Entropy and decisions

CSC401/2511 – Natural Language Computing – Spring 2020
Lecture 3, Frank Rudzicz and Sean Robertson
University of Toronto

This lecture

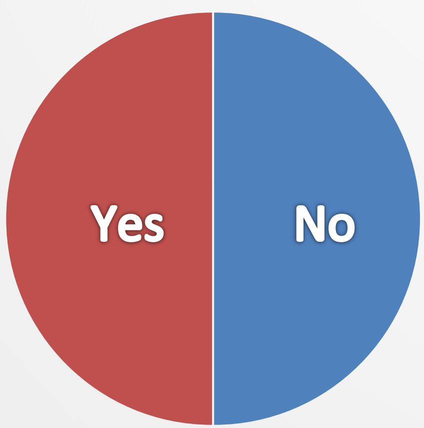
- Information theory and entropy.
- Decisions.
  - Classification.
  - Significance.

Can we quantify the statistical structure in a model of communication?
Can we quantify the <u>meaningful</u> difference between statistical models?

Information

- Imagine Darth Vader is about to say either "yes" or "no" with **equal** probability.
    - You don't know what he'll say.

- You have a certain amount of **uncertainty** – a lack of information.



Darth Vader is © Disney

And the prequels and Rey/Finn Star Wars suck

Star Trek is better than Star Wars

UNIVERSITY OF TORONTO

Information

- Imagine you then **observe** Darth Vader saying "no"
- Your uncertainty is **gone**; you've **received** **information**.
- **How much** information do you **receive** about event E when you observe it?

$$I(E) = \underbrace{\log_2}_{\text{For the units of measurement}} \underbrace{\frac{1}{P(E)}}_{\text{For the inverse}}$$

$$I(no) = \log_2 \frac{1}{P(no)} = \log_2 \frac{1}{1/2} = \underline{1 \text{ bit}}$$

UNIVERSITY OF TORONTO

Information

- Imagine Darth Vader is about to roll a **fair** die.
- You have **more uncertainty** about an event because there are **more possibilities**.
    - You **receive** more information when you observe it.

$$I(5) = \log_2 \frac{1}{P(5)}$$
$$= \log_2 \frac{1}{1/6} \approx 2.59 \text{ bits}$$

UNIVERSITY OF TORONTO

Information is additive

- From k **independent**, **equally** likely events E,

$$I(E^k) = \log_2 \frac{1}{P(E^k)} = \log_2 \frac{1}{P(E)^k} \qquad I(k \text{ binary decisions}) = \log_2 \frac{1}{\left({}^{1}/_{2}\right)^k} = \underline{k \text{ bits}}$$

- For a **unigram** model, with each of 50K words w **equally** likely,

$$I(w) = \log_2 \frac{1}{{}^{1}/_{50000}} \approx 15.61 \text{ bits}$$

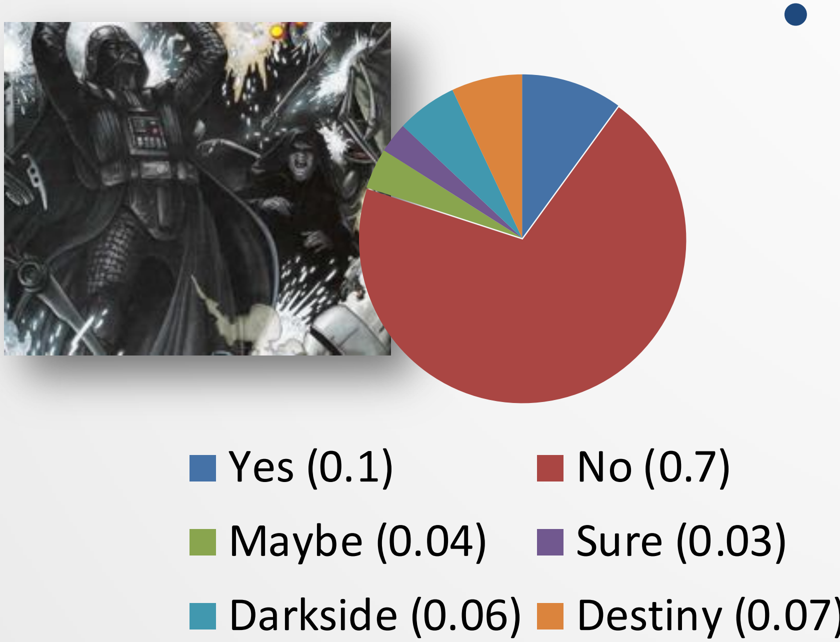
and for a **sequence** of 1K words in that model,

$$I(w^k) = \log_2 \frac{1}{\left({}^{1}/_{50000}\right)^{1000}} \approx \boxed{???}$$

UNIVERSITY OF TORONTO

Information with unequal events

- An information **source** S **emits** symbols **without memory** from a **vocabulary** $\{w_1, w_2, \dots, w_n\}$. **Each** symbol has its **own** probability $\{p_1, p_2, \dots, p_n\}$



- What is the ***<u>average</u>*** amount of information we get in **observing** the **output** of source S ?
  - You **still** have 6 events that are possible – **but** you're fairly sure it will be '*No*'.

Entropy

- **Entropy**: *n.* the **average** amount of information we get in observing the output of source S.

$$H(S) = \sum_i p_i I(w_i) = \sum_i p_i \log_2 \frac{1}{p_i}$$

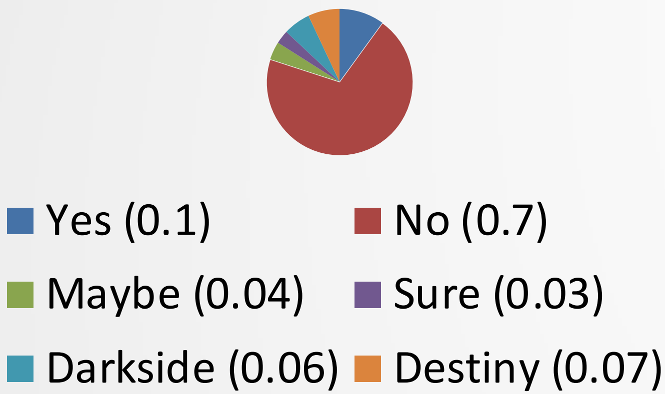
ENTROPY

Note that this is ***very*** similar to how we define the expected value (i.e., 'average') of something:

$$E[X] = \sum_{x \in X} p(x)\, x$$

UNIVERSITY OF TORONTO

Entropy – examples

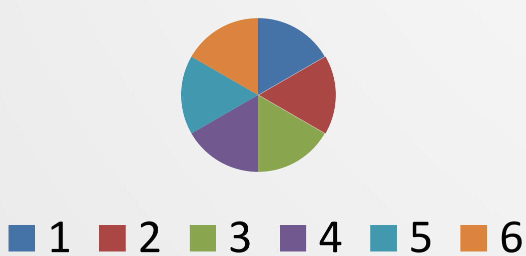


$$H(S) = \sum_i p_i \log_2 \frac{1}{p_i}$$

$$= 0.7 \log_2(1/0.7) + 0.1 \log_2(1/0.1) + \cdots$$

$$\boxed{= 1.542 \text{ bits}}$$

There is **less** average uncertainty when the probabilities are 'skewed'.



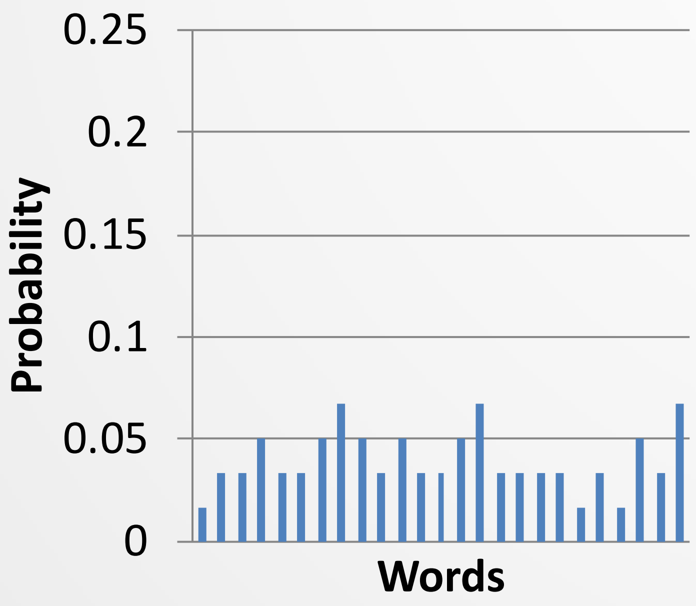
$$H(S) = \sum_i p_i \log_2 \frac{1}{p_i} = 6\left(\frac{1}{6} \log_2 \frac{1}{1/6}\right)$$

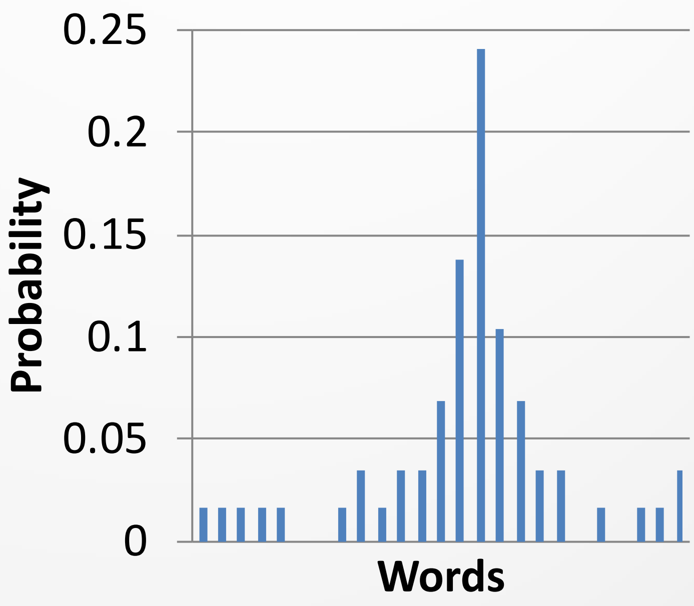
$$\boxed{= 2.585 \text{ bits}}$$

UNIVERSITY OF TORONTO

Entropy characterizes the distribution

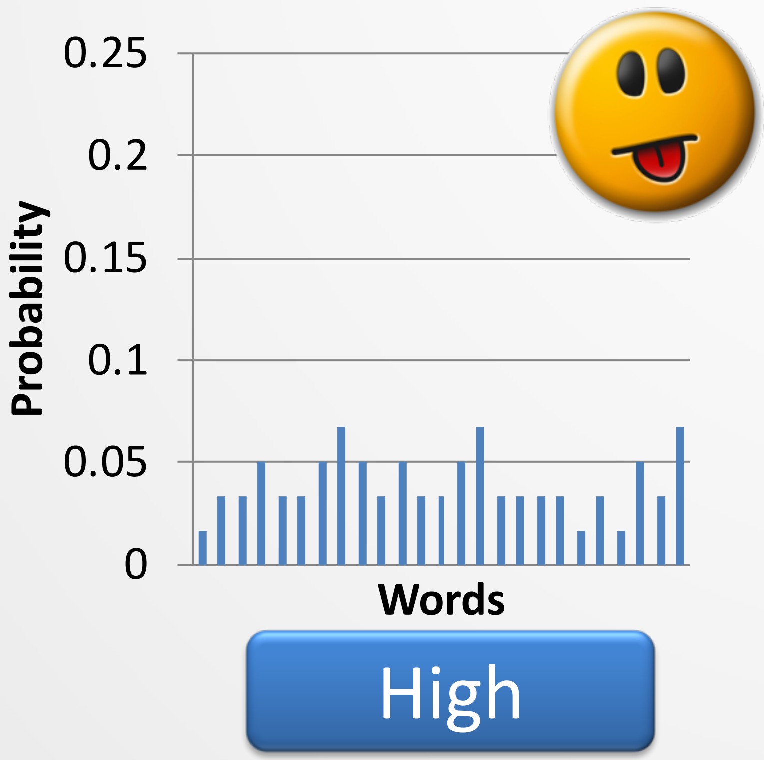
- '**Flatter**' distributions have a **higher** entropy because the choices are **more equivalent**, on average.
  - So which of these distributions has a **lower** entropy?

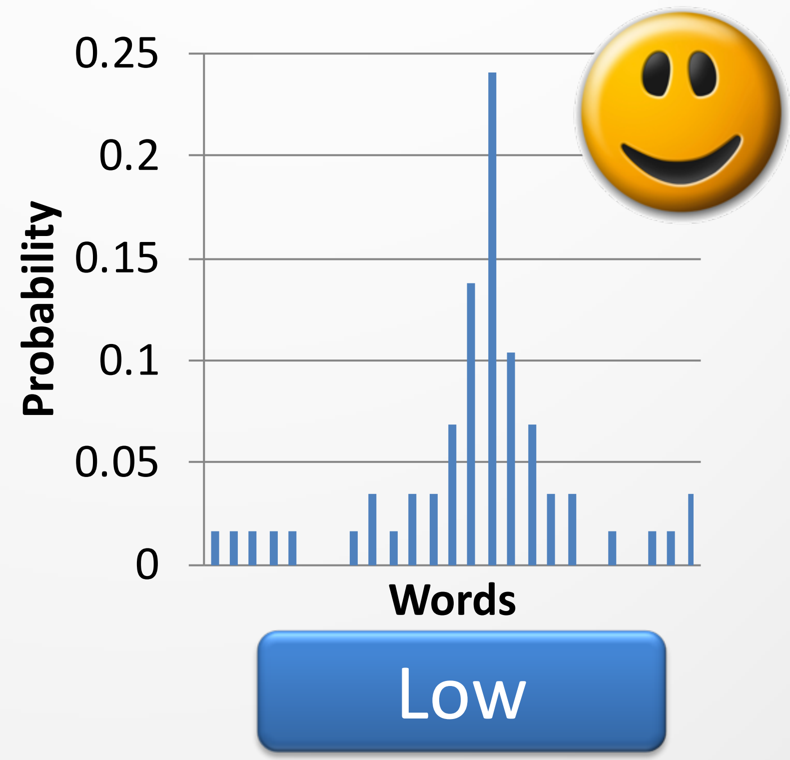




Low entropy makes decisions easier

- When predicting the next word, e.g., we'd like a distribution with **lower** entropy.
  - Low entropy ≡ less uncertainty





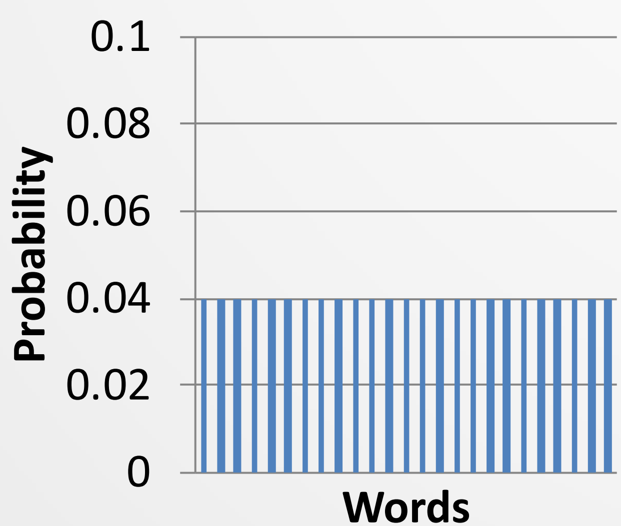
UNIVERSITY OF TORONTO

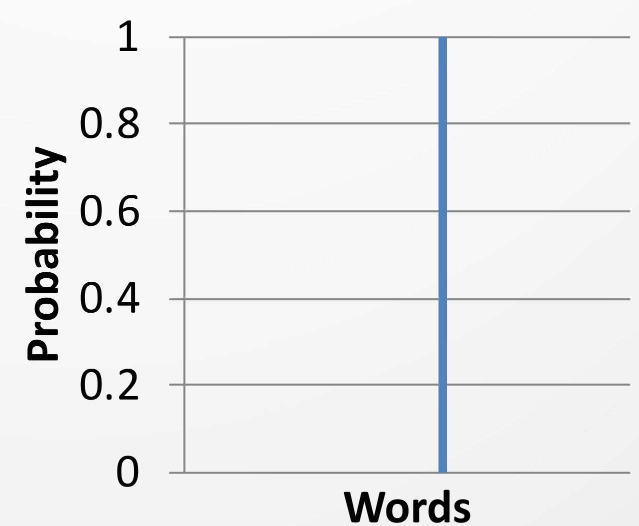
Bounds on entropy

- **Maximum**: uniform distribution S_1. Given M choices,

$$H(S_1) = \sum_i p_i \log_2 \frac{1}{p_i} = \sum_i \frac{1}{M} \log_2 \frac{1}{1/M} = \mathbf{\log_2 M}$$

- **Minimum**: only one choice, $H(S_2) = p_i \log_2 \frac{1}{p_i} = 1 \cancelto{0}{\log_2} 1 = \mathbf{0}$

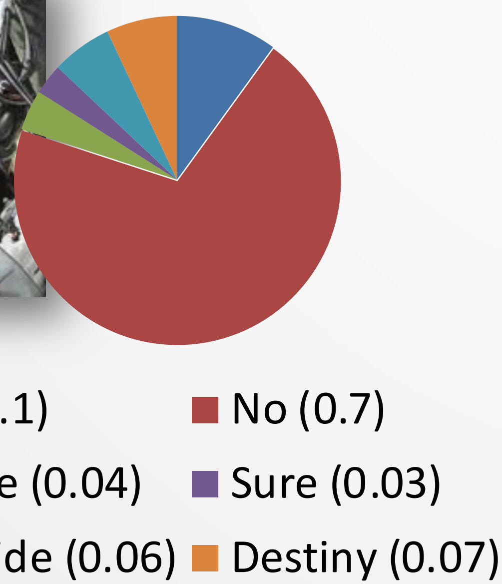




UNIVERSITY OF TORONTO

Coding symbols efficiently

- If we want to **transmit** Vader's words **efficiently**, we can **encode** them so that **more probable words** require **fewer bits**.
  - On **average**, fewer bits will need to be transmitted.

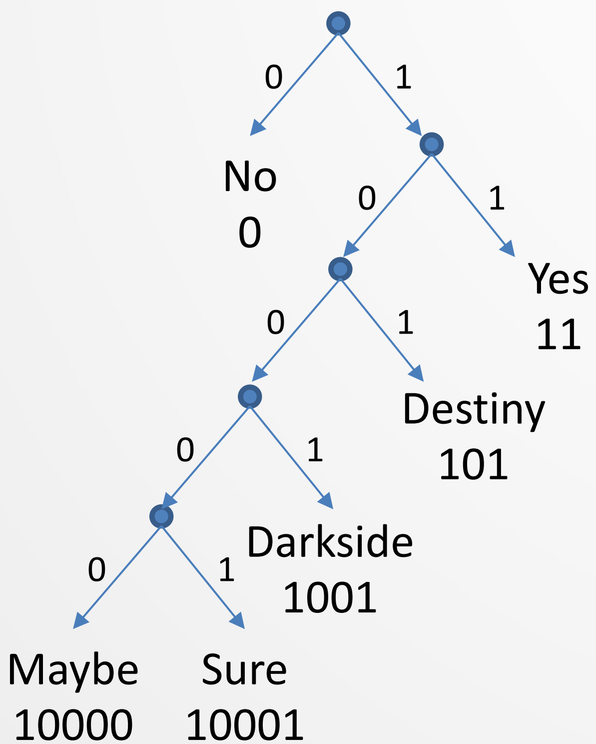


Word (sorted)	Linear Code	Huffman Code
No	000	0
Yes	001	11
Destiny	010	101
Darkside	011	1001
Maybe	100	10000
Sure	101	10001

UNIVERSITY OF TORONTO

Coding symbols efficiently

- Another way of looking at this is through the (binary) Huffman tree (*r*-ary trees are often flatter, all else being equal):



Word (sorted)	Linear Code	Huffman Code
No	000	0
Yes	001	11
Destiny	010	101
Darkside	011	1001
Maybe	100	10000
Sure	101	10001

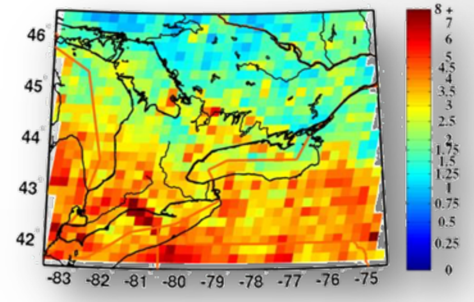
Alternative notions of entropy

- Entropy is **equivalently**:
  - The **average** amount of **information provided** by symbols in a vocabulary,
  - The **average** amount of **uncertainty** you have **before** observing a symbol from a vocabulary,
  - The **average** amount of '**surprise**' you receive when observing a symbol,
  - The number of bits needed to communicate that alphabet
    - Aside: Shannon showed that you **cannot** have a **coding scheme** that can communicate the vocabulary **more efficiently** than $H(S)$

Entropy of several variables

- Joint entropy
- Conditional entropy
- Mutual information

Entropy of several variables



- Consider the vocabulary of a meteorologist describing ***T***emperature and ***W***etness.
  - ***T***emperature = {*hot, mild, cold*}
  - ***W***etness = {*dry, wet*}

$$P(W = dry) = 0.6,$$
$$P(W = wet) = 0.4$$

$$\boldsymbol{H}(\boldsymbol{W}) = 0.6 \log_2 \frac{1}{0.6} + 0.4 \log_2 \frac{1}{0.4} = \mathbf{0.970951\ bits}$$

$$P(T = hot) = 0.3,$$
$$P(T = mild) = 0.5,$$
$$P(T = cold) = 0.2$$

$$\boldsymbol{H}(\boldsymbol{T}) = 0.3 \log_2 \frac{1}{0.3} + 0.5 \log_2 \frac{1}{0.5} + 0.2 \log_2 \frac{1}{0.2} = \mathbf{1.48548\ bits}$$

But W and T are *not* independent, $P(W,T) \neq P(W)P(T)$

Example from Roni Rosenfeld

Joint entropy

- **Joint Entropy**: *n.* the **average** amount of information needed to specify **multiple** variables **simultaneously**.

$$\boxed{H(X,Y) = \sum_x \sum_y p(x,y) \log_2 \frac{1}{p(x,y)}}$$

- **Hint**: this is *very* similar to univariate entropy – we just replace univariate probabilities with joint probabilities and sum over everything.

Entropy of several variables

- Consider joint probability, $P(W,T)$

	cold	mild	hot	
dry	0.1	0.4	0.1	**0.6**
wet	0.2	0.1	0.1	**0.4**
	0.3	**0.5**	**0.2**	**1.0**

- **Joint entropy**, $H(W,T)$, computed **as a sum over the space of joint events** $(W = w, T = t)$

$$H(W,T) = 0.1 \log_2 {}^1/_{0.1} + 0.4 \log_2 {}^1/_{0.4} + 0.1 \log_2 {}^1/_{0.1} \\ +0.2 \log_2 {}^1/_{0.2} + 0.1 \log_2 {}^1/_{0.1} + 0.1 \log_2 {}^1/_{0.1} = \mathbf{2.32193\ bits}$$

Notice $H(W,T) \approx 2.32 < 2.46 \approx H(W) + H(T)$

UNIVERSITY OF TORONTO

Entropy given knowledge

- In our example, **joint entropy** of two variables together is **lower** than the **sum** of their **individual** entropies
  - $H(W, T) \approx 2.32 < 2.46 \approx H(W) + H(T)$

- **Why**?

- Information is **shared** among variables
  - There are **dependencies**, e.g., between temperature and wetness.
  - E.g., if we knew **exactly** how **wet** it is, is there **less confusion** about what the **temperature** is … ?

Conditional entropy

- **Conditional entropy**: *n.* the **average** amount of information needed to specify one variable given that you know another.
  - A.k.a '**equivocation**'

$$\boxed{H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x)}$$

- **Hint**: this is *very* similar to how we compute expected values in general distributions.

Entropy given knowledge

- Consider **conditional** probability, $P(T|W)$

$P(W,T)$	T = cold	mild	hot	
W = dry	0.1	0.4	0.1	**0.6**
wet	0.2	0.1	0.1	**0.4**
	0.3	0.5	0.2	1.0

$$P(T|W) = P(W,T)/P(W)$$

$P(T\,|W)$	T = cold	mild	hot	
W = dry	0.1/0.6	0.4/0.6	0.1/0.6	1.0
wet	0.2/0.4	0.1/0.4	0.1/0.4	1.0

Entropy given knowledge

- Consider **conditional** probability, $P(T|W)$

$P(T\|W)$	$T=$ cold	mild	hot	
$W=$ dry	1/6	2/3	1/6	1.0
wet	1/2	1/4	1/4	1.0

- $\boldsymbol{H(T|W=dry)} = H\left(\left\{\frac{1}{6},\frac{2}{3},\frac{1}{6}\right\}\right) = \mathbf{1.25163}$ **bits**
- $\boldsymbol{H(T|W=wet)} = H\left(\left\{\frac{1}{2},\frac{1}{4},\frac{1}{4}\right\}\right) = \mathbf{1.5}$ **bits**
- **Conditional entropy** combines these:

$$\begin{aligned}&\boldsymbol{H(T|W)}\\&= [\overset{0.6}{p(W=dry)}H(T|W=dry)] + [\overset{0.4}{p(W=wet)}H(T|W=wet)]\\&= \mathbf{1.350978}\ \textbf{bits}\end{aligned}$$

UNIVERSITY OF TORONTO

Equivocation removes uncertainty

- Remember $H(T) = 1.48548$ bits
- $H(W, T) = 2.32193$ bits
- $H(T|W) = 1.350978$ bits

Entropy (i.e., confusion) about temperature is **reduced** if we **know** how wet it is outside.

- How much does W tell us about T?
  - $H(T) - H(T|W) = 1.48548 - 1.350978 \approx 0.1345$ bits
  - Well, a little bit!

Perhaps T is more informative?

- Consider **another** conditional probability, $P(W|T)$

$P(W\|T)$	T = cold	mild	hot
W = dry	0.1/0.3	0.4/0.5	0.1/0.2
wet	0.2/0.3	0.1/0.5	0.1/0.2
	1.0	1.0	1.0

- $H(W|T = cold) = H\left(\left\{\frac{1}{3}, \frac{2}{3}\right\}\right) = 0.918295$ bits
- $H(W|T = mild) = H\left(\left\{\frac{4}{5}, \frac{1}{5}\right\}\right) = 0.721928$ bits
- $H(W|T = hot) = H\left(\left\{\frac{1}{2}, \frac{1}{2}\right\}\right) = 1$ bit

- $\boldsymbol{H(W|T) = 0.8364528}$ **bits**

UNIVERSITY OF TORONTO

Equivocation removes uncertainty

- $H(T) = 1.48548$ bits
- $H(W) = 0.970951$ bits
- $H(W, T) = 2.32193$ bits
- $H(T|W) = 1.350978$ bits
- $\boldsymbol{H(T) - H(T|W)} \approx \boxed{\mathbf{0.1345\ bits}}$

Previously computed

- How much does T tell us about W on average?
  - $\boldsymbol{H(W) - H(W|T)} = 0.970951 - 0.8364528$ $\approx \boxed{\mathbf{0.1345\ bits}}$
  - Interesting … is that a coincidence?

UNIVERSITY OF TORONTO

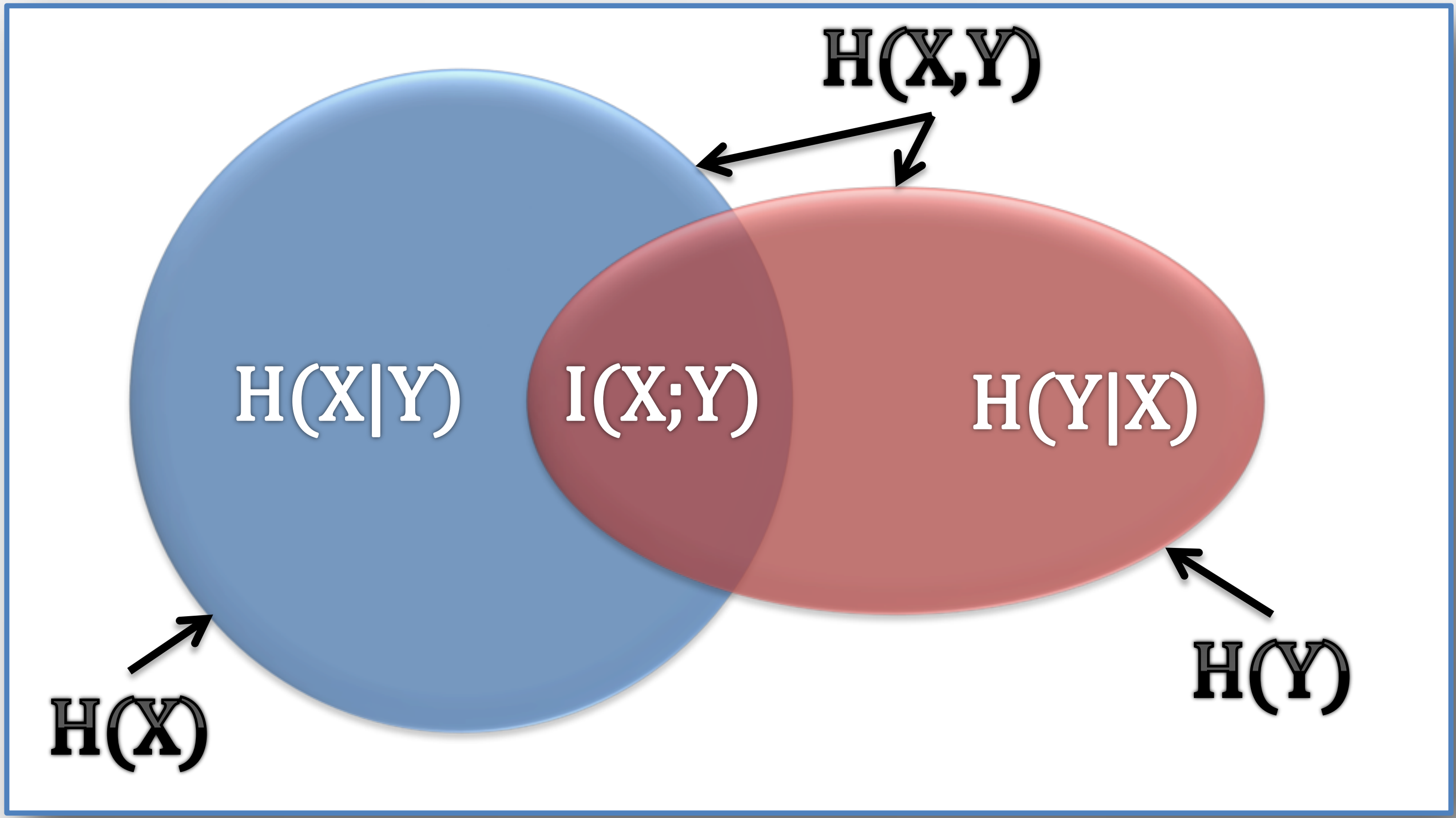
Mutual information

- **Mutual information**: *n.* the **average** amount of information **shared** between variables.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$
$$= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

- **Hint**: The amount of uncertainty **removed** in variable X if you know Y.
- **Hint2**: If X and Y are **independent**, $p(x,y) = p(x)p(y)$, then $\log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 1 = 0 \ \forall x, y$ – **there is no mutual information**!

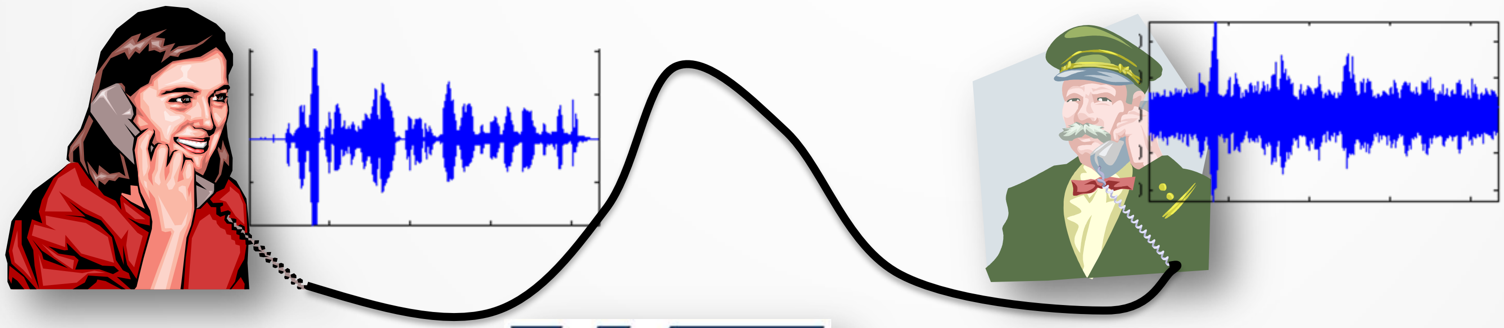
Relations between entropies



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

Reminder – the noisy channel

- Messages can get **distorted** when passed through a **noisy** conduit – <u>*how much information is lost/retained*</u>?
  - Signals



  - Symbols

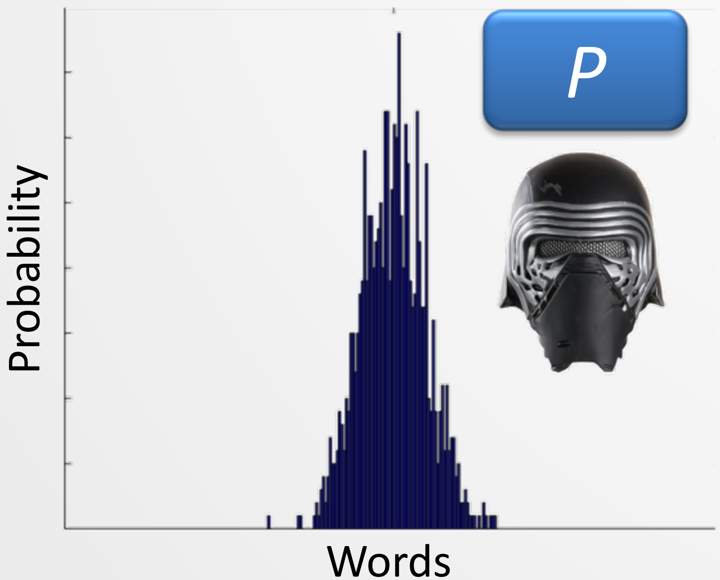
  - Languages

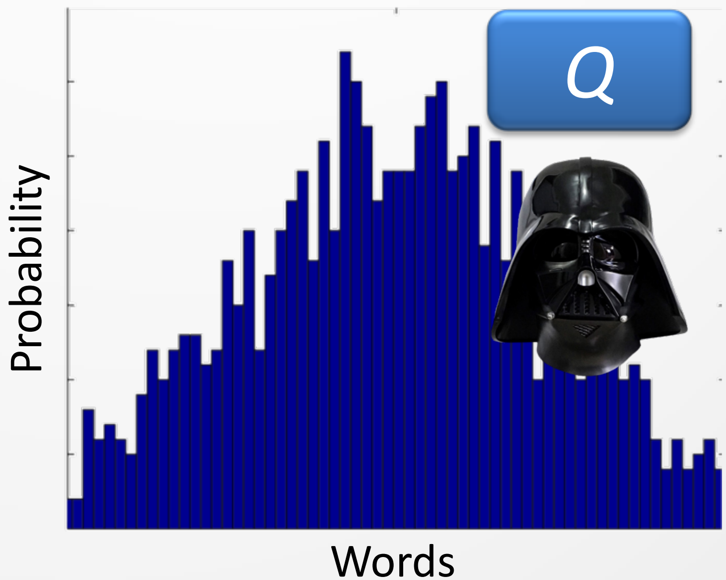
UNIVERSITY OF TORONTO

Relating corpora

Relatedness of two distributions

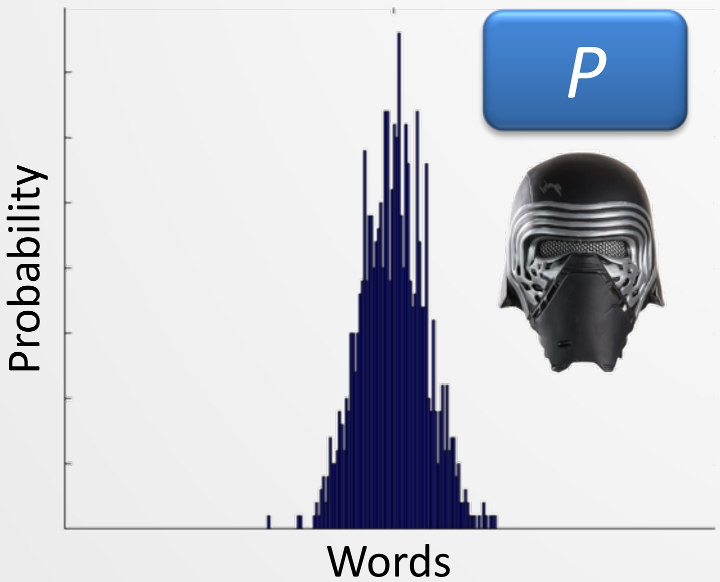
- How **similar** are two probability distributions?
  - e.g.,   Distribution *P* learned from *Kylo Ren*
              Distribution *Q* learned from *Darth Vader*

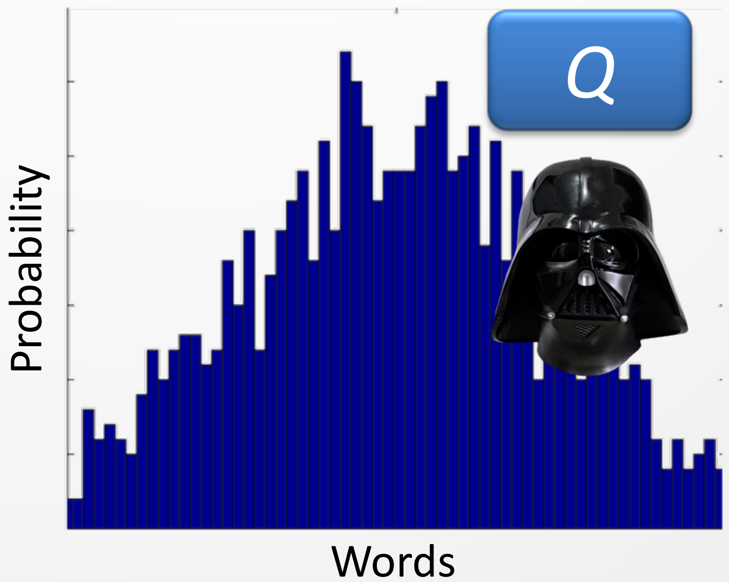




Relatedness of two distributions

- A Huffman code based on Vader (*Q*) instead of Kylo (*P*) will be less *efficient* at coding symbols that Kylo will say.
- *What is the **average number of extra bits** required to code symbols from P when using a code based on Q?*

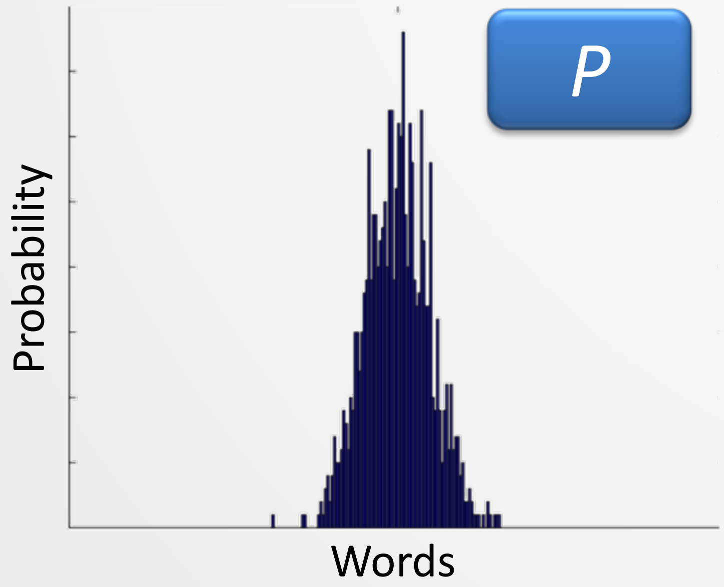


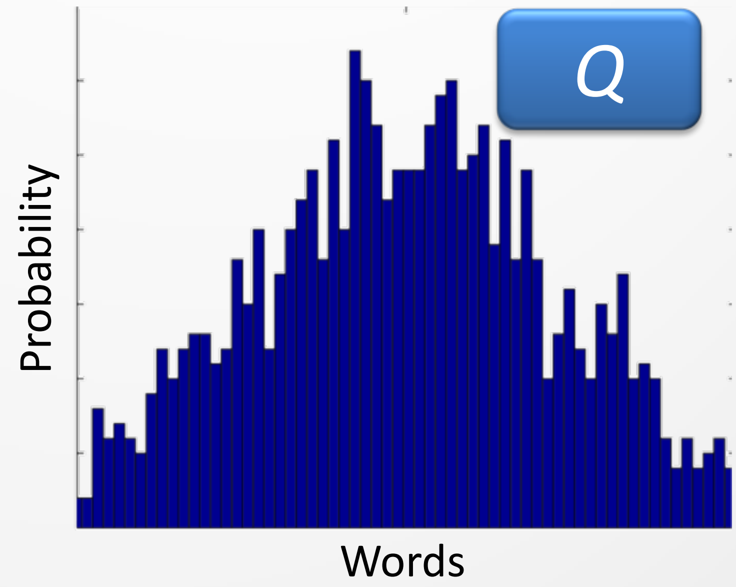


Kullback-Leibler divergence

- **KL divergence**: *n.* the **average log difference** between the distributions *P* and *Q*, relative to *Q*. a.k.a. **relative entropy**.

  caveat: we assume $0 \log 0 = 0$





Kullback-Leibler divergence

$$\boxed{D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}}$$

- Why $\log \frac{P(i)}{Q(i)}$?
- $\log \frac{P(i)}{Q(i)} = \log P(i) - \log Q(i) = \log\left(\frac{1}{Q(i)}\right) - \log\left(\frac{1}{P(i)}\right)$
- If word w_i is less probable in Q than P (i.e., it carries more information), it will be Huffman encoded in more bits, so when we see w_i from P, we need $\log \frac{P(i)}{Q(i)}$ more bits.

Kullback-Leibler divergence

- KL divergence:
  - is *somewhat* like a ‘**distance**’ :
    - $D_{KL}(P||Q) \geq 0 \quad \forall P, Q$
    - $D_{KL}(P||Q) = 0$ iff P and Q are identical.
  - is **not symmetric**, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$

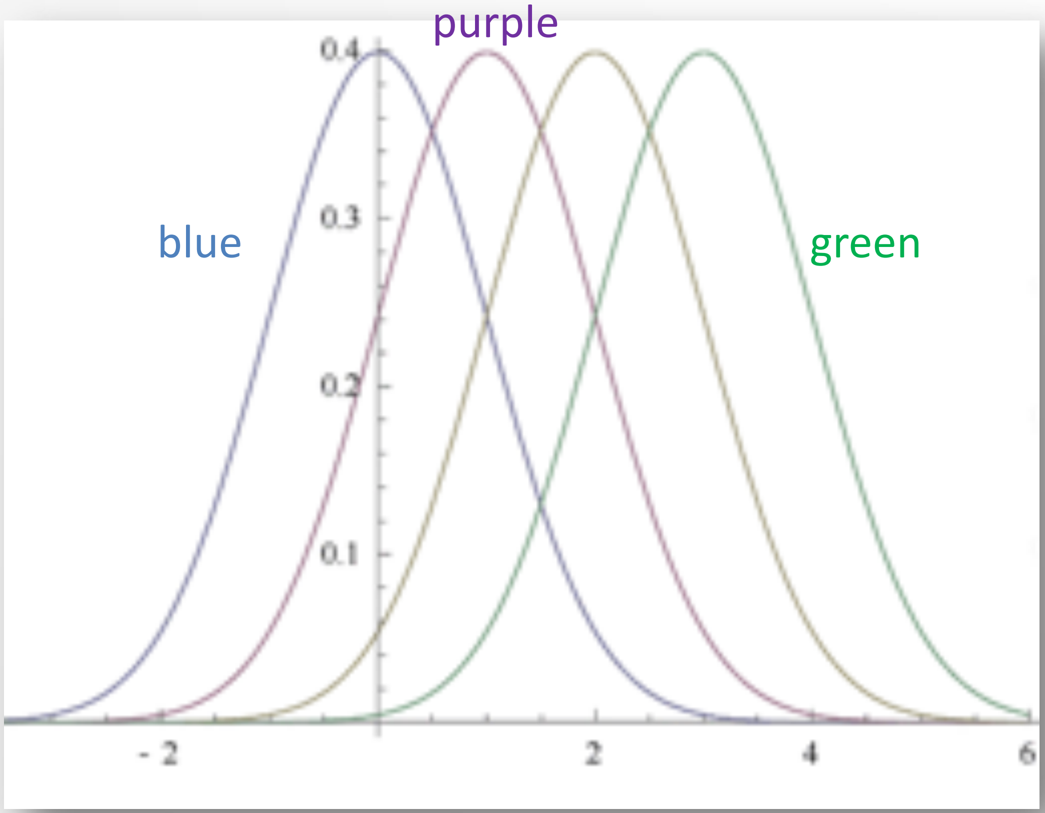
- Aside:

$$I(P;Q) = D_{KL}\big(P(X,Y)||P(X)P(Y)\big)$$

UNIVERSITY OF TORONTO

Kullback-Leibler divergence

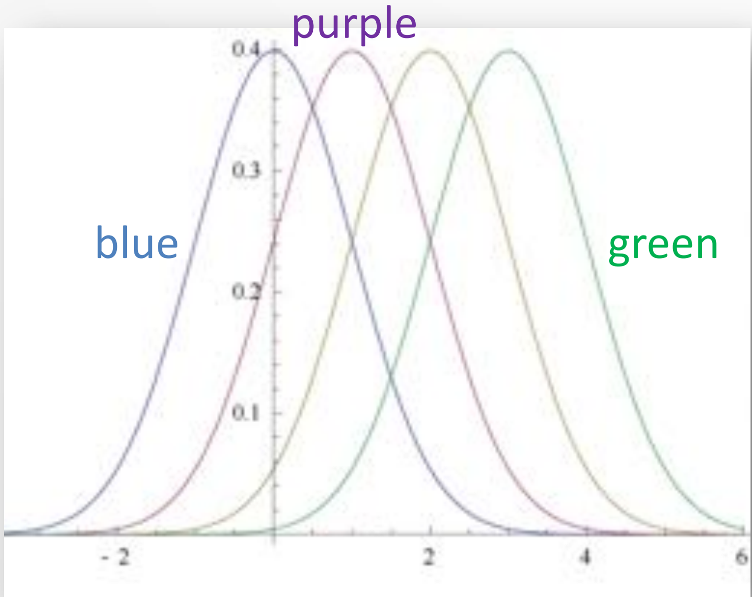
- KL divergence generalizes to **continuous** distributions.
- Below, $D_{KL}(\boldsymbol{blue}||\boldsymbol{green}) > D_{KL}(\boldsymbol{blue}||\boldsymbol{purple})$



UNIVERSITY OF TORONTO

Applications of KL divergence

- Often used towards some **other purpose**, e.g.,
  - In **evaluation** to say that ***purple*** is a **better** model than ***green*** of the ***true* distribution *blue***.
  - In **machine learning** to adjust the parameters of ***purple*** to be, e.g., less like ***green*** and more like ***blue***.



Entropy as intrinsic LM evaluation

- **Cross-entropy** measures how difficult it is to encode an event drawn from a ***true* probability** p given a **model** based on a distribution q.

- What if we don't know the ***true* probability** p?
  - We'd have to estimate the CE using a test corpus C:

$$H(p,q) \approx -\frac{\log_2 P_q(C)}{\|C\|}$$

- What's the probability of a corpus $P_q(C)$?

UNIVERSITY OF TORONTO

Probability of a corpus?

- The probability $P(C)$ of a **corpus** C requires similar **assumptions** that allowed us to compute the probability $P(s_i)$ of a **sentence** s_i.

	Sentence	Corpus
Chain rule	$P(s_i) = P(w_1)\prod_{t=2}^{n} P(w_t \mid w_{1:(t-1)})$	$P(C) = P(w_1)\prod_{t=2}^{\lVert C \rVert} P(w_t \mid w_{1:(t-1)})$
Approx.	$P(s_i) \approx \prod_t P(w_t)$	$P(C) \approx \prod_i P(s_i)$

- **Regardless** of the LM used for $P(s_i)$, we can assume **complete independence** between sentences.

Intrinsic evaluation – Cross-entropy

- **Cross-entropy** of a LM M and a *new* test corpus C with size $\|C\|$ (total number of words), where sentence $s_i \in C$, is *approximated* by:

$$H(C;M) = -\frac{\log_2 P_M(C)}{\|C\|} = -\frac{\sum_i \log_2 P_M(s_i)}{\sum_i \|s_i\|}$$

- **Perplexity** comes from this definition:

$$PP_M(C) = 2^{H(C;M)}$$

UNIVERSITY OF TORONTO

Decisions

Deciding what we know

- **Anecdotes** are often useless except as proofs by contradiction.
  - E.g., *"I saw Google used as a verb"* does **not** mean that *Google* is **always** (or even **likely** to be) a verb, just that it is **not always** a noun.

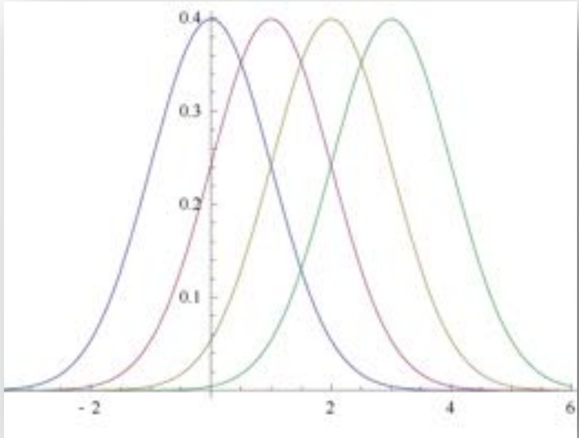
- **Shallow statistics** are often not enough to be truly meaningful.
  - E.g., *"My ASR system is 95% accurate on my test data. Yours is only 94.5% accurate, you horrible knuckle-dragging idiot."*
    - What if the test data was **biased** to favor my system?
    - What if we only used a **very small** amount of data?

- Given all this potential ambiguity, we need a **test** to see if our statistics actually **mean** something.

UNIVERSITY OF TORONTO

Differences due to sampling

- We saw that **KL divergence** essentially measures how **different** two distributions are from each other.
- But what if their difference is due to **randomness** in **sampling**?
- How can we tell that a distribution is ***really*** different from another?



Hypothesis testing

- Often, we assume a **null hypothesis**, H_0, which states that the **two distributions are *the same*** (i.e., come from the same underlying model, population, or phenomenon).

- We **reject** the null hypothesis if the probability of it being true is too small.
  - This is often our goal – e.g., if my ASR system beats yours by 0.5%, I want to show that this difference is **not** a random accident.
  - I assume it *was* an accident, then show how nearly *impossible* that is.

  - As scientists, we have to be very **careful** to not reject H_0 too hastily.
    - How can we ensure our **diligence**?

Confidence

- We **reject** H_0 if it is **too improbable**.
  - How do we determine the value of 'too'?

- **Significance level** α $(0 \leq \alpha \leq 1)$ is the **maximum** probability that two distributions are **identical** allowing us to **disregard** H_0.
  - In practice, $\alpha \leq 0.05$. Usually, it's much lower.
  - **Confidence level** is $\gamma = 1 - \alpha$
  - E.g., a confidence level of **95%** $(\alpha = 0.05)$ implies that we expect that our decision is correct 95% of the time, **regardless of the test data**.

Confidence

- We will briefly see three types of **statistical tests** that can tell us how **confident** we can be in a claim:

1. A ***t*-test**, which usually tests whether the **means** of two models are the same. There are many types, but most assume **Gaussian** distributions.
2. An **analysis of variance** (**ANOVA**), which generalizes the *t*-test to more than two groups.
3. The χ^2 **test**, which evaluates **categorical** (discrete) outputs.

1. The *t*-test

- The ***t*-test** is a method to compute if distributions are significantly different from one another.

- It is based on the mean $(\bar{x})$ and variance (σ) of N samples.
- It compares $\bar{x}$ and σ to H_0 which states that the samples are drawn from a distribution with a **mean** μ.

- If $\boxed{t = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/N}}}$ (the "t-statistic") is large enough, we can reject H_0.

An example would be nice…

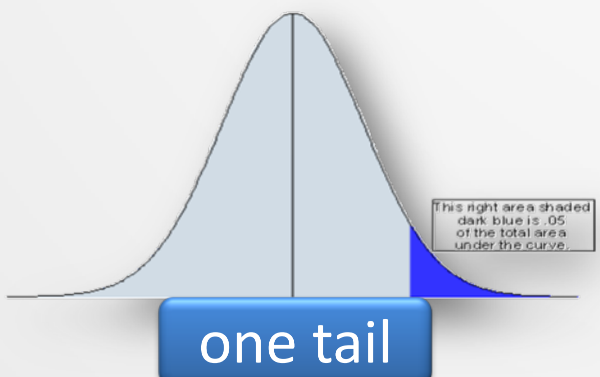
There are actually **several types** of *t*-tests for different situations…

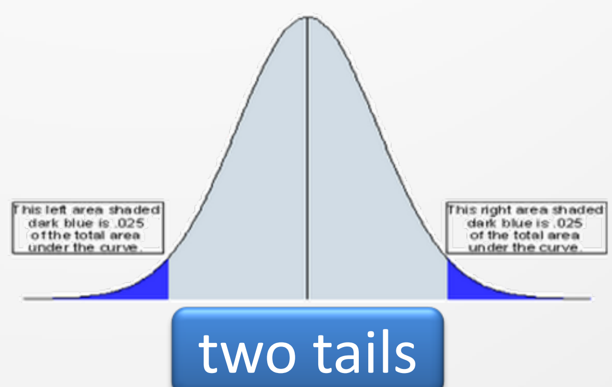
UNIVERSITY OF TORONTO

Example of the *t*-test: tails

- Imagine the average tweet length of a McGill 'student' is $\mu = 158$ chars.
- We sample $N = 200$ UofT students and find that our average tweet is $\bar{x} = 169$ chars (with $\sigma^2 = 2600$).
- Are UofT tweets significantly **longer** than much worse McGill tweets?

- We use a '**one-tailed**' test because we want to see if UofT tweet lengths are significantly **higher**.
  - If we just wanted to see if UofT tweets were significantly **different**, we'd use a **two-tailed** test.





UNIVERSITY OF TORONTO

Example of the *t*-test: freedom

- Imagine the average tweet length of a McGill 'student' is $\mu = 158$ chars.
- We sample $N = 200$ UofT students and find that our average tweet is $\bar{x} = 169$ chars (with $\sigma^2 = 2600$).
- Are UofT tweets significantly **longer** than much worse McGill tweets?

- **Degrees of freedom (d.f.):** *n.pl.* In *this t*-test, this is the sum of the number of observations, minus 1 (the number of sample sets).

- In our example, we have $N_{UofT} = 200$ for UofT students, meaning

$$d.f. = 199$$

  - (this example is adapted from Manning & Schütze)

Example of the *t*-test

- Imagine the average tweet length of a McGill 'student' is $\mu = 158$ chars.
- We sample $N = 200$ UofT students and find that our average tweet is $\bar{x} = 169$ chars (with $\sigma^2 = 2600$).
- Are UofT tweets significantly **longer** than much worse McGill tweets?

- So $t = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/N}} = \frac{169-158}{\sqrt{2600/200}} \approx 3.05$

- In a ***t*-test table**, we look up the minimum value of t necessary to reject H_0 at $\alpha = 0.005$ (we want to be quite confident) for a 1-tailed test…

Example of the *t*-test

- So $t = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/_N}} = \frac{169-158}{\sqrt{2600/_{200}}} \approx 3.05$

- In a ***t*-test table**, we look up the minimum value of t necessary to reject H_0 at $\alpha = 0.005$, and find 2.576 (using $d.f. = 199 \approx \infty$)
  - Since $3.05 > 2.576$, we can reject H_0 at the 99.5% level of confidence ($\gamma = 1 - \alpha = 0.995$) ; **<u>UofT students are significantly more verbose</u>**.

	α (one-tail)	0.05	0.025	0.01	0.005	0.001	0.0005
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
	∞	1.645	1.960	2.326	**2.576**	3.091	3.291

Example of the *t*-test

- Some things to observe about the *t*-test table:
  - We need **more evidence**, *t*, if we want to be **more confident** (left-right dimension).
  - We need **more evidence**, *t*, if we have **fewer measurements** (top-down dimension).
- A common criticism of the *t*-test is that picking α is ad-hoc. There are ways to correct for the selection of α.

	α (one-tail)	0.05	0.025	0.01	0.005	0.001	0.0005
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
	∞	1.645	1.960	2.326	**2.576**	3.091	3.291

UNIVERSITY OF TORONTO

Another example: collocations

- **Collocation**: *n.* a 'turn-of-phrase' or usage where a sequence of words is '**perceived**' to have a meaning '**beyond**' the sum of its parts.

- E.g., '*disk drive*', '*video recorder*', and '*soft drink*' **are** collocations. '*cylinder drive*', '*video storer*', '*weak drink*' **are not** despite some near-synonymy between alternatives.

- Collocations are **not** just highly frequent bigrams, otherwise '*of the*', and '*and the*' would be collocations.

- How can we test if a bigram is a collocation or not?

Hypothesis testing collocations

- For collocations, the **null hypothesis** H_0 is that there is **no association** between two given words **beyond pure chance**.
  - I.e., the bigram's **actual** distribution and pure chance are the **same**.
  - We compute the probability of those words occurring together if H_0 were true. If that probability **is too low**, we **reject** H_0.

  - E.g., we expect '*of the*' to occur together, because they're both likely words to draw randomly
    - We could probably **not** reject H_0 in that case.

Example of the *t*-test on collocations

- Is '*new companies*' a collocation?
- In our corpus of 14,307,668 word tokens, *new* appears 15,828 times and *companies* appears 4,675 times.
- Our **null hypothesis**, H_0 is that they are **independent**, i.e.,

$$\mathrm{H}_0\text{: } P(new\ companies) = P(new)P(companies)$$
$$= \frac{15828}{14307668} \times \frac{4675}{14307668}$$
$$\approx 3.615 \times 10^{-7}$$

Example of the *t*-test on collocations

- The Manning & Schütze text claims that if the process of randomly generating bigrams follows a **Bernoulli distribution**.
  - i.e., assigning 1 whenever *new companies* appears and 0 otherwise gives $\bar{x} = p = P(new\ companies)$
  - For Bernoulli distributions, $\sigma^2 = p(1-p)$. Manning & Schütze claim that we can assume $\sigma^2 = p(1-p) \approx p$, since for most bigrams, p is very small.

Example of the *t*-test on collocations

- So, $\mu = 3.615\times10^{-7}$ is the expected mean in H_0.
- We **actually count** 8 occurrences of *new companies* in our corpus
  - $\bar{x} = \frac{8}{14307667} \approx 5.591\times10^{-7}$

There is 1 fewer bigram instance than word tokens in the corpus

$\therefore \sigma^2 \approx p = \bar{x} = 5.591\times10^{-7}$

- So $t = \frac{\bar{x}-\mu}{\sqrt{\sigma^2/N}} = \frac{5.591\times10^{-7}-3.615\times10^{-7}}{\sqrt{5.591\times10^{-7}/14307667}} \approx \mathbf{0.9999}$

- In a ***t*-test table**, we look up the minimum value of t necessary to reject H_0 at $\alpha = 0.005$, and find $\mathbf{2.576}$.
  - Since $\mathbf{0.9999} < \mathbf{2.576}$, we cannot reject H_0 at the 99.5% level of confidence.
    - We **don't have enough evidence** to think that *new companies* is a collocation (we can't say that it definitely ***isn't***, though!).

2. Analysis of variance (aside)

- **Analyses of variance (ANOVAs)** (there are several types) can be:
  - A way to **generalize *t*-tests** to more than two groups.
  - A way to **determine which** (if any) of several **variables** are **responsible** for the **variation** in an observation (and the interaction between them).

- E.g., we measure the **accuracy** of an ASR system for different settings of **empirical parameters** M (# components) and Q (# states).

Accuracy (%)	$\boldsymbol{M = 2}$	$\boldsymbol{M = 4}$	$\boldsymbol{M = 16}$
$\boldsymbol{Q = 2}$	53.33	66.67	53.33
	26.67	53.33	40.00
	0.00	40.00	26.67
$\boldsymbol{Q = 5}$	93.33	26.67	100.00
	66.67	13.33	80.00
	40.00	0.00	60.00

H_0: no effect of source variables.

Source	$d.f.$	p value	
Q	1	0.179	Accept H_0
M	2	0.106	Accept H_0
interaction	2	0.006	Reject H_0 at $\alpha = 0.01$

A completely fictional example

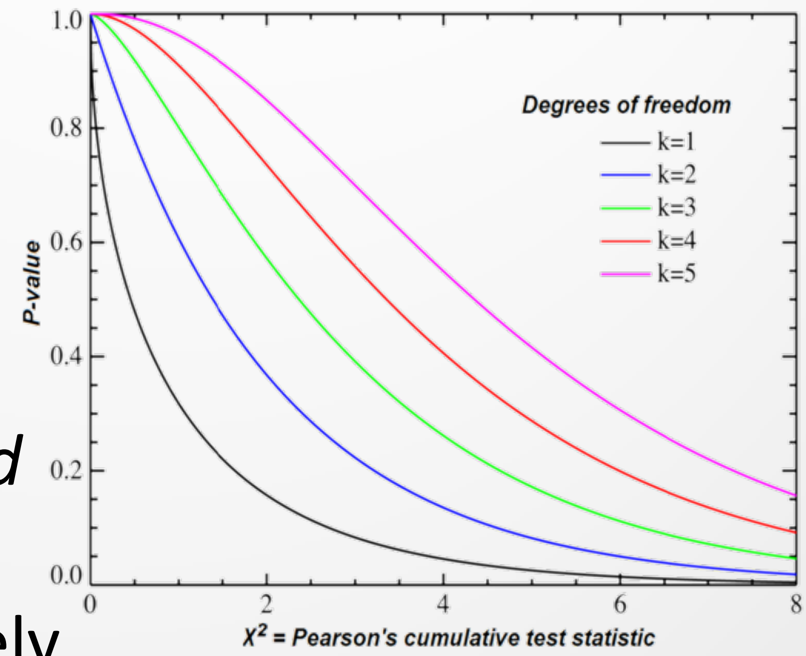
UNIVERSITY OF TORONTO

3. Pearson's χ^2 test (details aside)

- The χ^2 test applies to **categorical** data, like the output of a **classifier**.
- Like the t-test, we decide on the degrees of freedom (number of categories minus number of parameters), compute the test-statistic, then look it up in a table.
- The test statistic is:

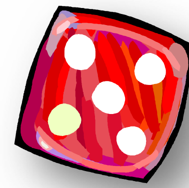
$$\chi^2 = \sum_{c=1}^{C} \frac{(O_c - E_c)^2}{E_c}$$

where O_c and E_c are the *observed* and *expected* number of observations of type c, respectively.



UNIVERSITY OF TORONTO

3. Pearson's χ^2 test



- For example, is our die from Lecture 2 fair or not?
- Imagine we throw it 60 times. The expected number of appearances of each side is 10.

c	O_c	E_c	$O_c - E_c$	$(O_c - E_c)^2$	$(O_c - E_c)^2/E_c$
1	5	10	-5	25	2.5
2	8	10	-2	4	0.4
3	9	10	-1	1	0.1
4	8	10	-2	4	0.4
5	10	10	0	0	0
6	20	10	10	100	10
				Sum (χ^2)	**13.4**

- With $df = 6 - 1 = 5$, the critical value is 11.07<**13.4**, so we throw away H_0: the die is biased.
- We'll see χ^2 again soon…

Feature selection

Determining a good set of features

- **Restricting** your feature set to a proper subset quickens **training** and reduces **overfitting**.

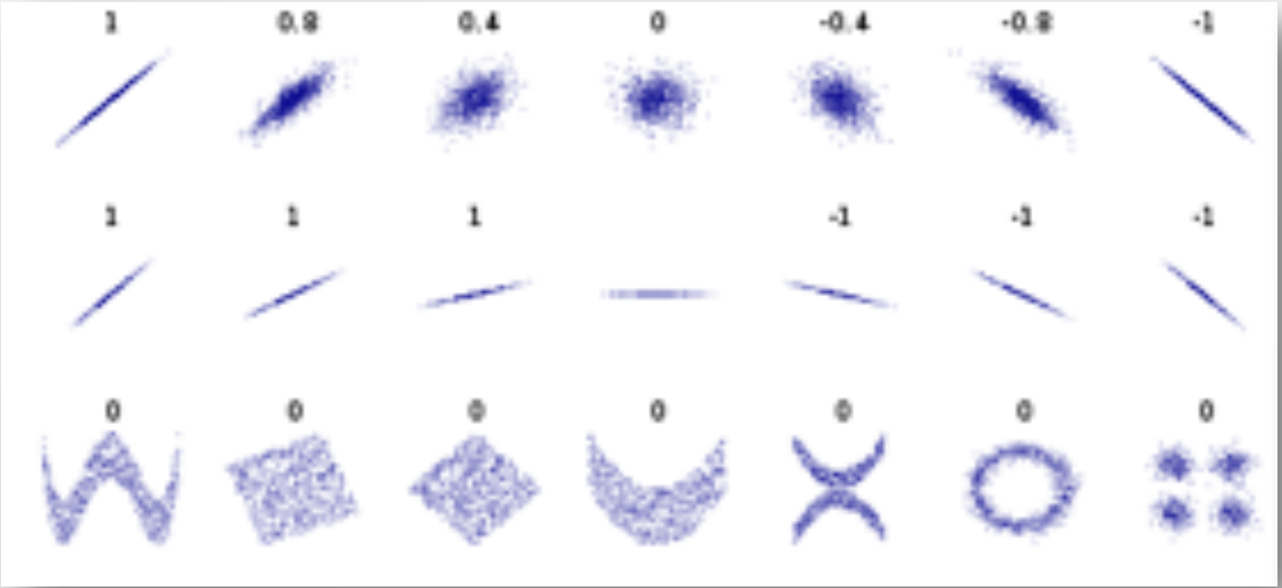
- There are a few methods that select good features, e.g.,
  1. Correlation-based feature selection
  2. Minimum Redundancy, Maximum Relevance
  3. χ^2

1. Pearson's correlation

- **Pearson** is a measure of **linear** dependence

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- Does not measure '**slope**' nor **non-linear** relations.



1. Spearman's correlation

- **Spearman** is a non-parametric measure of **rank** correlation, $r_{cX} = r(c, X)$.
  - It is basically Pearson's correlation, but on 'rank variables' that are monotonically increasing integers.
  - If the class c can be **ordered** (e.g., in any binary case), then we can compute the correlation between a feature X and that class.

1. Correlation-based feature selection

- 'Good' features should correlate **strongly** (+ or -) with the ***predicted variable*** but **not** with other ***features***.
- S_{CFS} is some set S of k features f_i that maximizes this ratio, given class c:

$$S_{CFS} = \underset{S}{\operatorname{argmax}} \frac{\sum_{f_i \in S} r_{c f_i}}{\sqrt{k + 2\sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \rho_{f_i f_j}}}$$

2. mRMR feature selection

- **Minimum-redundancy-maximum-relevance (mRMR)** can use **correlation**, **distance** scores (e.g., D_{KL}) or **mutual information** to select features.

- For feature set S of features f_i, and class c,
  $\boldsymbol{D(S,c)}$: a measure of **relevance** S has for c, and
  $\boldsymbol{R(S)}$: a measure of the **redundancy** within S,

$$\boxed{S_{mRMR} = \underset{S}{\operatorname{argmax}} \left[D(S,c) - R(S) \right]}$$

2. mRMR feature selection

- Measures of **relevance** and **redundancy** can make use of our familiar measures of *mutual information*,

- $D(S, c) = \frac{1}{\|S\|} \sum_{f_i \in S} I(f_i; c)$

- $R(S) = \frac{1}{\|S\|^2} \sum_{f_i \in S} \sum_{f_j \in S} I(f_i; f_j)$

- mRMR is **robust** but doesn't measure **interactions** of features in estimating c (for that we could use ANOVAs).

3. χ^2 method

- We adapt the χ^2 method we saw when testing whether distributions were significantly different:

$$\chi^2 = \sum_{c=1}^{C} \frac{(O_c - E_c)^2}{E_c} \Rightarrow \chi^2 = \sum_{c=1}^{C} \sum_{f_i=f}^{F} \frac{\left(O_{c,f} - E_{c,f}\right)^2}{E_{c,f}}$$

where $O_{c,f}$ and $E_{c,f}$ are the observed and expected number, respectively, of times the class c occurs together with the (discrete) feature f.
  - The expectation $E_{c,f}$ assumes c and f are **independent**.

- Now, **every feature has a *p*-value**. A lower *p*-value means c and f are *less likely* to be independent.
- Select the *k* features with the lowest *p*-values.

Multiple comparisons

- If we're just **ordering** features, this χ^2 approach is (mostly) fine.
- But what if we get a 'significant' *p*-value (e.g., $p < 0.05$)? Can we claim a significant effect of the class on that feature?

- Imagine you're flipping a coin to see if it's fair. You claim that if you get 'heads' in 9/10 flips, it's biased.
- Assuming H_0, the coin is fair, the probability that a fair coin would come up heads $\geq$ 9 out of 10 times is:

$$(\underbrace{10}_{\text{Number of ways 9 flips are heads}} + \underbrace{1}_{\text{Number of ways all 10 flips are heads}})\times 0.5^{10} = 0.0107$$

Multiple comparisons

- But imagine that you're simultaneously testing **173** coins – you're doing **173** (**multiple**) **comparisons**.
- If you want to see if *a specific chosen* coin is fair, you still have only a 1.07% chance that it will give heads $\geq \frac{9}{10}$ times.
- **But** if you don't preselect a coin, what is the probability that *none* of these fair coins will accidentally appear biased?

$$(1 - 0.0107)^{\mathbf{173}} \approx 0.156$$

- If you're testing 1000 coins?

$$(1 - 0.0107)^{\mathbf{1000}} \approx 0.0000213$$

Multiple comparisons

- The more features you evaluate with a statistical test (like χ^2), the more likely you are to accidentally find spurious (incorrect) significance **accidentally**.

- Various compensatory tactics exist, including **Bonferroni correction**, which basically divides your level of significance required, by the number of comparisons.
  - E.g., if $\alpha = 0.05$, and you're doing **173** comparisons, each would need $p < \frac{0.05}{173} \approx 0.00029$ to be considered significant.

Reading

- Manning & Schütze: 2.2, 5.3-5.5

Entropy and decisions

- **Information theory** is a vast ocean that provides statistical models of communication at the heart of **cybernetics**.
  - We've only taken a first step on the beach.
  - See the ground-breaking work of Shannon & Weaver, e.g.
- So far, we've mainly dealt with **random variables** that the world provides – e.g., words tokens, mainly.
- What if we could transform those inputs into new random variables, or **features**, that are directly engineered to be useful to decision tasks…