H}) \int_{\mathcal{H}} d\pi_0(\theta|\mathcal{H})\, 2^{-h(\theta)}}{\int d\pi_0(\theta)\, 2^{-h(\theta)}},$$

in which $h(\theta)$ is the differential entropy rate $h(\theta) \coloneqq \lim_{\tau\to\infty} h(Y_{\theta,1},\ldots,Y_{\theta,\tau})/\tau$, where $h(Y_{\theta,1},\ldots,Y_{\theta,\tau})$ is the differential entropy of $Y_{\theta,1},\ldots,Y_{\theta,\tau}$ for every $\theta \in \Theta$. In analogy with L, the function $\theta \mapsto \widetilde{L}(\theta) = c2^{-h(\theta)}$ for any $c > 0$ is called the *differential whetted likelihood function*.

Example 5. Suppose $Y_{\theta,1}, Y_{\theta,2},\ldots$ are independent variates drawn from the normal distribution with unknown mean μ and unknown standard deviation σ, where $\theta = (\mu,\sigma)$. Invariant measures that are limits of prior probability density functions include the left Haar measure and the right Haar measure, with densities proportional to σ^{-2} and σ^{-1}, respectively (Kass and Wasserman, 1996). Since $2^{h(\mu,\sigma)} \propto \sigma$ (Michalowicz et al., 2013), the whetted likelihood is $\widetilde{L}(\mu,\sigma) \propto \sigma^{-1}$. Thus, if $\widetilde{\pi}_0$ is the right Haar measure, then $\widetilde{\pi}$ is the left Haar measure, having a density proportional to $\widetilde{L}(\mu,\sigma)\,\sigma^{-1} \propto \sigma^{-2}$. Using that $\widetilde{\pi}$ as the prior, the posterior predictive distribution of the next observation after observing n observations of sample mean $\widehat{\mu}$ and unbiased sample variance $\widehat{\sigma}^2$ is the t distribution with parameters $\widehat{\mu}$, $\left(1 - n^{-1}\right)\widehat{\sigma}^2$, and $n-1$ (Held and Sabanés Bové, 2014, p. 301). Since the posterior distribution has a mean of $\widehat{\mu}$, the Bayes estimate of μ under squared error loss is $\widehat{\mu}$ since that minimizes the posterior expected loss using $\widetilde{\pi}$ as the prior. Like π of Example 3, this $\widetilde{\pi}$ applies to small n as well as to large n.

The differential whetted probability is a limit of the following whetted probabilities that are themselves limits in the sense of Theorem 1. $\mathcal{Y}_1, \mathcal{Y}_2, \ldots$ is a sequence of interval subsets of $\mathcal{Y}_\infty \coloneqq \mathcal{Y}$ such that $\mathcal{Y}_i \subset \mathcal{Y}_j$ for $i < j$, and $\mathfrak{Y}_i(\Delta)$ is a partition of $\mathcal{Y}_i$ into intervals of width $\Delta > 0$ for $i = 1, 2, \ldots$ and $i = \infty$. Denote by $X_{\theta,i,\Delta,t}$ a random element with probability masses $Q_{\theta,t}(\mathcal{Y}'|\mathcal{Y}_i)$ for all $\mathcal{Y}' \in \mathfrak{Y}_i(\Delta)$, $\theta \in \Theta$, and $t = 1, 2, \ldots$. The entropy $H(X_{\theta,i,\Delta,1},\ldots,X_{\theta,i,\Delta,\tau})$ and the whetted distribution for $X_{\theta,i,\Delta,1}, X_{\theta,i,\Delta,2}\ldots$ are abbreviated by $H(\theta,i,\Delta,\tau)$ and $\pi_{i,\Delta}$, respectively. That whetted distribution converges to $\widetilde{\pi}$, the *differential whetted distribution*, in the following sense.

Proposition 1. *If* $\lim_{i\to\infty} H(\theta,i,\Delta,\tau) = H(\theta,\infty,\Delta,\tau)$ *for all* $\Delta > 0$ *and* $\tau = 1, 2, \ldots$ *and*

$$\lim_{\Delta\to 0}\lim_{i\to\infty}\lim_{\tau\to\infty} H(\theta,i,\Delta,\tau) + \log\Delta = \lim_{\tau\to\infty}\lim_{\Delta\to 0}\lim_{i\to\infty} H(\theta,i,\Delta,\tau) + \log\Delta$$

for all $\theta \in \Theta$, *then, for all* $\mathcal{H} \in \mathfrak{H}$,

$$\lim_{\Delta\to 0}\lim_{i\to\infty} \pi_{i,\Delta}(\mathcal{H}) = \widetilde{\pi}(\mathcal{H}).$$

Proof. Since $g_{\theta,\tau}$ is Riemann integrable for all $\theta \in \Theta$ and $\tau = 1, 2, \ldots$,

$$\lim_{\Delta \to 0} H(\theta, \infty, \Delta, \tau) + \log \Delta = h(Y_{\theta,1}, \ldots, Y_{\theta,\tau})$$

(Cover and Thomas, 2006, Theorem 8.3.1). Thus, $H(\theta, i, \Delta, \tau) + \log \Delta \to h(Y_{\theta,1}, \ldots, Y_{\theta,\tau})$ and, with $H(\theta, i, \Delta, \tau) / \tau \to H(\theta, i, \Delta)$, which is the entropy rate of the process $X_{\theta,i,\Delta,1}, X_{\theta,i,\Delta,2}, \ldots$,

$$\begin{aligned}
\lim_{\Delta \to 0} \lim_{i \to \infty} H(\theta, i, \Delta) + 0 &= \lim_{\Delta \to 0} \lim_{i \to \infty} \lim_{\tau \to \infty} H(\theta, i, \Delta, \tau) / \tau + \lim_{\Delta \to 0} \lim_{\tau \to \infty} (\log \Delta) / \tau \\
&= \lim_{\tau \to \infty} \left(\lim_{\Delta \to 0} \lim_{i \to \infty} H(\theta, i, \Delta, \tau) + \log \Delta \right) / \tau \\
&= \lim_{\tau \to \infty} h(Y_{\theta,1}, \ldots, Y_{\theta,\tau}) / \tau = h(\theta).
\end{aligned}$$

Substituting $\lim_{\Delta \to 0} \lim_{i \to \infty} H(\theta, i, \Delta)$ for $H(\theta)$ in equation (3) completes the proof. $\square$

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada and by the Faculty of Medicine of the University of Ottawa.

References

Billingsley, P., 1999. Convergence of Probability Measures. Wiley Series in Probability and Statistics. Wiley, New York.

Cover, T., Thomas, J., 2006. Elements of Information Theory. John Wiley & Sons, New York.

Edwards, A. W. F., 1992. Likelihood. Johns Hopkins Press, Baltimore.

Held, L., Sabanés Bové, D., 2014. Applied Statistical Inference 10.

Horibe, Y., 2003. A note on Kolmogorov complexity and entropy. Applied Mathematics Letters 16 (7), 1129–1130.

Hutter, M., 2006. Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Texts in Theoretical Computer Science. An EATCS Series. Springer Berlin Heidelberg.

Jeffreys, H., 1948. Theory of Probability. Oxford University Press, London.

Kass, R. E., Wasserman, L., 1996. The selection of prior distributions by formal rules. Journal of the American Statistical Association 91, 1343–1370.

Li, M., Vitányi, P., 2008. An Introduction to Kolmogorov Complexity and Its Applications. Texts in Computer Science. Springer New York.

Michalowicz, J. V., Nichols, J. M., Bucholtz, F., 2013. Handbook of Differential Entropy. CRC Press, New York.

Rathmanner, S., Hutter, M., 2011. A philosophical treatise of universal induction. Entropy 13 (6), 1076–1136.

Robert, C., Christensen, R., Johnson, W., Branscum, A., Hanson, T., 2012. Bayesian Ideas and Data Analysis.

Serfling, R. J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.

Solomonoff, R., 1978. Complexity-based induction systems: Comparisons and convergence theorems. IEEE Transactions on Information Theory 24 (4), 422–432.

Solomonoff, R. J., 2008. The probability of 'undefined' (non-converging) output in generating the universal probability distribution. Information Processing Letters 106 (6), 238–240.

Uspensky, V. A., 1992. Complexity and entropy: An introduction to the theory of Kolmogorov complexity. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 85–102.

Uspensky, V. A., Shen, A., 1996. Relations between varieties of Kolmogorov complexities. Mathematical systems theory 29 (3), 271–292.