A Computational Theory of Surprise

Pierre Baldi
Department of Information and Computer Science
California Institute for Telecommunications and Information Technology
University of California, Irvine
Irvine, CA 92697-3425
pfbaldi@ics.uci.edu

February 24, 2002

Abstract

While eminently successful for the *transmission* of data, Shannon's theory of information does not address semantic and subjective dimensions of data, such as relevance and surprise. We propose an observer-dependent computational theory of surprise where surprise is defined by the relative entropy between the prior and the posterior distributions of an observer. Surprise requires integration over the space of models in contrast with Shannon's entropy, which requires integration over the space of data. We show how surprise can be computed exactly in a number of discrete and continuous cases using distributions from the exponential family with conjugate priors. We show that during sequential Bayesian learning, surprise decreases like $1/N$ and study how surprise differs and complements Shannon's definition of information.

Keywords: Information, Surprise, Relevance, Bayesian Probabilities, Entropy, Relative Entropy.

1 Introduction

The notion of information is central to science, technology, and many other human endeavors. While several approaches for quantifying information have been proposed, the most successful one so far has been Claude Shannon's definition introduced over half a century ago [20, 18, 4, 8]. According to Shannon, the information contained in a data set D is given by $-\log P(D)$, and the average information over all possible data sets $\mathcal{D}$ is the entropy $H(P(D)) = -\int_{\mathcal{D}} P(D) \log P(D) dD$.

Although it has been eminently successful for the development of modern telecommunication and computer technologies, Shannon's definition does not capture all aspects of

information and comes with a number of shortcomings that may in part explain why the theory has not been as successful as one would have hoped in other areas of science such as biology, psychology, or economics.

A first concern is that it fails to account how data can have different significance for different observers. This is rooted in the origin of the probabilities used in the definition of information. These probabilities are defined according to an observer or a model M which Shannon does not describe explicitly so that the information in a data set is rather the negative log-likelihood

$$I(D, M) = -\log P(D|M) \tag{1}$$

and the corresponding entropy is the average over all data sets

$$I(\mathcal{D}, M) = H(P(D|M)) = -\int_{\mathcal{D}} P(D|M) \log P(D|M) dD \tag{2}$$

As pointed out by Edward Jaynes ([14]), this observer is essentially the communication engineer designing the communication system and, as such, M is fixed. However, not only information ought to be a property of the data, it should also be highly dependent on the observer, because the same data may carry completely different meanings for different observers. Consider for instance the genomic DNA sequence of the AIDS virus. It is a string of about 10,000 letters over the 4-letter DNA alphabet, of great significance to researchers in the biological or medical sciences, but utterly uninspiring to a layman. Within Shannon's framework, one could consider two observers $\mathcal{O}_1$ and $\mathcal{O}_2$ (or two model classes $\mathcal{M}_1$ and $\mathcal{M}_2$) with models M_1 and M_2 and assign information $-\log P(D|M_1)$ and $-\log P(D|M_2)$ to the data relative to each model. This however remains unsatisfactory. In particular, even if the two likelihoods were the same, the data D could carry different amounts of information for $\mathcal{O}_1$ and $\mathcal{O}_2$ depending on their expectations. Thus information ought to depend on the observer and also on his expectations.

Indeed Shannon's theory of information explicitly ignores any notions of relevance or semantics in the data. As pointed out in the title of Shannon's seminal article, it is a theory of *communication*, in the sense of transmission rather than information. It concentrates on the problem of "reproducing at one point either exactly or approximately a message selected at another point" regardless of the relevance of the message. But there is clearly more to information than data reproducibility and somehow information ought to depend also on the model or hypothesis M, or rather on the class $\mathcal{M}$ of such models.

Shannon's theory also produces a well-known paradoxical effect that is often puzzling to new students in information theory. How is it that "white snow", the most boring of all television programs, carries the most Shannon information? On one hand, it is clear that the uniform distribution has the highest entropy and reproducing a snow pattern *exactly* requires a very large number of bits. On the other hand, producing "snow-like" patterns is very easy. How can we reconcile the two viewpoints in a rigorous way? Notice that this paradox has nothing to do with the complexity of the generative model being used. A high

order Markov model of the television images would still make snow highly improbable and therefore highly informative from Shannon's standpoint.

In short, there seems to be room for developing concepts of information that complement or extend Shannon's definition. The main purpose here is to develop a computational theory of subjective information surprise, or surprise. Surprise, no matter how one defines it, is obviously related to Shannon's information: a rare event is in general surprising and ought to carry a great deal of Shannon information due to its low probability. But beyond this obvious relationship, a theory of surprise should to be able to measure information surprise that is contained in data (1) in an observer-dependent way; (2) related to his changes in expectation; (3) through a definition that clearly establishes a connection with the foundations of probability theory; and (4) clarifies the "white snow" paradox and related concerns.

If such a definition exists, it must first of all be related to the foundations of the notion of probability, which can be approached from a frequentist or subjectivist, also called Bayesian, point of view[3, 6]. Here we follow the Bayesian approach which has been prominent in recent years and has led to important developments in many fields [12, 10]. The definition we propose stems directly from the Bayesian foundation of probability theory, and the relation given by Bayes theorem between the prior and posterior probabilities of an observer (see also [23]). The amount of surprise in the data for a given observer can be measured by looking at the change that has taken place in going from the prior to the posterior probabilities.

2 Information and Surprise

In the subjectivist framework, degrees of belief or confidence are associated with hypotheses or models. It can be shown that under a small set of reasonable axioms, these degrees of belief can be represented by real numbers and that when rescaled to the [0,1] interval these degrees of confidence must obey the rules of probability and in particular Bayes theorem [9, 19, 15]. Specifically, if an observer has a model M for the data, associated with a prior probability $P(M)$, the arrival of a data set D leads to a reevaluation of the probability in terms of the posterior distribution

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \tag{3}$$

The effect of the information contained in D is clearly to change the belief of the observer from $P(M)$ to $P(M|D)$. Thus, a complementary way of measuring information carried by the data D is to measure the distance between the prior and the posterior. To distinguish it from Shannon's communication information, we call this notion of information the surprise information or *surprise*

$$S(D, \mathcal{M}) = d[P(M), P(M|D)] \tag{4}$$

where d is a distance or similarity measure. There are different ways of measuring a distance between probability distributions. In what follows, for standard well known theoretical rea-

sons (including invariance with respect to reparameterizations), we use the relative entropy or Kullback-Liebler [17] divergence K which is not symmetric and hence not a distance. This lack of symmetry, however, does not matter in most cases and in principle can easily be fixed by symmetrization of the divergence. The surprise then is

$$\begin{aligned} S(D,\mathcal{M}) = K(P(M),P(M|D)) &= \int_{\mathcal{M}} P(M)\log\frac{P(M)}{P(M|D)}dM \\ &= -H(P(M)) - \int P(M)\log P(M|D)dM \\ &= \log P(D) - \int_{\mathcal{M}} P(M)\log P(D|M)dM \end{aligned} \tag{5}$$

Alternatively, we can define the single model surprise by the log-odd ratio

$$S(D,M) = \log\frac{P(M)}{P(M|D)} \tag{6}$$

and the surprise by its average

$$S(D,\mathcal{M}) = \int_{\mathcal{M}} S(D,M)P(M)dM \tag{7}$$

taken with respect to the prior distribution over the model class. In statistical mechanics terminology, the surprise can also be viewed as the free energy of the negative log-posterior at a temperature $t = 1$, with respect to the prior distribution over the space of models [2].

Note that this definition addresses the "white snow" paradox. At the time of snow onset, the image distribution we expect and the image we perceive are very different and therefore the snow carries a great deal of both surprise and Shannon's information. Indeed snow may be a sign of storm, earthquake, toddler's curiosity, or military putsch. But after a few seconds, once our model of the image shifts towards a snow model of random pixels, television snow perfectly fits the prior and hence becomes boring. Since the prior and the posterior are virtually identical, snow frames carry 0 surprise although megabytes of Shannon's information.

The similarities and differences of surprise with Shannon's information should now be clear—in particular, surprise is a dual notion that requires integration over the space of models rather than the space of data. In the next sections, we show how this integration can be carried analytically in simple cases. As is the case for Bayesian inference, however, integration over the space of models in general is likely to require Monte Carlo methods. Information is also bound to vary when data is received sequentially, i.e. during learning processes. Therefore we also show how the proposed definition is related to the behavior of learning curves. Details of calculations are provided in the Appendix.

3 Computation of Surprise

Here we consider a data set $D = \{x_1, \ldots, x_N\}$ containing N points. Surprise can be calculated exactly in a number of interesting cases. For simplicity, although this does not correspond to any restriction of the general theory, we consider only the case of conjugate priors, where the prior and the posterior have the same functional form. In this case, in order to compute the surprise defined by Equation 5, we need only to compute general terms of the form

$$F(P_1, P_2) = \int P_1 \log P_2 dx \tag{8}$$

where P_1 and P_2 have the same functional form. The surprise is then given by

$$S = F(P_1, P_1) - F(P_1, P_2) \tag{9}$$

where P_1 is the prior and P_2 is the posterior. Note also that in this case the symmetric divergence can easily be computed using $F(P_1, P_1) - F(P_1, P_2) + F(P_2, P_2) - F(P_2, P_1)$. Details for the calculation of $F(P_1, P_2)$ in the examples below are given in the Appendix. It should also be clear that in simple cases, for instance for certain members of the exponential family [7] of distributions, the posterior depends entirely on the sufficient statistics and therefore we can expect surprise also to depend only on sufficient statistics in these cases.

3.1 Discrete Data and Dirichlet Model

Consider the case where x_i is binary. The simplest class of models for D is then $M(p)$, the first order Markov models with a single parameter p representing the probability of emitting a 1. The conjugate prior on p is the Dirichlet prior (or beta distribution in the 2-D case)

$$D_1(a_1, b_1) = \frac{\Gamma(a_1 + b_1)}{\Gamma(a_1)\Gamma(b_1)} x^{a_1 - 1}(1 - x)^{b_1 - 1} = C_1 x^{a_1 - 1}(1 - x)^{b_1 - 1} \tag{10}$$

with $a_1 \geq 0$, $b_1 \geq 0$, and $a_1 + b_1 > 0$. The expectation is $a_1/(a_1 + b_1), b_1/(a_1 + b_1)$. With n sucesses in the sequence D, the posterior is a Dirichlet distribution $D_2(a_2, b_2)$ with [2]

$$a_2 = a_1 + n \quad \text{and} \quad b_2 = b_1 + (N - n) \tag{11}$$

The surprise can be computed exactly

$$S(D, \mathcal{M}) = K((D_1, D_2)) = \log \frac{C_1}{C_2} + n[\Psi(a_1 + b_1) - \Psi(a_1)] + (N - n)[\Psi(a_1 + b_1) - \Psi(b_1)] \tag{12}$$

where Ψ is the derivative of the logarithm of the Gamma function (see Appendix). When $N \to \infty$, and $n = pN$ with $0 < p < 1$ we have

$$S(D, \mathcal{M}) \approx N\mathrm{K}(p, a_1) \tag{13}$$

where $K(p, a_1)$ represents the Kullback-Liebler divergence distance between the empirical distribution $(p, 1-p)$ and the expectation of the prior $(a_1/(a_1+b_1), b_1/(a_1+b_1))$. Thus asymptotically surprise information grows linearly with the number of data points with a proportionality coefficient that depends on the discrepancy between the expectation of the prior and the observed distribution. The same relationship can be expected to be true in the case of a multinomial model. In the case of a symmetric prior $(a_1 = b_1)$, a slightly more precise approximation is provided by:

$$S(D_1, D_2) \approx N[\sum_{k=a_1}^{2a_1-1} \frac{1}{k} - H(p)] \tag{14}$$

For instance, when $a_1 = 1$ then $R(D_1, D_2) \approx N(1 - H(p))$, and when $a_1 = 5$ then $R(D_1, D_2) \approx N[0.746 - H(p)]$.

These results provide a clear explanation for the television "snow" effect. With a uniform symmetric prior, the empirical distribution with maximal entropy brings the least information. If we expect snow, the Kullback-Liebler divergence between the prior and the posterior is 0 and therefore there is essentially no surprise in the signal. As pointed out, this is not the case, however, at the time of onset of the snow where the divergence may even be large.

3.2 Continuous Data: Unknown Mean/Known Variance

When the x_i are real, we can consider first the case of unknown mean with known variance. We have a family $M(\mu)$ of models, with a Gaussian prior $G_1(\mu_1, \sigma_1^2)$. If the data has known variance σ^2, then the posterior distribution is Gaussian $G_2(\mu_2, \sigma_2^2)$ with parameters given by [10]

$$\mu_2 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{N\bar{m}}{\sigma^2}}{\frac{1}{\sigma_1^2} + \frac{N}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{N}{\sigma^2} \tag{15}$$

where $\bar{m}$ is the observed mean. In the general case

$$\begin{aligned} S(D, \mathcal{M}) = KG_1, G_2) &= \log \frac{\sigma}{\sqrt{\sigma^2 + N\sigma_1^2}} + N\frac{\sigma_1^2}{2\sigma^2} + \frac{N^2\sigma_1^2(\mu_1 - \bar{m})^2}{2\sigma^2(\sigma^2 + N\sigma_1^2} \\ &\approx \frac{N}{2\sigma^2}[\sigma_1^2 + (\mu_1 - \bar{m})^2] \end{aligned} \tag{16}$$

the approximation being valid for large N. In the special case where the prior has the same variance has the data $\sigma_1 = \sigma$ then the formula simplify a little and yield

$$S = K(G_1, G_2) = \frac{N}{2} - \frac{1}{2}\log(N+1) + \frac{N^2(\mu_1 - \bar{m})^2}{2(N+1)\sigma^2} \approx \frac{N}{2\sigma^2}[\sigma^2 + (\mu_1 - \bar{m})^2] \tag{17}$$

when N is large. In any case, surprise grows linearly with N with a coefficient that is the sum of the prior variance and the square difference between the expected mean and the empirical mean scaled by the variance of the data.

3.3 Continuous Data: Unknown Variance/Known Mean

When the x_i are real, we can then consider the case of unknown variance with known mean. We have a family $M(\sigma^2)$ of models, with a conjugate scaled inverse gamma prior

$$\Gamma_1(\nu_1, s_1) = \frac{(\frac{\nu_1}{2})^{\nu_1/2} s_1^{\nu_1}}{\Gamma(\frac{\nu_1}{2})} (\sigma^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{\nu_1 s_1^2}{2\sigma^2}} d\sigma^2 = C_1 (\sigma^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{\nu_1 s_1^2}{2\sigma^2}} d\sigma^2 \tag{18}$$

The posterior is then a scaled inverse gamma distribution [10] with

$$\nu_2 = \nu_1 + N \quad \text{and} \quad s_2^2 = \frac{\nu_1 s_1^2 + N\bar{\sigma}^2}{\nu_1 + N} \tag{19}$$

Here $\bar{\sigma}^2 = \sum (x_i - m)^2 / N$ is the observed variance, based on the known mean m. The surprise

$$S(D, \mathcal{M}) = K(\Gamma_1, \Gamma_2) = \log \frac{C_1}{C_2} - \frac{N}{2}[\Psi(\frac{\nu}{2}) + \log \frac{2}{\nu_1 s_1^2}] + \frac{N\bar{\sigma}^2}{2s_1^2} \tag{20}$$

For large values of N,

$$S = K(\Gamma_1, \Gamma_2) \approx \frac{N}{2}[\frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi(\frac{\nu_1}{2})] \tag{21}$$

Thus surprise information scales linearly with N, with a coefficient of proportionality that typically depends mostly on the ratio of the empirical variance to the scale parameters s_1^2, which is roughly the expectation of the prior [the expectation of the prior is $\nu_1 s_1^2/(\nu_1 - 2)$ provided $\nu_1 > 2$]. The effects of very large of very small values of $\bar{\sigma}$, or ν_1 can also be seen in the formula above. In particular, surprise is largest when the empirical variance $\bar{\sigma}^2$ goes to 0 or infinity, i.e. is very different from the prior expectation.

3.4 Continuous Data: Unknown Mean/Unknown Variance

When the x_i are real, we can finally consider the case of unknown mean with unknown variance. We have a family $M(\mu, \sigma^2)$ of models, with a conjugate prior $G_1\Gamma_1 = P(\mu|\sigma^2)P(\sigma^2) = G_1(\mu_1, \sigma^2/\kappa_1)\Gamma_1(\nu_1, s_1)$, product of a normal with a scaled inverse gamnma distribution. Thus the prior has four parameters $(\mu_1, \kappa_1, \nu_1, s_1)$, with $\kappa_1 > 0$, $\nu_1 > 0$, and $s_1 > 0$. The conjugate posterior has the same form, with similar parameters $(\mu_2, \kappa_2, \nu_2, s_2)$ satisfying (see for instance [10])

$$\mu_2 = \frac{\kappa_1}{\kappa_1 + N}\mu_1 + \frac{N}{\kappa_1 + N}\bar{m} \tag{22}$$

$$\kappa_2 = \kappa_1 + N \tag{23}$$
$$\nu_2 = \nu_1 + N \tag{24}$$
$$\nu_2 s_2^2 = \nu_1 s_1^2 + (N-1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1 + N}(\bar{m} - \mu_1)^2 \tag{25}$$

with $\bar{m} = \sum x_i/N$ and $\bar{\sigma}^2 = \sum (x_i - \bar{m})^2/(N-1)$. The surprise is

$$\begin{aligned} S(D, \mathcal{M}) &= K(G_1\Gamma_1, G_2\Gamma_2) = \frac{1}{2}\log\frac{\kappa_1}{\kappa_1 + N} + \frac{N}{2\kappa_1} + \frac{\kappa_1 + N}{2}\left[\frac{N(\bar{m} - \mu_1)}{(\kappa_1 + N)s_1}\right]^2 + \\ \log\frac{C_1}{C_2} &+ -\frac{N}{2}[\Psi(\frac{\nu_1}{2}) + \log\frac{2}{\nu_1 s_1^2}] + \frac{(N-1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1 + N}(\bar{m} - \mu_1)^2}{2s_1^2} \end{aligned} \tag{26}$$

For large values of N,

$$R(G_1\Gamma_1, G_2\Gamma_2) \approx \frac{N}{2}[\frac{1}{\kappa_1} + \frac{\bar{\sigma}^2}{s_1^2} + \log\frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi(\frac{\nu_1}{2}) + \frac{(\bar{m} - \mu_1)^2}{s_1^2}] \tag{27}$$

Surprise information is linear in N with a coefficient that is essentially the sum of the coefficients derived in the unknown mean and unknown variance partial cases.

4 Learning and Surprise

There is an immediate connection between surprise and computational learning theory. If we imagine that data points from a training set are presented sequentially, we can consider that the posterior distribution after the N-th point becomes the prior for the next iteration (sequential Bayesian learning). In this case we can expect on average surprise to decrease after each iteration, since as a system learns what is relevant in a data set, new data points become less and less surprising. This can be quantified precisely, at least in simple cases.

4.1 Learning Curves: Discrete Data

Consider first a sequence of 0-1 examples $D = (d_N)$. The learner starts with a Dirichlet prior $D_0(a_0, b_0)$. With each example d_N, the learner updates its Dirichlet prior $D_N(a_N, b_N)$ into a Dirichlet posterior $D_{N+1}(a_{N+1}, b_{N+1})$ with $(a_{N+1}, b_{N+1}) = (a_N + 1, b_N)$ if $d_{N+1} = 1$, and $(a_{N+1}, b_{N+1}) = (a_N, b_N + 1)$ otherwise. When $d_{N+1} = 1$, the corresponding surprise is easily computed using Equations 45 and 48. For simplicity, and without much loss of generality, let us assume that a_0 and b_0 are integers, so that a_N and b_N are also integers for any N. Then if $d_{N+1} = 1$ the relative surprise is

$$S(D_N, D_{N+1}) = \log\frac{a_N}{a_N + b_N} + \sum_{k=0}^{b_N - 1}\frac{1}{a_N + k} \tag{28}$$

and similarly in the case $d_{N+1} = 0$ by interchanging the role of a_N and b_N. Thus, in this case,

$$0 \leq S(D_N, D_{N+1}) \leq \frac{1}{a_N} + \log(1 - \frac{1}{a_N + b_N}) \tag{29}$$

Asymptotically we have $a_N \approx a_0 + pN$ and therefore

$$0 \leq S(D_N, D_{N+1}) \leq \frac{1-p}{pN} \tag{30}$$

Thus surprise decreases in time with the number of examples as $1/N$.

4.2 Learning Curves: Continuous Data

In the case of continuous Gaussian data with, for instance, known variance σ^2, the learner starts with a Gaussian prior $G_0(\mu_0, \sigma_0^2)$ on the mean. With each example d_N, the learner updates its Gaussian prior $G_N(\mu_N, \sigma_N^2)$ into a Gaussian posterior $G_{N+1}(\mu_{N+1}, \sigma_{N+1}^2)$ with

$$\mu_{N+1} = \frac{\frac{\mu_N}{\sigma_N^2} + \frac{d_{N+1}}{\sigma^2}}{\frac{1}{\sigma_N^2} + \frac{1}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_{N+1}^2} = \frac{1}{\sigma_N^2} + \frac{1}{\sigma^2} \tag{31}$$

From Equation 16, the relative surprise is

$$S(G_N, G_{N+1}) = \log \frac{\sigma}{\sqrt{\sigma^2 + \sigma_n^2}} + \frac{\sigma_N^2}{2\sigma^2}(1 + \frac{(\mu_N - d_{N+1})^2}{\sigma^2 + \sigma_N^2}) \tag{32}$$

Asymptotically

$$S(G_N, G_{N+1}) \leq \frac{\sigma_N^2}{2\sigma^2} \tag{33}$$

From Equation 15, we have $\frac{1}{\sigma_{N+1}^2} = \frac{1}{\sigma_0^2} + \frac{(N+1)}{\sigma^2}$, or $\sigma_{N+1}^2 = \frac{\sigma_0^2 \sigma^2}{\sigma^2 + (N+1)\sigma_0^2}$. Thus

$$0 \leq S(G_N, G_{N+1}) \leq \frac{1}{2(N+1)} \tag{34}$$

Thus in this case surprise decreases in time with the number of examples also as $1/N$.

5 Surprise, Evidence, and Mutual Information

To measure the effect of the data on the prior and the posterior, one could have envisioned using the difference between the entropy of the prior and the entropy of the posterior. However, unlike surprise which is always positive, the difference between these two entropies can be either positive or negative and therefore is not a suitable measure.

In the formula given above for the surprise (Equation 5), we have introduced the *evidence* $P(D) = P(D|\mathcal{M}) = \int_{\mathcal{M}} P(M, D)dM$. The evidence plays a key role in Bayesian analyis and is the hinge that leads to the next cycle of Bayesian analysis beyond the class of models $\mathcal{M}$. Shannon's information could be defined with respect to the evidence in the form

$$I(D, \mathcal{M}) = -\log P(D|\mathcal{M}) \tag{35}$$

with the associated evidence entropy

$$I(\mathcal{D}, \mathcal{M}) = -\int_{\mathcal{D}} P(D|\mathcal{M}) \log P(D|\mathcal{M})dD \tag{36}$$

For a fixed data set D, the surprise is

$$S(D, \mathcal{M}) = -I(D, \mathcal{M}) + \int_{\mathcal{M}} P(M)I(D, M)dM \tag{37}$$

is therefore the difference between the average Shannon information per model, taken with respect to the prior, and the Shannon information based on the evidence.

If we integrate the surprise with respect to the evidence

$$\int_{\mathcal{D}} P(D)S(D, \mathcal{M})dD = \int_{\mathcal{D},\mathcal{M}} P(D)P(M) \log \frac{P(D)P(M)}{P(D, M)} dDdM \tag{38}$$

we get the Kullback-Liebler divergence $K(P(D)P(M), P(D, M))$ which is the symmetric inverse of the mutual information MI between $\mathcal{D}$ and $\mathcal{M}$ $MI(\mathcal{D}, \mathcal{M}) = K(P(D, M), P(D)P(M))$.

6 Discussion and Extensions

Surprise is different from other definitions of information that have been proposed [1] as alternatives to Shannon's entropy. Most alternative definitions, such as Rényi's entropies, are actually algebraic variations on Shannon's definition rather than conceptually different approaches. While Shannon's definition fixes the model and varies the data, surprise fixes the data and varies the model. Surprise is a measure of dissimilarity between the prior and posterior distributions and as such it lies close to the axiomatic foundation of Bayesian probability.

In a number of cases, surprise can be computed analytically both in terms of exact and asymptotic formula. The analytical results presented here could be extended in several directions including non-conjugate and other prior distributions, more complex multidimensional distributions (e.g. multinomial, inverse Wishart), and more general families of distributions (e.g. exponential family [7]). In general, however, the computation of surprise can be expected to require Monte Carlo methods to approximate integrals over model spaces. In this respect, the computation of surprise should benefit from progress in this active area of research as well as increase in computing power.

While applications remain to be developed, a theory of surprise could be used in areas as diverse as game theory, machine learning, Internet commerce, and the design of sensory systems. Consider, for instance, the design of artificial sensory systems or the reverse engineering of natural ones. Clearly, attention mechanisms play a fundamental role allowing perceptual systems to shift their resources and bring them to bear on the most surprising region of the input space. In both natural systems and some of their artificial cousins, expectations could be generated by top down connections and compared in real time with input streams generated by bottom up connections [13]. Mismatches between input and expectations could be computed using surprise theory and lead to saliency maps. These maps in turn could guide attentional mechanisms, whereby additional processing resources are dynamically allocated to the regions of the input field that are the most surprising, i.e. which carry the highest amount of information with respect to the expectations.

Likewise, we have only touched upon the connection between surprise and machine learning [22] by showing that surprise decreases as $1/N$ during sequential learning in simple cases. This analysis could be extended to more complex settings, such as artificial neural networks.

But the notion of surprise has its own limitations. In particular, it does not capture all the semantics/relevance aspects of data. When the degree of surprise of the data with respect to the model class becomes low, the data is no longer informative for the given model class. This, however, does not necessarily imply that one has a good model since the model class itself could be unsatisfactory and in need of a complete overhaul. The process by which we decide a model is unsatisfactory in an alternative free setting, the open-ended aspect of inference, remains elusive to modeling.

Conversely, highly surprising data could be a sign that learning is required or that the data is irrelevant. If while surfing the web in search of a car one stumbles on a picture of Marilyn Monroe, the picture may carry a low degree of relevance, a high degree of surprise, and a low-to-high amount of Shannon information depending on the pixel structure. Thus, relevance, surprise, and Shannon's entropy are three different facets of information that can be present in different combinations. The notion of *relevance* in particular seems to be the least understood although there have been several attempts [16, 21]. A possible direction is to consider, in addition to the space of data and models, a third space $\mathcal{A}$ of actions or interpretations and define relevance as the relative entropy between the prior $P(A)$ and the posterior $P(A|D)$ distributions over $\mathcal{A}$. Whether this approach simply shifts the problem into the definition of the set $\mathcal{A}$ remains to be seen. In any event, the quest to understand the nature of information is unlikely to be over.

Appendix A: Discrete Case

In the two-dimensional case, consider two Dirichlet distributions $D_1 = D_(a_1, b_1)(x) = C_1 x^{a_1-1}(1-x)^{b_1-1}$ and $D_2 = D(a_2, b_2)(x) = C_2 x^{a_2-1}(1-x)^{b_2-1}$, with $C_1 = \Gamma(a_1+b_1)/\Gamma(a_1)\Gamma(b_1)$, and similarly for C_2. To calculate the relative entropy in the two dimensional case, we use the formula ([11]

$$\int_0^1 x^{u-1}(1-x)^{v-1}\log x dx = B(u,v)[\Psi(u)-\Psi(u+v)] \tag{39}$$

where $B(u,v)$ is the beta function $B(u,v)=\int_0^1 x^{u-1}(1-x)^{v-1}dx = \Gamma(u)\Gamma(v)/\Gamma(u+v)$ and $\Psi(x)$ is the derivative of the logarithm of the gamma function $\Psi(x) = d(\log\Gamma(x))/dx$. A cross term of the form $F(D_1,D_2)$

$$F(D_1,D_2) = \int_0^1 C_1 x^{a_1-1}(1-x)^{b_1-1}[\log C_2 + (a_2-1)\log x + (b_2-1)\log(1-x)] \tag{40}$$

is equal to

$$F(D_1,D_2) = \log C_2 + (a_2-1)[\Psi(a_1)-\Psi(a_1+b_1)] + (b_2-1)[\Psi(b_1)-\Psi(a_1+b_1)] \tag{41}$$

using the fact that $C_1 B(a_1,b_1)=1$. In particular, the entropy of a two-dimensional Dirichlet distribution such as D_1 is obtained by taking: $-F(D_1,D_1)$. With some algebra, the Kullback-Liebler divergence between any two Dirichlet distributions is finally given by:

$$K(D_1,D_2) = \log\frac{C_1}{C_2} + (a_1-a_2)[\Psi(a_1)-\Psi(a_1+b_1)] + (b_1-b_2)[\Psi(b_1)-\Psi(a_1+b_1)] \tag{42}$$

With n successes in the sequence D, the posterior is a Dirichlet distribution $D_2(a_2,b_2)$ with [2]

$$a_2 = a_1 + n \quad \text{and} \quad b_2 = b_1 + (N-n) \tag{43}$$

Using this relation between the prior and the posterior, we get the surprise

$$S(D_1,D_2) = \log\frac{C_1}{C_2} + n[\Psi(a_1+b_1)-\Psi(a_1)] + (N-n)[\Psi(a_1+b_1)-\Psi(b_1)] \tag{44}$$

Using the general fact that $\Psi(x)-\Psi(y) = \sum_{k=0}^{\infty}(\frac{1}{y+k}-\frac{1}{x+k})$, which implies $\Psi(x+n)-\Psi(x) = \sum_{k=0}^{n-1}\frac{1}{x+k}$ when n is an integer, we get

$$S(D_1,D_2) = \log\frac{C_1}{C_2} + n(\sum_{k=0}^{\infty}\frac{1}{a_1+k} - \frac{1}{a_1+b_1+k}) + (N-n)(\sum_{k=0}^{\infty}\frac{1}{b_1+k} - \frac{1}{a_1+b_1+k}) \tag{45}$$

Now we have

$$\sum_{k=0}^{\infty}(\frac{1}{a_1+k} - \frac{1}{a_1+b_1+k}) = \sum_{k=0}^{\lfloor b_1\rfloor -1}(\frac{1}{a_1+k}) + \text{Rest} \tag{46}$$

where

$$0 \leq \text{Rest} = \sum_{k=0}^{\infty}(\frac{1}{a_1 + \lfloor b_1 \rfloor + k} - \frac{1}{a_1 + b_1 + k}) \leq (b_1 - \lfloor b_1 \rfloor) \sum_{k=0}^{\infty} \frac{1}{(a_1 + \lfloor b_1 \rfloor + k)^2} \tag{47}$$

and similarly for the symmetric term. The rest is exactly 0 when a_1 and b_1 (and hence a_2 abd b_2) are integers, and in general decreases with the size of a_1 and b_1. This yields the approximation

$$S(D_1, D_2) \approx \log \frac{C_1}{C_2} + n(\sum_{k=0}^{\lfloor b_1 \rfloor - 1} \frac{1}{a_1 + k}) + (N - n)(\sum_{k=0}^{\lfloor a_1 \rfloor - 1} \frac{1}{b_1 + k}) \tag{48}$$

This approximation is *exact* when a_1 and b_1 are integers. Now for $x > 0$ we have $\log((x + n)/x) < \sum_{k=0}^{n-1} 1/(x+k) < \log((x+n-1)/x)+1/x$ or $0 < \sum_{k=0}^{n-1} 1/(x+k) - \log((x+n)/x) < 1/x$. Thus,

$$S(D_1, D_2) \approx \log \frac{C_1}{C_2} + n \log \frac{a_1 + b_1}{a_1} + (N - n) \log \frac{a_1 + b_1}{b_1} \tag{49}$$

Now we have,

$$\begin{aligned} \log \frac{C_1}{C_2} &= \log \frac{\Gamma(a_1 + b_1)\Gamma(a_2)\Gamma(b_2)}{\Gamma(a_1)\Gamma(b_1)\Gamma(a_2 + b_2)} \\ &= \log \frac{(a_1 + n - 1)(a_1 + n - 2) \ldots a(b_1 + N - n - 1) \ldots b_1}{(a_1 + b_1 + N - 1)(a_1 + b_1 + N - 2) \ldots (a_1 + b_1)} \end{aligned} \tag{50}$$

We can use bounds of the form $\log a + \int_{a_1}^{a_1+n1} \log x dx < \log a_1 + \ldots \log(a_1 + n - 1) \leq \log a_1 \int_{a_1+1}^{a_1+n} \log x dx$ to estimate this term. Alternatively, one can assume that a_1 and b_1 are integers and use binomial coefficient approximations, such as those in [5]. In all cases, neglecting constant terms and terms of order $\log N$, if we let $n = pN$ $(0 < p < 1)$ and N go to infinity we have

$$\log \frac{C_1}{C_2} \approx -\log \binom{N}{n} \approx -NH(p)) \tag{51}$$

where $H(p)$ is the entropy of the (p, q) distribution with $q = 1 - p$. Thus when $N \to \infty$, and $n = pN$ with $0 < p < 1$ we have

$$S(D_1, D_2) \approx N[p \log \frac{a_1 + b_1}{a_1} + q \log \frac{a_1 + b_1}{b_1} - H(p)] \approx N\text{K}(p, a_1) \tag{52}$$

where $K(p, a_1)$ is the relative entropy between the empirical distribution (p, q) and the expectation of the prior $(\frac{a_1}{a_1+b_1}, \frac{b_1}{a_1+b_1})$. Thus, asymptotically surprise grows linearly with the number of data points with a proportionality coefficient that depends on the discrepancy between the expectation of the prior and the observed distribution. The same relationship can be expected to be true in the case of a multinomial model.

Symmetric Prior ($a_1 = b_1$)

Consider now the case of a symmetric prior, then

$$S(D_1, D_2) = \log \frac{C_1}{C_2} + N[\Psi(2a_1) - \Psi(a_1)] \tag{53}$$

Using formulas in [11], $\Psi(2a_1) - \Psi(a_1) = \sum_{k=0}^{\infty} \frac{(-1)^k}{2a_1+k} + \log 2$ thus

$$S(D_1, D_2)) = \log \frac{C_1}{C_2} + N \sum_{k=0}^{\infty} \frac{(-1)^k}{2a_1 + k} + \log 2 \approx N[\sum_{k=0}^{\infty} \frac{(-1)^k}{2a_1 + k} + \log 2 - H(p)] \tag{54}$$

the approximation being in the regime $n = pN$ and $N \to \infty$. When a_1 is an integer, we also have $\Psi(2a_1) - \Psi(a_1) = \sum_{k=1}^{2a_1-1} (-1)^{k+1}/k = \sum_{k=a_1}^{2a_1-1} 1/k$. Thus when a_1 is an integer

$$S(D_1, D_2) = N[\sum_{k=a_1}^{2a_1-1} \frac{1}{k}] + \log \frac{(2a_1 - 1)\binom{2a_1-2}{a_1-1}}{(2a_1 + N - 1)\binom{N+2a_1-2}{n+a_1-1}} \tag{55}$$

As $N \to \infty$ with $0 < p < 1$

$$S(D_1, D_2) \approx N[\sum_{k=a_1}^{2a_1-1} \frac{1}{k}] - \log \binom{N + 2a_1 - 2}{n + a_1 - 1} \approx N[\sum_{k=a_1}^{2a_1-1} \frac{1}{k}] - \log \binom{N}{n} \tag{56}$$

and therefore

$$S(D_1, D_2) \approx N[\sum_{k=a_1}^{2a_1-1} \frac{1}{k} - H(p)] \tag{57}$$

For instance, when $a_1 = b_1 = 1$, this gives:

$$S(D_1, D_2) = N - \log(N + 1) - \log \binom{N}{n} \tag{58}$$

with the asymptotic form

$$S(D_1, D_2) \approx N(1 - H(p)) + \log \frac{\sqrt{2N\pi pq}}{N + 1} \approx N(1 - H(p)) \tag{59}$$

With a uniform symmetric prior, the empirical distribution with maximal entropy brings the least information. When $a_1 = b_1 = 5$ this gives $R(D_1, D_2) \approx N[0.746 - H(p)]$. As we increase $a_1 + b_1$, keeping $a_1 = b_1$, the constant $\sum_{a_1}^{2a_1-1}(1/k)$ decreases to its asymptotic value $\log 2$ which corresponds to the asymptotic form $S(D_1, D_2) \approx NK(p, 0.5)$. The stronger the strength of the uniform prior (the larger $a_1 + b_1$), the smaller the surprise created by a die with maximum entropy.

Appendix B: Continuous Case

Unknown Mean/Known Variance

Consider now two Gaussians $G_1(\mu_1, \sigma_1)$ and $G_2(\mu_2, \sigma_2)$. Then, after some algebra, the cross term is given by

$$F(G_1, G_2) = \int_{-\infty}^{+\infty} G_1 \log G_2 dx = -\frac{1}{2}\log(2\pi\sigma_2^2) - \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} \tag{60}$$

here using for simplicity natural logarithms. $F(G, G) = \frac{1}{2}\log[2\pi e\sigma^2] = H(G)$ is the entropy. The Kullback-Liebler divergence can then be obtained

$$K(G_1, G_2) = -\frac{1}{2} + \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} \tag{61}$$

Consider now a data set with N points $x_1, \ldots, x_N$ with empirical mean $\bar{m}$. If the data has known variance σ^2, then the posterior parameters are given by:

$$\mu_2 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{N\bar{m}}{\sigma^2}}{\frac{1}{\sigma_1^2} + \frac{N}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_2^2} = \frac{1}{\sigma_1^2} + \frac{N}{\sigma^2} \tag{62}$$

In the general case

$$S(G_1, G_2) = \log\frac{\sigma}{\sqrt{\sigma^2 + N\sigma_1^2}} + N\frac{\sigma_1^2}{2\sigma^2} + \frac{N^2\sigma_1^2(\mu_1 - \bar{m})^2}{2\sigma^2(\sigma^2 + N\sigma_1^2)} \approx \frac{N}{2\sigma^2}[\sigma_1^2 + (\mu_1 - \bar{m})^2] \tag{63}$$

when N is large. In the special case where the prior has the same variance has the data $\sigma_1 = \sigma$ then the formula simplifies a little and yields

$$S(G_1, G_2) = \frac{N}{2} - \frac{1}{2}\log(N+1) + \frac{N^2(\mu_1 - \bar{m})^2}{2(N+1)\sigma^2} \approx \frac{N}{2\sigma^2}[\sigma^2 + (\mu_1 - \bar{m})^2] \tag{64}$$

when N is large. In any case, surprise grows linearly with N with a coefficient that is the sum of the prior variance and the square difference between the expected mean and the empirical mean scaled by the variance of the data.

Unknown Variance/Known Mean

In the case of unknown variance and known mean, we have a family $M(\sigma^2)$ of models with a conjugate prior for σ^2 that is a scaled inverse gamma distribution

$$\Gamma_1(\nu_1, s_1) = \frac{(\frac{\nu_1}{2})^{\nu_1/2} s_1^{\nu_1}}{\Gamma(\frac{\nu_1}{2})}(\sigma^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{\nu_1 s_1^2}{2\sigma^2}} d\sigma^2 = C_1(\sigma^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{\nu_1 s_1^2}{2\sigma^2}} d\sigma^2 \tag{65}$$

with $\nu_1 > 0$ degrees of freedom and scale $s_1 > 0$. F can be computed expanding the integrals and using the fact that $\int_0^{+\infty} x^{\nu/2-1} e^{-x} \log x = \Gamma(\frac{\nu}{2})\Psi(\frac{\nu}{2})$. This yields:

$$F(\nu_1, s_1; \nu_2, s_2) = \log \frac{(\nu_2/2)^{\nu_2/2} s_2^{\nu_2}}{\Gamma(\frac{\nu_2}{2})} + (\frac{\nu_2}{2} + 1)[\Psi(\frac{\nu_1}{2}) + \log \frac{2}{\nu_1 s_1^2}] - \frac{\nu_2 s_2^2}{2 s_1^2} \tag{66}$$

The posterior is then a scaled inverse gamma distribution [10] with

$$\nu_2 = \nu_1 + N \quad \text{and} \quad s_2^2 = \frac{\nu_1 s_1^2 + N\bar{\sigma}^2}{\nu_1 + N} \tag{67}$$

where $\bar{\sigma}^2$ is the empirical variance $\bar{\sigma}^2 = \sum_i (x_i - m)^2/N$, based on the known mean m. The surprise is given by

$$S(\Gamma_1, \Gamma_2) = \log \frac{C_1}{C_2} - \frac{N}{2}(\Psi(\frac{\nu_1}{2}) + \log \frac{2}{\nu_1 s_1^2}) + \frac{N\bar{\sigma}^2}{2 s_1^2} \tag{68}$$

For large values of N, taking only the leading terms

$$\begin{aligned} S(\Gamma_1, \Gamma_2) &\approx \frac{N}{2}(\frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2} - \Psi(\frac{\nu_1}{2})) \\ &+ \log \Gamma(\frac{\nu_1 + N}{2}) - \frac{\nu_1 + N}{2} \log \frac{\nu_1 + N}{2} - \frac{(\nu_1 + N)}{2} \log \frac{\nu_1 s_1^2 + N\bar{\sigma}^2}{\nu_1 + N} \end{aligned} \tag{69}$$

$$S(\Gamma_1, \Gamma_2) \approx \frac{N}{2}[\frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi(\frac{\nu_1}{2})] \tag{70}$$

Thus surprise information scales linearly with N, with a coefficient of proportionality that typically depends mostly on the ratio of the empirical variance to the scale parameters s_1^2, which is roughly the expectation of the prior [the expectation of the prior is $\nu_1 s_1^2/(\nu_1 - 2)$ provided $\nu_1 > 2$]. The effects of very large of very small values of $\bar{\sigma}$, or ν_1 can also be seen in the formula above. In particular, surprise is largest when the empirical variance $\bar{\sigma}^2$ goes to 0 or infinity, i.e. is very different from the prior expectation.

Unknown Mean/Unknown Variance

In the case of unknown mean and unknown variance, we have a family $M(\mu, \sigma^2)$ of models with a conjugate prior of the form $G_1\Gamma_1 = P(\mu|\sigma^2)P(\sigma^2) = G_1(\mu_1, \sigma^2/\kappa_1)\Gamma_1(\nu_1, s_1)$. Thus the prior has four parameters $(\mu_1, \kappa_1, \nu_1, s_1)$, with $\kappa_1 > 0$, $\nu_1 > 0$, and $s_1 > 0$. The conjugate posterior has the same form, with similar parameters $(\mu_2, \kappa_2, \nu_2, s_2)$ satisfying (see for instance [10])

$$\mu_2 = \frac{\kappa_1}{\kappa_1 + N}\mu_1 + \frac{N}{\kappa_1 + N}\bar{m} \tag{71}$$

$$\kappa_2 = \kappa_1 + N \quad (72)$$

$$\nu_2 = \nu_1 + N \quad (73)$$

$$\nu_2 s_2^2 = \nu_1 s_1^2 + (N-1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1 + N}(\bar{m} - \mu_1)^2 \quad (74)$$

with $\bar{m} = \sum x_i/N$ and $\bar{\sigma}^2 = \sum (x_i - \bar{m})^2/(N-1)$. Computation of $F = F(\mu_1, \kappa_1, \nu_1, s_1; \mu_2, \kappa_2, \nu_2, s_2)$ is similar to the two cases treated above and yields:

$$\begin{aligned} F(\mu_1, \kappa_1, \nu_1, s_1; \mu_2, \kappa_2, \nu_2, s_2) &= -\frac{1}{2}[\log \frac{2\pi}{\kappa_2} + \frac{\kappa_2}{\kappa_1} + \log \frac{\nu_1 s_1^2}{2} - \Psi(\frac{\nu_1}{2}) + \kappa_2(\mu_2 - \mu_1)^2 s_1^{-2}] \\ &+ \frac{\log(\frac{\nu_2}{2})^{\nu_2/2} s_2^{\nu_2}}{\Gamma(\frac{\nu_2}{2})} + (\frac{\nu_2}{2} + 1)[\Psi(\frac{\nu_1}{2}) + \log \frac{2}{\nu_1 s_1^2}] - \frac{\nu_2 s_2^2}{2 s_1^2} \end{aligned} \quad (75)$$

From Equation 75, we can derive the surprise

$$\begin{aligned} S(G_1\Gamma_1, G_2\Gamma_2) &= \frac{1}{2}[\log \frac{\kappa_1}{\kappa_2} - 1 + \frac{\kappa_2}{\kappa_1} + \kappa_2(\mu_2 - \mu_1)^2 s_1^{-2}] + \log \frac{C_1}{C_2} \\ &+ (\frac{\nu_1 - \nu_2}{2})[\Psi(\frac{\nu_1}{2}) + \log \frac{2}{\nu_1 s_1^2}] + \frac{\nu_2 s_2^2 - \nu_1 s_1^2}{2 s_1^2} \end{aligned} \quad (76)$$

Substituting the value of the posterior parameters

$$\begin{aligned} S(G_1\Gamma_1, G_2\Gamma_2) &= \frac{1}{2}\log \frac{\kappa_1}{\kappa_1 + N} + \frac{N}{2\kappa_1} + \frac{\kappa_1 + N}{2}\left[\frac{N(\bar{m} - \mu_1)}{(\kappa_1 + N)s_1}\right]^2 + \log \frac{C_1}{C_2} \\ &+ -\frac{N}{2}[\Psi(\frac{\nu_1}{2}) + \log \frac{2}{\nu_1 s_1^2}] + \frac{(N-1)\bar{\sigma}^2 + \frac{\kappa_1 N}{\kappa_1 + N}(\bar{m} - \mu_1)^2}{2 s_1^2} \end{aligned} \quad (77)$$

For simplicity, we can consider the case where $\mu_1 = \bar{m}$. Then

$$S(G_1\Gamma_1, G_2\Gamma_2) = \frac{1}{2}\log \frac{\kappa_1}{\kappa_1 + N} + \frac{N}{2\kappa_1} + \log \frac{C_1}{C_2} - \frac{N}{2}[\Psi(\frac{\nu_1}{2}) + \log \frac{2}{\nu_1 s_1^2}] + \frac{(N-1)\bar{\sigma}^2}{2 s_1^2} \quad (78)$$

In all cases, for large values of N we always have the approximation

$$S(G_1\Gamma_1, G_2\Gamma_2) \approx \frac{N}{2}[\frac{1}{\kappa_1} + \frac{\bar{\sigma}^2}{s_1^2} + \log \frac{\nu_1 s_1^2}{2\bar{\sigma}^2} - \Psi(\frac{\nu_1}{2}) + \frac{(\bar{m} - \mu_1)^2}{s_1^2}] \quad (79)$$

Surprisee is linear in N with a coefficient that is essentially the sum of the coefficients derived in the unknown mean and unknown variance partial cases.

Acknowledgements

The work of PB is supported by a Laurel Wilkening Faculty Innovation Award and grants from the NIH and Sun Microsystems at UCI.

References

[1] J. Aczel and Z. Daroczy. *On measures of information and their characterizations.* Academic Press, New York, 1975.

[2] P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach.* MIT Press, Cambridge, MA, 2001. Second edition.

[3] J. O. Berger. *Statistical decision theory and Bayesian analysis.* Springer-Verlag, New York, 1985.

[4] R. E. Blahut. *Principles and practice of information theory.* Addison-Wesley, Reading, MA, 1987.

[5] Bela Bollobas. *Random Graphs.* Academic Press, London, 1985.

[6] G. E. P. Box and G. C. Tiao. *Bayesian inference in statistical analysis.* John Wiley and Sons, New York, 1992. (First Edition in 1973).

[7] L. D. Brown. *Fundamentals of Statistical Exponential Families.* Institute of Mathematical Statistics, Hayward, CA, 1986.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, New York, 1991.

[9] R. T. Cox. Probability, frequency and reasonable expectation. *Am. J. Phys.*, 14:1–13, 1964.

[10] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, London, 1995.

[11] I. S. Gradshteyn and I. M. Ryzhik. *Table of integrals, series, and products.* Academic Press, New York, 1980.

[12] S.F. Gull. Developments in maximum entropy data analysis. In J. Skilling, editor, *Maximum entropy and Bayesian methods*, pages 53–71. Kluwer, Dordrecht, 1989.

[13] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995.

[14] E. T. Jaynes. *Probability Theory: The Logic of Science.* 1996. Unpublished.

[15] E.T. Jaynes. Bayesian methods: General background. In J.H. Justice, editor, *Maximum Entropy and Bayesian Methods in Statistics*, pages 1–25. Cambridge University Press, Cambridge, 1986.

[16] G. Jumarie. *Relative information.* Springer Verlag, New York, 1990.

[17] S. Kullback. *Information theory and statistics.* Dover, New York, 1968. (First Edition in 1959).

[18] R. J. McEliece. *The Theory of Information and Coding.* Addison-Wesley Publishing Company, Reading, MA, 1977.

[19] L. J. Savage. *The foundations of statistics.* Dover, New York, 1972. (First Edition in 1954).

[20] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[21] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In B. Hajek and R. S. Sreenivas, editors, *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377. University of Illinois, 1999.

[22] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.

[23] A. Zellner. Optimal information processing and Bayes's theorem. *The American Statistician*, 42(4):278–284, 1998.