Information Theory - introduction

What is information?

- Provides answers to a question
- Reduces uncertainty
- Related to data and knowledge

- Measure of surprise or unexpectedness or uncertainty

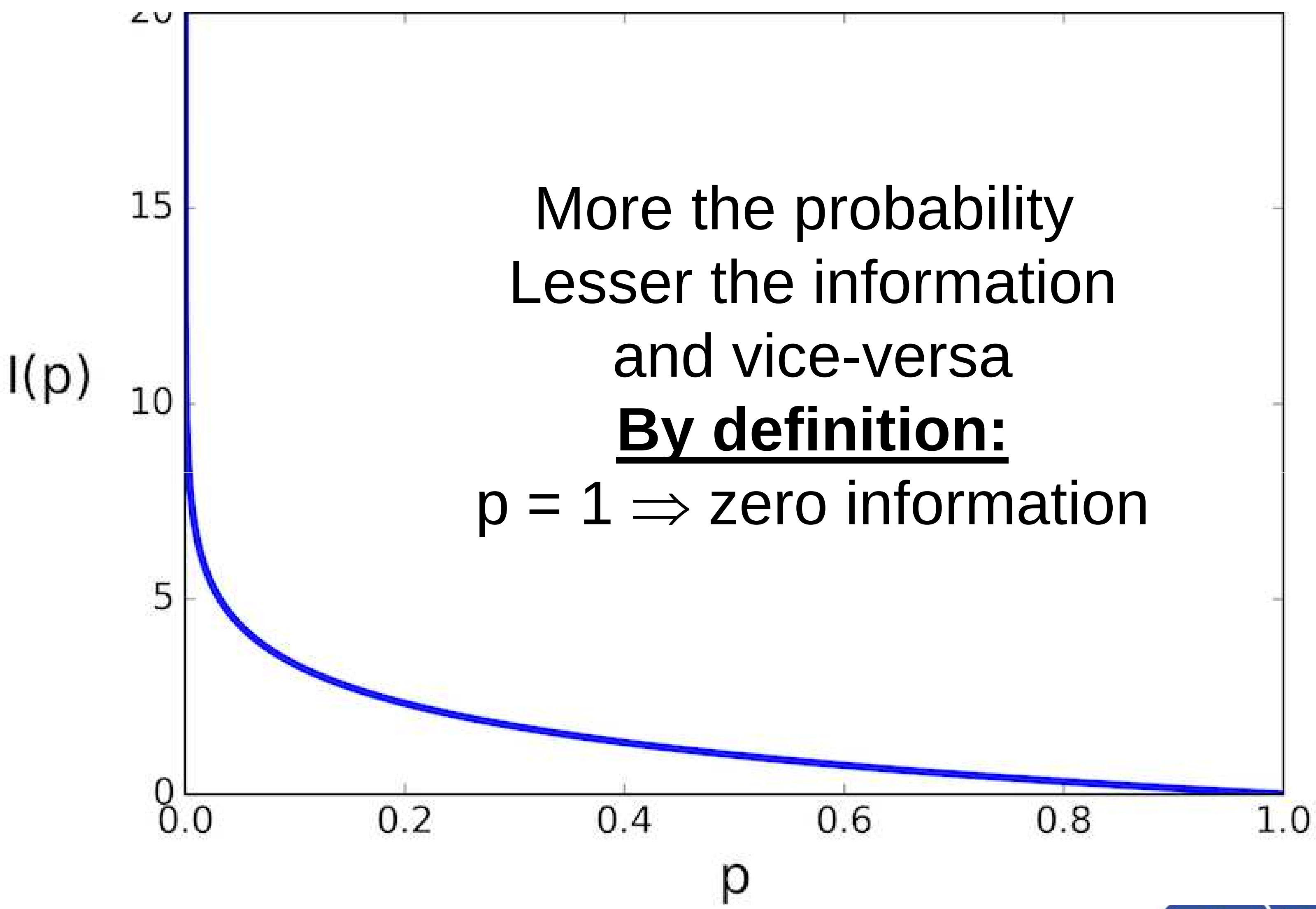
More and less information

Statements	Does it contain information?
ITA class is boring	Known fact – No info
Students are sleeping when RS is taking class	Known fact – No info
RS is sleeping while taking class	Some info
One student in ITA class shouted @ RS for the boring lecture	Can happen – still some info
One student in ITA class was admitted in Kilpak mental hospital after attending ITA course	May happen – compared to previous cases more info

Measure the Information

- Defined as: $\log_2(1/p)$ or $-\log_2(p)$
- Sometimes called **self information**

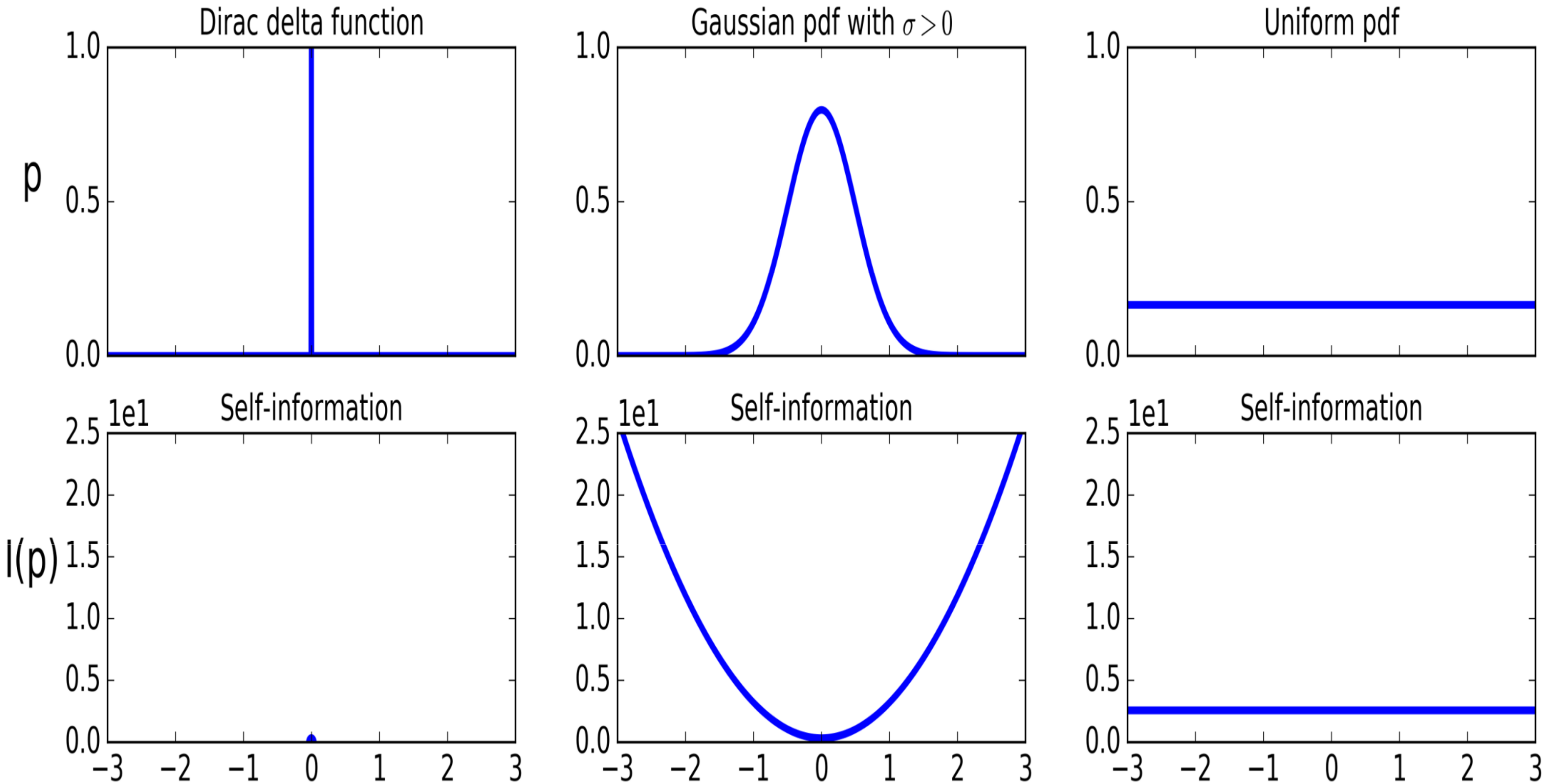
Information versus probability



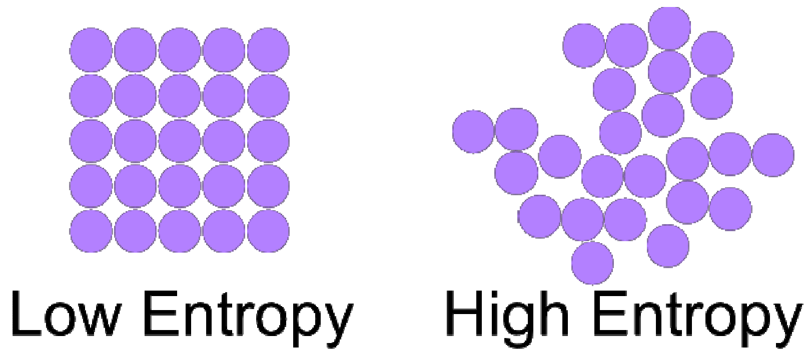
More the probability
Lesser the information
and vice-versa

By definition:

$p = 1 \Rightarrow$ zero information


Dirac delta function
Gaussian pdf with $\sigma > 0$
Uniform pdf
p
1.0
0.5
0.0
Self-information
Self-information
Self-information
1e1
I(p)
2.5
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3

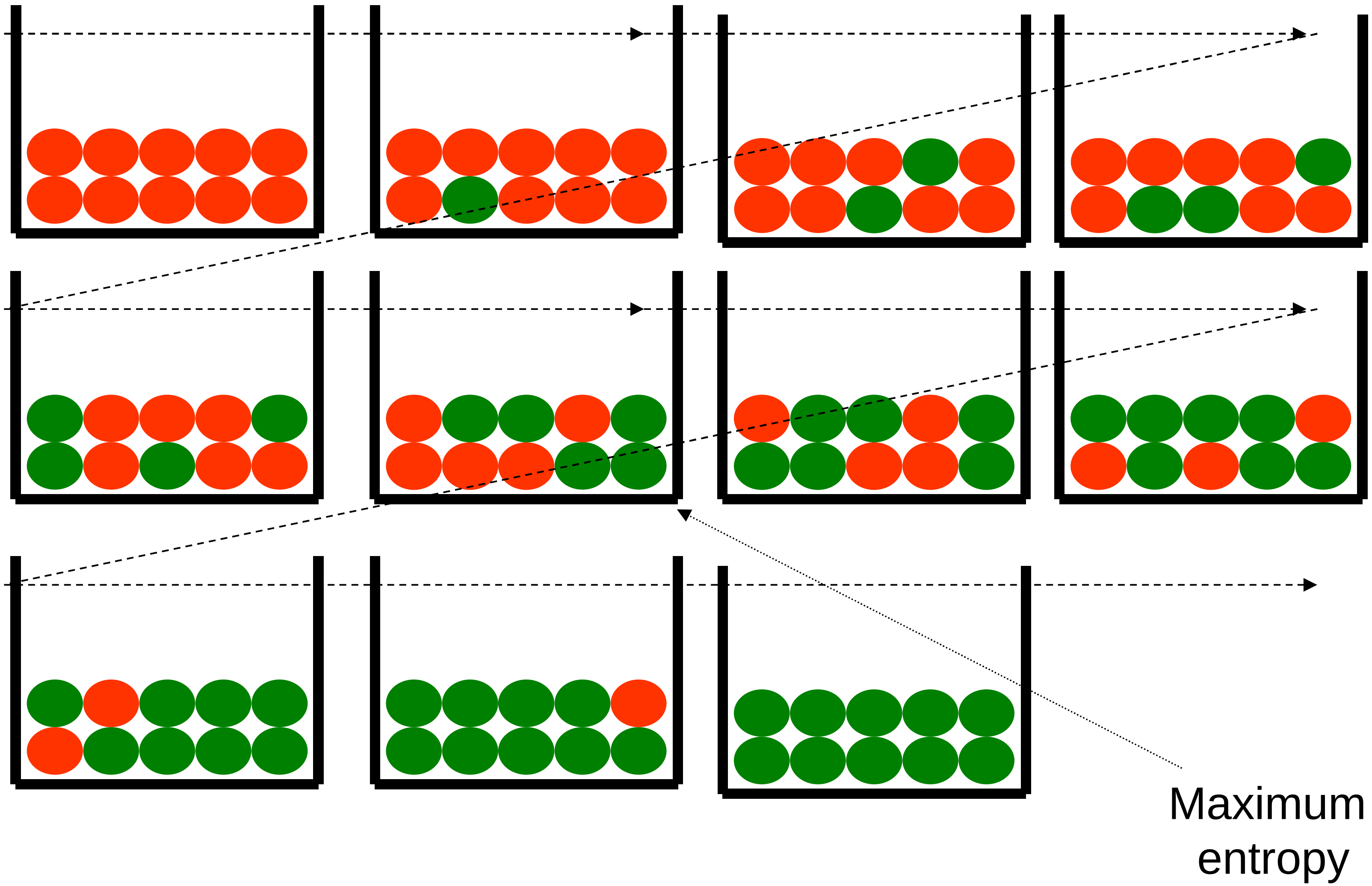
More entropy more disorder



Entropy maps the degree of order and disorder:

Higher entropy indicates more disorder

We are born organized, age with increasing disorder, and die in maximum disorder



Entropy increases and then decreases

High and Low entropy

Higher Entropy	Lower Entropy
Random	Non-random
Disorganized	Organized
Disordered	Ordered
Configurational Variety	Restricted Arrangements
Freedom of Choice	Constraint
Uncertainty	Reliability
Higher Error Probability	Fidelity

How to measure uncertainty?

Hartley's measure

$I = \log_b r$

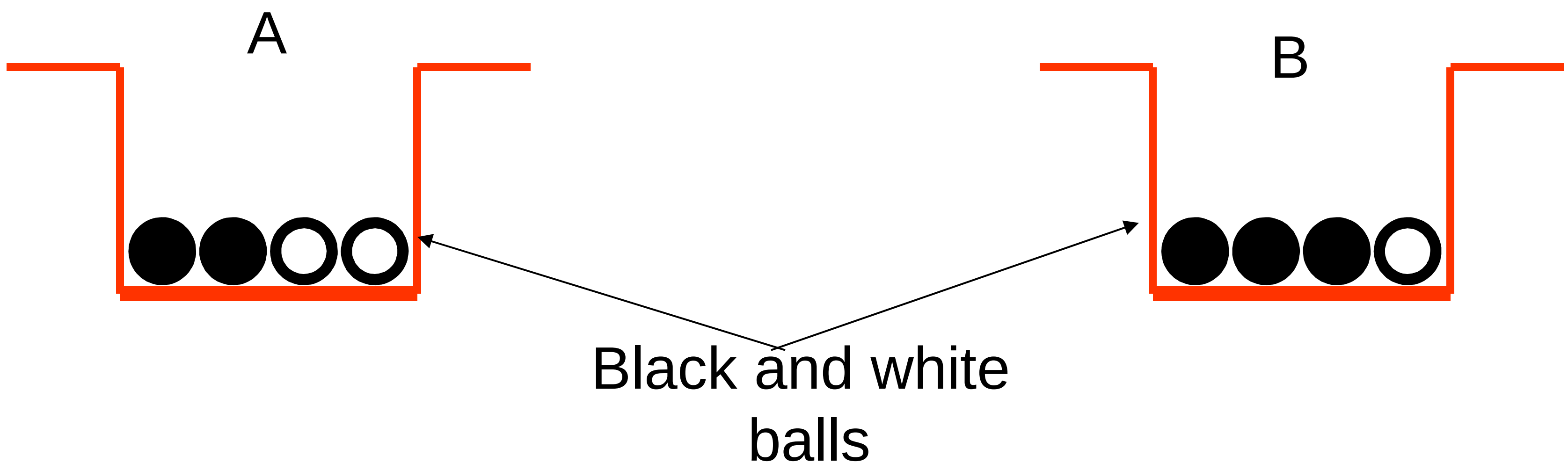
r - number of all possible outcomes of a random message U

E.g.

- Coin flip experiment
- Two possible outcomes

$I = \log_2 2 = 1$ bit

Motivating e.g. for a better measure



Possible outcomes = 2
{Black, White}
$I = \log_2 2 = 1$ bit

Possible outcomes = 2
{Black, White}
$I = \log_2 2 = 1$ bit

- From A&B we choose a random ball, and we got black balls from both
- Getting a black ball from B is somewhat expected i.e. we already anticipated the result of B i.e. the information content is low for B

Lesson

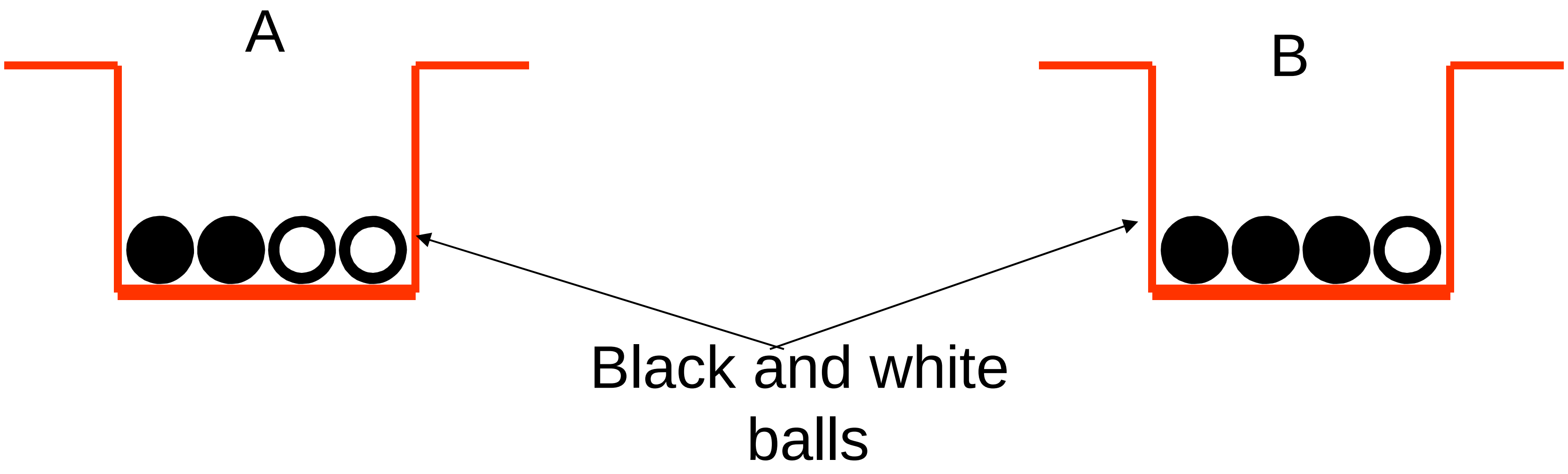
- Measure needs to consider the possible probabilities of the various events

Shannon's measure

$I = \log_2(1/p_i) = -\log_2(p_i)$

- p_i denotes the probability of the i^{th} possible outcome

Motivating e.g. for a better measure



- From A&B we choose a random ball, and we got black balls from both

Possible outcomes = 2
{Black, White}
p(black) = p(white) = ½
$I = -\log_2 2 = 1$ bit

Possible outcomes = 2
{Black, White}
p(black) = ¾ ; p(white) = ¼
$I = -\log_2 (3/4) = 0.415$ bit

Shannon's measure - average

$$H(U) \triangleq -\sum_{i=1}^{r} p_i \log_b p_i$$

Assumption: Exclude all indices *i* with $p_i = 0$

$$\sum_{i=1}^{r} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{r} p_i \log_2 p_i$$

Entropy is positive

P_i is 0 to 1 $\Rightarrow$ $(1/P_i)$ is greater than 1 $\Rightarrow$ $\log(1/P_i)$ is greater than or equal to 0

Example

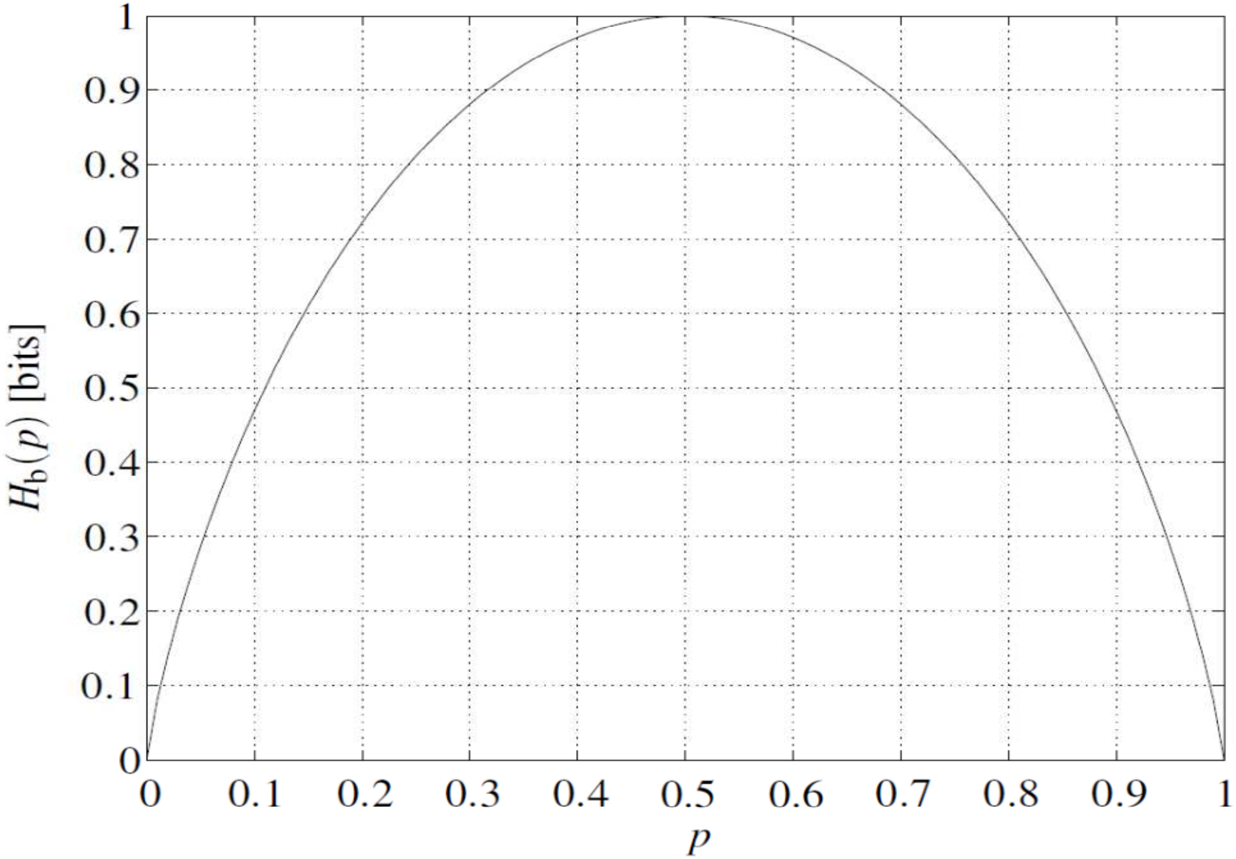
$$X = \begin{cases} a & \text{with probability } \frac{1}{2}, \\ b & \text{with probability } \frac{1}{4}, \\ c & \text{with probability } \frac{1}{8}, \\ d & \text{with probability } \frac{1}{8}. \end{cases}$$

The entropy of X is

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4} \text{ bits.}$$

Binary entropy function (e.g. coin flipping)

$$H_{\mathrm{b}}(p) \triangleq -p \log_2 p - (1-p)\log_2(1-p), \qquad p \in [0,1]$$



Entropy - a to z

i	a_i	p_i	$h(p_i)$
1	a	.0575	4.1
2	b	.0128	6.3
3	c	.0263	5.2
4	d	.0285	5.1
5	e	.0913	3.5
6	f	.0173	5.9
7	g	.0133	6.2
8	h	.0313	5.0
9	i	.0599	4.1
10	j	.0006	10.7
11	k	.0084	6.9
12	l	.0335	4.9
13	m	.0235	5.4
14	n	.0596	4.1
15	o	.0689	3.9
16	p	.0192	5.7
17	q	.0008	10.3
18	r	.0508	4.3
19	s	.0567	4.1
20	t	.0706	3.8
21	u	.0334	4.9
22	v	.0069	7.2
23	w	.0119	6.4
24	x	.0073	7.1
25	y	.0164	5.9
26	z	.0007	10.4
27	–	.1928	2.4
$\sum_i p_i \log_2 \frac{1}{p_i}$			4.1

- A random character is picked from an English document. Denote the output by x.
- Outcome x = z has a Shannon information content of 10.4 bits
- x = e has an information content of 3.5 bits

Property # 1

- Addition of an impossible event does not change the entropy

$H_{N+1}(p_1, p_2, p_3, \ldots p_N, 0) = H_N(p_1, p_2, p_3, \ldots p_N)$

Property#2

- Entropy vanishes when one outcome is certain to happen

$H_N(p_1, p_2, p_3,\ldots p_N) = 0$

if $p_i=1$, $p_j = 0$ with $i \neq j$

Property#3

- Two independent probability distributions $P_X = \{p_1, . . . , p_N\}$ & $Q_Y = \{q_1, . . . , q_M\}$

$$H = -\sum_{i=1}^{N} \sum_{j=1}^{M} p(i, j) \log (p(i, j))$$

$$p(i, j) = p(i) q(j)$$

$$H = -\sum_{i=1}^{N} \sum_{j=1}^{M} p(i) q(j) \log (p(i) q(j))$$

$$= -\sum_{i=1}^{N} \sum_{j=1}^{M} p(i) q(j) [\log p(i) + \log q(j)]$$

$$= -\sum_{i=1}^{N} \sum_{j=1}^{M} p(i) q(j) \log p(i) + -\sum_{i=1}^{N} \sum_{j=1}^{M} p(i) q(j) \log q(j)$$

$$= -\sum_{i=1}^{N} p(i) \log p(i) \{q_1 + \ldots + q_M\} + -\sum_{j=1}^{M} q(j) \log q(j) \{p_1 + \ldots + p_N\}$$

$$\{q_1 + \ldots + q_M\} = \{p_1 + \ldots + p_N\} = 1$$

$$H = -\sum_{i=1}^{N} p(i) \log p(i) + -\sum_{j=1}^{M} q(j) \log q(j)$$

$$= H_{P_X} + H_{Q_Y}$$

Discrete versus Continuous pdf

Discrete case

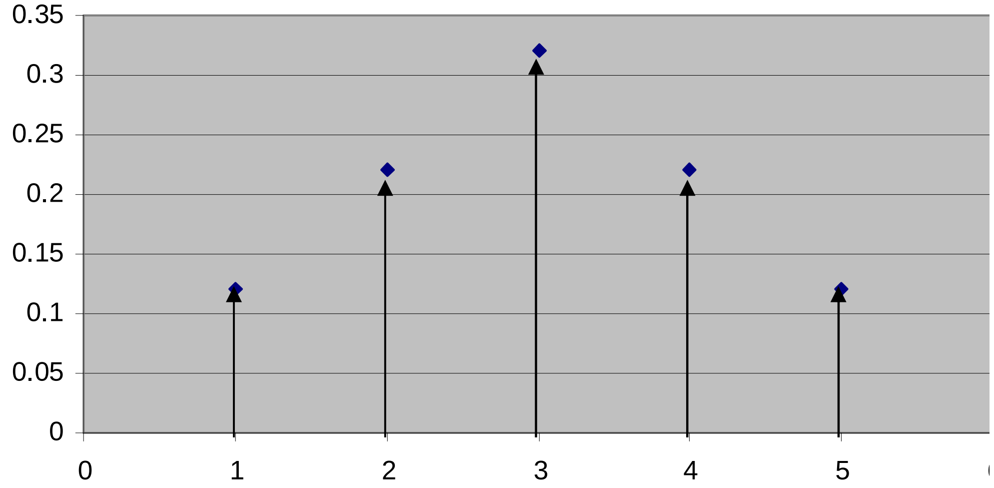
- Pdf values @ a point < 1
- $\sum_i p_i = 1$

Continuous case

- Pdf values @ a point can be >1
- $\int p(x)\, dx = 1$

Find entropy-discrete pdf - example

- p(1) = 0.12; p(2)=0.22; p(3)=0.32; p(4)=0.22; p(5)=0.12

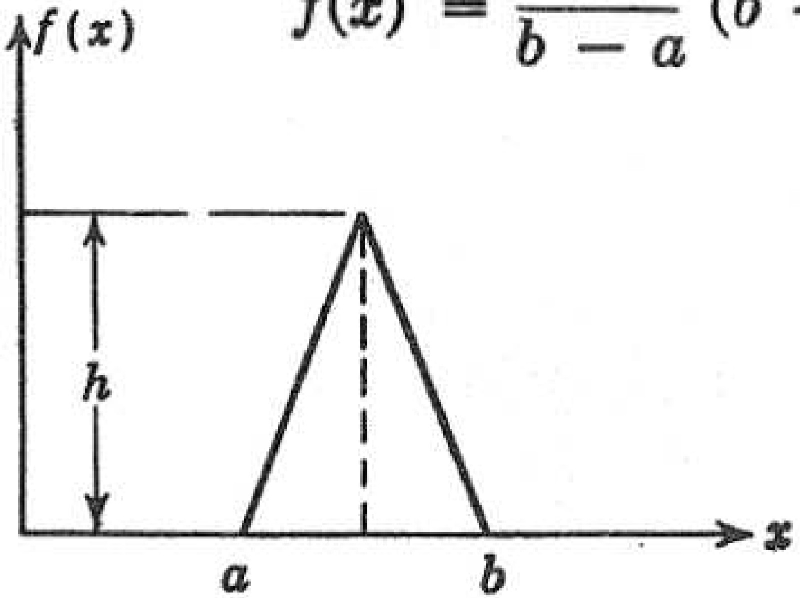


- H=0.12*log2(1/0.12) + 0.22*log2(1/0.22) + 0.32*log2(1/0.32) + 0.22*log2(1/0.22) + 0.12*log2(1/0.12) = 2.22 bits

Find entropy-continuous pdf - example

$$f(x) = \frac{2h}{b-a}(x-a) \qquad \text{for } a \le x \le \frac{a+b}{2}$$

$$f(x) = \frac{2h}{b-a}(b-x) \qquad \text{for } \frac{a+b}{2} \le x \le b$$



$$H(X) = -\int_a^{(a+b)/2} \frac{2h}{b-a}(x-a)\ln\frac{2h}{b-a}(x-a)\,dx - \int_{(a+b)/2}^{b} \frac{2h}{b-a}(b-x)\ln\frac{2h}{b-a}(b-x)\,dx$$

$$\int x \ln \lambda x \, dx = \frac{x^2}{2}\ln \lambda x - \frac{x^2}{4}$$

$$= \frac{-h}{b-a}\left[(x-a)^2 \ln\frac{2h}{b-a}(x-a) - \frac{(x-a)^2}{2}\right]_a^{(a+b)/2} + \frac{h}{b-a}\left[(b-x)^2 \ln\frac{2h}{b-a}(b-x) - \frac{(b-x)^2}{2}\right]_{(a+b)/2}^{b}$$

$$= \frac{-h}{b-a}\left\{\left[\frac{(b-a)^2}{4}\ln h - \frac{(b-a)^2}{8}\right] + \left[\frac{(b-a)^2}{4}\ln h - \frac{(b-a)^2}{8}\right]\right\}$$

$$= \frac{h(b-a)}{2}\left(-\ln h + \frac{1}{2}\right)$$

$$\frac{h(b-a)}{2} = \int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$H(X) = -\ln h + 1/2$$

Entropy becomes negative

$$
\begin{aligned}
H(X) &> 0 \qquad \text{for } h < \sqrt{e} \\
H(X) &= 0 \qquad \text{for } h = \sqrt{e} \\
H(X) &< 0 \qquad \text{for } h > \sqrt{e}
\end{aligned}
$$

Entropy - example

- A fair coin is flipped until the first head occurs. Let *X* denote the number of flips required. Find entropy H(X).
- Let p and q denote head and tail probability = 0.5
- Getting head in 1^{st} toss = 0.5
- Getting head in 2^{nd} toss = q . p
- Getting head in 3^{rd} toss = q .q. p
- Getting head in n^{th} toss = $q^{(n-1)}.p$

$$H(X) = -\sum_{n=1}^{\infty} pq^{(n-1)} \log\left(pq^{(n-1)}\right)$$

$$= -\left[\sum_{n=1}^{\infty} pq^{(n-1)} \log(p) + \sum_{n=1}^{\infty} pq^{(n-1)} \log\left(q^{(n-1)}\right)\right]$$

$$= -\left[\sum_{n=0}^{\infty} pq^{(n)} \log(p) + \sum_{n=0}^{\infty} pq^{(n)} \log\left(q^{(n)}\right)\right]$$

$$= -\left[\sum_{n=0}^{\infty} pq^{(n)} \log(p) + \sum_{n=0}^{\infty} npq^{(n)} \log\left(q\right)\right]$$

$$we \quad know \;, \sum_{n=0}^{\infty} r^{n} = \frac{1}{1-r} \quad and \quad \sum_{n=0}^{\infty} nr^{n} = \frac{r}{(1-r)^2}$$

$$= -p\log(p).\frac{1}{1-q} - p\log(q).\frac{q}{(1-q)^2}$$

$$= -\frac{p\log(p)}{p} - \frac{pq\log(q)}{p^2} = \frac{-(p\log(p) + q\log(q))}{p}$$

$$= 2 \;\; bits$$

Entropy decomposition

- A random variable $x \in \{0, 1, 2\}$ is created by
  - first flipping a fair coin to determine whether $x = 0$
  - if $x \neq 0$, flipping a fair coin a second time to determine whether x is 1 or 2
- $P(x=0) = ½$; $P(x=1) = ¼$; $P(x=2) = ¼$

Entropy decomposition

1st time toss (getting head implies x=0)	2nd time toss(only when we didn't get head 1st time) H→1 T→2	P(x=0)= get head 1st time	P(x=1)= get tail 1st time x get head 2nd time = ½ x ½	P(x=2)= get tail 1st time x get tail 2nd time = ½ x ½
H		½	¼	¼
T	H			
	T			

- $P(x=0) = ½$; $P(x=1) = ¼$; $P(x=2) = ¼$

1st method

$H(X) = \frac{1}{2} \cdot \log 2 + \frac{1}{4} \cdot \log 4 + \frac{1}{4} \cdot \log 4 = 1.5$ bits

2nd method

1st toss: $H(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2} \cdot \log 2 + \frac{1}{2} \cdot \log 2 = 1$ bit

2nd toss: $H(\frac{1}{2}, \frac{1}{2}) = \frac{1}{2} \cdot \log 2 + \frac{1}{2} \cdot \log 2 = 1$ bit

But the 2nd toss happens only half of the time

$H = H(1^{st} \text{ toss}) + \frac{1}{2} \cdot H(2^{nd} \text{ toss})$

$\quad = 1 + \frac{1}{2} \cdot \frac{1}{2} = 1.5$ bits

Example

- A source produces a character x from the alphabet A = {0, 1,…., 9; a, b,…., z}
  – With probability 1/3, x is a numeral (0, 1, 2, …,9)
  – with probability 1/3, x is a vowel (a, e, i, o, u)
  – with probability 1/3 it's one of the 21 consonants
- All numerals are equiprobable, and the same goes for vowels and consonants. Estimate the entropy of X

Two successive events

1st event: Getting numerals or vowels or consonants

2nd event: Distribution with in them – but this will happen 1/3 of time only

$H(1^{st} \text{ event}) = 1/3 \,.\, \text{Log } 1/3 + 1/3 \,.\, \text{Log } 1/3 + 1/3 \,.\, \text{Log } 1/3 = \log 3$

$H(2^{nd} \text{ event:numerals}) = \log 10$

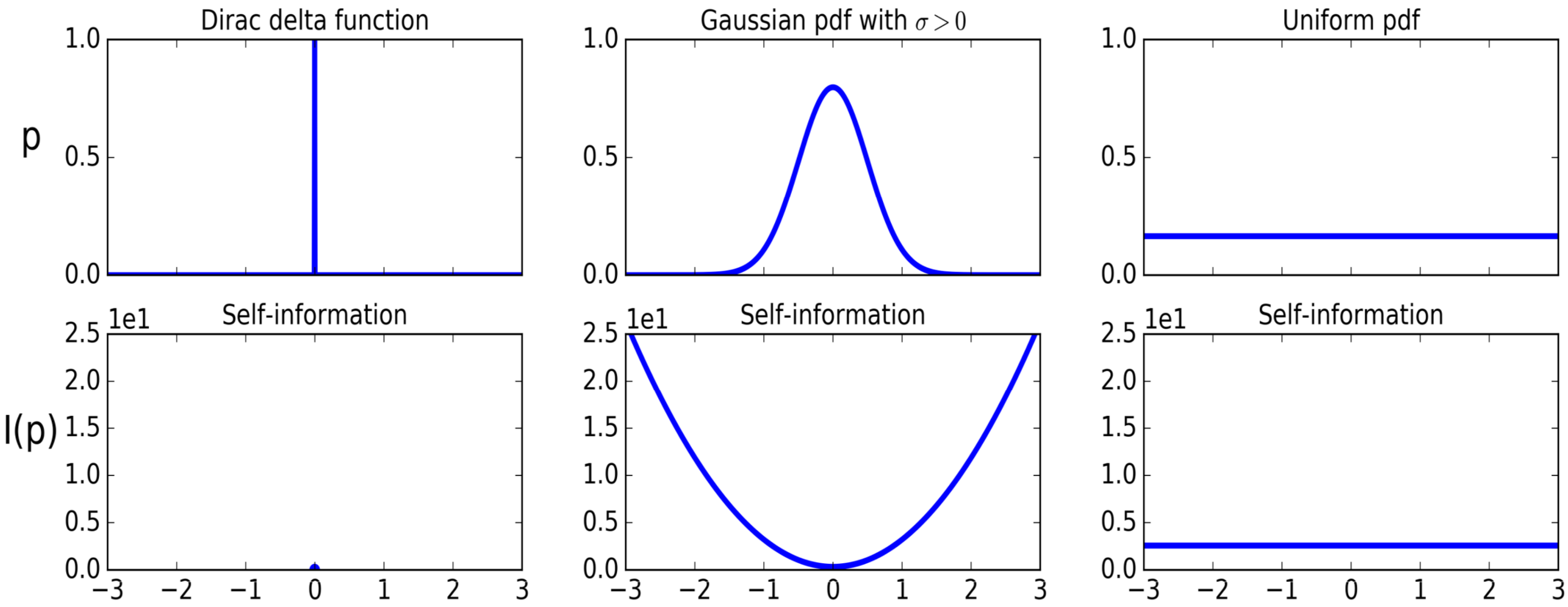
$H(2^{nd} \text{ event:vowels}) = \log 5$

$H(2^{nd} \text{ event:consonants}) = \log 21$

$$H = \log 3 + \frac{1}{3}\left(\log 10 + \log 5 + \log 21\right)$$

$$= 1.585 + \frac{1}{3}\left(3.322 + 2.322 + 4.392\right)$$

$$= 4.93 \ \ bits$$



Entropy = 0 Entropy = 174 Entropy = 431

- Broader the distribution higher the entropy
- Gaussian becomes broader with increasing σ (std dev.)
- Higher the σ, more the entropy and viceversa
- In the limit, $\sigma \to \infty$, Gaussian → uniform distribution
- In the limit, $\sigma \to 0$, Gaussian → delta distribution

Expectation

- Consider a random variable
- **Probability function**

$[x_1, x_2, \ldots\ldots, x_n]$

$[p_1, p_2, \ldots\ldots, p_n]$

- Average of X

$$\overline{X} = \sum_{k=1}^{n} p_k x_k$$

$$\overline{X} = \sum_{k=1}^{n} p_k x_k$$

More generic form,

$$\overline{\psi(X)} = \sum_{k=1}^{n} p_k . \psi(x_k)$$

Entropy as expectation

What is entropy?

$H=-\{p_1.\log(p_1)+p_2.\log(p_2)+\ldots+p_n.\log(p_n)\}$

- What we are summing?
  - Log of probabilities
- What are their weighing factor?
  - Their probability values
- Entropy = -E(log(P))