Logistics
Review
Complexity
Kolmogorov
Algorithmic Randomness
Universal Prob.
Scratch

EE515A – Information Theory II
Spring 2012

Prof. Jeff Bilmes

University of Washington, Seattle
Department of Electrical Engineering
Spring Quarter, 2012
http://j.ee.washington.edu/~bilmes/classes/ee515a_spring_2012/

Lecture 29 - May 4th, 2012

Outstanding Reading

• Read all chapters assigned from IT-I (EE514, Winter 2012).
• Read chapter 8 in the book.
• Read chapter 9 in the book.
• Read chapter 10 in the book (chapter on rate distortion theory).
• Read chapter 14 in the book (Kolmogorov complexity)
Announcements, Assignments, and Reminders

- Please do use our discussion board (https://catalyst.uw.edu/gopost/board/bilmes/27386/) for all questions, comments, so that all will benefit from them being answered.

On Final Presentations

- Your task is to give a 15-20 minute presentation that summarizes 2-3 related and significant papers that come from IEEE Transactions on Information Theory (or a very related area).
- The papers must not be ones that we covered in class, although they can be related.
- You need to do the research to find the papers yourself (i.e., that is part of the assignment).
- The papers must have been published in the last 10 years (so no old or classic papers).
- Your grade will be based on how clear, understandable, and accurate your presentation is.
- This is a real challenge and will require significant work! Many of the papers are complex. To get a good grade, you will need to work very hard to present very complex ideas in an extremely simple yet still precise way.
- Again, don’t expect this to be easy, you might need to try a few topics until you find one that is suitable.
Final Presentation Milestones

All submissions done in PDF file format via our dropbox (https://catalyst.uw.edu/collectit/dropbox/bilmes/21171)

- Wed, May 2nd: Candidate proposed papers submitted. Include short at most 1-page write up: 1) why you chose these papers; 2) why they are important to pure IT; and 3) how they are fundamental and/or deep, and 4) how will you summarize them in a simple and precise way.
- Friday, May 11th: Updated list of proposed papers decided, based on feedback. Updated write up.
- Friday, May 18th: short write up on more details of how you will present the ideas in a simple fashion.
- Friday, May 25th: updated short write up on more details of how you will present the ideas in a simple fashion.
- Final presentations: Monday, June 4th in the afternoon late/evening (currently scheduled for 8:30am but that is too early). What to turn in: your slides and a short at most 4 page summary of the papers.

Communication Theory and Entropy are Intertwined

General model of communication

Source ——> Encoder ——> Channel ——> Decoder ——> Receiver

General model of communication expanded:

- Source
- Source Coder - compress, remove redundancy
- Channel Coder - expand, add redundancy
- Noisy Channel - distort
- Channel Decoder - compress, remove redundancy, correct errors
- Source Decoder - expand, add redundancy
- Receiver
Entropy and its Relationships

\[ H(X) = EI(X) = - \sum_x p(x) \log p(x) \]  \hspace{1cm} (1)

\[ H(X, Y) = - \sum_{x,y} p(x, y) \log p(x, y) \]  \hspace{1cm} (2)

\[ H(Y|X) = - \sum_{x,y} p(x, y) \log p(y|x) \]  \hspace{1cm} (3)

\[ H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \]  \hspace{1cm} (4)

\[ 0 \leq H(X) \leq \log n, \text{ where } n \text{ is } X'\text{s alphabet size.} \]  \hspace{1cm} (5)

Mutual Information

- The mutual information (MI) is the average amount of information that r.v. \( X \) has about \( Y \), and vice versa.

**Definition 2.1 (mutual information)**

\[ I(X; Y) = E_{p(x,y)} \log \frac{p(x|y)}{p(x)} = E_{p(x,y)} \log \frac{p(x|y)p(y)}{p(x)p(y)} \]  \hspace{1cm} (6)

\[ = E_{p(x,y)} \log \frac{p(x, y)}{p(x)p(y)} = \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]  \hspace{1cm} (7)
Mutual Information and Entropy - Venn Diagram

Note, these are not sets in the standard sense. Rather the area of the regions convey “degree of information” and the overlapped region correspond to the overlap in information. I.e., the intersection consists of information that is, on average, revealed by both $X$ and $Y$.

Mutual Information and Entropy - Block Diagram

We have the following relationships between entropy, conditional entropy, joint entropy, and mutual information.
Entropy

- Information - we use entropy to measure information.
- The communication theory model
- Surprise of an event \( \{X = x\} \) is measured \( \log \frac{1}{p(x)} \), and there are reasons for using \( \log \).
- Entropy is lower bound on min number of guesses (on average) to guess the value of a random variable.
- Entropy is the minimum number of bits to compress a source.
- Entropy is the optimal coding “length”, of a random source.
- Entropy is the minimum description length (MDL) of a random source that can be achieved without probability of error.

Why log?

- We defined the information in event \( \{X = x\} \) as \( I(\{X = x\}) = I(x) = \log \frac{1}{p(x)} \), but why log?
- Intuition says we want the information about an event to be inversely related to the probability, but there are many such relationships that might be useful.
- E.g., other possible functions include \( I(x) = (p(x))^{-1/n} \) for some \( n > 0 \).
- Another example. \( I(x) = \text{number of prime factors in } \lceil 1/p(x) \rceil \)
- But log (as well will see again below), has a number of attractions.
**Why log?**

- For a distribution on \( n \) symbols with probabilities \( p = (p_1, p_2, \ldots, p_n) \), let \( H(p) = H(p_1, p_2, \ldots, p_n) \) be the entropy of that distribution.

- Consider any information measure, say \( \mathcal{H}(p) \) on \( p \), and consider the following three natural and desirable properties.
  1. \( \mathcal{H}(p) \) takes its largest value when \( p_i = 1/n \) for all \( i \).
  2. If we define the conditional information as
     \[
     \mathcal{H}(Y|X) \equiv \sum_x p(x) \mathcal{H}(Y|X = x) \tag{8}
     \]
     then we wish to have additivity
     \[
     \mathcal{H}(X,Y) = \mathcal{H}(X) + \mathcal{H}(Y|X) \tag{9}
     \]
  3. For a distribution on \( n + 1 \) symbols, then if the probability of one is zero, we wish for \( \mathcal{H}(p_1, p_2, \ldots, p_n, 0) = \mathcal{H}(p_1, p_2, \ldots, p_n) \).

---

**Theorem 2.2 (Khinchin’s Theorem)**

If \( \mathcal{H}(p_1, \ldots, p_n) \) satisfies the above 3 properties for all \( n \) and for all \( p \) such that \( p_i \geq 0, \forall i \) and \( \sum_i p_i = 1 \) (i.e., all probability distributions), then

\[
\mathcal{H}(p_1, \ldots, p_n) = -\lambda \sum_i p_i \log p_i \tag{10}
\]

for \( \lambda \) a positive constant.

- Thus, we get entropy for some logarithmic base.
- But other than trying other functions inversely related to \( p(x) \), are there other ways to measure inherent complexity of an object?
Strings

Consider the strings:

- Consider 0101011. While this is an infinitely long string, and while there is no probability distribution governing it (it is one string after all), it is somehow simple to describe. For example, we can just say “Print 01 forever” which is a 16 byte description of an infinitely long string. Alternatively, we can say “01101” which is a, perhaps, 5 byte description of the string.

- Consider 11.0010010000111110110101010001000010110100011000100011010011… Actually, this is a binary representation of \(\pi\). Is it hard to describe it? There is a simple algorithm to generate this string to any length, and the algorithm is fixed length. We could think of the algorithm itself as a code for this string. Whenever we want to send it, we just send the algorithm to generate it. The computer then, in some sense, becomes the “codebook.”

- Consider \(b_1b_2b_3\ldots b_n\) where this was generated via stochastic process \(p(b_i = 1) = 1 - p(b_i = 0) = 0.5\) and \(b_i\) are i.i.d. Is there a short way to describe this string? No, this string is incompressible. The only way to send the string (or compress it) is to send (store) it directly.
String Complexity

Consider the complexity of a string to be the length of an algorithm needed to generate that string. I.e., for the examples above, we’d have:

- Repeat: Print 01; which is very short
- Print $\pi$; which is longer, but still relatively simple.
- Print $b_1b_2b_3\ldots$; The program is long since the string needs to be embedded as an immediate operand to the program.

On Program Length

- We don’t care about what language we use since, if we have a universal computing device, all computers can simulate each other.
- A Turing machine (a mathematical model for computation) consists of an input stream (tape, read sequentially), a working tape of infinite length (memory), and an output stream (tape also).

Properties: 1) An infinite “tape” (or memory); 2) read/write head (can read and write to memory); 3) machine has a finite number of states (FSM); 4) a transition function $\delta$ which does one of:

$$\delta(\text{current state, read input}) = \begin{cases} 
\text{read/write on work tape} \\
\text{write in output} \\
\text{move to a new state}
\end{cases}$$ (11)
Turing Machine

- All existing standard computing models can be reduced to Turning machine.
- Church Thesis: A Turing machine can compute any function commutable by a “reasonable” computing device with only polynomial cost over any other.
- Reasonable does not include quantum models of computation.
- Thus, we make the following assumption:

**Lemma 3.1**

*All computing devices (at least that we deal with) are equivalent to a universal computer. \( \mathcal{U} \) will designate this universal computer.*

- We’ll use \( p \) in this context to designate a program and the length of this program, as measured by \( \mathcal{U} \) will be \( \ell(p) \). \( \mathcal{U}(p) \) is \( \mathcal{U} \) run on program \( p \) which will produce some output (say \( x \)).

Kolmogorov Complexity

- We are now ready to define another notion of complexity of an object that is entirely non-probabilistic.

**Definition 4.1 (Kolmogorov Complexity)**

The **Kolmogorov Complexity** of a string \( x \) measured according to \( \mathcal{U} \) is defined as:

\[
K_{\mathcal{U}}(x) = \min_{p : \mathcal{U}(p) = x, \text{and } p \text{ halts}} \ell(p)
\]  

- So \( K_{\mathcal{U}}(x) \) is the shortest algorithmic description length of the string \( x \) over all algorithmic descriptions that generate \( x \) by computer \( \mathcal{U} \).
Non-Halting

- The definition requires that \( p \) halts, meaning that it eventually will terminate (even if it takes a long time).
- Not all programs will halt and there are reasons (as we will see) for excluding these non-halting problems.
- Consider the problem of computing the smallest natural number that is not the sum of \( n \) squares (\( 0^2 = 0 \) included).
- So if \( p \) is a program that does this iteratively (by trying out each number), then \( p(0) = 1, p(1) = 2, p(2) = 3, \) and \( p(3) = 7 \) (since \( 8 = 0^2 + 2^2 + 2^2 \)). Thus, for these arguments, \( p \) will halt.
- What is \( p(4) \)? If we present \( p \) to a turning machine then \( U(p, 4) \) does not halt, since Lagrange showed that every number is the sum of four square numbers.
- Thus \( p \), when given 4 as input, does not halt.

Example

- Consider sequences/strings of large length, e.g., \( \ell(x) \approx 10^{15} \) (overloading \( \ell \) here but there should be no confusion).
- One possible case 1: Expand \( e \) to \( 10^{15} \) bits. Is this complex (via \( K_U \))? No, since we can write a short program to do this.
- Case 2: are most strings of length \( 10^{15} \) complex or simple (by \( K_U \))? A: Complex. Why? How many simple strings can there be of a given length? No more than the number of short programs, and there can’t be too many of those.
- Intuition: AEP. Recall \( P_n \leq (n + 1)|X| \),
  \[ Q^n(x) = 2^{-n(D(P_x||Q) + H(P_x))}, \quad |T(P)| = 2^{nH(P)}, \]
  \[ Q^n(T(P)) \approx 2^{-nD(P||Q)}. \]
- So for \( x \) a string with type \( P_x \), we have
  \[ |T(P_x)| = 2^{nH(P_x)} \quad (13) \]
  meaning that there are exponentially more sequences of higher entropy than there are of lower entropy. But this is just entropy . . .
**Conditional Kolmogorov Complexity**

**Definition 4.2**
The conditional Kolmogorov complexity is defined for a string $x$ given $y$ as input that is not charged in the length. I.e.,

$$K_U(x|y) = \min_{p : U(p,y) = x} \ell(p)$$

(14)

- So here, the program can use $y$ as input, and the length of $y$ is not charged against the length of the program $p$.
- $K_U(x|x) = \text{constant}$, since if we have $x$ we can just have a program that says print my input.
- $K_U(x|y)$ when $y$ is i.i.d. and $x$ is also i.i.d. will be such that $K_U(x|y) = K_U(x)$.

**Definition 4.3**
The length conditional Kolmogorov complexity is defined for a string $x$ as follows:

$$K_U(x|\ell(x)) = \min_{p : U(p,\ell(x)) = x} \ell(p)$$

(15)

- So program has more info available to it, but only the length of what it must generate.
- Does this increase or decrease Kolmogorov Complexity? More soon ...
Universality

Theorem 4.4

Kolmogorov complexity is universal. I.e., given universal computer $U$ and another computer, say $A$, then

$$K_U(x) \leq K_A(x) + c_A$$

(16)

for all $x \in \{0,1\}^*$, and where $c_A$ is a constant and does not depend on $x$.

- Consider program $s_A$ to simulate $A$ by $U$ (so that $U(s_A, p_A) = A(p_A)$).
- Given $p_A$ s.t. $A(p_A) = x$, then $U(s_A p_A) = x$.
- But $\ell(s_A p_A) = \ell(s_A) + \ell(p_A) = c_A + \ell(p_A)$
- Thus, since $K_U(x) \leq \ell(s_A p_A)$ for all $p_A : A(p_A) = x$, we have

$$K_U(x) = \min_{p : U(p) = x} \ell(p) \leq \min_{p : A(p) = x} (\ell(p) + c_A) = K_A(x) + c_A$$

(17)

- Hence, we ignore which computer it is (unless otherwise stated).

Conditioning

Theorem 4.5

$$K(x|\ell(x)) \leq \ell(x) + c$$

(18)

where $c$ is constant.

Proof.

- The computer knows $\ell(x)$, so consider the program:

  print the sequence $x_1: \ell(x)$

(19)

- This program has length $\ell(x) + c$
- No additional bits are needed to store the value $\ell(x)$ itself, since that is an input to the program.

Often not a good upper bound, as this is really the worst case.
Comparing w. Conditioning

**Theorem 4.6**

\[ K(x) \leq K(x|\ell(x)) + 2 \log \ell(x) + c \]  

(20)

*where c is constant.*

**Proof.**

- In worst case, for \( K(x) \), need a way of representing \( \ell(x) \) in program.
- Can’t just use \( \log n \) bits to store a number \( n \), still need to know how many bits we use to store a number, a meta regress problem.
- Instead, express representation of \( \ell(x) \) using \( 2\lceil \log n \rceil \) bits, duplicating bits, and then use 01 as code for an “end” symbol.
- I.e., 5 = 101 gets represented as 11001101 where the final 01 is the end symbol.

Comparing w. Conditioning

**Theorem 4.7**

\[ K(x) \leq K(x|\ell(x)) + \log^* \ell(x) + c \]  

(21)

*where \( \log^* n = \log n + \log \log n + \log \log \log n + \ldots \) and this continues until just before the term goes negative.*

- This is just another way to encode a number.
- We use \( \log n \) bits to store a number \( n \), but we need to know many bits are being used to store \( n \),
- Then we use \( \log \log n \) bits store a number \( \log n \), but we need to know how many bits are being used to store \( \log n \),
- etc.
Complexity and counting

- So complexity is measured based on the shortest program that can generate it.
- There are many more (exponentially more) long programs than there are short programs, simply since short programs, themselves, take up few bits.
- Thus, there can't be more strings that have short programs then there are short programs to begin with,
- Thus, unfortunately, most strings are complex, which we formalize next.

Theorem 4.8

The number of strings $x$ with $K(x) < k$ is

$$|\{x \in \{0,1\}^* : K(x) < k\}| < 2^k$$

(22)

Proof.

- How many program are there with length $< k$? A: $= 2^k - 1 < 2^k$.
- I.e., there is 1 program with length 0; there are 2 program (0 and 1) with length 1; there are 4 programs (00, 01, 10, 11) with length 2; 8 program with length 3, etc.
- In general, $2^k$ program with length $k$
- $\sum_{l=1}^{k-1} 2^k = 2^k - 1 < 2^k$
- Each program (at best) outputs a unique string $x$.
- Thus, there can be at most $2^k$ programs with complexity $< k$. 
Note, if we look at all $2^k$ programs of length $< k$, and don’t find one such that $U(p) = x$, then $K(x) \geq k$.

Finding the shortest program in this way has computational cost exponential in $K(x)$, but we’ll discuss this more soon.

Examples

- $K(00\ldots 0 | n) = c$ for all $n$.
- $K(\underbrace{\pi_1 \pi_2 \ldots \pi_n}_{n\times} | n) = c$ for all $n$
  bit expansion of $\pi$
- $K(01\overline{01} | n) = c$, i.e., “Print 01 $n/2$ times”
Examples: Weather prediction (does it rain or not).

- Given binary string $x_1:n$ with $x_i = 1$ iff it rains, otherwise $x_i = 0$.
- We model this as a Markov chain, and learn (from $x$) a conditional histogram $p(x_i|x_{i-1})$ which requires 4 table entries.
- Consider histogram of counts of pairs, numerator of histogram is at most $n$, so $\log n$ bits needed. Thus, $O(\log n)$ bits for $p(x_i|x_{i-1})$.
- We need $\log 1/p(x_i|x_{i-1})$ bits to describe symbol $x_i$ in context of $x_{i-1}$ (e.g., arithmetic coding).
- Thus, the Kolmogorov Complexity is:
  \[
  K(x_1:n|n) = Hn + O(\log n) + c
  \]  
  (23)
  $Hn$ bits are needed for the string ($H$ is the per symbol entropy), $O(\log n)$ bits needed for the table, and $c$ is a constant.
- Seattle case:
  \[
  K(x_1:n,n,\text{Seattle}) = c
  \]  
  (24)
  i.e., program is “while(1) { print rain; }”

Example: Fractal Image

- For all $c \in \mathbb{C}$ image where $c$ is a complex number indexing a pixel element, find $n : z_{n+1} = z_n^2 + c$, and $z_{n+1} > \tau$. Resulting $n$ for $c$ then indexes the color of pixel $c$.
- Resulting images looks very complex (r.h.s. is other form of fractal).

  ![Fractal Image](image1)

  ![Fractal Image](image2)

- but, $K(\text{fractal image}|\text{num. pixels}) = \text{const.}$
- Natural images? Is there a simple algorithm to generate them?
More Examples

- \( K(n) \leq \log^* n + c \) for integer \( n \) (again, \( \log n \) bits to store \( n \), \( \log \log n \) bits to store \( \log n \), etc.).

- Any \( x_{1:n} \) s.t. \( \sum_i x_i = k \), number of strings with \( k \) 1's. The question is: how can we encode this?
  - Consider all strings of length \( n \) with \( k \) 1's in lexical order. Given index \( i \), can write a program with constant length to print out the \( i \)'th one, i.e., “print out the \( i \)'th string if the \( i \)'th one = \( x_{1:n} \).”
  - So we need to represent both \( k \) and \( i \).
  - For \( k \), need \( 2 \log k \) bits (or could use \( \log^* k \) bits, or even \( \log n \) bits).
  - For \( i \in [1, \ldots, \binom{n}{k}] \), need \( \log \binom{n}{k} \) bits. Why no 2? Since we have both \( n \) (as input) and \( k \), so we know number of bits needed for \( i \).
  - So the length of the program is:
    \[
    \ell(p) = c + 2 \log k + \log \binom{n}{k} \leq c + 2 \log k + nH(k/n)
    \]
    (25)
  - since \( \binom{n}{k} \leq 2^{nH(k/n)} \)
  - Thus,
    \[
    K(x_{1:n}) \leq 2 \log k + nH(k/n) + c
    \]
    (27)
  - Note, if \( k = n/2 \), this is as long as the string itself.
$K$ vs. $H$

- As we saw, in a few examples, the expression for $K(x)$ involved the entropy. How do they compare more generally?
- First point: $H$ requires a distribution while $K$ does not (but then neither does Lempel-Ziv compression which, we can show, while algorithmic will converge to the entropy).
- First, an important theorem.

### Theorem 4.9

\forall \mathcal{U} \ (\text{meaning under any “reasonable” computing device}),

\[
0 \leq \Omega \triangleq \sum_{p \in \mathcal{U} \ (p) \text{ halts}} 2^{-\ell(p)} \leq 1 \tag{28}
\]

**Proof.**

If $p$ halts, then it can’t be a prefix of any other program. Therefore, halting programs satisfy the prefix property, and we immediately have countably infinite Kraft inequality (from Lecture 9).

---

### Theorem 4.10

Let $\{X_i\}_i$ be i.i.d. drawn $\sim p(x)$ for $x \in \mathcal{X}$. Also, $p(x_{1:n}) = \prod p(x_i)$. Then $\exists c$ s.t.

\[
H(X) \leq \frac{1}{n} \sum_{x_{1:n}} p(x_{1:n}) K(x_{1:n}|n) \leq H(X) + \frac{|\mathcal{X}|}{n} \frac{c}{n} \tag{29}
\]

Also,

\[
E \frac{1}{n} K(X_{1:n}|n) \rightarrow H(X) \tag{30}
\]

So, $K(x_{1:n}|n)$ is, on average, and when $n$ gets big, no better than the codeword lengths suggested by $\log 1/p(x_{1:n})$. 
**Proof of Theorem 4.10.**

- **Lower bound:** Consider error-free coding \( x_{1:n} \) with codewords of length \( \ell(x_{1:n}) \).
- We know in such case \( \sum_{x_{1:n}} p(x_{1:n}) \ell(x_{1:n}) \geq H(X_{1:n}) \)
- Let’s code \( x_{1:n} \) with \( \text{argmin}_{p:U(p,n)=x} \ell(x) = p^*(x) \), so that \( \ell(p^*(x)) = K(x|n) \) which as we saw above satisfies Kraft inequality.
- Therefore, \[
\sum_{x_{1:n}} p(x_{1:n}) K(x_{1:n}|n) \geq H(X_{1:n}) = nH(X) \tag{31}
\]

- **Upper bound:** We encode the “type” of the string, and then encode which (in lexicographic order) of the strings it is with that type.

Consider the type of \( x_{1:n} \):

\[
P_{x_{1:n}} = \left( \frac{n(a_1)}{n}, \frac{n(a_2)}{n}, \ldots, \frac{n(a_{|X|})}{n} \right) \tag{32}
\]

- So, for the type, we need \( |X| \log n \) bits to describe the type.
- Then, we enumerate which element is is within the type class, and use the result \( |T(P_x)| = 2^{nH(P_x)} \leq 2^n \log(P_x) \)
- Thus, need \( nH(P_x) \) bits for encoding which entry in type class.
- Thus, we have

\[
K(x_{1:n}|n) \leq nH(P_x) + |X| \log n + c \tag{33}
\]

\[
\Rightarrow \quad EK (X_{1:n}|n) \leq nE[H(P_X)] + |X| \log n + c \tag{34}
\]
Proof of Theorem 4.10.

- Note that \( n_i(x_1:n) = n_i = \sum_{j=1}^{n} 1_{\{x_j=i\}} \) so that \( E_n(x_1:n) = p_i n. \)
- Also, recall \( H(p) \) is concave in \( p \), so \( E[H(P)] \leq H(E[P]). \)
- The last part of the proof follows from the following derivations:

\[
EH(P_X) = -\sum_{x_1:n} p(x_1:n) \left( \sum_i \frac{n_i}{n} \log \frac{n_i}{n} \right) 
\]
\[
\leq -\sum_i \sum_{x_1:n} p(x_1:n) n_i \log \frac{n_i}{n} \sum_{x_1:n} p(x_1:n) n_i 
\]
\[
= -\sum_i \frac{p_i n}{n} \log \frac{p_i n}{n} 
\]
\[
= H(X) 
\]

Algorithmic Randomness

- What is randomness? Normally, we define randomness via the presumption of the existence of certain probabilities. But the existence of probabilities, or what they mean, is not universally agreed upon (e.g., frequentist vs. Bayesian interpretation).
- We can dispense with probabilities and define algorithmic randomness of \( x \) to be an inability to find a short program to print \( x \).
- We can relate algorithmic randomness to our “normal” notion of randomness such as entropy, proportion of outcomes (typical sets), etc.
- How many simple sequences are there? Not too many. How many complicated sequences are there? Lots! We can look at this via randomness as, in some sense, a surrogate for counting.
Bernoulli model of source

**Theorem 5.1**

Let $X_1, \ldots, X_n$ be $\sim$ Bernoulli$(1/2)$. Then

$$Pr(K(X_1:n|n) < n - k) < 2^{-k}$$

(39)

**Proof.**

$$Pr(K(X_1:n|n) < n - k) = \sum_{x_1:n : K(x_1:n|n) < n - k} Pr(x_1:n)$$

(40)

$$= \sum_{x_1:n : K(x_1:n|n) < n - k} 2^{-n}$$

(41)

$$= |\{x_1:n : K(x_1:n|n) < n - k\}| 2^{-n}$$

(42)

$$< 2^{n-k} 2^{-n} = 2^{-k}$$

(43)

Bernoulli model of source and Algorithmically Random

- So, as we increase $k$ (i.e., compress $x_1:n$ by $k$ bits, from $n$ bits to $n - k$ bits), the probability of being able to do so decreases exponentially fast!
- Multiply the probabilities by $n$, and we get that the number of sequences of length $n$ that can be compressed by $k$ bits decreases exponentially fast in $k$!

**Definition 5.2**

$x_1:n$ is **algorithmically random** if

$$K(x_1:n) \geq n - c_n \text{ with } c_n \to 0 \text{ as } n \to \infty$$

(44)

- So this is really a definition of a sequence of strings $\{x_1:n\}_n$ (or alternatively, an infinite length string that is considered at each point truncated to length $n$), such that, when $n$ gets big, the string becomes ever less compressible.
Compressibility

Definition 5.3

A string $x_1, x_2, \ldots$, infinite in length, is incompressible if:

$$\lim_{n \to \infty} \frac{K(x_1, x_2, \ldots, x_n | n)}{n} = 1$$

(45)

- So this is like $K(x_{1:n} | n) = n + c$.
- We know that if a random string is incompressible in an entropic sense, there are likely to be about half 0’s and half 1’s (i.e., the probability is close to 1/2). Is there an analogous notion for Kolmogorov complexity?

Theorem 5.4 (strong law)

If $x_{1:\infty}$ is incompressible, then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i = 1/2$$

(46)

Like a series of i.i.d. variables where $p(X_i = 1) = 1/2$, but note that we’ve said nothing about probabilities, rather we’ve only mentioned its algorithmic complexity.

Proof.

- Define $\theta_n = \frac{1}{n} \sum_{i=1}^{n} x_i$, then from Eq (33),

$$n - c_n \leq K(x_{1:n} | n) \leq nH(\theta_n) + 2\log n + c'$$

(47)

with $c_n/n \to 0$.
- Therefore,

$$H(\theta_n) \geq 1 - \frac{2\log n + c_n + c'}{n}$$

(48)
Compressibility

...proof continued.

- But $H(\theta_n) \leq 1$ and

$$\frac{2\log n + c_n + c'}{n} \to 0$$  \hspace{1cm} (49)

- Thus, eventually $H(\theta_n) \to 1$ meaning $\theta_n \to 1/2$.

- So, note we’ve taken an non-random sequence and used entropy (which is defined on a probability distribution). Is this circular?
- No. We are using the distribution only as a combinatorial “counting” quantity, it is normalized so that we can still take advantage of the capabilities we know (and love 😊) about entropy.
- This is akin to what we did with the method of types, and is also akin to the various proofs of purely combinatorial and/or matrix bounds using entropic quantities for assistance.

Also, truly algorithmically incompressible sequences all look random, and satisfy all statistical tests for randomness.

- Why all? If not all, we can use the test to reduce the description length. I.e., suppose for certain $i$, $x_i$ and $x_{i+k}$ are more likely to be 11 than 00, 01, 10, then we can group sequences into two equal sized groups with different probabilities $\Rightarrow$ compressible.
- Next, for random strings, we can get a direct relationship between Kolmogorov complexity and entropy, as we do next:
Theorem 5.5

Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli($\theta$). Then

$$
\frac{1}{n} K(X_{1:n}|n) \rightarrow H(\theta) \text{ in probability}
$$

(50)

meaning $\Pr(|\frac{1}{n} K - H| > \epsilon) \rightarrow 0$ $\forall \epsilon > 0$

We’ll show:

- A: $\Pr(K > n(H + \epsilon)) \rightarrow 0$ as $n \rightarrow \infty$ for any $\epsilon > 0$, and
- B: $\Pr(K < n(H - c)) \rightarrow 0$ as $n \rightarrow \infty$ for any $c > 0$ ($c$ is used as an $\epsilon$ like quantity but we use symbol $c$ since we’ll also use $\epsilon$ in the same context below).

Proof.

proof of Theorem 5.5

- First (A), define $\bar{X}_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i$, the probability of 1’s in $X_{1:n}$.
- From Eq (33), we have that

$$
K(x_{1:n}|n) \leq n H(\bar{x}_{1:n}) + 2 \log n + c
$$

(51)

or

$$
\frac{K}{n} - \frac{2 \log n + c}{n} \leq H(\bar{x})
$$

(52)
Compressibility

Proof.

proof of Theorem 5.5

We then get the following:

\[ \Pr(|\bar{X} - \theta| > \epsilon) \to 0 \text{ as } n \to 0 \quad \forall \epsilon > 0 \]  \hfill (53)

\[ \Rightarrow \Pr(|H(\bar{X}) - H(\epsilon)| > \epsilon) \to 0 \]  \hfill (54)

\[ \Rightarrow \Pr(-\epsilon + H(\Theta) < H(\bar{X}) < \epsilon + H(\theta)) \to 1 \]  \hfill (55)

\[ \Rightarrow \Pr\left(\frac{K}{n} - \frac{2\log n + c}{n} < \epsilon + H(\theta)\right) \to 1 \]  \hfill (56)

\[ \Rightarrow \Pr\left(\frac{K}{n} - H(\theta) \geq \epsilon \right) \to 0 \]  \hfill (57)

\[ \Rightarrow \Pr(K > n(H + \epsilon)) \to 0 \]  \hfill (58)

Which is the first part.

...
Compressibility

Proof.

proof of Theorem 5.5 continued

\[ \leq \epsilon + 2^{n(H(\theta)-\epsilon)} 2^{-n(H(\theta)-\epsilon)} = \epsilon + e^{-n(c-\epsilon)} \]  

(63)

which can be made arbitrarily small for appropriate \(\epsilon\), \(c\), and \(n\).

- Note, in the above case \(B\), we first choose a \(c > 0\), and then choose an \(0 < \epsilon < c\) so that for large enough \(n\) the above will \(\to 0\).

Universal Probability

- Suppose \(H = 1\), so that the number of typical sequences of length \(n\) is \(2^{nH} = 2^n\), so all sequences are equally likely.
- Thus, \(p(0, 0, \ldots, 0) = \frac{1}{2^n}\) where we have \(L\) 0’s.
- Also, \(p(1011011011101\ldots01101) = \frac{1}{2^n}\).
- Both the all zeros, and the seemingly “random” sequence are in the typical set and are thus considered equally likely?
- Which one is more surprising to you intuitively? Which one is more simple? Which one is easier to describe?
- Is Shannon’s theory lacking? It only gives complexity to an ensemble of messages (or a diversity of messages, not to the message contents). Shannon’s approach does not ask: How complex is the message \(x_{1:n}\) itself?
- Shannon says that the “complexity” of the single message is the coding length \(\log \frac{1}{p(x_{1:n})}\), irregardless of \(x_{1:n}\).
- But how can we get to coding length without probabilities? We appeal to philosophy and Occam’s razor.
An Appeal to Philosophy

- Occam’s razor: “If there are multiple explanations for a phenomenon, then all other things being equal, we should select the simplest one”
- “simplicity of an object” is the same as “the object has the shortest effective description”
- Occam’s razor says that the short explanations, or causes, of objects are the more probable ones.
- This is really a prior probability on nature itself.

We arrange in our thought all possible events in various classes; and we regard as extraordinary those classes which include a very small number. In the game of heads and tails, if heads comes up a hundred times in a row then this appears to us extraordinary, because the almost infinite number of combinations that can arise in a hundred throws are divided:

1. in regular sequences, or those in which we observe a rule that is easy to grasp, and
2. in irregular sequences, that are incomparably more numerous.

*Philosophical Essay on Probabilities, Laplace, 1819*

Of course, Laplace wrote this 150 years before Kolmogorov complexity was formalized. How do we define “regular”? Does this exhibit a real-world bias, a bias towards the type of regular events we see in our universe?
An Appeal to Philosophy: On Occam’s razor, by Laplace

The regular combinations occur more rarely only because they are less numerous. If we seek a cause wherever we perceive symmetry, it is not that we regard the symmetrical event as less possible than the others, but since this event ought to be the effect of

1. a regular cause [or short program] or
2. that of chance,

the first of these suppositions is more probable than the second. On a table we see letters arranged in this order “Constantinople” and we judge that this arrangement is not the result of chance, not because it is less possible than others, for if this word were not employed in any language we would not suspect it came from any particular cause, but this word being in use among us, it is comparably more probable that some person [program] has thus arranged the aforesaid letters than that this arrangement is due to chance.

Philosophical Essay on Probabilities, Laplace, 1819

Regular Sequences

- 00001111000011110000111100001111 This is just repeat 4 0’s and 4 1’s for ever. Is this likely to happen by pure chance?
- 2 3 5 7 11 13 17 19 23 29 31 37 41 43 47 . . . This is is just the primes. Is this likely to happen by chance?
- 0 1 2 3 1 1 3 3 5 9 1 2 1 2 5 8 7 . . . Let $p_i$ be the $i$’th prime. Then this sequence is $n_i = p_i \mod i$. Is this likely to happen by chance?
- So it seems perhaps intuitive that simple explanations are more probable for a given phenomena, if they exist, than the alternative that they happen by chance.
- Can we derive a probability distribution for this?
Prefix-free programs

- We can quantify this by defining a probability on random programs that are prefix-free (halting). That is, for program $p$, define:

$$\Pr(p) = 2^{-\ell(p)}$$

so that $\sum_p \Pr(p) \leq 1$ by Kraft inequality.
- This is an assumed fundamental property of nature.
- Short programs are much more probable than long ones.
- Most sequences $x_{1:n}$ of length $n$ are complex (since there are just not enough short programs to go around to cover all the long sequences, of which there are many).
- When short programs produce long strings, they produce simple ones (by definition).

Universal Probability

- We can use this to produce a probability distribution on strings that is universal, where simple strings are more probable than complex ones of the same length.

**Definition 6.1**

The universal probability of a string $x$ is given by:

$$P_U(x) = \sum_{p : U(p) = x} 2^{-\ell(p)} = \Pr(U(P) = x)$$

where on the r.h.s. $P$ is a random program variable under $\Pr(P = p)$.

- This is also equal to the probability that random fair coin flips will create a program to print out $x$. 
The Meaning Of Life?

- $x$ could be natural images, natural scenes, or even sentience.
- And $p$ the program is what causes $x$ to occur.
- If nature acts as a grand computer in response to some random input, and if short programs are more probable than long ones, then this could cause $x$ to have structure even if $x$ is very big itself.
- Fractals might be an example of this but it might be only the most mundane of examples.

Computing Kolmogorov Complexity

- The Kolmogorov Complexity of a string $x$ is:

$$K_U(x) = \min_{p : U(p) = x} \ell(p)$$

(66)

- We have found bounds on $K_U(x)$, but how to compute it?
- We could try all short programs and see which one of them (the first one) generates $x$.
- Some short programs might take a long time to halt.
- How do you know if a program will halt or not, since if it doesn’t halt we needn’t bother with this program.
- On the other hand, if it hasn’t halted yet, how do you know if we just haven’t waited long enough?
- We thus need a way to tell if $p$ will halt.
Halting Theorem

Theorem 6.2

There is no program $h(p, y)$ that takes a program $p$ with input $y$, where $h(p, y) = 1$ if $p$ halts on input $y$, and $h(p, y) = 0$ otherwise.

Proof.

- Suppose, to the contrary, that $\exists$ such an $h$,
- and given $h$ we can easily construct program $\Psi(x)$ as follows:

$$
\Psi(x) = \begin{cases} 
1 & \text{if } h(x, x) = 0 \quad \text{i.e., } x \text{ does not halt on } x. \\
\text{spin forever} & \text{if } h(x, x) = 1 \quad \text{i.e., } x \text{ halts on } x.
\end{cases}
$$

(67)

$\Psi(x)$ asks: does $x$ not halt on itself as input? If it does halt, we spin forever.

Ex Contradictions: Statement A: “Statement B is false.” Statement B: “Statement A is true.”
Computing Kolmogorov Complexity

- Thus, $K_U(x)$ is not computable, since we cannot solve the halting problem.
- Suppose, just for fun, that we could compute $\Omega$.

$$0 \leq \Omega \triangleq \sum_{p \notin \mathcal{U}(\text{halts})} 2^{-\ell(p)} \leq 1$$  \hspace{1cm} (69)

again which follows from the Kraft inequality.
- Key point: $0 < \Omega \leq 1$ actually exists (it is merely a number between 0 and 1) but we can’t compute it due to the halting theorem (this is actually an instance of Gödel’s theorem).
- But what if we could compute $\Omega$? What would this number tell us?
- Note, $\Omega$ is the halting probability, the probability that a program (drawn from the universal probability of programs) will halt.

Proving $n$-bit theorems

- What could we do, given $\Omega$?
- Given $\Omega$, we could prove any $n$-bit theorem?
- How? To do this for $n$-bit theorems, compute $\Omega_n = \lfloor \Omega \rfloor_n$ which is $\Omega$ truncated to $n$ bits.
- Next, run all programs simultaneously, $n$ cycles at a time.
- That is, given enumerable (countable number of) functions $p_1, p_2, p_3, \ldots$ run:

```plaintext
1 for i = 1, \ldots do
2     for j = 1, \ldots, i do
3         run i more cycles of $p_j$;
```
Proving $n$-bit theorems

Once the sum of programs that have halted are such that

$$\sum_{p \text{ halted so far}} 2^{-\ell(p)} \geq \Omega_n \tag{70}$$

then no other program of length $< n$ could halt, and we have solved the halting problem for any program of arbitrary length (up to length $n$).

How to use the halting problem to solve any hard conjecture?

Here’s how to prove that $P = NP$ or $P \neq NP$.

Let $p$ be a program to search all programs that solve some NP-complete problem and halt only if the solution is polynomial time.

Then if $\text{halt}(p)$ is true, then $P = NP$.

Compressibility of $\Omega$

Even if we could compute $\Omega$, it is incompressible.

**Theorem 6.3**

$\Omega$ cannot be compressed by more than a constant. I.e., there exists a constant $c$ s.t.

$$K(\Omega_n) = K(\omega_1 \omega_2 \ldots \omega_n) \geq n - c \quad \forall n \tag{71}$$

**Proof.**

- Given $\Omega_n$, we can solve the halting problem for any program of length $\leq n$ bits.
- We can compute $\Omega_n$ with a length $K(\omega_1 \omega_2 \ldots \omega_n) = K(\Omega_n)$-bit program.
- Generate list of all programs length $\leq n$ that halt, and their output:

$$\left( p_{h_1}, x_{h_1} \right), \left( p_{h_2}, x_{h_2} \right), \ldots, \left( p_{h_m}, x_{h_m} \right) \tag{72}$$
Compressibility of $\Omega$

...proof continued.

- Find the shortest string $x_0$ not in this list of output strings and output $x_0$.
- The above “program” (i.e., the enumeration of programs and output of $x_0$) has length $K(\Omega_n) + c$ (i.e., $K(\Omega_n)$ to compute $\Omega_n$ and $c$ to do the enumeration.
- We can’t have that $K(x_0) \leq n$ since we have enumerated all programs of length $\leq n$ that halt, and $x_0$ was not among them, so $K(X_0) > n$.
- The above program must also have length $\geq K(x_0)$ since it is a program to compute $x_0$. Thus,
  \[ K(\Omega_n) + c \geq K(x_0) > n \tag{73} \]
- Or $K(\Omega_n) > n - c$ and $\Omega_n$ can’t be compressed by more than a constant.

Universality

- We have a bound on the universal probability of a string $p_U(x)$ as follows:
  \[ p_U(x) = \sum_{p : U(p) = x} 2^{-\ell(p)} \geq 2^{-\ell(p^*)} = 2^{-K(x)} \tag{74} \]
  where $p^* = \text{argmin}_{p : U(p) = x} \ell(p)$
- In fact, we have that:
  \[ 2^{-K(x)} \leq p_U(x) \leq c 2^{-K(x)} \tag{75} \]
  or
  \[ K(x) - c' \leq \log \frac{1}{p_U(x)} \leq K(x) \tag{76} \]
  so that $\frac{1}{p_U(x)}$ is the Shannon “information” of $x$.
- So, Shannon lengths are as natural as $K(x)$ when using $p_U(x)$ as a distribution.
Also, the entropy of \( p_U(x) \) is \( E_K(X) \) where \( X \) is a random variable drawn via \( p_U \).

So, we can think of this as the universal entropy or the universal complexity of the universe.

All, assuming, the Occam’s razor assumption as being the guiding principle of the universe.