Notes on ensembles as a model of theory choice

by Duncan K. Foley
Department of Economics
New School for Social Research
6 East 16th St, New York, NY 10003
foleyd@newschool.edu
External Professor, Santa Fe Institute

These notes are a continuation of a long and productive conversation with David Oliver on the sources of randomness in the natural world. I'd like to acknowledge helpful conversations with Murray Gell-Mann and Seth Lloyd, and helpful comments on earlier versions from Gregor Semeniuk and Ellis Scharfenaker.

This paper examines the foundations of the concept of effective complexity proposed by Murray Gell-Mann and Seth Lloyd using the methods of Bayesian inference. Given a data string, \( x \in \{0, 1\}^\omega \) of length \( |x| \), an ensemble, \( \mathcal{E} = \{X_\omega, \omega_\omega\} \), assigns a frequency, \( \mathcal{E}(x') \) to strings \( x' \) with \( |x'| = |x| \) of the same length as the data string. Programs to compute ensemble frequencies partition (or coarse-grain) the set of strings into exclusive subsets

\[
X_\omega = \{X_\omega^1, \ldots, X_\omega^k\},
\]

and assign a coarse-grained frequency distribution, \( \omega_\omega = \{\omega_\omega^1, \ldots, \omega_\omega^k\} \).\( \omega_\omega^j \geq 0, \sum_j \omega_\omega^j = 1 \) to the subsets. Let \( j(x') \) be the index of the subset containing the string \( x' \). The ensemble frequency of every string \( x' \in X_\omega^j \) is \( \mathcal{E}(x') = \omega_\omega^j \frac{|x'|}{|X_\omega^j|} \), where \( |X_\omega^j| \) is the number of strings in the subset \( X_\omega^j \). According to Bayes’ theorem, given a prior assignment of probabilities over a class of ensembles, such as the class of computable ensembles, \( P[\mathcal{E}] \), the log posterior probability of any ensemble given the data string is \( \log[P[\mathcal{E}] + \log[P(x | \mathcal{E}] \), the sum of the log of the ensemble’s prior probability and the log of the data string’s conditional probability given the ensemble (its likelihood in Bayesian terminology). The ensemble specifies the location and coarse-grained frequency of the subset containing the data string, \( \omega_{\mathcal{E}, j(\omega)} \), so we have \( \log[P(x | \mathcal{E})] = -\log[\frac{|x'|}{|X_\omega^j|}] \). An ensemble’s

Kolmogorov algorithmic information content (AIC), \( K[\mathcal{E}] \), is the length of the shortest program string that leads a prefix-free universal Turing machine to compute the ensemble frequencies when the string and a precision \( n \) are presented to it as input. With a universal prior over ensembles that is inversely proportional to Kolmogorov algorithmic information content, \( \log[P[\mathcal{E}] = -K[\mathcal{E}] \), the log posterior probability of an ensemble given the data string is \( -K[\mathcal{E}] + \log[P(x | \mathcal{E}] \), the sum of the negative of the ensemble AIC and the frequency of the data string under the ensemble; this sum is also the negative of the length of the two-part code for the data string using the ensemble as the model. Energy-like constraints may limit the set of admissible ensembles. The AIC of an ensemble is the sum of the AIC of the partition and the bits required to specify the coarse-grained frequencies, \( K[\mathcal{E}] = K[X_\omega] + K[\omega_\omega] \), the total information of the ensemble. The term \( K[\omega_\omega] \) depends on the prior distribution over frequency distributions that corresponds to the coding system used to specify the frequency distribution. If, for example, the prior probability of a coarse-grained frequency distribution is inversely proportional to its entropy relative to the uniform distribution, the Shannon code length for a coarse-grained frequency distribution is \( K[\omega_\omega] = \log[\omega_\omega] - H[\omega_\omega] \). With this coding of ensemble frequencies the log posterior probability of an ensemble is \( -K[X_\omega] - \log[\omega_\omega] + H[\omega_\omega] - \log[\frac{|x'|}{|X_\omega^j|}] \). For any partition of the strings, with this particular coding of frequency distributions maximization of the posterior probability leads to a maximum entropy coarse-grained frequency distribution subject to the constraints. Other coding systems can result in minimum entropy frequency distributions or frequency distributions that maximize or minimize other entropy-like quantities. In general partitions that put the data string in a small subset require more subsets, a longer program, and a lower coarse-grained entropy to meet the constraints.

The problem

An arbitrary object can be represented as a finite (binary) data string, \( x \), in the set of all finite strings, \( \{0, 1\}^\omega \). Every string in this countably infinite set has a finite length denoted \( |x| \). An ensemble is a computable function, \( \mathcal{E} \), that assigns frequencies, \( \mathcal{E}(x') \geq 0 \), to each string, with \( \sum_{x' \in \{0,1\}^{|x|}} \mathcal{E}(x') = 1 \). The Kolmogorov algorithmic information content (AIC) of an ensemble, \( K[\mathcal{E}] \), is the length of the shortest program, which, when supplied as input to a (prefix-free) Universal Turing Machine (UTM) along with a string and a precision \( n \), halts with \( n \) bits of \( \mathcal{E}(x) \) on its tape. (See Gell-Mann and Lloyd.)
The problem of choosing an ensemble to represent a given string is an abstract model of theory choice in science. The given string, \( x \), stands for the data we have about the world, and the ensemble \( \mathcal{E} \) stands for a theory that explains the data.

These notes consider this problem from the point of view of Bayesian inference with a prior probability over ensembles, \( P(\mathcal{E}) \), which represents the methodological and philosophical principles, such as parsimony, on which theory choice rests.

**Bayesian framework**

One way of associating ensembles with a given string is to use Bayes’ Theorem to assign posterior probabilities to a set of hypotheses on the basis of observed data. The Bayesian (or inverse probability) method, which Laplace and Jaynes championed and Cox showed is the only consistent system of probabilistic inference, exploits the identity:

\[
P(H \mid D) = \frac{P(H \mid D) P(D)}{P(D)}
\]

where \( P(H) \) is the prior probability assigned to each hypothesis in the absence of any conditioning data, \( P(D \mid H) \) is the likelihood, the conditional probability of the data on the basis of the hypothesis, \( P(H \mid D) \) is the posterior probability of the hypothesis given the data, and \( P(D) \) is a normalizing factor, sometimes called the evidence, equal to the probability of the data under the whole collection of hypotheses. Given \( H \), the likelihood, \( P(D \mid H) \), is a normalized probability distribution over possible data observations. Taking \( D \) as given, on the other hand, the likelihood is an unnormalized weighting of the set of hypotheses.

In the context of the version of algorithmic information theory considered here, the string \( x \) plays the role of the data \( D \), and the ensemble \( \mathcal{E} \) plays the role of the hypothesis, \( H \).

The advantage of working with the Bayesian formalism is that it assigns a posterior probability to every ensemble for which the likelihood of the data string is positive, instead of singling out one ensemble to represent the string. Ensembles which give a very low but nonzero frequency to the data string in question will be assigned a low, but not zero, posterior probability by the Bayesian formalism.

The Bayesian setting for this problem also underlines the central point that typically two normalized distributions are involved in this scenario. Any given ensemble, \( \mathcal{E} \), assigns frequencies to the strings and to the coarse-grained partition. The prior and posterior probability distributions, \( P(\mathcal{E}) \) and \( P(\mathcal{E} \mid x) \), on the other hand, assign probabilities to the ensembles themselves. I will reserve the term “probability” for the Bayesian probabilities \( P(\cdot) \), and use “frequency”, for the ensemble distribution, \( \mathcal{E} \).

Given any prior \( P \) over ensembles, and a data string \( x \), the Bayesian formalism assigns a posterior probability to each ensemble on the basis of the likelihood of the string in the ensemble, \( P(x \mid \mathcal{E}) \):

\[
P(\mathcal{E} \mid x) \propto P(x \mid \mathcal{E}) P(\mathcal{E})
\]

It is simpler to work with the logarithm of the posterior probability, which can be written in the following form, where \( C = -\log[P(x)] \) is the logarithm of the normalizing constant \( P(x) \).

\[
\log[P(\mathcal{E} \mid x)] = \log[P(\mathcal{E})] + \log[P(x \mid \mathcal{E})] + C
\]

The two-part code transmits the data string by first transmitting a description of the ensemble (serving as a model for the data), which takes \(-\log[P(\mathcal{E})]\) bits using the Shannon code based on the prior probabilities, and then transmitting a Shannon code word for the data string based on the frequency assigned to it by the ensemble, which requires \(-\log[P(x \mid \mathcal{E})]\) bits. Thus the first two terms of the log posterior probability are the negative of the length of the two-part code to transmit the string \( x \) using the ensemble \( \mathcal{E} \) as the model. The posterior favors ensembles that provide a short two-part code length, thus incorporating the minimum description length principle of Rissanen and Wallace and Boulton, since an ensemble is a model in Rissanen’s terminology, and the log prior probability of the ensemble can be interpreted as its description length.

**Constraints**

The context in which data is collected and encoded can be represented by constraints on the ensembles to be considered. (These constraints could be incorporated in the prior probability over ensembles, but it is useful to make them explicit.) In general the constraints can be represented by:

\[
F(\mathcal{E}) = 0
\]

Here \( F(\cdot) \) is a (possibly vector-valued) computable function. For example, in statistical mechanics applications constraints are imposed in the form of expectations of particular functions, such as energy, under the ensemble frequencies. Only ensembles that satisfy these constraints are admissible.

\[
< f(\cdot) >_{\mathcal{E}} = \sum_{\mathcal{X}^t} \mathcal{E}(\mathcal{X}^t) f(\mathcal{X}^t) = \sum_{\mathcal{X}^t} \omega(\mathcal{X}^t) f(\mathcal{X}^t) = \mathcal{F}
\]

The expected value of the energy-like function can be conceived of either at the fine-grained level by summing over individual strings, or at the coarse-grained level by summing over the average value of the energy in each subset of the partition.
Maximization of the (log) posterior probability then gives rise (ignoring constants) to the programming problem:

\[ \begin{align*}
\max_{(\mathcal{E}, x')} & \quad \log[P(\mathcal{E} \mid x)] = \log[P(\mathcal{E})] + \log[P(x \mid \mathcal{E})] \\
\text{subject to } F[\mathcal{E}] &= 0
\end{align*} \]

(6)

The “Universal” prior

Following Solomonoff, a “universal” prior over ensembles makes the prior probability of the ensemble inversely proportional to its AIC:

\[ \begin{align*}
p^U[\mathcal{E}] &= 2^{-K[\mathcal{E}]} \\
\log[p^U[\mathcal{E}]] &= -K[\mathcal{E}]
\end{align*} \]

(7)

With the universal prior, the maximum posterior probability program becomes:

\[ \begin{align*}
\max_{(\mathcal{E}, x')} & \quad \log[p^U(\mathcal{E} \mid x)] = -K[\mathcal{E}] + \log[P(x \mid \mathcal{E})] \\
\text{subject to } F[\mathcal{E}] &= 0
\end{align*} \]

(8)

Now consider each of the terms in this expression.

Computing ensembles

A program to compute ensemble frequencies effectively has two subprograms. The first subprogram sorts the strings (of the same length as the data string) into \( k_x \) subsets or bins, \( X_\mathcal{E} = \{ X_1^\mathcal{E}, ..., X_k^\mathcal{E} \} \), which constitute a coarse-graining of the set of strings. We define \( j_x[\mathcal{E}] \) to be the index of the subset or bin the ensemble \( \mathcal{E} \) puts the string \( x \) in.

The second program assigns non-negative coarse-grained frequencies to the bins \( \omega_x = \{ \omega_1^x, ..., \omega_k^x \} \geq 0 \) with \( \sum \omega_x^i = 1 \). Thus the AIC of an ensemble is the sum of the length of the two minimal subprograms:

\[ K[\mathcal{E}] = K[X_\mathcal{E}] + K[\omega_\mathcal{E}] \]

(9)

Shannon’s Theorem shows that, given a finite set of \( k \) messages that will be transmitted with frequencies \( \{ p_1, ..., p_k \} \), it is possible to devise a coding system such that the code assigned to the \( j \)th message has a length very close to \( -\log[p_j] \). (If logarithms are taken with base 2, as will be the convention in the rest of this note, code word lengths will be expressed in bits.) The informational (Boltzmann-Gibbs-Shannon) (neg)entropy of the message ensemble, \( H[\{ p_1, ..., p_k \}] = -\sum p_j \log[p_j] \), is the average code word length over many transmissions.

- The AIC of the coarse-grained partition, \( K[X_\mathcal{E}] \)

Under the universal prior, the code for the program to sort the strings into the coarse-grained partition bins is \( -\log[P^U(x \mid \mathcal{E})] = K[X_\mathcal{E}] \) bits long.

- The AIC of the ensemble frequency distribution, \( K[\omega_\mathcal{E}] \)

The specification of the frequencies the ensemble assigns to the coarse-grained partition is computationally independent of the computation of the partition itself. Thus to complete the program necessary to compute the ensemble frequencies of individual strings, a coding of the frequency distribution to be assigned to the coarse-grained partition is necessary. Any prior probability over frequency distributions can be the basis of the Shannon code for the coarse-grained frequencies.

- Maximum entropy coding of frequency distributions

The maximum entropy principle is based on the procedure of minimizing the relative entropy (or Kullbach-Leibler distance) between a frequency distribution \( \omega \) over a set of \( k \) outcomes and the uniform distribution, \( \{ \frac{1}{k}, ..., \frac{1}{k} \} \), which is

\[ \sum \omega^i \log \left[ \frac{\omega^i}{\frac{1}{k}} \right] = \log[k] + \sum \omega^i \log[\omega^i] = \log[k] - H[\omega]. \]

It is possible to base a prior probability distribution over frequency distributions on this principle by assigning \( \log[P(\omega)] \propto -\log[k] + H[\omega] \). The Shannon code derived from this prior gives the the shortest code word (one bit, in fact) to the uniform distribution, and longer code words (equal in length to the difference between the entropy of the uniform distribution and the entropy of the distribution) to distributions of lower entropy. With this prior, the number of bits required to specify ensemble frequencies will be \( K[\omega_\mathcal{E}] = \log[k_\mathcal{E}] - H[\omega_\mathcal{E}] \). Thus adopting the entropy prior over frequency distributions, the first part of the two part code has a length

\[ K[\mathcal{E}] = K[X_\mathcal{E}] + \log[k_\mathcal{E}] - H[\omega_\mathcal{E}] \]

(10)
It would be possible, on the other hand, to code the ensemble frequencies in the opposite order, with the lowest entropy frequency distributions assigned the shortest code words. With such a coding, the maximum posterior probability would favor maximizing the entropy of the ensemble frequency distribution subject to the constraints. (Brand discusses the role of minimum entropy distributions in recovering information from images.) Other entropy functions could provide further alternative codes for frequency distributions. See Gorban, et al, for a survey.

### The Likelihood, $-\log P(x | E)$

What is the length of the second part of the two-part code, $-\log P(x | E)$?

A computable ensemble must assign the same frequency to all of the strings of a subset of the coarse-grained partition, since it can’t tell them apart without a more sophisticated computation, which would require a longer program. Using the notation $X_E'$ for the number of strings in the subset $X_E'$, the ensemble assigns the uniform probability $\omega_1 / X_E'$ to each string in the subset $X_E'$. Since the ensemble $E$ specifies both which subset the data string is in, and the coarse-grained frequency of that subset, $P[x | E] = 1 / X_E^{(d)}$, or $-\log P[x | E] = \log \frac{X_E^{(d)}}{X_E'}$. The log posterior probability of an ensemble under the universal prior, with a maximum entropy prior on frequency distributions, is

$$-K[X_E] - \log[k_E] + H[\omega_E] - \log X_E^{(d)}$$

The constrained maximum log posterior probability program is:

$$\max_{\{E, x'\}} \log P[E | x] = -K[E] + \log P[x | E] = -K[X_E] - \log[k_E] + H[\omega_E] - \log X_E^{(d)}$$

subject to $F[E] = 0$

We can view the log posterior probability of an ensemble as the negative of the total information of the ensemble, the sum of the AIC of the coarse grained partition, $K[X_E]$ and a residual entropy $\log[k_E] - H[\omega_E] + \log X_E^{(d)}$. The latter term is the length of the message necessary to identify the data string. The coarse-grained probability of the bin containing the data string is contained in the coding of the coarse-grained distribution, which requires $\log[k_E] - H[\omega_E]$ bits, and the identification of the data string among the strings in its bin requires $\log X_E^{(d)}$ bits. The coarse-grained entropy message length is minimized by choosing the maximum entropy frequency distribution that meets the constraints. Taking this maximization into account, the coarse-grained frequency distribution is a function of the constraints and the coarse-grained partition, so we can think of the log posterior probability of the ensemble as a function of its coarse-grained partition and the constraints:

$$\max_X -K[X] - \log[k[X]] + H[X, F] - \log X_E^{(d)}$$

Given a coarse-graining and the maximum entropy prior over frequency distributions, this program maximizes the entropy of the coarse-grained frequency subject to the constraints, thus reproducing the derivations of statistical mechanics. In general all four terms will play a role in determining the maximum posterior probability ensemble. An ensemble that assigns the data string to a very small bin will generally require a longer program to coarse-grain the data, incur a larger number of bins, and lower maximum coarse-grained entropy in order to meet the constraints.

### Visualizing the tradeoffs

The Figure below, adapted from Ay et al., illustrates the situation. Any coarse-graining $X$ corresponds to a point in $(H, K)$ space with the entropy, $\log[k[X]] - H[X, F] + \log \frac{X_E^{(d)}}{X_E'}$, plotted on the horizontal axis and $K[X]$, the AIC of the coarse-grained partition, on the vertical axis. The dark curve represents (schematically) the boundary of computable ensembles that satisfy the constraints for some coarse-grained frequency distribution. The figure is drawn to emphasize the tradeoff between maximizing entropy subject to the constraints and AIC. Minimizing total information (or two-part code length) leads to the ensemble labeled “Min Total Info”, which is also the maximum posterior probability ensemble, and as we have seen, maximizes the entropy for its coarse-graining. For reference, the figure visualizes two other ensembles, $\{(x), (x', x)\}$ is the “Dirac” ensemble that puts the data string as the only member of one subset of its partition, and assigns probability 1/2 to it. The residual entropy of this ensemble is zero, but it is difficult in general to compute, because the program has to be able to recognize the specific data string,$\{(x')\}$ is the ensemble that puts all data strings of length $|x|$ in a single class and assigns equal probabilities to all these strings. The AIC of this ensemble is effectively zero, since it requires only the length of the string to calculate the probabilities, and its entropy is $|x|$, since there are $2^{|x|}$ strings of length $|x|$. 


Maximizing posterior probability under the universal prior, the maximum entropy principle, and constraints involves choosing a coarse-graining \( X \) that minimizes the length of the two-part code to transmit the data string, \( K[X] + \log K[X] + \log [X_{D}^{k}] = \max_{X_{D}} H_{D}[X_{D}] \). A schematic frontier represents the tradeoff between the AIC of the coarse-graining on the vertical axis, and the entropy on the horizontal axis. The sum is the total information. Two coarse grainings that do not in general meet the constraints, the partition that puts the data string as the only member of one subset and all the other strings in another subset, which has minimum entropy, and the uniform coarse graining that puts all strings in one subset, which has minimum AIC, are indicated.

- **Maximum posterior probability with both data and constraints**

Maximizing posterior probability under the universal prior and the maximum entropy principle requires trading off the conflicting goals of minimizing the AIC of the coarse-grained partition and minimizing the residual uncertainty of the data string given the coarse-grained partition. Coarse-grainings that put the data string in a very small set will generally require long programs, and a large sacrifice of entropy to meet the constraints. The Figure shows two extreme cases. In the first, the coarse-grained partition puts the data string alone in one bin, and all the rest of the strings in another bin. This coarse-graining, which will not in general be consistent with the constraints, assigns equal probability to the two bins, so that \( \log k = H = 0 \), and also requires no bits to distinguish the data string in the bin, so its residual entropy is zero. The other extreme case is the coarse-graining that puts all the strings in one bin. This has \( K[X] = 0 \), since the program does not have to sort the strings at all, \( \log k[x] = H = 0 \), but \( \log [X_{D}^{k}] = \log [x] \), which is the maximum residual entropy possible. Gell-Mann and Lloyd argue that the complexity of an object is best represented by the complexity of the minimum total information (in the sense of these notes also the maximum posterior probability) ensemble given the data string describing the object. If the data string is completely random with no features that allow an effective coarse-graining, the best we can do is to use the uniform ensemble, or the most uniform ensemble that meets the constraints, to represent it, and thus it has very low effective complexity. If the data string is constructed from an easily recognized pattern, the minimal program to compute an ensemble that puts the string in a very small bin is quite short, and the string will also have low effective complexity. Thus as in Adriaans, 2011, neither very structured strings (which can be recognized by short programs) nor very unstructured strings represent effectively complex objects, and thus interesting problems for theoretical study.

**References**


