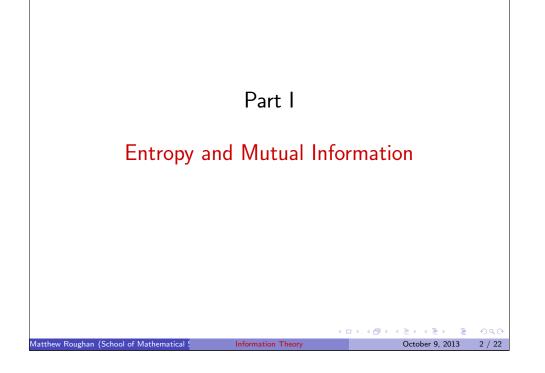
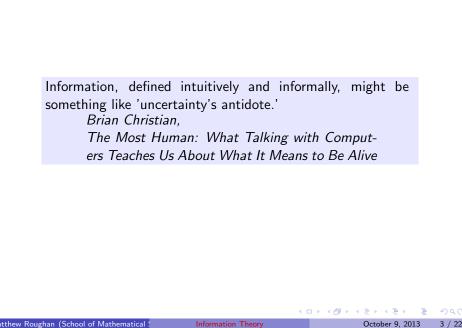
Information Theory and Networks Lecture 6: Entropy and Mutual Information

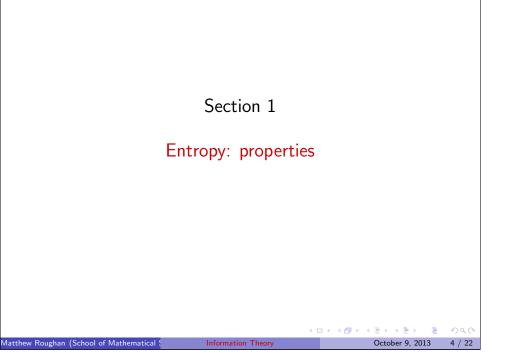
Matthew Roughan <matthew.roughan@adelaide.edu.au> http://www.maths.adelaide.edu.au/matthew.roughan/ Lecture_notes/InformationTheory/

> School of Mathematical Sciences, University of Adelaide

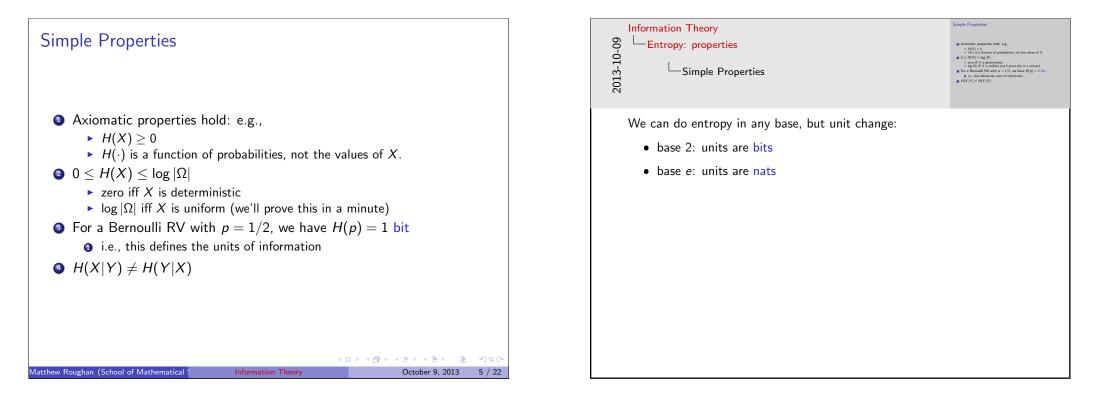
> > October 9, 2013

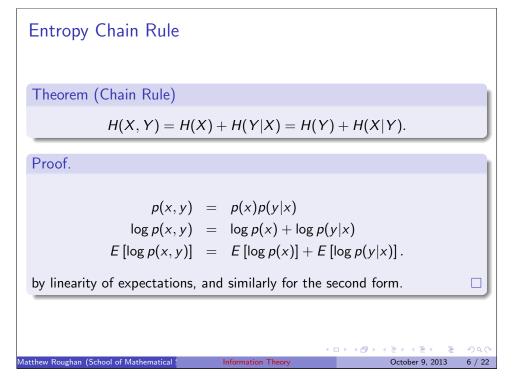


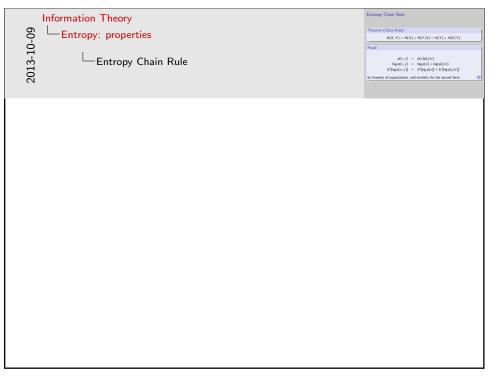


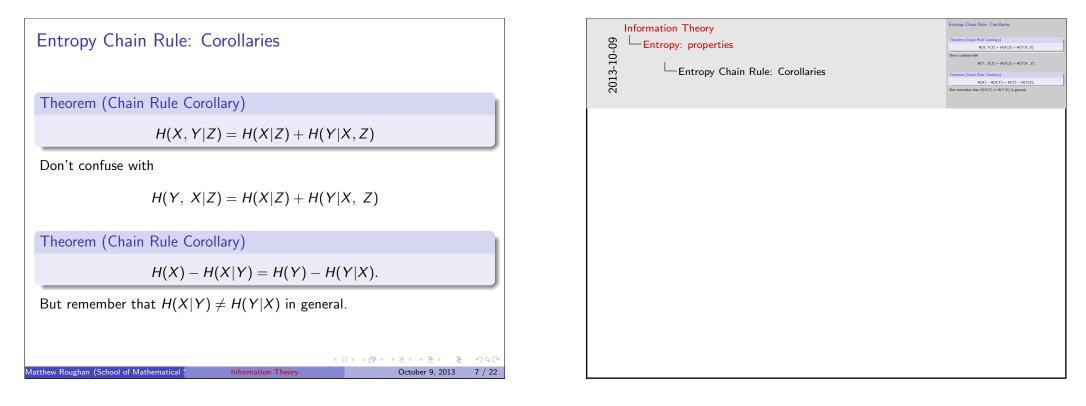


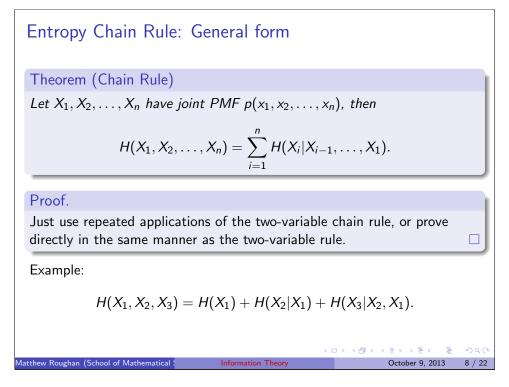
3 / 22

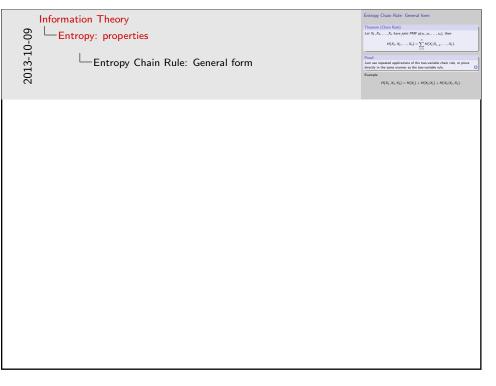


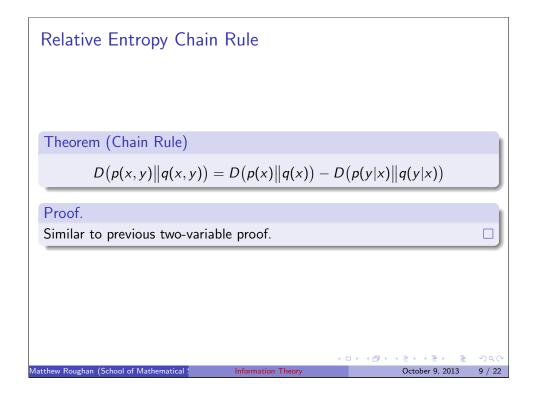












Relative Entropy Properties
Theorem
$D(p\ q)\geq 0$
with equality only iff $p(x) = q(x)$ for all x.
Proof.
$-Dig(pig\ qig) = E\left[-\lograc{p(X)}{q(X)} ight] \leq -\log E\left[rac{p(X)}{q(X)} ight],$
by Jensen's inequality, as $-\log$ is strictly convex, and so equality arises only when p/q is a constant (in this case 1 when $p = q$ for all x). Next
$-D(p \ q) \leq \log E\left[\frac{q(X)}{p(X)}\right] = \log \sum_{x} p(x) \frac{q(x)}{p(x)} = \log \sum_{x} q(x) = \log 1 = 0$
< □ > < 图 > < 言 > 〈 言 > 〈 言 > 〉 見 · うへ
Matthew Roughan (School of Mathematical : Information Theory October 9, 2013 10 / 2

Information Theory		Relative Entropy Chain Rule
66 <u>Entropy: proper</u> 67 Poper 67 Poper 68 Poper 60 Poper 69 Poper 60 P	<mark>ties</mark> Entropy Chain Rule	$\begin{tabular}{ l l l l l l l l l l l l l l l l l l l$



We can prove the result even more directly using Gibbs' Inequality (or the related log-sum inequality).

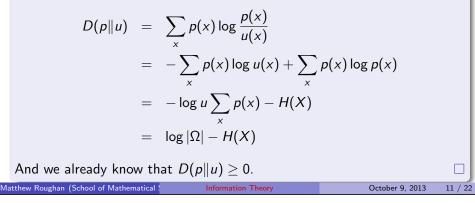
Corollary

Theorem

 $H(X) \leq \log |\Omega|.$

Proof.

Take distributions p(x) and compare it to the uniform distribution $u(x) = 1/|\Omega|$:



Convexity of relative entropy

Theorem

The relative entropy D(p||q) is a convex function of (p, q), i.e., for two pairs of distributions $(p^{(1)}, q^{(1)})$ and $(p^{(2)}, q^{(2)})$.

$$egin{aligned} & \mathcal{D}\Big(\lambda p^{(1)} + (1-\lambda)p^{(2)}\Big\|\lambda q^{(1)} + (1-\lambda)q^{(2)}\Big) \ & \leq & \lambda Dig(p^{(1)}\|q^{(1)}ig) + (1-\lambda)Dig(p^{(2)}\|q^{(2)}ig) \end{aligned}$$

for all $0 \leq \lambda \leq 1$.

Proof.

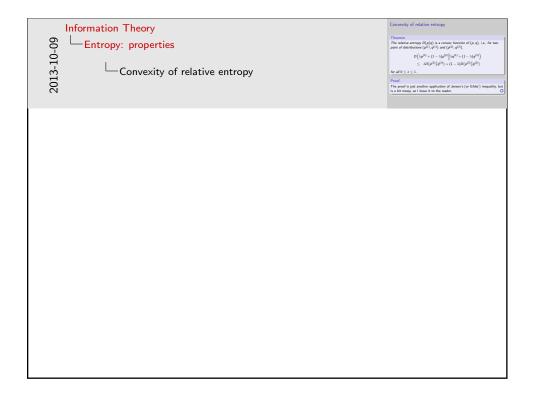
The proof is just another application of Jensen's (or Gibbs') inequality, but is a bit messy, so I leave it to the reader. $\hfill \Box$

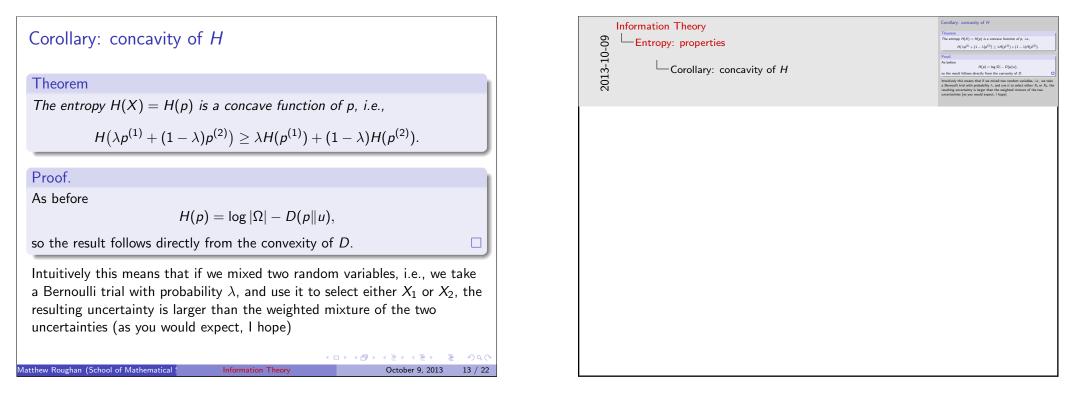
Information Theory

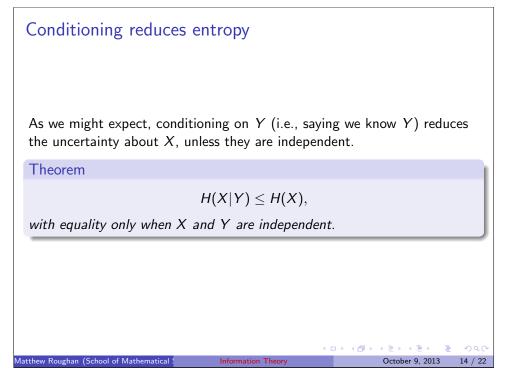
October 9, 2013

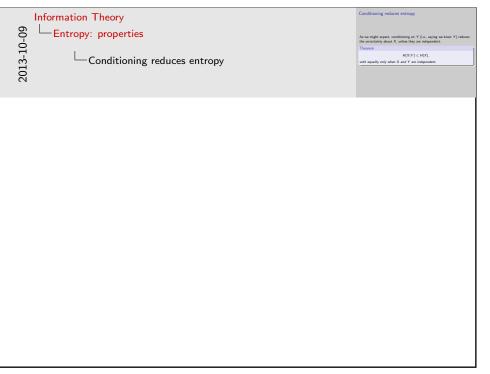
12 / 22

Information Theory Entropy: properties Corollary Cor









Conditioning reduces entropy

Proof.

Given p(x, y) define $q(x, y) = p_X(x)p_Y(y)$, where $p_X(x)$ and $p_Y(y)$ are the marginal distributions of X and Y respectively. Now define

$$I(X;Y) = D(p(x,y)||q(x,y)) = E\left[\log\frac{p(X|Y)}{p_X(X)}\right],$$

By definition of conditional probabilities

atthew Roughan (School of Mathematical :

$$E\left[\log \frac{p(X,Y)}{p_X(X)p_Y(Y)}\right] = E\left[\log \frac{p(X|Y)}{p_X(X)}\right] = E\left[\log p(X|Y)\right] - E\left[\log p_X(X)\right],$$
So

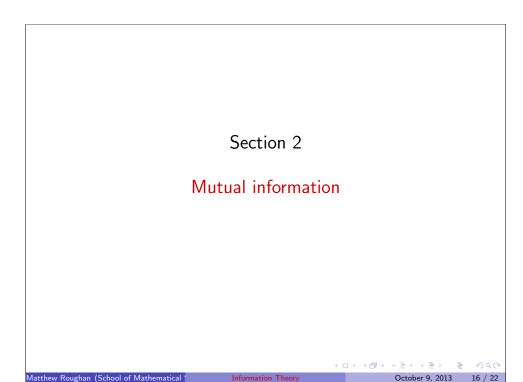
$$H(X;Y) = -H(X|Y) + H(X),$$

but we also know that I(X; Y) is defined in terms of relative entropy, and hence $I(X; Y) \ge 0$, and hence the result.

Information Theory

15 / 22

October 9, 2013





The quantity I(X; Y) is called the mutual information, and we will get to that in a moment. In particular, we'll use the result

$$I(X;Y) = -H(X|Y) + H(X)$$

again so keep you eye on it.

Information Theory Mutual information	Section 2 Metual information

Motivation

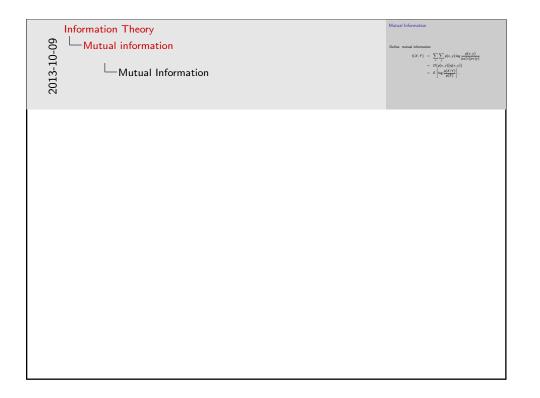
- We created an "information" metric before, based on a single probability, but found that entropy was a more useful idea.
- Now lets return to trying to say something useful about information
- The mutual information is a measure of the information that we learn about one random variable from another.



Information Theory Mutual information Motivation Notivation	Motivation • We coated as "information" motor before, based on a single probability, but floods that entropy uses a new and/a idea. • New last sceners to trying to surgurate that water the second of a structure of inversion is a measure of the information that was last about one random variable from another.

Mutual Information

$$I(X; Y) = \sum_{x} \sum_{y} p(x, y) \log \frac{p(x, y)}{p_X(x)p_Y(y)}$$
 $= D(p(x, y) || q(x, y))$
 $= E\left[\log \frac{p(X|Y)}{p(X)}\right],$



Relationship between entropy and mutual information

We already showed that

$$I(X;Y) = H(X) - H(X|Y).$$

- So the mutual information is the reduction in uncertainty in X given knowledge of Y.
- By symmetry

latthew Roughan (School o<u>f Mathematical</u> :

$$I(X; Y) = H(Y) - H(Y|X).$$

• Also the "self-information"

$$I(X; X) = H(X) - H(X|X) = H(X).$$

・ロト・(四)・(日)・(日)・(日)・

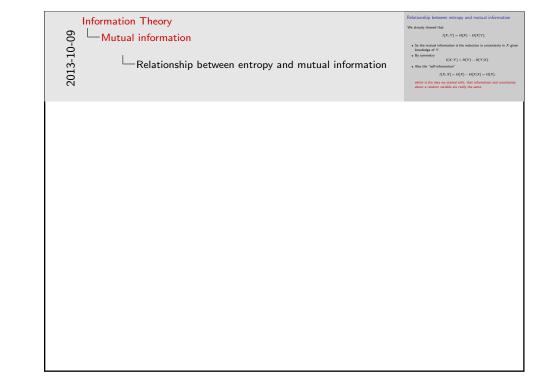
October 9, 2013 19 / 22

which is the idea we started with, that information and uncertainty about a random variable are really the same.

Information Theory

Mutual Information Properties
Mutual Information is non-negative, and is zero, iff X and Y are independent (see proof of previous theorem)
Mutual Information has a conditional form (see [CT91, p.22] for details.)
Mutual Information has a chain rule (see [CT91, p.22] for details.)

Information Theory



Information Theory Mutual information Hutual Information Properties	Mutual Information Properties • Matual Information Properties • Matual Information has a conditional from (see [TTR, p.22] for data.), • Matual Information has a chain rule (see [TTR, p.22] for data.),

October 9, 2013

20 / 22

Assignment

There are lots of practice problems in [CT91, Chapter 1], which is available in electronic form in our Library. I recommend you have a go, but I won't mark these.

The assignment is to calculate the entropy of Morse code symbols, given standard frequencies of English letters. Hints:

- Remember Morse code really has four symbols:
 - dot
 - dash
 - letter-break
 - word-break

latthew Roughan (School of Mathematical S

- Model the frequencies of word-breaks as well as just letters.
 - ▶ you may need to make your own measurements of text lots is available, e.g., at http://www.gutenberg.org/

Information Theory

◆□▶ ◆□▶ ◆三▶ ◆三▶ ・三 ・ つへの

October 9, 2013 21 / 22

```
2013-10-09
              Assignment
```

Information Theory

-Mutual information

Reading in text can be done in Matlab, but you may find some easier approaches. For instance, in R http://www.r-bloggers.com/ text-mining-the-complete-works-of-william-shakespeare/ I personally prefer to use Perl for tasks like this, and I can pretty much guarantee that a Perl implementation will be faster to run, and faster to write (if you learn a bit about Perl), but its up to you.

Further reading I		
Thomas M. Cover and Joy , and Sons, 1991.	A. Thomas, <i>Elements</i> (of information theory, John Wiley
Matthew Roughan (School of Mathematical s	Information Theory	 (전) (전) (전) (전) (전) (전) (전) (전) (전) (전)

