Some Applications of Information and Learning to Philosophy...

Or, Barwise Inverse Relation Principle, Bayesian Surprise, Boosting,
and Other Things that Begin with the Letter B

Ely M. Spears and Miguel Aljácen

December 15, 2011

1 Introduction

A central discussion in the philosophy of science is the debate over what information is, what it consists of, and how one can reliably measure or detect it. As with many branches of philosophy, theories of information have been largely influenced by modern mathematical and physical results and in this case Shannon's theory of mathematical communication has proved centrally important. As Weaver put it, however,

> the word information [in Shannon's sense] relates not so much to what you do say, as to what you could say. The mathematical theory of communication deals with the carriers of information, symbols and signals, not with information itself. That is, information is the measure of your freedom of choice when you select a message [15].

Some have remarked that *information* is not perhaps the right term to use when describing Shannon's results. In fact, it was Von Neumann who suggested that Shannon should call his primary mathematical quantity *entropy*:

> You should call it entropy for two reasons: first, the function is already in use in thermodynamics under the same name; second, and more importantly, most people don't know what entropy really is, and if you use the word entropy in an argument you will win every time.[11]

It is therefore no surprise that scholars have found much room to argue over the connections and merits of Shannon's theory in the philosophical setting.

In this paper, we aim to describe a few of the ways that Shannon's theory was used in philosophy and a few of the central debates that prevent it from being more widely applicable. We then introduce some elementary concepts from information theory and offer some additional arguments in favor of their connection to philosophy. First we describe the construction of the Typical Set from Shannon's theory and argue that certain properties of this set have ramifications for the use of Occam's razor in practical settings. We then discuss relative entropy and the recently popularized notion of Bayesian surprise. By following some recent work in convex optimization, we point out that the update procedure found in the popular method of Boosting is really just constrained minimization of Bayesian surprise and discuss interpretations of Boosting weights in this light. Finally, we present the influential result of [7] that strong and weak learning are equivalent in certain settings and investigate some possible implications of this fact for the classical problem of the external world.

1.1 Historical Context

Floridi gives a substantial bibliography [2] showing the convergence of information theories to the issue of being *data based*. To this end, three criteria are proposed for what is known as the *General Definition of Information* (GDI): (1) information

consists of one or more data; (2) the data are well-formed[1]; (3) the data have meaning[2]. The general understanding of data itself appears to be widely debated, but the general notion of a *lack of uniformity* is a consensus view. Such a lack of uniformity could be ontologically basic, or just a product of perceptual differences between physical states or symbols. In any case, we prefer not to speculate on the merits of different definitions of data and instead adopt the broader notion that data are any lack of uniformity that can be exploited to achieve some goal. The general name given to information that arises from data satisfying the conditions (1)-(3) is *semantic content* and we will adopt this term to keep it distinct from information in the Shannon setting.

Since Shannon's entropy deals primarily with the amount of information that one *could* transmit, rather than what useful information might be carried in any particular string of symbols, traditional philosophy of semantic information was not directly amenable to Shannon's theory [1]. Two extremes emerged where at one end of the spectrum, philosophers compared the relationship between semantic information and Shannon's information to the relationship between mechanical engineering and Newtonian physics – the former is essentially over-determined by the latter but we still attribute expressive and creative aspects to the former. The other end took essentially the opposite point of view, that Shannon's theory underdetermines semantic information in much the same way that Newtonian physics can hardly be said to "determine" a tennis match (leaving aside questions of universal determinism). In the next two sections, we explore some of the philosophical ideas that were developed in the wake of Shannon's theory and how these led to the current philosophical status of information theory.

1.2 The Barwise Inverse Relation Principle

When Shannon's theory was new, philosophers tended to believe it offered strong constraints on semantic information, but in the decades since, their position has gradually faded to believing that it only offers weak constraints [1]. In particular, one specific result that motivated skeptical treatment of Shannon's work is called the Inverse Relation Principle (IRP) [4]. This idea roughly states that the probability p of a symbol x is *inversely* related to the amount of semantic information contained in x. The intuition behind this is that the semantic content of an object x (symbol or instance of information) should be determined by the set of states that are inconsistent with the occurence of x. That is, if x is seen to be the case, then an observer gains information that eliminates all states of affairs in which x fails to occur.

To mimic the mathematical language used by Shannon, Barwise expressed this notion analytically as follows. Suppose that for some bit x we have a prior probability $p(x)$ regarding what symbol x will actually be. Then the *semantic content* $CONT(x)$ is given by

$$CONT(x) = 1 - p(x)$$

(i.e. the probability of the complement of the union of sets in which x occurs) and the *informativeness* of x as

$$INF(x) = \log \frac{1}{1 - CONT(x)} = -\log p(x).$$

Barwise summed up this idea with a useful comment: "Whenever there is an increase in available information there is a decrease in possibilities, and vice versa" [5]. This line of thinking is suggestive when considered with questions of free will, but to a mathematician the above symbol pushing may appear trivial at first. The real work comes, however, in constructing or interpreting the probability $p(x)$ in the semantic content setting. In Shannon's setting, this is common knowledge about the occurance of symbols to be used for sending messages, but no attention is paid to the semantic meaning of those messages. When semantic meaning is part of the problem, the important approaches include the modal

[1] This can mean several things in a philosophical setting. The most widely accepted idea is that well-formed data are data that respect the syntax of a particular framework. For example, the rules of 3D perspective might constrain the form that pictorial data can take if it is meant to convey certain information visually. If a certain picture conforms to those expectations then it would be considered a well-formed piece of data, and thus information could consists of that picture, at least in part.

[2] This too can mean many things in a philosophical setting. Like the well-formed criterion, the spirit of this constraint is that the data must obey the semantic rules of a particular framework. Using the same pictorial example as in footnote 1, a drawing of how to use a desk chair might be well-formed if the linear perspective, colors, and lighting are all sensical for expressing a visual image, but if the desk chair is resting on top of a person's head, this would be meaningless in the framework of visual instructions for using the chair. Even this can be disputed, of course, and some have argued that this is circular and assumes a concept of information in order to define information. However, it is still the most widely accepted foundational view for philosophy of information.

approach [6], the systemic approach [4] and the inferential approach. A description of these approaches is beyond the scope of this paper. However, in the next section we present an interesting paradox that arises when considering this problem and discuss the conclusions one is forced to choose from when dealing with inconsistencies in the light of semantic information.

1.3 The Bar-Hillel-Carnap Paradox

A major result noted by Bar-Hillel and Carnap [16] is that if $p(x) = 0$ for some information x, then x has maximum semantic content. In particular, impossible symbols and contradictions maximize semantic content and informativeness (spurring Bar-Hillel and Carnap to quip that such probability-zero objects are "too informative to be true.") Similarly, tautological pieces of information have probability 1 and therefore have no content nor informativeness. This apparently puzzling conclusion that contradictions have maximal semantic content goes by the name Bar-Hillel-Carnap Paradox (BCP).

BCP forces one to address the information content of probability zero events. In decision theory, these events are effectively excluded from consideration when describing information [17, 18] because consistency is a required prerequisite to most of the rest of the theory. This still seems inadequate because one could have a sequence of events, each of which has probability greater than zero, but whose limit is a zero probability event. In this sense, BCP suggests that rational decision making may not behave well under the limit operation. That is, the semantic content of the limit of a sequence of events may not be well-defined even if each of the events in the sequence has well-defined semantic content. If one attempts to *define* the semantic content to be zero for any inconsistent or probability-zero event, as advocated in [21], it seems one still suffers this limit operation problem. In fact, in this case one not only gives up well-behaved limits, but also continuity because semantic content increases indefinitely as probability goes to zero, but then suddenly drops to zero when probability is zero. These mathematical inconveniences are now mostly accepted as unpleasant but unavoidable facts about any quantitative approach to semantic content.

Lastly, it seems reasonable that we want a concept of semantic content that corresponds to what is factually true about the world. Thus, if we want the notion of semantic content to be both factual and also related to the set of possibilities that are excluded, we really do run into a problem with inconsistencies and probability-zero events. Either these events are factual by virtue of having lots of semantic content (and thus we would be forced to believe that inconsistencies *could* be factually true about the world) or else there cannot be a concept of semantic information that adequately corresponds with our notion of factual truth. It is stated well in [6]:

> Assigning less than maximal content to a contradiction would require one to either adopt a (non-semantic) dialetheist position (the world is possibly inconsistent), or to reject the factuality requirement on semantic information.

It appears that this result in particular has caused many professional philosophers of information to regard Shannon's theory as more of a weak set of constraints than a quantitative foundation for semantic content.

These historical connections between the philosophical theory of semantic content and Shannon's information theory are salient, but surely not exhaustive. For more details on these topics and their connection to information theory, see [2]. In the next section we recap some of the basic constructions we need from Shannon's theory for later arguments that the mathematical theory of communication has additional, perhaps more direct, roles in philosophy.

1.4 Elements of Shannon's Theory

The two primitives that we need from Shannon's theory are the *entropy* and *relative entropy*. For a discrete random variable X that is distributed according to probability mass function p on domain Ω, the entropy is given by

$$H(X) = -\sum_{x\in\Omega} p(x)\log p(x) = E_p\left[\frac{1}{\log p(X)}\right].$$

For two probability mass functions P and Q each defined on Ω, the relative entropy or Kullback-Leibler Divergence from P to Q is given by

$$\Delta(P,Q) = E_P\left[\log \frac{P(X)}{Q(X)}\right].$$

These basic concepts are explained more fully in [10] and in particular, Chapter 3 of [10] develops an important property called the Asymptotic Equipartition Property (AEP).

Really, the AEP is nothing more than the Law of Large Numbers. If $(x_1, \cdots, x_n)$ is an n-tuple of points drawn i.i.d. from a fixed probability distribution $P(x)$, then under mild conditions the Law of Large Numbers says that the empirical mean of the tuple of points converges on the average, $E_P[X]$. The AEP just replaces the random variable under consideration with the new random variable $-\log P(X)$. Thus,

$$-\frac{1}{n}\log P(x_1, \cdots, x_n) = -\frac{1}{n}\sum_{i=1}^{n} \log P(x_i) \to -E_P[\log P(X)] = H(X).$$

Motivated by this result, the Typical Set with respect to the distribution P and dimension n with tolerance ϵ is defined as

$$A_\epsilon^{(n)} = \{(x_1, \cdots, x_n) \in \Omega^n | \quad 2^{-n(H(X)+\epsilon)} \leq P(x_1, \cdots, x_n) \leq 2^{-n(H(X)-\epsilon)}\}.$$

This set has interesting properties related to forming a communication code for the data. See Chapter 3 of [10] for all of the properties and their proofs. We'll revisit a few of these properties when we discuss an extension of this set using Kolmogorov complexity.

Recent results in [14] have yielded a particular interpretation of the relative entropy in the case when the distributions P and Q are a posterior and prior belief distribution, respectively. In this case, the relative entropy from the posterior to the prior is called *Bayesian surprise*, although perhaps just *surprise* is a better term because the intuition behind the concept applies equally well outside of the Bayesian setting. For concreteness though, suppose we have some observed data D. Let M be the class of all models we are willing to entertain as explanations for D. Then let $P(M)$ be a prior probability distribution over the choice of explanatory model. Bayes' Theorem then tells us that

$$P(M|D) = \frac{P(D|M)}{P(D)} \cdot P(M).$$

Thus, the relative entropy $\Delta(P(M|D), P(M))$ is in some sense a measure of how much the act of observing the data D has propelled us away from the prior belief $P(M)$. If this is large, we would say that the data D is *surprising* given the way in which it led to a posterior belief that is far away (in the information-theoretic sense) from the prior belief.

In [14], the primary focus is on developing the use of Bayesian surprise as a focus-of-attention mechanism in mammalian brains. Evidence is given that suggests that involuntary movements of human eyes (saccades) are well-predicted by considering the Bayesian surprise of image content with respect to biologically inspired models of image features being exploited in the optic pathway[3].

In the next two sections, we apply the Typical Set and relative entropy / Bayesian surprise to questions regarding Occam's razor.

2 The Typical Set, Entropy, and Kolmogorov Complexity

In [22] Kolmogorov complexity is investigated extensively for its use as a potential criterion identifying correct hypotheses or theories from a pool of options. In the setting of information theory and learning, Kolmogorov complexity shows up under the heading *minimum description length* which is a method for selecting succinct models that describe a set of data.

[3] These include blue-yellow color difference, red-green color difference, luminance, optical flow, and other visual cues believed to play important roles in the early stages of visual processing in mammalian brains. See [14] for more specific details about how a model of Bayesian surprise is developed from these features.

Often it is advised that prediction accuracy should be traded off with simplicity of a model to avoid overfitting[4]. One central question about this use of Occam's razor is whether or not one is justified to believe that simplicity is a reliable indicator of generalization accuracy.

Let X be a random variable with domain Ω distributed according to the computable probability mass function P with entropy $H(X) = -\sum_{x\in\Omega} P(x)\log P(x)$ and let $K(x)$ be the prefix free Kolmogorov complexity. Then, a theorem from [13] says:

$$0 \leq \sum_{x\in\Omega} P(x)K(x) - H(X) \leq c_P$$

where c_P is a constant depending only on the length of the program necessary to compute $P(x)$. Restating this as

$$H(X) \leq E_P[K(X)] \leq H(X) + c_P$$

we can define our own modified version of the Typical Set that is based on the Kolmogorov complexity. Recall that this set was defined for n-tuples of data points, $(x_1, \cdots, x_n)$, where each x_i is assumed to be drawn i.i.d. according to $P(X)$. The condition defining the set $A_\epsilon^{(n)}$ is that

$$H(X) - \epsilon \leq -\frac{1}{n}\log P(x_1, \cdots, x_n) \leq H(X) + \epsilon.$$

Directly utilizing the theorem from [13], we propose the Kolmogorov Typical Set $B_\epsilon^{(n)}$ consisting of n-tuples for which

$$(E_P[K(X)] - c_P) - \epsilon \leq -\frac{1}{n}\log P(x_1, \cdots, x_n) \leq E_P[K(X)] + \epsilon.$$

Because we are enlarging the original Typical Set inequality in each direction, $A_\epsilon^{(n)} \subset B_\epsilon^{(n)}$ and some properties are inherited from the Typical Set.

1. $P(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large, thus $P(B_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large.
2. Following the same derivation as in [10],

$$(1-\epsilon)2^{n(E_P[K(X)]-c_P-\epsilon)} \leq |B_\epsilon^{(n)}| \leq 2^{n(E_P[K(X)]+\epsilon)}$$

By construction then, the Kolmogorov Typical Set has probability very close to 1, all of the elements are nearly equiprobable, and the cardinality of this set is very close to $2^{nE_P[K(X)]}$ (depending on the constant c_P).

Our view is that this newly constructed Kolmogorov Typical Set offers a connection to Occam's Razor by virtue of the term c_P. Recall that this term is a constant based on the Kolmogorov complexity of the underlying distribution P (see [13] for specific details on the exact constant). Here is the setting we propose: a person is obtaining data that is produced by some data generating process that obeys a fixed but unknown probability law P. One of the chief goals of inductive reasoning is to learn P by studying the example data believed to be produced by P. In particular, as we obtain more samples ($n \to \infty$) we believe that our understanding of P should become more accurate.

But in terms of models, P is always something that has to be hypothesized and tested. Thus, scientists wield some control over selecting hypothetical data generating processes that have simple descriptions (small c_P) or complex descriptions (large c_P). This has a direct impact on the set $B_\epsilon^{(n)}$. For complex models, c_P will be larger, hence property 2 from the list above tells us that for any fixed tolerance level ϵ and any fixed number of data points n, $B_\epsilon^{(n)}$ will be a larger set if complex models are assumed over simple models. From property 1, we know that the probability of a high-dimensional collection of data being a member of this typical set is very high and that members of this set tend to have complexity very close to the average complexity $E_P[K(X)]$.

[4]Overfitting is the phenomenon where a model "artificially" produces high accuracy by virtue of having enough free parameters to account for the data by "brute force" in some sense. A simpler model that performs nearly as well might be preferred because the simpler model yields greater generalization accuracy or explanatory power.

For the moment, let us take it as a premise that we can use the Typical Set as a proxy for falsifiability. If a probability model accurately describes the world, we would like it to be difficult to falsify. One way of interpreting falsification is that we observe data which appear atypical under our current model. In our setting, this directly means that if we have chosen an accurate probability model P, then most high-dimensional data sets should have close to the same Kolmogorov complexity, namely the expected complexity under P. If we have chosen our probability model P^* different from the true probability P, then one way to view falsifiability is that we will see high-dimensional collections of data with a different average complexity: training data will not belong to the Kolmogorov typical set as they should.

This gives an additional reason for favoring Occam's razor. Choosing simpler data generation models P implies that the Kolmogorov Typical Set will be more tightly centered around the expected Kolmogorov complexity (by virtue of the smaller constant c_P). This will make it comparatively easier to falsify than a complex model P^*, which might give the appearance of enclosing most data in the Typical Set merely due to the parochial fact that c_{P^*} is much larger than c_P. Of course, since c_P is not computable, this does not give rise to a practical method for determining "objective" simplicity of a model. But it does give a theoretical reason for believing that complex models have an intrinsically larger Kolmogorov Typical Set, and hence will be harder to falsify just by virtue of their complexity. If a simple theory turns out to be inaccurate, it will at least be easier to falsify than a comparable but more complex theory that might suffer similar inaccuracies.

3 The Information Geometry of Boosting

Boosting [8] is a mathematical technique for improving the quality of classification of a modestly successful classifier[5]. The modestly performing classifier is known as a *weak learner* and can consist of any learning algorithm that non-trivially performs better than chance. Boosting iteratively creates new instances of the weak learner that are successively better able to classify the tough corner cases that stumped its immediate predecessor. Though there are many variants of the boosting algorithm, AdaBoost is the most ubiquitous.

AdaBoost Algorithm (for binary classification case):
Initialize: Training Data $\{x_1, \cdots, x_m\}$, Training Labels $\{y_1, \cdots, y_m\}$, Weights $D_1(i) = 1/m$, and Iterations T.
Note: Each datum $x_i \in \mathbb{R}$, each label $y_i \in \{-1, 1\}$.

`For` $k := 1$ `to` T

1. Train weak learner/classifier C_k using samples drawn from training data according to distribution $D_k(i)$.
2. Set E_k to the weighted training error of C_k on the data (using weights $D_k(i)$).
3. Set $\alpha_k = \frac{1}{2} \log\left[\frac{1 - E_k}{E_k}\right]$ (see below for details on this choice for α_k).
4. Update $D_{k+1}(i) = D_k(i) \cdot \exp\{-\alpha_k y_i h_k(x_i)\}$, (where $h_k(x)$ is the class label assigned to x by C_k).
5. Set $Z_k = \sum_{i=1}^{m} D_{k+1}(i)$ (the normalizing constant).
6. Normalize the new weights to $D_{k+1}(i) \leftarrow \frac{D_{k+1}(i)}{Z_k}$.

`return` C_k and α_k for $k \in \{1, \cdots, T\}$.

When presented with a new datum x to classify, the composite classification rule resulting from this algorithm is to compute the sign of $G(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$. $G(x) \in [-1, 1]$ is known as the *margin* and its absolute value yields a confidence in the classification. One of the main reasons why boosting was adopted is that a useful upper bound on the composite

[5]The term *hypothesis* is more commonly used in place of *classifier* in the computer science literature, while classifier is more common in engineering literature. We stick with classifier but the terms can be used essentially interchangeably.

probability of error E was found [8]. The result states that if the weak learner C_k resulting from iteration k of the algorithm has an error rate of E_k then by the assumption that C_k does better than chance, we can write $E_k = 0.5 - G_k$ for some positive number G_k. The bound obtained then states[6]

$$E = \prod_{k=1}^{T} \left[2\sqrt{E_k(1-E_k)} \right] \leq \exp\left\{ -2 \sum_{k=1}^{T} G_k^2 \right\}.$$

Thus, the training error decreases exponentially in the sum of the amounts G_k that each weak learner performs better than chance.

In fact, this upper bound on the training error can be written in another way that illuminates the optimization procedure going on behind the scenes and relates the boosting procedure to information theory in a suggestive way. As shown in the algorithm above, the weights are updated as

$$D_k(i) = \frac{D_{k-1}(i) \cdot \exp\left\{ -\alpha_{k-1} y_i h_{k-1}(x_i) \right\}}{Z_k}$$

but we can recursively expand the term $D_{k-1}(i)$ which yields a product of exponentials in the numerator and a product of the normalizing terms in the denominator:

$$D_k(i) = \frac{\exp\left\{ -y_i \sum_{t=1}^{k-1} \alpha_t h_t(x_i) \right\}}{m \prod_{t=1}^{k-1} Z_t}.$$

For a single datum x_i from the training data, we can trivially bound the probability of misclassification as:

$$P(\text{sgn } G(x_i) \neq y_i) \leq \exp\left\{ -y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) \right\}.$$

Then the expected probability of misclassification assuming a uniform draw from the training data is:

$$\begin{aligned}
\frac{1}{m} \sum_{i=1}^{m} P(\text{sgn } G(x_i) \neq y_i) &\leq \frac{1}{m} \sum_{i=1}^{m} \exp\left\{ -y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) \right\} \\
&= \sum_{i=1}^{m} \left(\prod_{t=1}^{T} Z_t \right) \cdot \frac{\exp\left\{ -y_i \sum_{t=1}^{T} \alpha_t h_t(x_i) \right\}}{m \prod_{t=1}^{T} Z_t} \\
&= \sum_{i=1}^{m} \left(\prod_{t=1}^{T} Z_t \right) \cdot D_{T+1}(i) \\
&= \prod_{t=1}^{T} Z_t.
\end{aligned}$$

The last inequality follows from the fact that the weights D_{T+1} are assumed to be a valid probability distribution. This result shows that the weights computed in AdaBoost are a function of the successive normalizing constants Z_t each of which only depends on the voting parameter α_t. Since the expected classification error is upper bounded by this quantity, one popular approach is to minimize the choice of $Z_t(\alpha_t)$ at each iteration and in fact this is how the value for α_t is obtained for use in the AdaBoost algorithm above.

[6] This notation is due to the derivation in Chapter 9 of [9]

The interesting point was expressed in [12] where boosting was recast as entropy projection. There it is shown that the primal approach to achieving optimal update weights $D_t(i)$ can be translated into the dual approach which involves constrained optimization of relative entropy. In this setting, [12] shows that the following equality holds:

$$\min_{\substack{\vec{D}_t(i)\in P_m \\ \vec{D}_t(i)\cdot u_{t-1}=0}} \Delta\left(\vec{D}_t(i), \vec{D}_{t-1}(i)\right) = \max_{\alpha_t\in\mathbb{R}} \quad [-\log Z_t(\alpha_t)]$$

where u_t is a vector that contains in component i the binary classification assigned to datum x_i by the weak learner C_t, and the notation $\vec{D}_t(i)$ indicates the vector of weight coefficients indexed by i (rather than just the secific weight for component i).

More simply, the above result shows that instead of choosing α_t to minimize the normalizing constant Z_t (or, equivalently maximize its negative log) and making the corresponding choice for the weights $D_t(i)$, it is equivalent to think of the weight selection in AdaBoost as minimizing the relative entropy Δ between the weights at iteration t and the weights at iteration $t+1$, subject to a constraint. Both this way of looking at the problem and the constraint are of philosophical interest.

Firstly, the constraint $\vec{D}_t(i) \cdot u_{t-1} = 0$ intuitvely means that we seek a new set of weights that is uncorrelated with the weak learner C_k at the current iteration. The constraint condition says that the expected classification of the current learner under the new weights should be zero, giving us no information about what the past weak learner knew about the data set. In this sense, AdaBoost forces the next weak learner to *learn something new* about the training data. In [12], this constraint is generalized to enforce zero correlation with *all* of the past weak learners. This approach is referred to as the *totally corrective* approach. While it sounds appealing to include constraints on all past learners, several problems are mentioned. First, you can run into situations where not all of the imposed zero-correlation constraints can be simultaneously satisfied and so there is a problem of how to relax them. Additionally, computational methods are much less efficient and in practice one must weigh this fact against performance gains.

Secondly, and more importantly, among all possible weightings on the training data, this approach seeks to minimize the relative entropy between the new weights and the old weights. If we use the language of posteriors and priors, then Bayesian surprise is immediately applicable. The old weight vector, $\vec{D}_k(i)$ can be viewed as the prior belief about the difficulty for learning, or salience, of a particular datum of the training set. In the beginning we assign this the maximum entropy prior $D_1(i) = 1/m$. As the algorithm progresses, *we seek to update our prior belief about the distribution of difficulty-to-learn in the least surprising way.*

This can immediately be understood in the light of Occam's razor. Starting from a maximum entropy prior, what are we to believe about the training data set? Well, on one hand we witness the performance of a weak learner, so we know something about which examples are hard to learn and which aren't. So we want to shift the weights around to re-prioritize our beliefs about the data set and we want the new weight to be uncorrelated with the most recent learner. But any choice we make in addition to this constraint will necessarily impose unjustified assumptions about the structure of the training data. To prevent this from happening, we select those weights which are least surprising (literally, in much the same spirit as Hume's advocacy that science should be "the least bit surprising" in his essay *An Enquiry Concerning Human Understanding*). The equivalence of the primal and dual approaches in this setting very straightforwardly suggests that minimizing training error is equivalent to selecting the least surprising distribution that will learn something new.

A full survey of the empirical success of AdaBoost and its success in reducing generalization error is beyond the scope of this paper. But one can find a summary of such results in [23], and so it is not a stretch to view AdaBoost as bolstering a particular use of Occam's razor. In the next few sections, we examine some additional ramifications of boosting, and in particular the limitations of weak learners, which lead to further philosophical speculation.

4 Weak Learnability and the Problem of the External World

The *problem of the external world* refers to a central issue in the epistemology of perception: how can we justify our beliefs about the physical world as conveyed by our perceptual faculties? [3] Contemporary non-skeptical views on this problem[7] fall on two camps: *representationalism* or *representative realism*, and *direct realism*[8]. Representationalists argue that sense-data arriving from experience form *representations* of the subjects of perception via some mechanism of inference. We can rely on such representations to form our beliefs about the external world, they argue, insofar as we are prepared to argue about the soundness of the perceptual inference mechanism that creates them. Representationalists point out at the *fragmentariness* of our immediate sensory experience (e.g. the sequence of frames of a tracking eye), together with its seeming *orderliness* of its overall character (e.g. the composite image of a scene), as evidence of an underlying mechanism of representation that is accurately conveying reality, and so we are warranted to rely on its representations. Adherents to this view, however, concede that there is a difficult circularity in this argument, in that it argues for the reliability of our perceptual faculties by appealing to *another* product of those very own faculties: namely, various features of our sensory experience.

Direct realists, on the other hand, deny the need to argue about the soundness of our means of perception in the first place. Instead, they contend that we have direct access to the objects of the external world, as manifest by the apparent immediacy of our sensory experience: when something becomes present in our mind, we become aware of it without any apparent conditioning, regardless of whatever mechanisms are involved in the formation of such perception. Such immediacy, they argue, is enough to warrant our belief in our perceptions. However, by appealing to a particular aspect of our sensory experience in order to justify its reliability, the direct realist account of perception fails to surpass the circularity problem that plagues the representationalist position. The issues point at a deeper difficulty acknowledged by adherents to both doctrines: there does not seem to be a non-circular way for arguing about the reliability of our perception on the basis of the subjective aspects of our perceptual experience alone, even though those very subjective experiences are our primal source of evidence about the nature of perception itself.

Earlier we mentioned the notion of *boosting* in computational learning theory, introduced by Schapire in 1989 [7]. This provides a speculative, potentially non-circular way for formulating and answering questions about the reliability of perception. First let us recall the basic definitions of the PAC framework [20]: a *concept* $c : \mathcal{X} \to \{0,1\}$ is a Boolean function defined over an instance space $\mathcal{X} = \{0,1\}^n$. A *concept class* $\mathcal{C}$ is a family of concepts. A *learner* L is an algorithm with access to an oracle $EX(c, \mathcal{D})$ which, given a concept c and a fixed unknown distribution $\mathcal{D}$, chooses an instance v from $\mathcal{X}$ according to $\mathcal{D}$ and returns a pair $(v, c(v))$ —an *example*. The learner's goal is to feasibly approximate some target concept c using the examples provided by $EX(c, \mathcal{D})$ by outputting a *hypothesis* $h : \mathcal{X} \to \{0,1\}$ from some arbitrary (but polynomial-time evaluable) *hypothesis class* $\mathcal{H}$. For a given hypothesis h for a target concept c under $\mathcal{D}$, we define its *error rate* $\mathrm{err}(h)$ as the probability that h misclassifies any given instance $v \in \mathcal{X}$, that is $\mathrm{err}(h) = \Pr_{v \in \mathcal{D}}[h(v) \neq c(v)]$. If the learner L is able to feasibly produce a hypothesis h for some $c \in \mathcal{C}$ with arbitrarily small error with arbitrarily high probability, we say $\mathcal{C}$ is *strongly learnable* (or *learnable*). More formally, an algorithm L *strongly* learns a concept class $\mathcal{C}$ using hypothesis class $\mathcal{H}$ if, for any $0 < \epsilon \leq 1$ and $0 < \delta \leq 1$, any concept $c \in \mathcal{C}$ of size s, any distribution $\mathcal{D}$, and access to an examples oracle $EX(c, \mathcal{D})$, L runs in time polynomial in s, $1/\epsilon$ and $1/\delta$, and outputs hypothesis $h \in \mathcal{H}$ with error rate at most ϵ, with probability at least $1 - \delta$.

In 1989, Kearns and Valiant [19] introduced a variant of this criterion, called *weak learnability*: consider a learning algorithm L' that instead of producing hypotheses with arbitrary accuracy, only produced hypotheses h with error slightly better than random guessing. More formally [20], an algorithm L' *weakly* learns a concept class $\mathcal{C}$ using hypothesis class $\mathcal{H}$ if there exists a polynomial $p(\cdot, \cdot)$ such that, for any concept $c \in \mathcal{C}$ of size s, any distribution $\mathcal{D}$, and all $0 < \delta \leq 1$, given access to an examples oracle $EX(c, \mathcal{D})$, L runs in time polynomial in s and $1/\delta$, and outputs hypothesis $h \in \mathcal{H}$ with error at most $1/2 - 1/p(n, s)$, with probability at least $1 - \delta$.

A natural question to ask is whether both criteria are equivalent. For learning under a specific distribution (as opposed to an arbitrary distribution), both criteria are *not* equivalent: for instance, monotone boolean formulae are weakly, but not

[7] More precisely: contemporary *internalist*, non-skeptical views. Externalist (reliabilist) accounts of the problem of the external world only concern themselves with the question of the 'objective' reliability of the mechanism of perception, and do not see the need to justify our perceptual beliefs in the first place. We only consider here the more interesting *internalist* accounts of the problem.

[8] The literature often distinguishes a third, historically prior view: *phenomenalism*. However, contemporary authors consider it largely untenable, and lies outside the scope of this paper [3].

strongly learnable [20]. However, in 1989 Schapire [7] proved the opposite: for distribution-free learning, he described a feasible mechanism to 'boost' the accuracy of a weak learner arbitrarily. The key insight is to exploit the distribution-free property of weak learnability: since the weak learner performs slightly better than random under any distribution, one can obtain extra accuracy by 'forcing' the weak learner to re-classify wrongly misclassified examples. One can then combine all such predictions into a single prediction rule, which results in a modest increase in accuracy (as shown below). The method can be then applied recursively in order to attain any arbitrarily-low error.

For simplicity, consider a weak learner L that *always* outputs hypothesis h for some concept $c \in \mathcal{C}$ under distribution $\mathcal{D}$ with some *fixed* error rate α and with high probability[9]. Now consider the following procedure [24]:

1. Run L with access to the examples oracle $EX(c, \mathcal{D})$. Let h_1 denote the output hypothesis; and let $\mathcal{S}_I$, $\mathcal{S}_C$ denote the sets of examples labeled incorrectly and correctly by h_1, respectively.
2. Run L again with access to an oracle $EX(c, \mathcal{D}_2)$, where $\mathcal{D}_2$ is a distribution such that, *ceteris paribus*, drawing an example from either $\mathcal{S}_I$ or $\mathcal{S}_C$ is equally likely[10]. Let h_2 denote the output hypothesis of this run.
3. Run L again with access to an oracle $EX(c, \mathcal{D}_3)$, where $\mathcal{D}_3$ is a distribution filtered such that it only contains examples v where $h_1(v) \neq h_2(v)$. Let h_3 denote the output hypothesis of this run.
4. Output as hypothesis h' for $c \in \mathcal{C}$ on $\mathcal{D}$ the majority vote of all three hypotheses: $h = \text{majority}(h_1, h_2, h_3)$.

Informally, on each ulterior simulation, this procedure forces L to yield new information on the existing data. Note that the distribution $\mathcal{D}_2$ is constructed so that h_1 performs as random choice. Therefore, since L produces hypotheses with non-negligible edge over random choice under arbitrary distributions, this distortion forces L to improve on h_1 in step 2. The purpose of $\mathcal{D}_3$ is also to uncover new information, this time by forcing L to learn from the examples on which h_1 and h_3 disagree.

Now, what is the accuracy gain obtained by this procedure? We have the following key result:

Lemma 1 (Schapire [7])
Let $g(\alpha) = 3\alpha^2 - 2\alpha^3$, and let learner L and hypothesis h be as defined above. Then $\text{err}(h) \leq g(\alpha)$.

Lemma 1 demonstrates that we can *always* improve the accuracy of a weak learning algorithm by reducing the error from α to $g(\alpha)$. Let γ denote the advantage or 'edge' of the weak learner over random guessing: $\alpha = 1/2 - \gamma$. Then, after invoking the boosting procedure, it is straightforward to verify that the edge γ' of the output hypothesis is $\gamma' = \gamma(\frac{3}{2} - 2\gamma^2)$, a clear improvement. We can then invoke the procedure recursively until attaining an arbitrarily high accuracy ϵ. Moreover, as the growth rate of the edge suggests, the size of the resulting ternary tree is a function polynomial on $1/\gamma$ and $\log(1/\epsilon)$ [24]; and the running time required to achieve arbitrarily high accuracy ϵ with probability $1 - \delta$ is polynomial in $1/\epsilon$ and $1/\delta$, and so constitutes a proper strong learner [20].

The relevance of the equivalence of strong and weak learnability to the problem of the external world is that it describes a plausible setup in which we can justifiably rely on our subjective perceptual beliefs about the external world in a non-circular way. If we are prepared to argue that our mechanisms of perception are at least *minimally* (but non-trivially) informative about the external world, then the equivalence of strong and weak learnability describes a large class of situations in which our composite representations of reality are *provably*[11] approximately correct with high confidence, despite the acknowledged subjectivity of our perceptual faculties. We can then justifiably rely on our experience of reality on the same grounds that we can justifiably rely on the soundness of induction[12].

[9] Boosting the reliability δ of a weak learner L is easily achieved by repeatedly simulating L on independent samples from $EX(c, \mathcal{D})$ for a number of times, until obtaining the desired reliability. It can be easily shown that one can increase confidence from $1 - \delta_0$ to $1 - \delta$ by simulating L a number of times bounded by $O(\log(1/\delta)/(1 - \delta_0))$.

[10] We can construct an oracle for $EX(c, \mathcal{D}_2)$ in several ways, for instance: on input $(c, \mathcal{D}, h_1)$, we flip a coin and output the first example from $EX(c, \mathcal{D})$ that h_1 classifies correctly (if heads) or incorrectly (if tails).

[11] Pun intended.

[12] In fact, one of the earliest (and most radical) critiques of representative realism in the epistemology of perception was articulated by Hume (1739-40) on the grounds of the untenability of induction: given that our only source of information about the workings of our perceptual faculties is our experience of the external world itself, since we cannot rely on inductive inference, then we cannot *ever* make any justifiable claims about the reliability of perception!

The assumption of weak learnability, however, is not trivial. Crucially, it relies on the *distribution-free* property of PAC learning: recall that a weak learner is guaranteed to offer an arbitrarily small edge over random choice for examples drawn *from a fixed arbitrary distribution.* This assumption is stronger than it looks: most learning algorithms known today only work for a few (albeit large) classes of distributions, and so most positive learnability results only hold for PAC learning under specific distributions such as the uniform–in which case the strong and weak learnability equivalence does not hold anymore. It is unclear how to even approach the weak learnability properties of our perceptual faculties, or for that matter whether the external world is PAC learnable at all; yet this setup offers a structured way to consistently formulate and examine these fuzzy aspects of the problem of perception.

5 Conclusions

Our efforts focused on providing context for the philosophy of information and paradoxes such as the Bar-Hillel-Carnap Paradox that lead to the modern view that Shannon's theory underdetermines the concept of semantic information. We then described the Typical Set and the extension to the Kolmogorov Typical Set by utilizing a result from [13]. If one accepts the premise that membership in the Kolomogorov Typical Set is a proxy for the falsifiability of a hypothesized probability model, then the fact that the Kolmogorov Typical Set will be larger for more complex models suggests that simpler theories should be favored on the grounds that they would be more genuinely falsifiable, echoing views of Popper and Deutsch. We also illustrated how the optimization process described by the AdaBoost algorithm directly relates the minimization of training error with the selection of the least surprising weights. From the perspective of Occam's razor, this suggests that seeking beliefs about observed data that impose no unjustified criteria really do yield beliefs with lower error.

Finally, we discussed the classical problem of the external world and whether any non-circular basis can be established for trusting one's perceptual faculties to produce true conclusions about the world. One only needs to be willing to believe that perceptual systems are a form of weak learning, and then the result of [7] suggests that boosting perceptual systems can lead to arbitrarily accurate depictions of the world. In this sense, the perceptual faculties act like a weak proxy to the true external world. It is interesting to ask whether one can realistically believe that human perceptual systems resulting from evolution can be understood as weak learners, but at least this framework removes the need for the perceptual systems themselves to be strong learners.

References

[1] L. Floridi. "Semantic Conceptions of Information." *The Stanford Encyclopedia of Philosophy.* Retrieved from file on 12/13/2011. http://plato.stanford.edu/entries/information-semantic/.

[2] L. Floridi. "Is Information Meaningful Data?" *Philosophy and Phenomenological Research*, 70(2): 2005.

[3] L. BonJour. "Epistemological Problems of Perception." *The Stanford Encyclopedia of Philosophy.* Retrieved from file on 12/13/2011. http://plato.stanford.edu/entries/perception-episprob/.

[4] J. Barwise and J. Seligman. *Information Flow: The Logic of Distributed Systems*, Cambridge: Cambridge University Press. 1997.

[5] J. Barwise. "Information and Impossibilities." *The Notre Dame Journal of Formal Logic*, Vol. 38, No. 4, 1997.

[6] P. Allo. "Formalising Semantic Information: Lessons from Logical Pluralism." *Computation, Information, Cognition: The Nexus and the Liminal.* G. Dodig Crnkovic and S. Stuart eds. Cambridge Scholars Publishing: 2007.

[7] R. Schapire. "The strength of weak learnability." *Machine Learning*, 5(2):197-227, 1990.

[8] Y. Freund and R. Schapire. "A short introduction to boosting." *Technical report.* 1999.

[9] R. Duda, P. Hart, and D. Stork. *Pattern Classification*, 2 ed. John Wiley and Sons. 2001.

[10] T. Cover and J. Thomas. *Elements of Information Theory*, 2 ed. John Wiley and Sons. 2006.

[11] A. Golan. "Information and Entropy Econometrics – Editor's View", *Journal of Econometrics*, 107(12): 115. 2002.

[12] J. Kivinen and M. K. Warmuth. "Boosting as entropy projection." *12th Annual Conference on Computational Learning Theory*, pp. 134-144, ACM, 1999.

[13] T. Cover, P. Gacs, and R.M. Gray. "Kolmogorov's contributions to information theory and algorithmic complexity." *The annals of probability*, Vol. 17, No. 3, pp. 840-865. 1989.

[14] P. F. Baldi and L. Itti. "Of bits and wows: A Bayesian theory of surprise with applications to attention," *Neural Networks*, Vol. 23, No. 5, pp. 649-666, 2010.

[15] W. Weaver. "The Mathematics of Communication", *Scientific American*, 181(1): 1115. 1949.

[16] Y. Bar-Hillel and R. Carnap. "An Outline of a Theory of Semantic Information," *Technical Report*. 1953. Reproduced in Y. Bar-Hillel. *Language and Information: Selected Essays on Their Theory and Application*. Reading, MA. London: Addison-Wesley. 1964.

[17] *Philosophical Aspects of Information Systems*. 1 ed. R. Winder, S. Probert, and I. A. Beeson eds. Taylor and Francis Ltd. London: 1997.

[18] D. Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. Cambridge: 2003.

[19] M. Kearns and L. Valiant. "Cryptographic limitations on learning boolean formulae and finite automata." *J. ACM*, 41(1):67-95, 1994.

[20] M. Kearns and U. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, 1994.

[21] J. Mingers. "The Nature of Information and Its Relationship to Meaning." Reproduced in [17]. 1997.

[22] R. de Wolf. *Philosophical Applications of Computational Learning Theory*. Master's thesis. Retrieved from file on 12/13/2011. http://homepages.cwi.nl/∼rdewolf/publ/philosophy/phthesis.pdf.

[23] E. Bauer and R. Kohavi. "An empirical comparison of voting classification algorithms: bagging, boosting, and variants." *Journal of Machine Learning* Vol. 36, Issue 1-2, 1999.

[24] A. Blum. "Boosting I: weak vs. strong learning, basic issues." *Online lecture post.* 2009. Retrieved from file on 12/13/2011. www.cs.cmu.edu/∼avrim/ML09/lect0209.txt.