COMP2610/6261 - Information Theory

Lecture 16: Arithmetic Coding (cont.)

Mark Reid and Aditya Menon

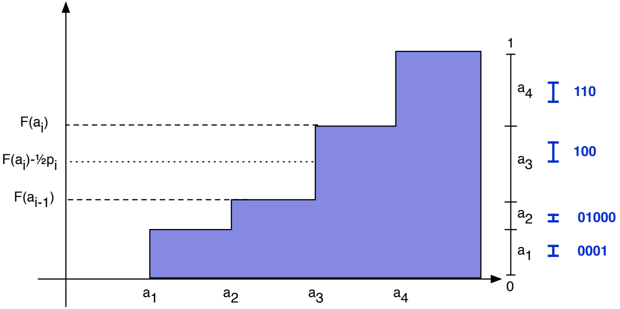
Research School of Computer Science
The Australian National University

September 24th, 2014

Interval Codes

Recall the **Shannon-Fano-Elias Coding** method from last lecture:

- Order the alphabet $\mathcal{A}$.
- Represent distribution $\mathbf{p}$ by cumulative distribution F
- Construct code by finding intervals of width $\frac{p_i}{2}$ that lie in each symbol interval $[F(a_{i-1}), F(a_i))$

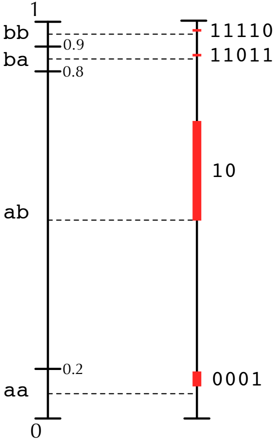


Interval Coding Blocks

What if we apply SFE coding to blocks?

Example: Let $\mathcal{A} = \{\texttt{aa}, \texttt{ab}, \texttt{ba}, \texttt{bb}\}$ with $\mathbf{p} = (0.2, 0.6, 0.1, 0.1)$.

x	p	$\bar{F}$	$\bar{F}_2$	ℓ	Code
aa	0.2	0.1	$0.0\overline{0011}_2$	4	0001
ab	0.6	0.5	0.1_2	2	10
ba	0.1	0.85	$0.110\overline{1100}_2$	5	11011
bb	0.1	0.95	$0.11\overline{1100}_2$	5	11110



Extend to longer sequences

This works but:

- Need $P(\mathbf{x})$ for all $\mathbf{x}$
- Total $|\mathcal{A}|^N$ values for length N
- Huffman has similar complexity but shorter codes.

Interval Coding Sequences

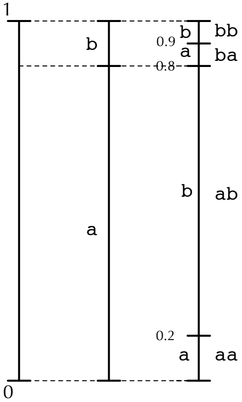
How might we build codes "on the fly"?

$$P(x_1 x_2) = P(x_2 | x_1) P(x_1)$$

Interval Coding Sequences

How might we build codes "on the fly"?

$$P(x_1 x_2) = P(x_2|x_1)P(x_1)$$



Example:

$$P(a) = 0.8 \qquad P(b) = 0.2$$
$$P(a|a) = 0.25 \qquad P(b|a) = 0.75$$
$$P(a|b) = 0.5 \qquad P(b|b) = 0.5$$

Intervals for $P(aa)$ and $P(ab)$:

- Lie inside interval for $P(a)$
- Lengths proportional to $P(a|a)$ and $P(b|a)$

Knowing $P(x_1 \dots x_{N-1})$ and $P(x|x_1 \dots x_{N-1})$ is enough to compute $P(\mathbf{x})$

Building the Code

Recall the *prefix property* of binary intervals:
If $\mathbf{b}'$ is a prefix of $\mathbf{b}$ the interval for $\mathbf{b}$ is contained in the interval for $\mathbf{b}'$.

$$01 \leftrightarrow [0.01, 0.1)_2 = [0.25, 0.5)_{10} \supset [0.375, 0.5)_{10} = [0.011, 0.1)_2 \leftrightarrow \underline{01}1$$

Building the Code

Recall the *prefix property* of binary intervals:

If $\mathbf{b}'$ is a prefix of $\mathbf{b}$ the interval for $\mathbf{b}$ is contained in the interval for $\mathbf{b}'$.

$$01 \leftrightarrow [0.01, 0.1)_2 = [0.25, 0.5)_{10} \supset [0.375, 0.5)_{10} = [0.011, 0.1)_2 \leftrightarrow \underline{01}1$$

This means:

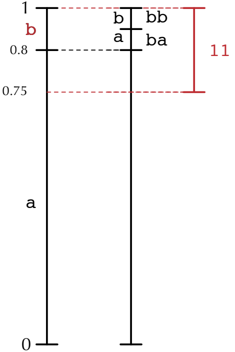
If interval for $\mathbf{x}$ is in the interval for code $\mathbf{b}$, the code for $\mathbf{x}$ starts with $\mathbf{b}$

Building the Code

Recall the *prefix property* of binary intervals:
If $\mathbf{b}'$ is a prefix of $\mathbf{b}$ the interval for $\mathbf{b}$ is contained in the interval for $\mathbf{b}'$.

$$01 \leftrightarrow [0.01, 0.1)_2 = [0.25, 0.5)_{10} \supset [0.375, 0.5)_{10} = [0.011, 0.1)_2 \leftrightarrow \underline{01}1$$

This means:
If interval for $\mathbf{x}$ is in the interval for code $\mathbf{b}$, the code for $\mathbf{x}$ starts with $\mathbf{b}$



Example:

- The probability $P(b) = 0.2$ so b has interval is $[0.8, 1.0)$.
- This lies inside the interval $[0.75, 1) = [0.11, 1.0)_2$ for 11.
- The intervals for ba and bb must also lie in $[0.11, 1.0)_2$
- So codes for ba and bb must start with 11.

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \dots x_n$

Arithmetic Coding of stream $x_1 x_2 \dots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \ldots x_n$

Arithmetic Coding of stream $x_1 x_2 \ldots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, F = c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
    1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \ldots x_n$

Arithmetic Coding of stream $x_1 x_2 \ldots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
   1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
   2. Update interval $[u, v)$ to $[u + (v - u)F(a_{i-1}), u + (v - u)F(a_i))$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \ldots x_n$

Arithmetic Coding of stream $x_1 x_2 \ldots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
    1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
    2. Update interval $[u, v)$ to $[u + (v - u)F(a_{i-1}), u + (v - u)F(a_i))$
    3. Find longest $\mathbf{b}'$ such that $[u, v) \subset [0.\mathbf{b}\mathbf{b}', 0.\mathbf{b}\mathbf{b}'\overline{1})_2$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \ldots x_n$

Arithmetic Coding of stream $x_1 x_2 \ldots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
   1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
   2. Update interval $[u, v)$ to $[u + (v - u)F(a_{i-1}), u + (v - u)F(a_i))$
   3. Find longest $\mathbf{b}'$ such that $[u, v) \subset [0.\mathbf{b}\mathbf{b}', 0.\mathbf{b}\mathbf{b}'\overline{1})_2$
   4. Output $\mathbf{b}'$ and update $\mathbf{b}$ to $\mathbf{b}\mathbf{b}'$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \dots x_n$

Arithmetic Coding of stream $x_1 x_2 \dots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
   1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
   2. Update interval $[u, v)$ to $[u + (v - u)F(a_{i-1}), u + (v - u)F(a_i))$
   3. Find longest $\mathbf{b}'$ such that $[u, v) \subset [0.\mathbf{bb}', 0.\mathbf{bb}'\overline{1})_2$
   4. Output $\mathbf{b}'$ and update $\mathbf{b}$ to $\mathbf{bb}'$
   5. Let $\mathbf{p} = G(\mathbf{x})$ and F be c.d.f. for $\mathbf{p}$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \dots x_n$

Arithmetic Coding of stream $x_1 x_2 \dots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
   1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
   2. Update interval $[u, v)$ to $[u + (v-u)F(a_{i-1}), u + (v-u)F(a_i))$
   3. Find longest $\mathbf{b}'$ such that $[u, v) \subset [0.\mathbf{b}\mathbf{b}', 0.\mathbf{b}\mathbf{b}'\overline{1})_2$
   4. Output $\mathbf{b}'$ and update $\mathbf{b}$ to $\mathbf{b}\mathbf{b}'$
   5. Let $\mathbf{p} = G(\mathbf{x})$ and F be c.d.f. for $\mathbf{p}$
2. Find the shortest $\mathbf{b}'$ so that $[0.\mathbf{b}\mathbf{b}', 0.\mathbf{b}\mathbf{b}'\overline{1})_2 \subset [u + \frac{1}{2}(v-u), v]$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \ldots x_n$

Arithmetic Coding of stream $x_1 x_2 \ldots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
    1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
    2. Update interval $[u, v)$ to $[u + (v - u)F(a_{i-1}), u + (v - u)F(a_i))$
    3. Find longest $\mathbf{b}'$ such that $[u, v) \subset [0.\mathbf{bb}', 0.\mathbf{bb}'\overline{1})_2$
    4. Output $\mathbf{b}'$ and update $\mathbf{b}$ to $\mathbf{bb}'$
    5. Let $\mathbf{p} = G(\mathbf{x})$ and F be c.d.f. for $\mathbf{p}$
2. Find the shortest $\mathbf{b}'$ so that $[0.\mathbf{bb}', 0.\mathbf{bb}'\overline{1})_2 \subset [u + \frac{1}{2}(v - u), v]$
3. Output $\mathbf{b}'$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \ldots x_n$

Arithmetic Coding of stream $x_1 x_2 \ldots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
    1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
    2. Update interval $[u, v)$ to $[u + (v-u)F(a_{i-1}), u + (v-u)F(a_i))$
    3. Find longest $\mathbf{b}'$ such that $[u, v) \subset [0.\mathbf{b}\mathbf{b}', 0.\mathbf{b}\mathbf{b}'\overline{1})_2$
    4. Output $\mathbf{b}'$ and update $\mathbf{b}$ to $\mathbf{b}\mathbf{b}'$
    5. Let $\mathbf{p} = G(\mathbf{x})$ and F be c.d.f. for $\mathbf{p}$
2. Find the shortest $\mathbf{b}'$ so that $[0.\mathbf{b}\mathbf{b}', 0.\mathbf{b}\mathbf{b}'\overline{1})_2 \subset [u + \frac{1}{2}(v-u), v]$
3. Output $\mathbf{b}'$

Arithmetic Coding: A Sketch

Suppose we have a guesser G that predicts a distribution $\mathbf{p}$ over $\mathcal{A}$ after seeing the sequence $\mathbf{x} = x_1 \dots x_n$

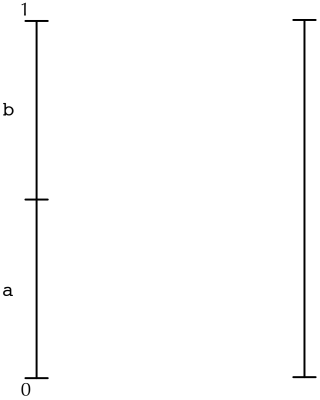
Arithmetic Coding of stream $x_1 x_2 \dots$ using G

Set $[u, v) = [0, 1)$, $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $\mathbf{p} = G(\epsilon)$, $F =$ c.d.f. for $\mathbf{p}$,

1. While there are more symbols in the stream:
    1. Append the next symbol $x = a_i \in \mathcal{A}$ to $\mathbf{x}$
    2. Update interval $[u, v)$ to $[u + (v - u)F(a_{i-1}), u + (v - u)F(a_i))$
    3. Find longest $\mathbf{b}'$ such that $[u, v) \subset [0.\mathbf{bb}', 0.\mathbf{bb}'\overline{1})_2$
    4. Output $\mathbf{b}'$ and update $\mathbf{b}$ to $\mathbf{bb}'$
    5. Let $\mathbf{p} = G(\mathbf{x})$ and F be c.d.f. for $\mathbf{p}$
2. Find the shortest $\mathbf{b}'$ so that $[0.\mathbf{bb}', 0.\mathbf{bb}'\overline{1})_2 \subset [u + \frac{1}{2}(v - u), v]$
3. Output $\mathbf{b}'$

- Interval $[u + (v - u)F(a_{i-1}), u + (v - u)F(a_i))$ is for a_i inside $[u, v]$.
- Adding $\overline{1}$ to end of binary string $\mathbf{b}$ is equivalent to incrementing.
- Code: `https://gist.github.com/mreid/eacfa4dc78c6b142c78f`

Input: ϵ
Output: ϵ

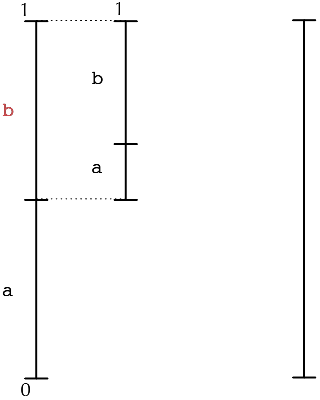


Steps:

- $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $[u, v) = [0, 1)$
  $G(\mathbf{x}) = (0.5, 0.5)$

Arithmetic Coding: Example

Input: b
Output: ϵ

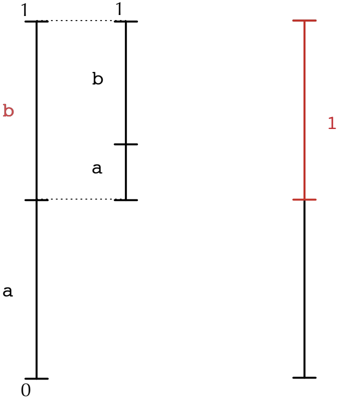


Steps:

- $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $[u, v) = [0, 1)$
  $G(\mathbf{x}) = (0.5, 0.5)$
- $\mathbf{x} = \text{b}$, $\mathbf{b} = \epsilon$, $[u, v) = [0.5, 1)$
  $G(\mathbf{x}) = (0.34, 0.66)$

Arithmetic Coding: Example

Input: b b
Output: 1

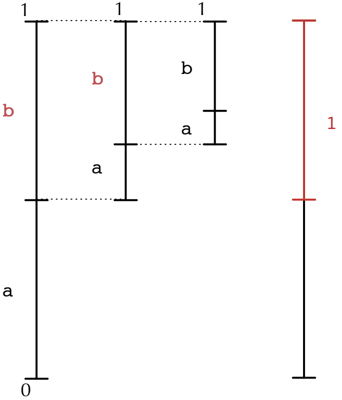


Steps:

- $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $[u, v) = [0, 1)$
  $G(\mathbf{x}) = (0.5, 0.5)$
- $\mathbf{x} = \text{b}$, $\mathbf{b} = 1$, $[u, v) = [0.5, 1)$
  $G(\mathbf{x}) = (0.34, 0.66)$

Arithmetic Coding: Example

Input: `b b`
Output: 1

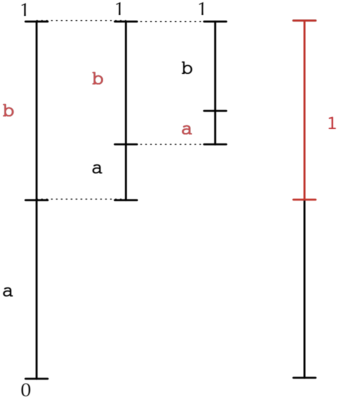


Steps:

- $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $[u, v) = [0, 1)$
  $G(\mathbf{x}) = (0.5, 0.5)$
- $\mathbf{x} = \text{b}$, $\mathbf{b} = 1$, $[u, v) = [0.5, 1)$
  $G(\mathbf{x}) = (0.34, 0.66)$
- $\mathbf{x} = \text{bb}$, $\mathbf{b} = 1$, $[u, v) = [0.67, 1)$
  $G(\mathbf{x}) = (0.25, 0.75)$

Arithmetic Coding: Example

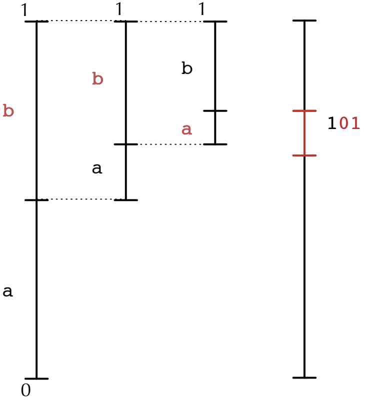
Input: b b a
Output: 1



Steps:

- $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $[u, v) = [0, 1)$
  $G(\mathbf{x}) = (0.5, 0.5)$
- $\mathbf{x} = \mathtt{b}$, $\mathbf{b} = 1$, $[u, v) = [0.5, 1)$
  $G(\mathbf{x}) = (0.34, 0.66)$
- $\mathbf{x} = \mathtt{bb}$, $\mathbf{b} = 1$, $[u, v) = [0.67, 1)$
  $G(\mathbf{x}) = (0.25, 0.75)$
- $\mathbf{x} = \mathtt{bba}$, $\mathbf{b} = 1$, $[u, v) = [0.67, 0.75)$
  $G(\mathbf{x}) = \ldots$

Input: b b a
Output: 1 0 1

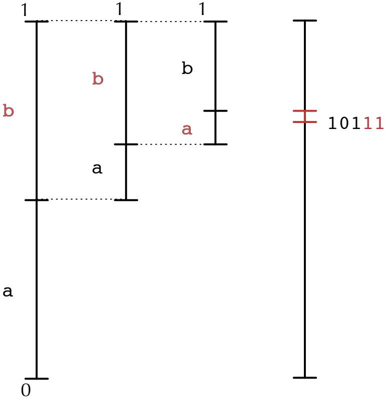


Steps:

- $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $[u, v) = [0, 1)$
  $G(\mathbf{x}) = (0.5, 0.5)$
- $\mathbf{x} = \text{b}$, $\mathbf{b} = 1$, $[u, v) = [0.5, 1)$
  $G(\mathbf{x}) = (0.34, 0.66)$
- $\mathbf{x} = \text{bb}$, $\mathbf{b} = 1$, $[u, v) = [0.67, 1)$
  $G(\mathbf{x}) = (0.25, 0.75)$
- $\mathbf{x} = \text{bba}$, $\mathbf{b} = 101$, $[u, v) = [0.67, 0.75)$
  $G(\mathbf{x}) = \ldots$

Arithmetic Coding: Example

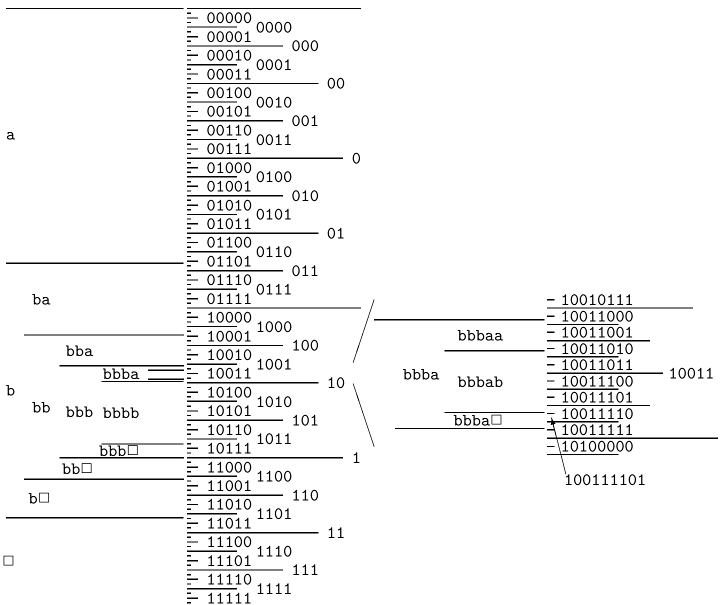
Input: b b a
Output: 1 0 1 1 1



Steps:

- $\mathbf{x} = \epsilon$, $\mathbf{b} = \epsilon$, $[u, v) = [0, 1)$
  $G(\mathbf{x}) = (0.5, 0.5)$
- $\mathbf{x} = \text{b}$, $\mathbf{b} = 1$, $[u, v) = [0.5, 1)$
  $G(\mathbf{x}) = (0.34, 0.66)$
- $\mathbf{x} = \text{bb}$, $\mathbf{b} = 1$, $[u, v) = [0.67, 1)$
  $G(\mathbf{x}) = (0.25, 0.75)$
- $\mathbf{x} = \text{bba}$, $\mathbf{b} = 101$, $[u, v) = [0.67, 0.75)$
  $G(\mathbf{x}) = \ldots$
- $\mathbf{x} = \text{bba}$, $\mathbf{b} = 10111$

Arithmetic Coding: Example (MacKay, Figure 6.4)



Decoding

How do we decode a sequence of bits?

Rough Sketch:

- Guesser G gives initial distribution $G(\epsilon)$
- Keep reading bits until binary interval in a symbol interval
- Output that symbol and pass to G for next distribution
- Repeat

Homework:

- Describe above decoding method in more detail
  (e.g., like previous sketch for coding)
- Running through decoding output of previous example

Building a Better Guesser

So far we assume the sequence of probabilities are given in advance.

In **Lecture 5**, you saw Bernoulli distribution for two outcomes

- Beta distribution $\text{Beta}(\theta|m_h, m_t)$ as a prior for $\text{Bern}(x|\theta)$
- The posterior after observing n_h heads and n_t tails is just $\text{Beta}(\theta|m_h + n_h, m_t + n_t)$
- The expected value of θ is

$$p(x = \mathtt{h}|\mathcal{D}, m_h, m_t) = \frac{m_h + n_h}{m_h + n_h + m_t + n_t}$$

Building a Better Guesser

So far we assume the sequence of probabilities are given in advance.

In **Lecture 5**, you saw Bernoulli distribution for two outcomes

- Beta distribution $\text{Beta}(\theta|m_h, m_t)$ as a prior for $\text{Bern}(x|\theta)$
- The posterior after observing n_h heads and n_t tails is just $\text{Beta}(\theta|m_h + n_h, m_t + n_t)$
- The expected value of θ is

$$p(x = \mathtt{h}|\mathcal{D}, m_h, m_t) = \frac{m_h + n_h}{m_h + n_h + m_t + n_t}$$

Example: Start with $m_h = m_t = 1$ and observe sequence $\mathtt{hht}$.

$$G(\epsilon) = (\tfrac{1}{2}, \tfrac{1}{2}),\ G(\mathtt{h}) = (\tfrac{2}{3}, \tfrac{1}{3}),\ G(\mathtt{hh}) = (\tfrac{3}{4}, \tfrac{1}{4})\ G(\mathtt{hht}) = (\tfrac{3}{5}, \tfrac{2}{5})$$

Laplace's Rule

Dirichlet Model

A **Dirichlet distribution** is a generalisation of the Beta distribution to more than two outcomes. Its parameter is a vector $\mathbf{m} = (m_1, \ldots, m_K)$ can be viewed as "virtual counts" for each symbol $a_1, \ldots, a_K$:

$$P(x = a_i | x_1 \ldots x_n) = \frac{\sharp(a_i) + m_i}{\sum_{k=1}^{K} \sharp(a_k) + m_k}$$

Can implement an adaptive guesser by just counting symbol occurrences.

Dirichlet Model

A **Dirichlet distribution** is a generalisation of the Beta distribution to more than two outcomes. Its parameter is a vector $\mathbf{m} = (m_1, \ldots, m_K)$ can be viewed as "virtual counts" for each symbol $a_1, \ldots, a_K$:

$$P(x = a_i | x_1 \ldots x_n) = \frac{\sharp(a_i) + m_i}{\sum_{k=1}^{K} \sharp(a_k) + m_k}$$

Can implement an adaptive guesser by just counting symbol occurrences.

Benefits:

- A deterministic guesser G specified by "virtual counts" $\mathbf{m}$
- Flexible
  - e.g., Choose $\mathbf{m}$ to be frequency of English letters
  - $\sum_k m_k$ Large = Stable; Small = Responsive

Summary and Reading

Main Points

- Arithmetic Coding:
  - Uses interval coding and conditional probability
  - Separates coding and prediction
  - No need to compute every code word
  - Sketch of coding algorithm
- Predictive distributions:
  - Update distribution after each symbol
  - Beta and Dirichlet priors = virtual counts

Reading

- Section 6.2 of MacKay