

Information Theory and Networks

Lecture 4: Uncertainty and Entropy

Matthew Roughan

<matthew.roughan@adelaide.edu.au>

[http://www.maths.adelaide.edu.au/matthew.roughan/
Lecture_notes/InformationTheory/](http://www.maths.adelaide.edu.au/matthew.roughan/Lecture_notes/InformationTheory/)

School of Mathematical Sciences,
University of Adelaide

September 18, 2013

Part I

Uncertainty and Entropy

In the beginning was the word ...

John 1:1

Think about the early chaps, who discovered writing. It must have been as close to magic as you can imagine. Suddenly, through strange marks on clay they could store and transmit numbers and words precisely.

Morse code test

- / / -. --- - / .- / -. .- .- .- .-

Morse code test

..... / / -. --- - / .- / -. .- .- .- .-

Things to think about:

- Note there are really 4 symbols: dot, dash, letter-break, word-break
- Are dots or dashes more common in the resulting code?
- How do you learn it? Shortest first? Does that work well?

Symbols

We take symbols for granted: we are taught them at an early age, and our entire consciousness is formed around language and symbols, so we don't really appreciate what they do for us.

- Start with symbols for things: pictograms
 - ▶ many to learn – one per word
 - ▶ specialised knowledge
- Alphabets code for small bits of words
 - ▶ anyone can learn
 - ▶ any language can be expressed in the one alphabet (almost)
 - ▶ but its a profound jump from pictograms
 - ▶ in turn this shapes language, and how powerful it can become
- What about numbers?
 - ▶ how recent is our notation?

Symbols

We take symbols for granted: we are taught them at an early age, and our entire consciousness is formed around language and symbols, so we don't really appreciate what they do for us.

- Start with symbols for things: pictograms
 - ▶ many to learn – one per word
 - ▶ specialised knowledge
- Alphabets code for small bits of words
 - ▶ anyone can learn
 - ▶ any language can be expressed in the one alphabet (almost)
 - ▶ but its a profound jump from pictograms
 - ▶ in turn this shapes language, and how powerful it can become
- What about numbers?
 - ▶ how recent is our notation?

- A for 'Orses
- B for Mutton
- C for Yourself
- D for Mation
- E for Brick
- F for Vescence
- G for Police
- H for Consent
- I for Lutin
- J for Orange
- K for Teria
- L for Leather
- M for Sis
- N for Mation
- O for A Muse of Fire
- P for Ate
- Q for A Song
- S for Something Else
- T for Two
- U for Mism
- V for La France
- W for Mism
- X for Breakfast
- Y for Lover
- Z (zee) for yourself

What was all that?

- We are going to be talking about transmission of information, so we need to know the form it takes:
 - ▶ some sequence of abstract symbols
- Even if a message is long, its information content may be small, e.g., I could say
 - ▶ “11111111111111111111111111111111” or “30 1s”or
 - ▶ “110010010000111111011010101000100” or ??so we need a better idea to express information
- More to the point, if I only ever send the two messages:

1010101010110 and 1111111111111

I could replace them with X and Y

- ▶ I still have 2 symbols, but messages are shorter
- ▶ So information isn't a function of the messages themselves!!!



What was all that?

We will assume we know a probability distribution on the messages, i.e., we know the PMF p_i

What was all that?
• We are going to be talking about transmission of information, so we need to know the form it takes:
• some sequence of abstract symbols
• Even if a message is long, its information content may be small, e.g., I could say
• “11111111111111111111111111111111” or “30 1s”
or
• “110010010000111111011010101000100” or ??
so we need a better idea to express information
• More to the point, if I only ever send the two messages:
1010101010110 and 1111111111111
I could replace them with X and Y
• I still have 2 symbols, but messages are shorter
• So information isn't a function of the messages themselves!!!

Uncertainty and Information and Surprise

- Information cancels out uncertainty
 - ▶ think of uncertainty as not knowing which symbol was transmitted
 - ▶ when you receive the signal (information) the uncertainty is removed
- Fundamentally, to understand information, we have to understand uncertainty
 - ▶ implicitly, the information of an event or message, depends on the ensemble of all possible events, e.g., how much information is there in
 - ★ $X = 1$ in the context of X always equals 1?
 - ★ $X = 1$ when it could take many other values?
- We might improve our intuition about information, if we think of high-information content messages as being more surprising:
 - ▶ e.g., how surprising is
 - ★ $X = 1$ in the context of X always equals 1?
 - ★ $X = 1$ when it could take many other values?
 - ▶ so information should be a function of probability of the message.



Uncertainty and Information and Surprise

Uncertainty and Information and Surprise
• Information cancels out uncertainty
• think of uncertainty as not knowing which symbol was transmitted
• when you receive the signal (information) the uncertainty is removed
• Fundamentally, to understand information, we have to understand uncertainty
• implicitly, the information of an event or message, depends on the ensemble of all possible events, e.g., how much information is there in?
• $X = 1$ in the context of X always equals 1?
• $X = 1$ when it could take many other values?
• We might improve our intuition about information, if we think of high-information content messages as being more surprising.
• e.g., how surprising is
• $X = 1$ in the context of X always equals 1?
• $X = 1$ when it could take many other values?
• so information should be a function of probability of the message.

Can we come up with some axioms?

What properties should a metric of information have?

Can we come up with some axioms?

Information

Lets think about information we get from an event

- Say the event has probability p , and I tell you it occurs, then say I am conveying information $I(p)$.
- Simple things
 - ▶ its a **metric**, and so should be number.
 - ▶ can't have negative information
 - ★ $I(p) \geq 0$
 - ▶ a small change in p leads to a small change in $I(p)$
 - ★ $I(p)$ is continuous
 - ▶ we want it to differentiate between cases so

$$I(p) = I(q) \text{ only if } p = q.$$

Lets think about information we get from an event

- Say the event has probability p , and I tell you it occurs, then say I am conveying information $I(p)$.
- Simple things
 - ▶ its a **metric**, and so should be number.
 - ▶ can't have negative information
 - ★ $I(p) \geq 0$
 - ▶ a small change in p leads to a small change in $I(p)$
 - ★ $I(p)$ is continuous
 - ▶ we want it to differentiate between cases so

$I(p) = I(q)$ only if $p = q$.

Information

Information

Lets think about information we get from a pair of events

- Say the event has probability p , and I tell you it occurs, then say I am conveying information $I(p)$.
- Events that are less likely convey more information
 - ▶ e.g., if I pick a card from a deck and tell you it is an ACE it conveys more than if I tell you its a SPADE
 - ▶ i.e.

$$\text{if } p \leq q, \text{ then } I(p) \geq I(q).$$



Information

Information
Lets think about information we get from a pair of events
• Say the event has probability p , and I tell you it occurs, then say I am conveying information $I(p)$.
• Events that are less likely convey more information
• e.g., if I pick a card from a deck and tell you it is an ACE it conveys more than if I tell you its a SPADE
• i.e.
if $p \leq q$, then $I(p) \geq I(q)$.

Information

Lets think about information we get from an event

- Say the event has probability p , and I tell you it occurs, then say I am conveying information $I(p)$.
- What happens with two messages or events?
 - ▶ e.g. imagine I tell you that
 - ★ a card is an ACE
 - ★ a card is a SPADE
 - ▶ reasonable hypothesis is that if the two events/messages are **independent**, then the information from the two adds, i.e.,

$$I(\text{ACE OF SPADES}) = I(\text{ACE}) + I(\text{SPADE})$$

- ▶ independent events have $P(A \cap B) = P(A)P(B)$ so

$$I(pq) = I(p) + I(q)$$

- ▶ Note that if we take $p = q = 1$, we get $I(1.1) = I(1) + I(1)$, so $I(1) = 0$.



Information

Information
Lets think about information we get from an event
• Say the event has probability p , and I tell you it occurs, then say I am conveying information $I(p)$.
• What happens with two messages or events?

- e.g. imagine I tell you that
 - a card is an ACE
 - a card is a SPADE
- reasonable hypothesis is that if the two events/messages are **independent**, then the information from the two adds, i.e.,
 $I(\text{ACE OF SPADES}) = I(\text{ACE}) + I(\text{SPADE})$
- independent events have $P(A \cap B) = P(A)P(B)$ so
 $I(pq) = I(p) + I(q)$
- Note that if we take $p = q = 1$, we get $I(1.1) = I(1) + I(1)$, so $I(1) = 0$.

$I(1) = 0$ makes sense, as there is no uncertainty in an event with a fixed outcome.

Information Axioms

For all $p, q \in (0, 1)$

- 1 Continuous, real-valued, non-negative function
- 2 Decreasing, and distinguishes values

if $p < q$, then $I(p) > I(q)$.

- 3 Independent events have

$$I(pq) = I(p) + I(q)$$

Theorem

The only function $I(\cdot)$ which satisfies the above axioms is

$$I(p) = -k \log(p), \text{ for some constant } k.$$

Sort of makes sense as we need $n = \log(m)$ bits to represent a number of size m (think of numbers as our possible messages), so if all numbers up to m were equally possible, then ...



Information Axioms

Information Axioms
 For all $p, q \in (0, 1)$
 1 Continuous, real-valued, non-negative function
 2 Decreasing, and distinguishes values
 if $p < q$, then $I(p) > I(q)$
 3 Independent events have
 $I(pq) = I(p) + I(q)$

Theorem
 The only function $I(\cdot)$ which satisfies the above axioms is
 $I(p) = -k \log(p)$, for some constant k .

Sort of makes sense as we need $n = \log(m)$ bits to represent a number of size m (think of numbers as our possible messages), so if all numbers up to m were equally possible, then ...

We chose $I(p) = -\log_2(p)$.
 Special cases:

- $I(0) = +\infty$
- $I(1) = 0$
- $I(2^{-n}) = n$

Theorem

The only function $I(\cdot)$ which satisfies the above axioms is

$$I(p) = -k \log(p), \text{ for some constant } k.$$

Proof.

It is easy to show $I(\cdot)$ satisfies the axioms, so other direction:

- 1 As noted $I(pq) = I(p) + I(q)$ implies $I(1) = 0$.
- 2 $I(pq) = I(p) + I(q)$ also implies

$$I(p^k) = k I(p),$$

which we can show by induction taking $p^{k+1} = p \times p^k$, so

$$I(p^{k+1}) = I(p) + I(p^k) = I(p) + k I(p) = (k + 1) I(p).$$



Theorem
 The only function $I(\cdot)$ which satisfies the above axioms is
 $I(p) = -k \log(p)$, for some constant k .

Proof.
 It is easy to show $I(\cdot)$ satisfies the axioms, so other direction:
 1 As noted $I(pq) = I(p) + I(q)$ implies $I(1) = 0$.
 2 $I(pq) = I(p) + I(q)$ also implies
 $I(p^k) = k I(p)$,
 which we can show by induction taking $p^{k+1} = p \times p^k$, so
 $I(p^{k+1}) = I(p) + I(p^k) = I(p) + k I(p) = (k + 1) I(p)$.

Proof (Cont.)

- Take a probability p , then for any positive integer r there exists a k such that

$$p^{k+1} \leq (1/2)^r < p^k$$

From monotonicity

$$I(p^{k+1}) \geq I(2^{-r}) > I(p^k)$$

And from previous result

$$(k+1)I(p) \geq rI(2^{-1}) > kI(p)$$

or

$$\frac{(k+1)}{r} \geq \frac{I(2^{-1})}{I(p)} > \frac{k}{r}$$

Proof (Cont.)

Take a probability p , then for any positive integer r there exists a k such that $p^{k+1} \leq (1/2)^r < p^k$

From monotonicity $I(p^{k+1}) \geq I(2^{-r}) > I(p^k)$

And from previous result $(k+1)I(p) \geq rI(2^{-1}) > kI(p)$

or $\frac{(k+1)}{r} \geq \frac{I(2^{-1})}{I(p)} > \frac{k}{r}$

Proof (Cont.)

$$\frac{(k+1)}{r} \geq \frac{I(2^{-1})}{I(p)} > \frac{k}{r}$$

and we also know from properties of logs and similar argument

$$\frac{(k+1)}{r} \geq \frac{\log(2^{-1})}{\log(p)} > \frac{k}{r}$$

So the two middle terms can differ by no more than $1/r$, i.e.,

$$\left| \frac{\log(2^{-1})}{\log(p)} - \frac{I(2^{-1})}{I(p)} \right| < \frac{1}{r}$$

We fixed p so take the limit as $r \rightarrow \infty$ and the two must converge so

$$I(p) = C \log(p)$$

and the constant $C = I(1/2)/\log(1/2)$ is determined by $I(1/2)$ which is arbitrary, depending on units, and implies the base of the log.

Proof (Cont.)

$$\frac{(k+1)}{r} \geq \frac{I(2^{-1})}{I(p)} > \frac{k}{r}$$

and we also know from properties of logs and similar argument $\frac{(k+1)}{r} \geq \frac{\log(2^{-1})}{\log(p)} > \frac{k}{r}$

So the two middle terms can differ by no more than $1/r$, i.e., $\left| \frac{\log(2^{-1})}{\log(p)} - \frac{I(2^{-1})}{I(p)} \right| < \frac{1}{r}$

We fixed p so take the limit as $r \rightarrow \infty$ and the two must converge so $I(p) = C \log(p)$

and the constant $C = I(1/2)/\log(1/2)$ is determined by $I(1/2)$ which is arbitrary, depending on units, and implies the base of the log.

We usually scale so that $I(1/2) = 1$, i.e., the units are bits.

Can we come up with some axioms?

What properties should a metric of uncertainty have?

- Obviously similar/related to information
- Our idea of information of a message isn't good enough because it is about one message, and we need to deal with all of the possible messages.
- What else can we say?

Can we come up with some axioms?

- What properties should a metric of uncertainty have?
- Obviously similar/related to information
 - Our idea of information of a message isn't good enough because it is about one message, and we need to deal with all of the possible messages.
 - What else can we say?

Uncertainty

Lets think from uncertainty viewpoint

- Its a **metric**, and so should be real, non-negative number.
- We are talking about (discrete) probabilistic systems, so lets make it a function of the PMF.

$$uncertainty = H(p_1, p_2, \dots, p_n)$$

- ▶ it doesn't depend on the messages themselves
- If two distributions are just reordered versions of each other, e.g., $(q_1, q_2) = (p_2, p_1)$, then that shouldn't change uncertainty.
- Our measure should increase as "uncertainty" increases
 - ▶ maybe we should make it continuous?
 - ▶ are there any other rules?

Uncertainty

- Lets think from uncertainty viewpoint
- Its a **metric**, and so should be real, non-negative number.
 - We are talking about (discrete) probabilistic systems, so lets make it a function of the PMF.
 $uncertainty = H(p_1, p_2, \dots, p_n)$
 - it doesn't depend on the messages themselves
 - If two distributions are just reordered versions of each other, e.g., $(q_1, q_2) = (p_2, p_1)$, then that shouldn't change uncertainty.
 - Our measure should increase as "uncertainty" increases
 - ▶ maybe we should make it continuous?
 - ▶ are there any other rules?

Can we come up with some axioms for uncertainty?

It should increase as “uncertainty” increases.

- Consider a Bernoulli trial with $\Omega = \{0, 1\}$, and probability p of success.
 - ▶ we are least uncertain when $p = 0$ or 1 because the outcome is fixed.
 - ▶ most uncertain when $p = 1/2$
 - ▶ so we need a function of $(p, 1 - p)$ with its min for $p = 0$ or 1 , and max for $p = 1/2$
 - ▶ also $H(p, 1 - p) = H(1 - p, p)$ so it has symmetry

Can we come up with some axioms for uncertainty?

It should increase as “uncertainty” increases.
• Consider a Bernoulli trial with $\Omega = \{0, 1\}$, and probability p of success:

- we are least uncertain when $p = 0$ or 1 because the outcome is fixed.
- most uncertain when $p = 1/2$
- so we need a function of $(p, 1 - p)$ with its min for $p = 0$ or 1 , and max for $p = 1/2$
- also $H(p, 1 - p) = H(1 - p, p)$ so it has symmetry

Can we come up with some axioms for uncertainty?

It should increase as “uncertainty” increases.

- If we have uniform distributions with M possibilities $p_i = 1/M$, then uncertainty should increase as M increases as there are more possible outcomes.
 - ▶ from previous discussion of information, it probably makes sense for it to increase logarithmically
 - ▶ we can get that again from assuming the distribution is uniform over $\{1, \dots, M\} \times \{1, \dots, L\}$, and noting there are ML possible events, but if we condition on one there are M or L left, and so we get the same type of sum we saw for information:

$$f(ML) = f(M) + f(L)$$

Can we come up with some axioms for uncertainty?

It should increase as “uncertainty” increases.
• If we have uniform distributions with M possibilities $p_i = 1/M$, then uncertainty should increase as M increases as there are more possible outcomes.

- from previous discussion of information, it probably makes sense for it to increase logarithmically
- we can get that again from assuming the distribution is uniform over $\{1, \dots, M\} \times \{1, \dots, L\}$ and noting there are ML possible events, but if we condition on one there are M or L left, and so we get the same type of sum we saw for information

$$f(ML) = f(M) + f(L)$$

Can we come up with some axioms for uncertainty?

The grouping axiom

- Imagine an experiment with M outcomes and PMF p_i
 - divide the outcomes into two groups

$$A = \{x_1, \dots, x_r\} \text{ and } B = \{x_{r+1}, \dots, x_M\}$$

- where

$$P(A) = \sum_{i=1}^r p_i \text{ and } P(B) = \sum_{i=r+1}^M p_i$$

- We could conduct the experiment two ways:
 - Randomly draw X using the PMF p_i
 - Randomly draw Y using $P(A)$ and $P(B)$ to determine the group and then, draw from the groups:
 - if $Y = A$, then using $q_i = P(X = x_i | Y = A) = p_i / P(A)$, for $i \in A$
 - if $Y = B$, then using $q_j = P(X = x_j | Y = B) = p_j / P(B)$, for $j \in B$



Can we come up with some axioms for uncertainty?

Can we come up with some axioms for uncertainty?
The grouping axiom

- Imagine an experiment with M outcomes and PMF p_i
 - divide the outcomes into two groups
- where

$$A = \{x_1, \dots, x_r\} \text{ and } B = \{x_{r+1}, \dots, x_M\}$$
- We could conduct the experiment two ways:
 - Randomly draw X using the PMF p_i
 - Randomly draw Y using $P(A)$ and $P(B)$ to determine the group and then draw from the groups:
 - if $Y = A$, then using $q_i = P(X = x_i | Y = A) = p_i / P(A)$, for $i \in A$
 - if $Y = B$, then using $q_j = P(X = x_j | Y = B) = p_j / P(B)$, for $j \in B$

Can we come up with some axioms for uncertainty?

The grouping axiom

- Two equivalent ways to do the experiment
 - so must have the same uncertainty
- If we revealed the group selected, the uncertainty would be
 - if $Y = A$, it would be $H(q_1, \dots, q_r)$
 - if $Y = B$, it would be $H(q_{r+1}, \dots, q_M)$
- The expected uncertainty of when the grouping is specified is

$$P(A)H(q_1, \dots, q_r) + P(B)H(q_{r+1}, \dots, q_M)$$

The uncertainty about the grouping is

$$H(P(A), P(B))$$

Total uncertainty of grouped experiment

$$H_{group} = H(P(A), P(B)) + P(A)H(q_1, \dots, q_r) + P(B)H(q_{r+1}, \dots, q_M)$$

- So the uncertainty calculated two ways should be

$$H(p_1, \dots, p_M) = H_{group}$$



Can we come up with some axioms for uncertainty?

Can we come up with some axioms for uncertainty?
The grouping axiom

- Two equivalent ways to do the experiment
 - so must have the same uncertainty
- If we revealed the group selected, the uncertainty would be
 - if $Y = A$, it would be $H(q_1, \dots, q_r)$
 - if $Y = B$, it would be $H(q_{r+1}, \dots, q_M)$
- The expected uncertainty of when the grouping is specified is

$$P(A)H(q_1, \dots, q_r) + P(B)H(q_{r+1}, \dots, q_M)$$
- The uncertainty about the grouping is

$$H(P(A), P(B))$$
- Total uncertainty of grouped experiment

$$H_{group} = H(P(A), P(B)) + P(A)H(q_1, \dots, q_r) + P(B)H(q_{r+1}, \dots, q_M)$$
- So the uncertainty calculated two ways should be

$$H(p_1, \dots, p_M) = H_{group}$$

Entropy

- The only function that satisfies all of these axioms is

$$H(p_1, \dots, p_n) = - \sum_i p_i \log p_i,$$

- ▶ we should be able to see that it is the expectation of the information function we defined earlier
- We call this the Shannon entropy because
 - ▶ given different axioms we might come up with a different function
 - ▶ entropy has a long history, but Shannon was the first to use it in the context of information

• The only function that satisfies all of these axioms is
 $H(p_1, \dots, p_n) = - \sum_i p_i \log p_i$
• we should be able to see that it is the expectation of the information function we defined earlier
• We call this the Shannon entropy because
• given different axioms we might come up with a different function
• entropy has a long history, but Shannon was the first to use it in the context of information

Entropy

Other forms of entropy:

- Hartley:

$$H_0(X) = \log |\Omega|$$

(which is the Shannon entropy on the uniform distribution)

- Min-entropy

$$H_\infty(X) = \min_i (-\log p_i)$$


(smallest possible (Rényi) entropy measure – like looking at the worst case which is needed for ensuring complete randomness, say for cryptography applications)


- Rényi entropy

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_i p_i^\alpha \right)$$

Generalisation that includes all the others as special cases (e.g., Shannon is limit as $\alpha \rightarrow 1$).

Further reading I

 Robert B. Ash, *Information theory*, Dover, 1995, Reprinted from John Wiley, 1965.

 Gjerrit Meinsma, *Data compression & information theory*, Mathematisch cafe, 2003, wwwhome.math.utwente.nl/~meinsmag/onzin/shannon.pdf.