03

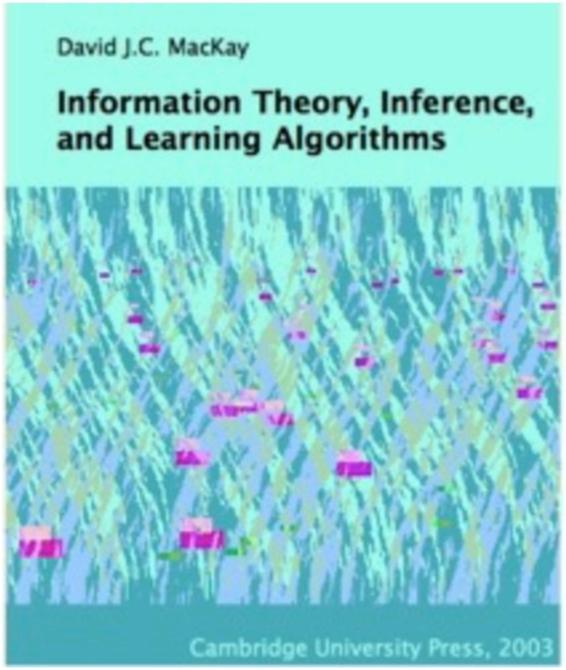
Entropy and related functions

Notice

- **Author**
  - **João Moura Pires (jmp@fct.unl.pt)**

Bibliography

- **Many examples are extracted and adapted from:**



Information Theory, Inference, and Learning Algorithms
David J.C. MacKay
2005, Version 7.2

- **And some slides were based on Iain Murray course**
  - **http://www.inf.ed.ac.uk/teaching/courses/it/2014/**

Table of Contents

- **Definition of Entropy and related Functions**
- **Decomposability of the entropy**
- **Gibbs' inequality**
- **Jensen's inequality for convex functions**
- **Designing informative experiments**

Definition of Entropy and related Functions

The Shannon information content of an outcome

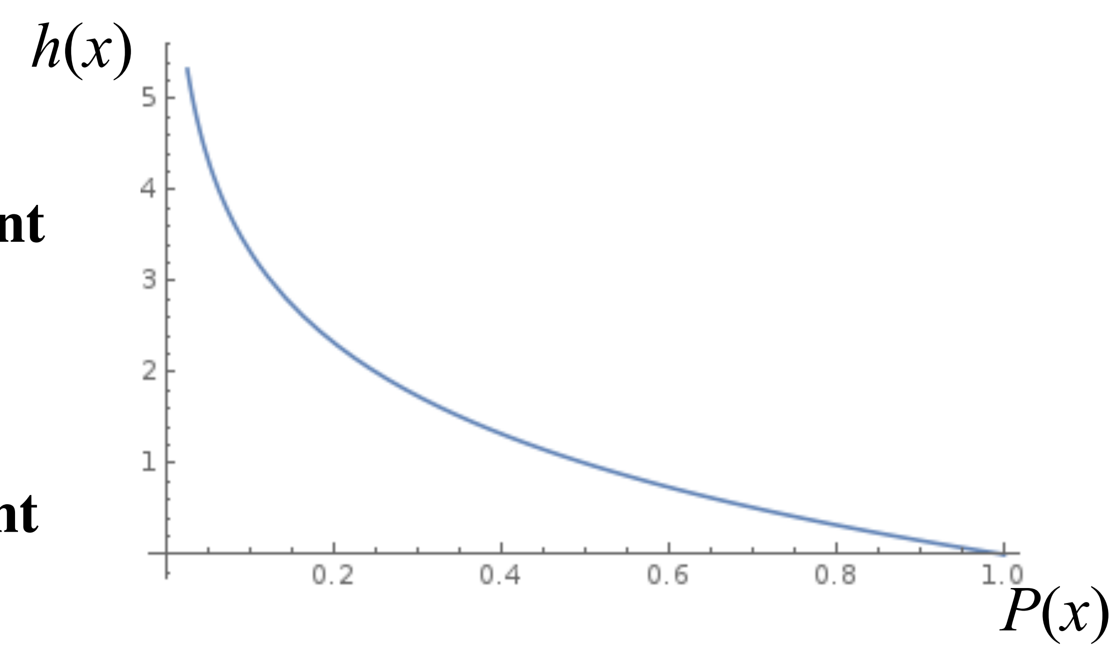
- The **Shannon information content** of an outcome x is defined to be

$$h(x) = \log_2 \frac{1}{P(x)} = -\log_2 P(x)$$

- It is measured in **bits**
  - The word bit is is also used to denote a variable whose value is 0 or 1 (**b**inary dig**it**)
- $h(a_i)$ is indeed a natural **measure of the information content** of the event $x = a_i$.
  - When a_i is **almost certain** ($P(a_i)$ near to 1)

    the occurrence of a has a **small information content**
  - When a_i is **very unlikely** ($P(a_i)$ near to 0)

    the occurrence of a has a **large information content**

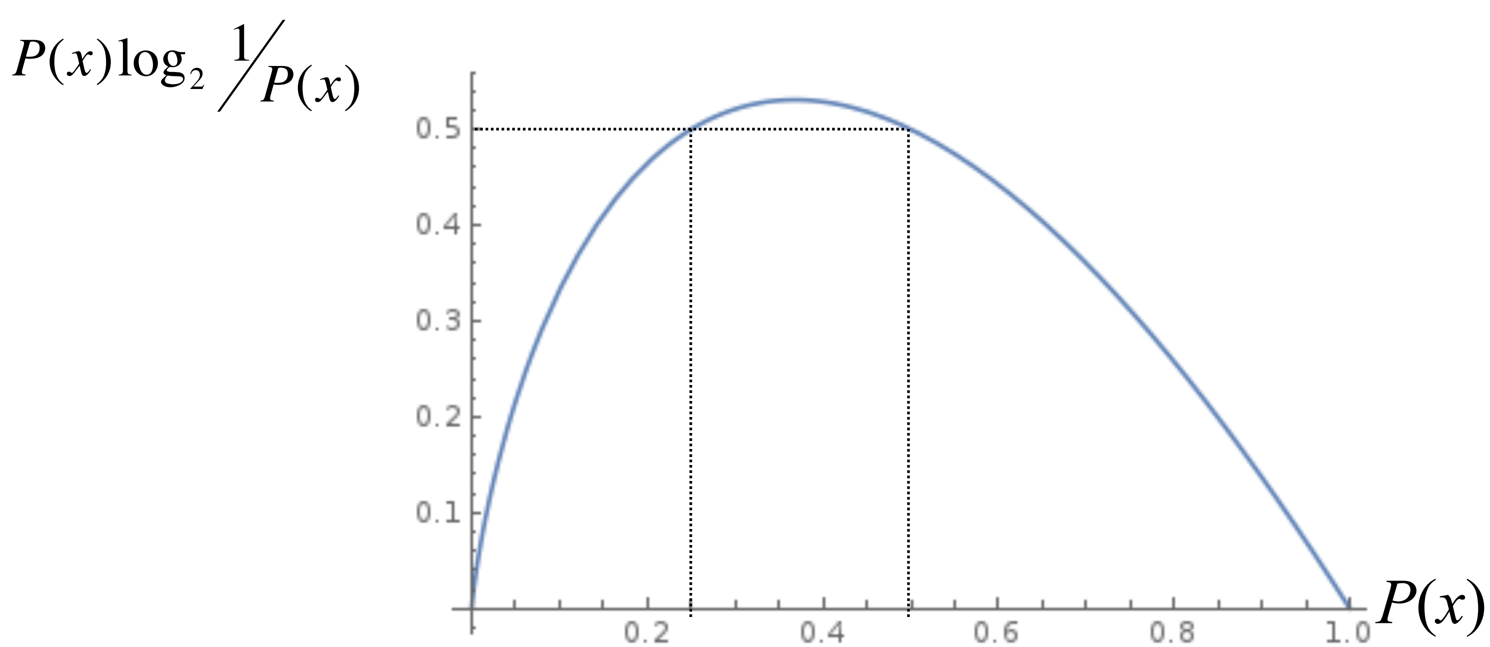


Entropy of an ensemble X

- The **entropy** of an ensemble X is defined to be the **average Shannon information content** of an outcome:

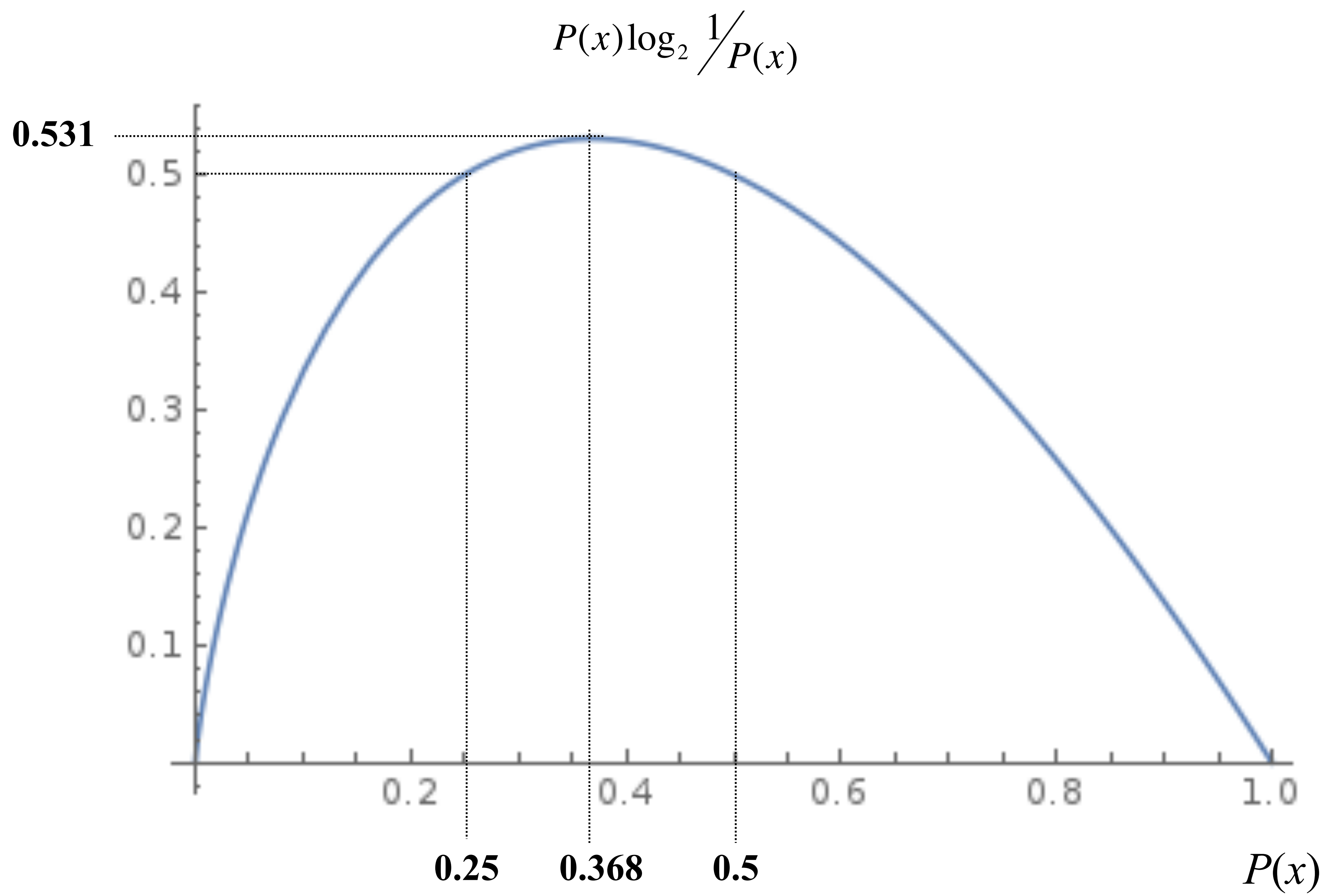
$$H(x) = \sum_{x \in A_X} P(x) \log_2 \frac{1}{P(x)} = -\sum_{x \in A_X} P(x) \log_2 P(x)$$

with the convention for $P(x) = 0$ that $0 \times \log 1/0 \equiv 0$, $\lim_{\theta \to 0+} \theta \log {}^{1}\!/\!_{\theta} = 0$



The contribution of each outcome x

- The contribution of each outcome x to the entropy of an ensemble X is $P(x)\log_2 {}^{1}\!/\!_{P(x)}$



An example

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

Shannon information contents of the outcomes a–z from a text.

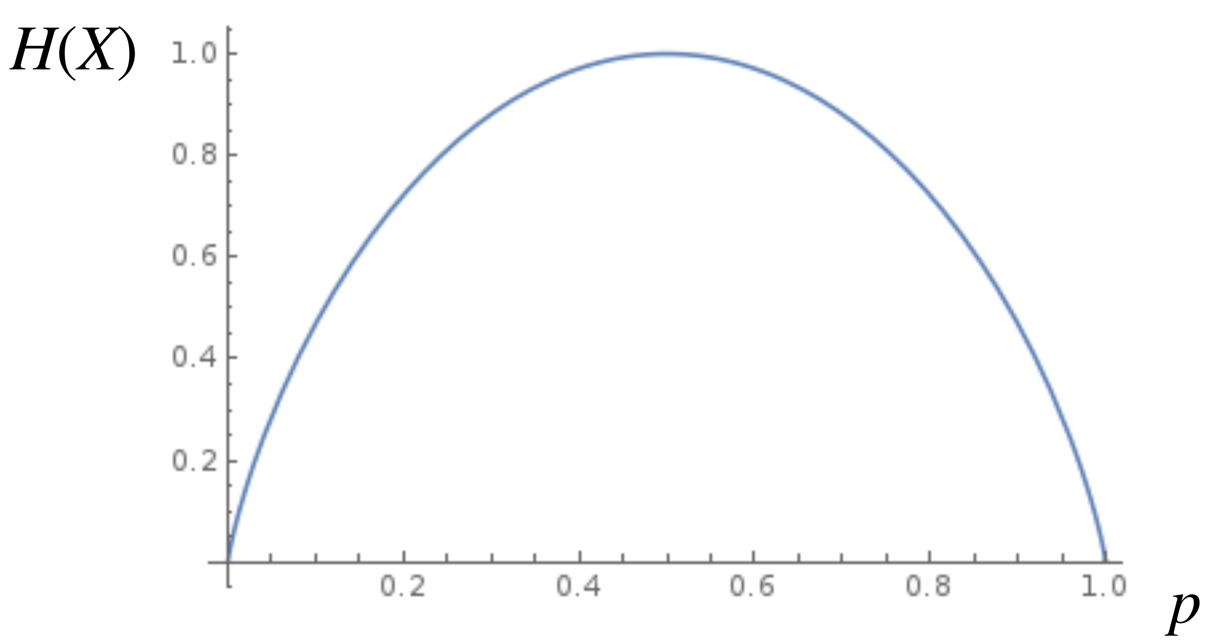
$$H(X) = 4.1 \text{ bits}$$

Some properties of $H(X)$

- $H(X) \geq 0$
  - $H(X) = 0$ if and only if $p_i = 1$ for one i.
- Entropy is maximized if ***p*** is uniform $H(X) \leq \log(|A_X|)$
  - $H(X) = \log(|A_X|)$ if and only if $p_i = \frac{1}{|A_X|}$ for all i

---

- Case of binary ensemble $A_X = \{a_1, a_2\}$ and $P(a_1) = p$ and consequently $P(a_2) = 1 - p$



$H(X) = 1$ bit

only when p = 1/2

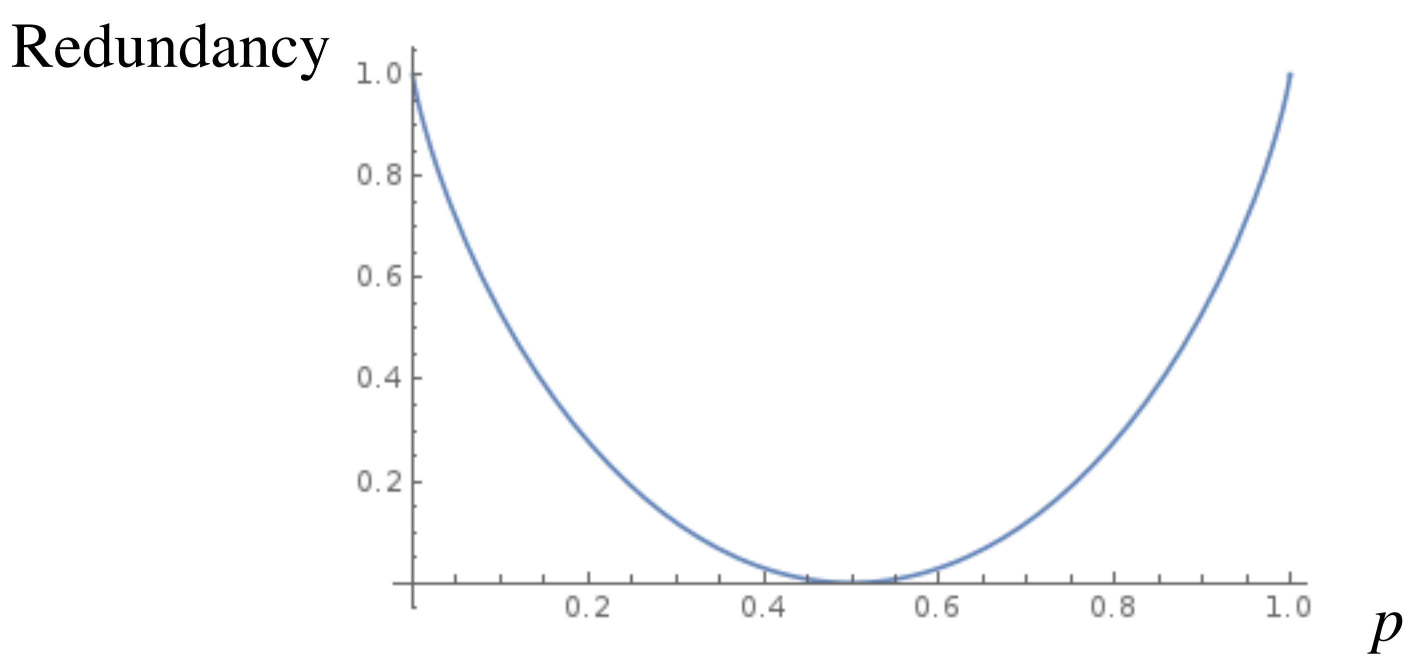
Redundancy of X

- The **redundancy** of X is:

$$1 - \frac{H(X)}{\log|A_X|}$$

  - When the entropy (or uncertainty) is maximal the redundancy is minimal
  - When the entropy (or uncertainty) is minimal the redundancy is maximal

- Case of binary ensemble $A_X = \{a_1, a_2\}$ and $P(a_1) = p$ and consequently $P(a_2) = 1 - p$



Joint Entropy of X, Y

- The **Joint Entropy** of X, Y

$$H(X,Y) = \sum_{x \in A_X, y \in A_Y} P(x,y) \log_2 \frac{1}{P(x,y)}$$

- Entropy is additive for independent random variables:

$$H(X,Y) = H(X) + H(Y) \quad \textit{iff} \quad P(x,y) = P(x)P(y)$$

Decomposability of the entropy

Decomposability of the entropy

- The entropy function satisfies a **recursive property** that can be very useful when computing entropies.
- We can write $H(X)$ as $H(\boldsymbol{p})$, where $\boldsymbol{p}$ is the **probability vector** associated with the ensemble X.

$A_X = \{0, 1, 2\}$

$P(x = 0) = 1/2$; $P(x = 1) = 1/4$; $P(x = 2) = 1/4$;

$H(X) = 1/2 \log 2 + 1/4 \log 4 + 1/4 \log 4 = 1.5$

$H(X) = H(1/2, 1/4, 1/4) = 1.5$

$$\boldsymbol{p} = [\boxed{1/2,}\ \boxed{1/4, 1/4}]$$

$H(X) = H(1/2, 1/2) + 1/2\, H(1/2, 1/2) = 1.5$

Decomposability of the entropy

- For any probability distribution $\boldsymbol{p} = \{p_1, p_2, \ldots, p_I\}$

$$H(\boldsymbol{p}) = H(p_1, 1-p_1) + (1-p_1)H\left(\frac{p_2}{1-p_1}, \frac{p_3}{1-p_1}, \ldots, \frac{p_I}{1-p_1}\right)$$

- And can be more generalized for

$$\begin{aligned} H(\mathbf{p}) \quad = \quad & H\left[(p_1 + p_2 + \cdots + p_m), (p_{m+1} + p_{m+2} + \cdots + p_I)\right] \\ & +(p_1 + \cdots + p_m) H\left(\frac{p_1}{(p_1 + \cdots + p_m)}, \ldots, \frac{p_m}{(p_1 + \cdots + p_m)}\right) \\ & +(p_{m+1} + \cdots + p_I) H\left(\frac{p_{m+1}}{(p_{m+1} + \cdots + p_I)}, \ldots, \frac{p_I}{(p_{m+1} + \cdots + p_I)}\right) \end{aligned}$$

Decomposability of the entropy

- And can be more generalized for

$$
\begin{aligned}
H(\mathbf{p}) &= H\left[(p_1 + p_2 + \cdots + p_m), (p_{m+1} + p_{m+2} + \cdots + p_I)\right] \\
&+ (p_1 + \cdots + p_m) H\left(\frac{p_1}{(p_1 + \cdots + p_m)}, \ldots, \frac{p_m}{(p_1 + \cdots + p_m)}\right) \\
&+ (p_{m+1} + \cdots + p_I) H\left(\frac{p_{m+1}}{(p_{m+1} + \cdots + p_I)}, \ldots, \frac{p_I}{(p_{m+1} + \cdots + p_I)}\right)
\end{aligned}
$$

$\sum = A$	$\sum = B$
$p_1, p_2, \ldots p_m$	$p_{m+1}, p_{m+2}, \ldots p_I$

$p'_i = \frac{p_i}{A}$ $\qquad$ $p''_j = \frac{p_j}{B}$

$$H(\boldsymbol{p}) = H(A,B) + AH(p'_1, p'_2, \ldots, p'_m) + BH(p''_{m+1}, p''_{m+2}, \ldots, p''_I)$$

Decomposability of the entropy: an example

- A source produces a character x from the alphabet $A = \{0, 1, \ldots, 9, a, b, \ldots, z\}$
  - With probability 1/3, x is a numeral (0,…,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.

$1/3$	$1/3$	$1/3$
5 vowels	10 numerals	21 consonants

$$H(X) = H\left(1/3, 1/3, 1/3\right) + \tfrac{1}{3}(\log 5 + \log 10 + \log 21)$$

Gibbs' inequality

Relative entropy or Kullback–Leibler divergence

- The **relative entropy** or Kullback–Leibler **divergence between two probability distributions** $P(x)$ and $Q(x)$ that are defined over the same alphabet A_X is

$$D_{KL}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- The relative entropy satisfies Gibbs' inequality

$$D_{KL}(P \| Q) \geq 0 \qquad\qquad D_{KL}(P \| Q) = 0 \text{ only if } P = Q$$

- In general $D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$

For more information read <u>here</u>

Relative entropy

$$D_{KL}(P \parallel Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

$P(x)$	$Q(x)$	$P(x)/Q(x)$	$P(x)\log_2(P(x)/Q(x))$
0,5	0,5	1,00	0,00
0,25	0,3	0,83	-0,07
0,25	0,2	1,25	0,08

$$D_{KL}(P \parallel Q) = 0{,}0147$$

$Q(x)$	$P(x)$	$Q(x)/P(x)$	$Q(x)\mathrm{Log}_2(Q(x)/P(x))$
0,5	0,5	1,00	0,00
0,3	0,25	1,20	0,08
0,2	0,25	0,80	-0,06

$$D_{KL}(P \parallel Q) = 0{,}0145$$

$P(x)$	$Q(x)$	$P(x)/Q(x)$	$P(x)\log_2(P(x)/Q(x))$
0,5	0,3333	1,50	0,29
0,25	0,3333	0,75	-0,10
0,25	0,3333	0,75	-0,10

$$D_{KL}(P \parallel Q) = 0{,}0850$$

$Q(x)$	$P(x)$	$Q(x)/P(x)$	$Q(x)\mathrm{Log}_2(Q(x)/P(x))$
0,3333	0,5	0,67	-0,19
0,3333	0,25	1,33	0,14
0,3333	0,25	1,33	0,14

$$D_{KL}(P \parallel Q) = 0{,}0817$$

$$H\left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right) = \log_2 3 = 1.585 bits$$

$$H(0.5, 0.25, 0.25) = 1.5 bits$$

$$H(0.5, 0.3, 0.20) = 1.485 bits$$

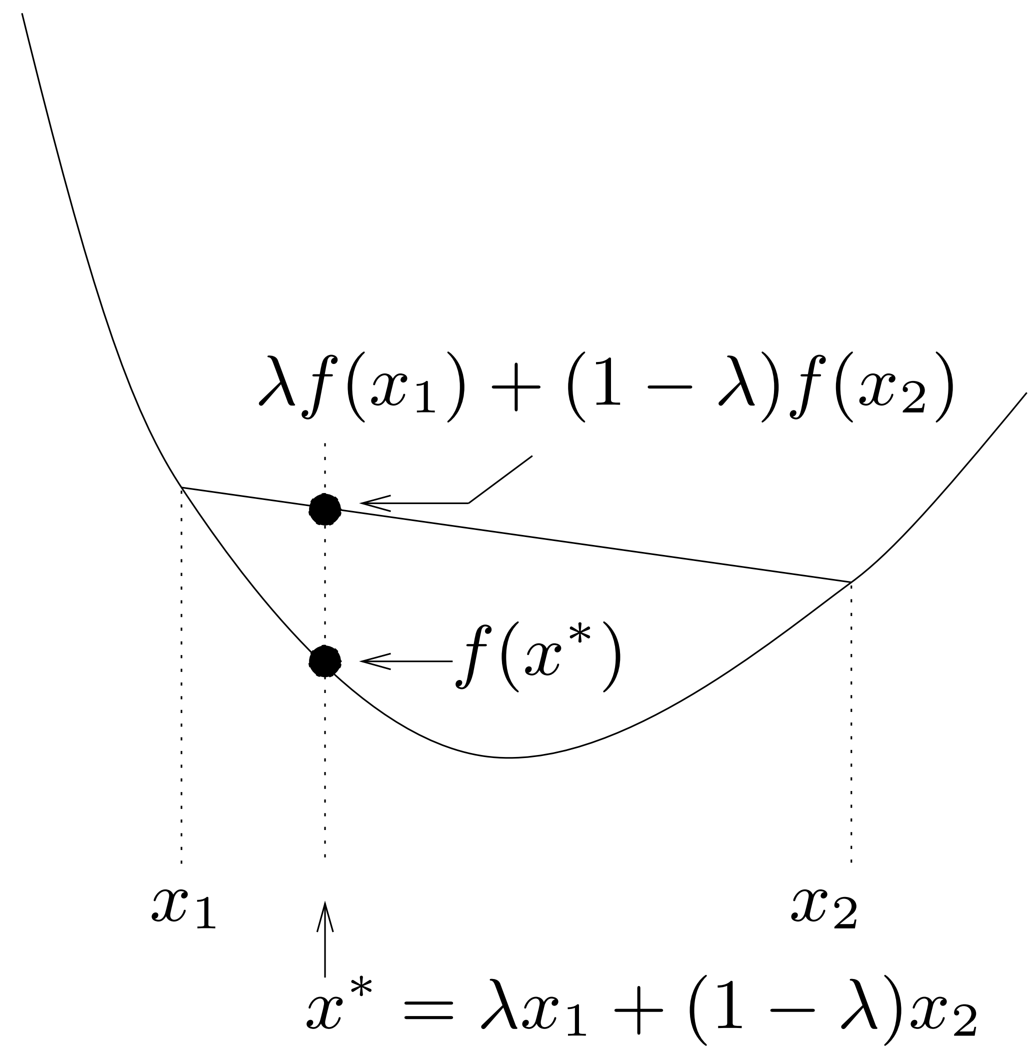
Jensen’s inequality for convex functions

Convex (and concave) functions

- **Convex ⌣ functions.** A function $f(x)$ **is convex ⌣ over (a, b)** if every chord of the function lies above the function, as shown in figure, that is, for all $x_1, x_2 \in (a, b)$ and $0 \le \lambda \le 1$,

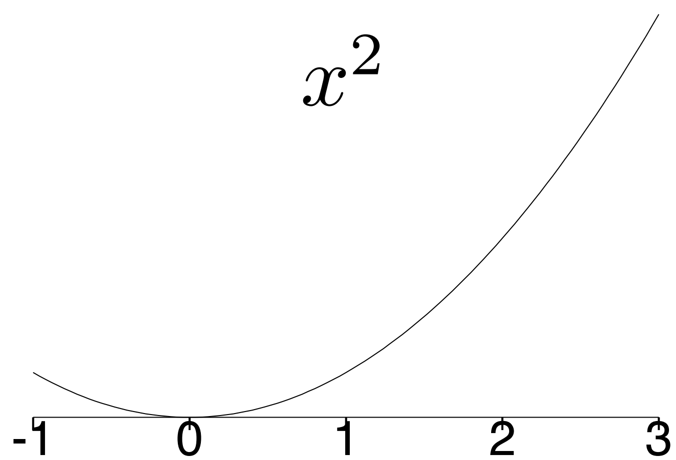
$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$$

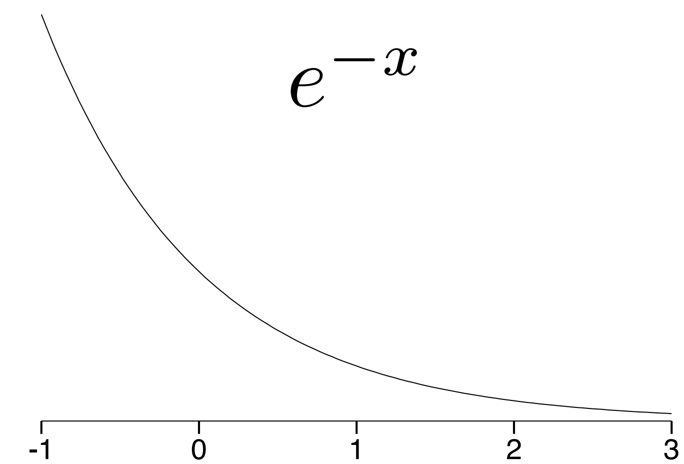
- A function is **strictly convex** ⌣ if, for all $x_1, x_2 \in (a, b)$ the equality holds only for $\lambda = 0$ and $\lambda = 1$.

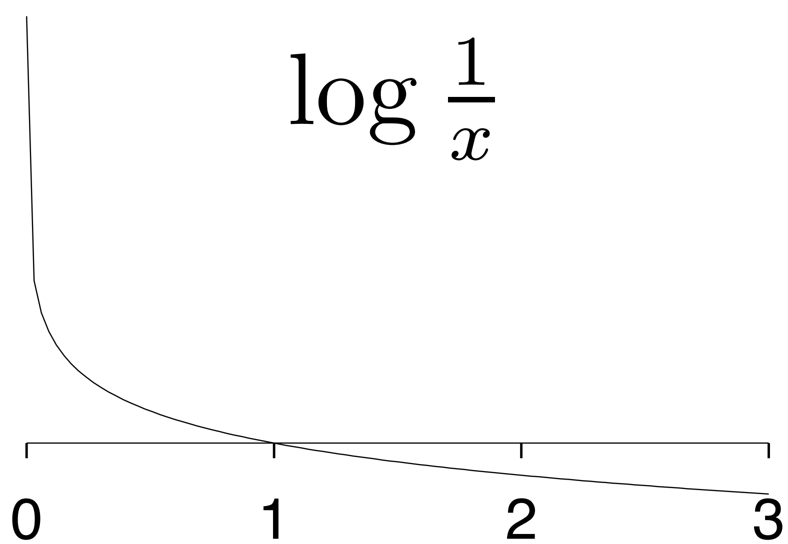


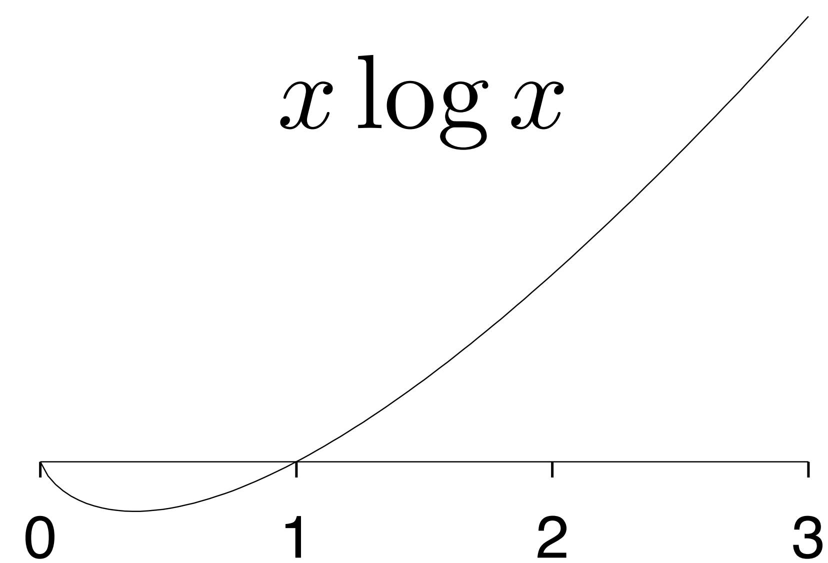
Similar definitions apply to concave ⌢ and strictly concave ⌢ functions.

Examples of convex functions









Jensen's inequality

- **Jensen's inequality.** If f is a convex $\smile$ function and x is a random variable then

$$\varepsilon\left[f(x)\right] \geq f(\varepsilon\left[x\right])$$

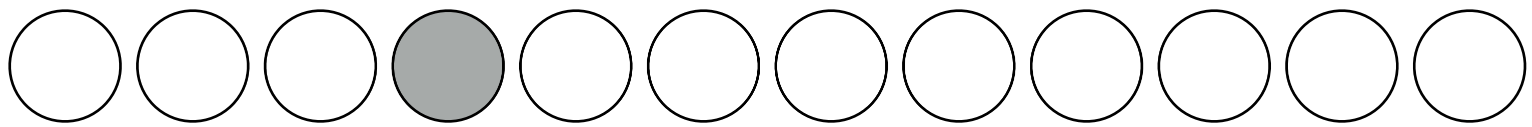
- If f is strictly convex $\smile$ and $\varepsilon\left[f(x)\right] = f(\varepsilon\left[x\right])$ then the random variable x is a constant.

- A Jensen's inequality can also be rewritten for a concave $\frown$ function, with the direction of the inequality reversed.

Designing informative experiments

The weighting problem

- You are given **12 balls**, all equal in weight except for **one that is either heavier or lighter**.



- **A two-pan balance to use**. In each use of the balance you may put any number of the 12 balls on the left pan, and the same number on the right pan.

there are three possible outcomes:

- the weights are equal,

- the balls on the left are heavier,

- the balls on the left are lighter

- Design a strategy to **determine which is the odd ball** and **whether it is heavier or lighter than the others** in **as few uses of the balance as possible**.

The weighting problem and the **measure of information**

- Consider the following questions:
  - How can one **measure *information***?
  - When you have identified the odd ball and whether it is heavy or light, how much **information have you gained**?
  - Once you have designed a strategy, draw a tree showing, for each of the possible outcomes of a weighing, what weighing you perform next. At each node in the tree, **how much information** have the outcomes **so far given you**, and **how much information remains to be gained**?

The weighting problem and the **measure of information**

- Consider the following questions (cont):
  - How much **information is gained** when you learn
    - the state of a flipped coin;
    - the states of two flipped coins;
    - the outcome when a four-sided die is rolled?
  - How much **information is gained** on the **first step of the weighing problem** if 6 balls are weighed against the other 6?
  - How much is gained if 4 are weighed against 4 on the first step, leaving out 4 balls?

The weighting problem: design a strategy

- What do you propose?
- Lets try to better understand the problem
  - What are the possible scenarios?
    - The odd ball is the ball n and is heavier or is lighter.
    - Let's say that $A_X = \{1^+, 2^+, \ldots, 12^+, 1^-, 2^-, \ldots, 12^-\}$

And all are equally probable

    - $|A_X| = 24$
- Lets try to better understand the available tool
  - **left heavier**: the odd ball is heavier and is on the left or the odd ball is lighter and is on the right
  - **right heavier**: the odd ball is lighter and is on the left or the odd ball is heavier and is on the right
  - **balanced**: the odd ball was not not the balance ! The ball is one not used in this measure

The weighting problem: design a strategy

1^+
2^+
3^+
4^+
5^+
6^+
7^+
8^+
9^+
10^+
11^+
12^+
1^-
2^-
3^-
4^-
5^-
6^-
7^-
8^-
9^-
10^-
11^-
12^-

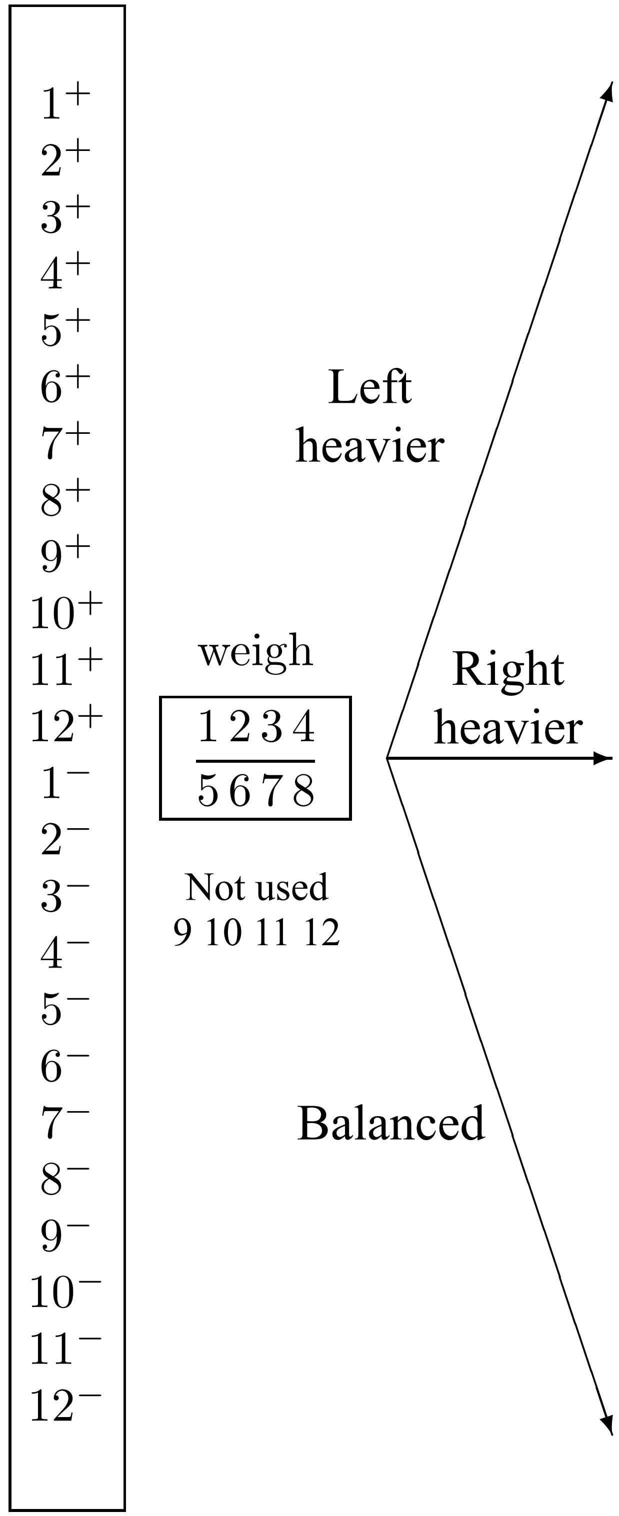
1^+
2^+
3^+
4^+
5^+
6^+
7^+
8^+
9^+
10^+
11^+
12^+
1^-
2^-
3^-
4^-
5^-
6^-
7^-
8^-
9^-
10^-
11^-
12^-

1^+
2^+
3^+
4^+
5^+
6^+
7^+
8^+
9^+
10^+
11^+
12^+
1^-
2^-
3^-
4^-
5^-
6^-
7^-
8^-
9^-
10^-
11^-
12^-

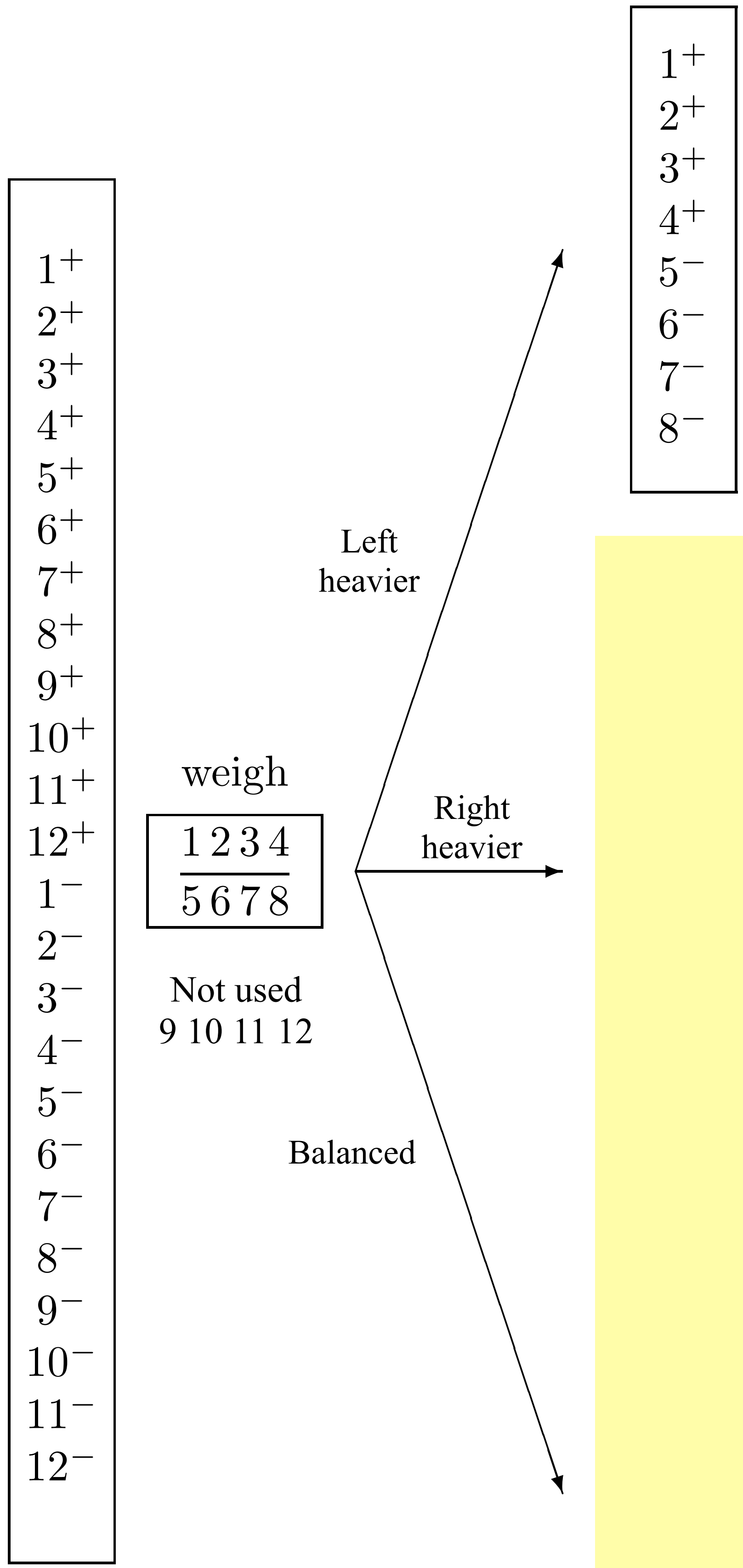
weigh

$$\frac{1\,2\,3\,4}{5\,6\,7\,8}$$

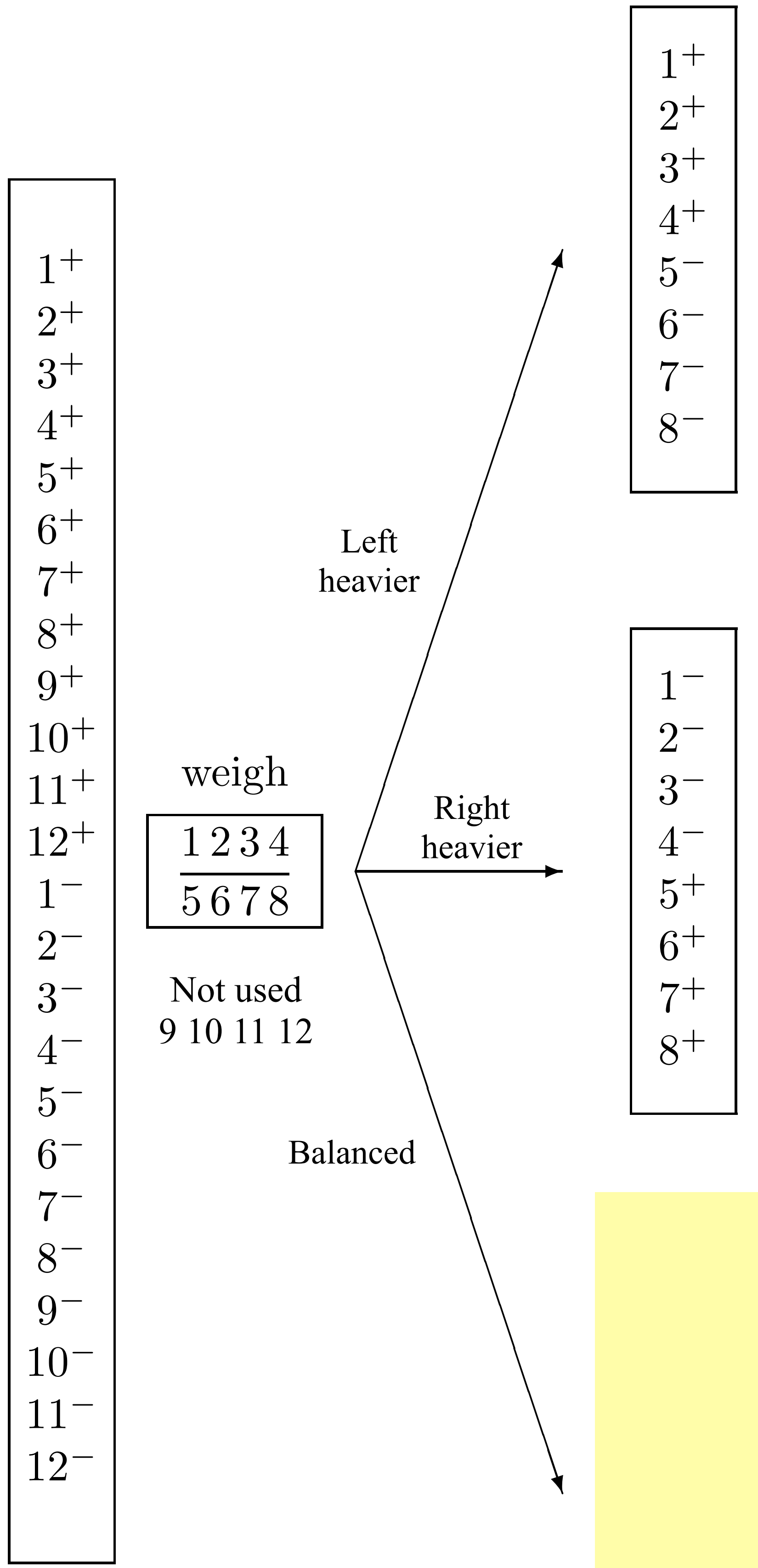
Not used
9 10 11 12



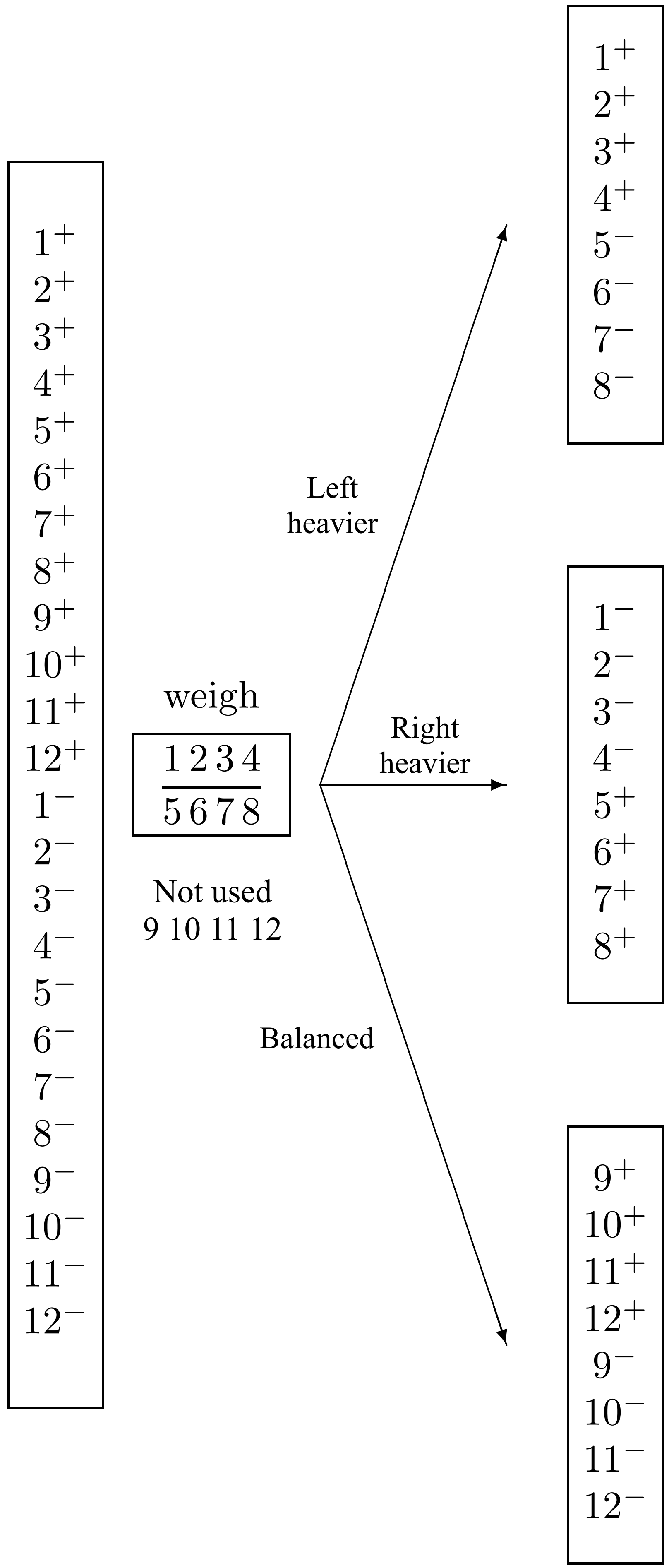
1+
2+
3+
4+
5+
6+
7+
8+
9+
10+
11+
12+
1−
2−
3−
4−
5−
6−
7−
8−
9−
10−
11−
12−
weigh
1 2 3 4
5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced



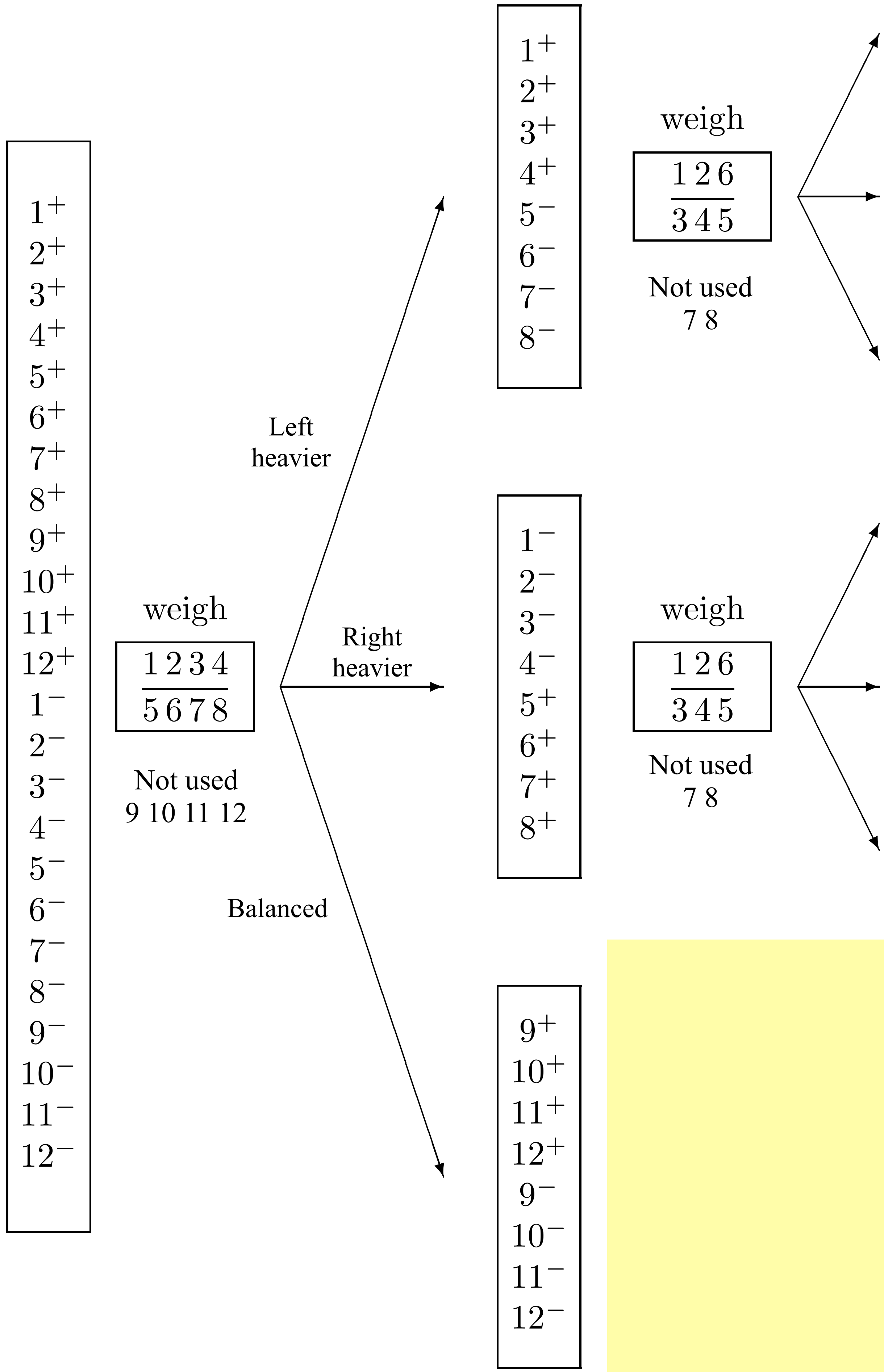
1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4
5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−

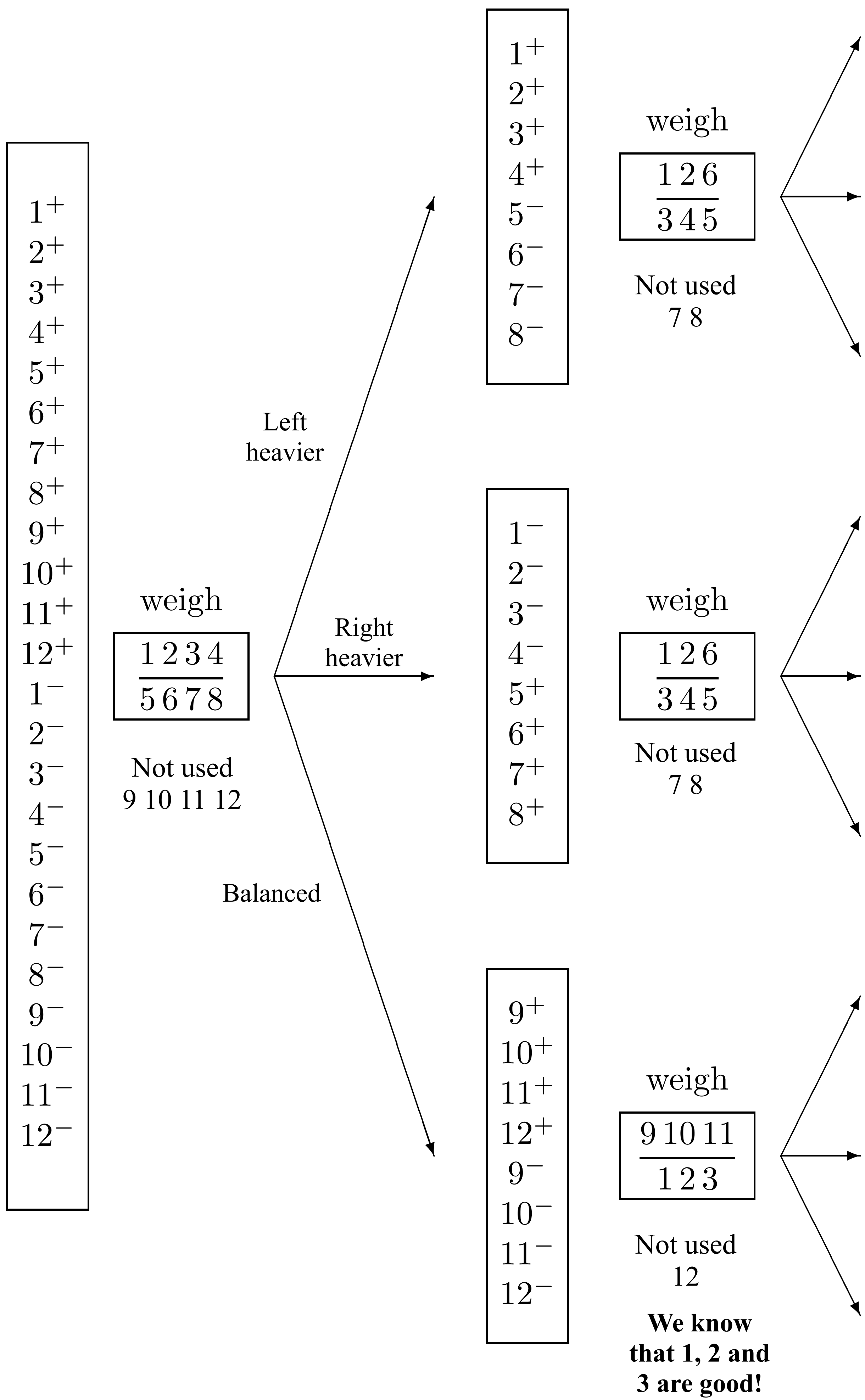


1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4
5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−
1− 2− 3− 4− 5+ 6+ 7+ 8+

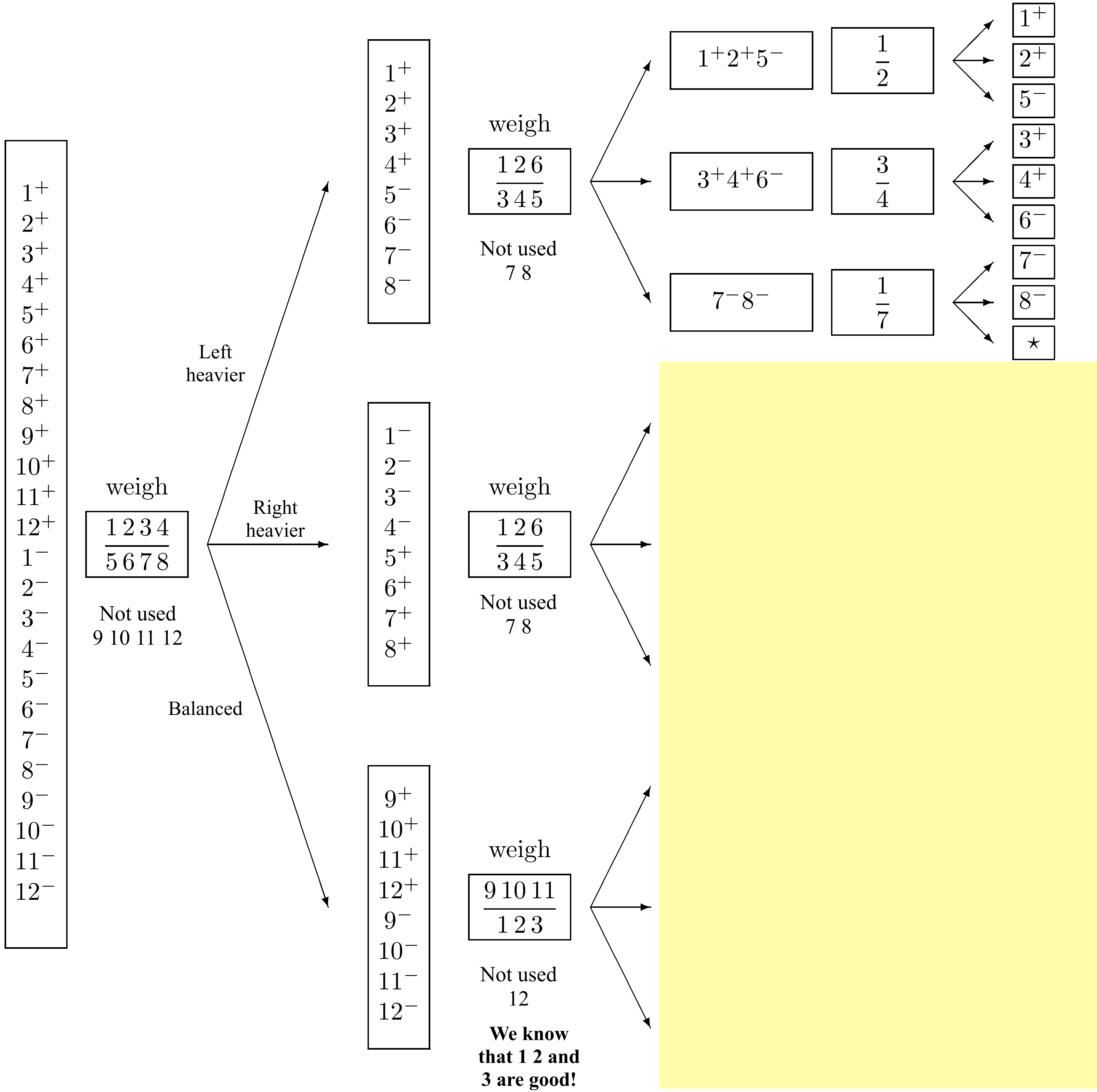


1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4
5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−
1− 2− 3− 4− 5+ 6+ 7+ 8+
9+ 10+ 11+ 12+ 9− 10− 11− 12−

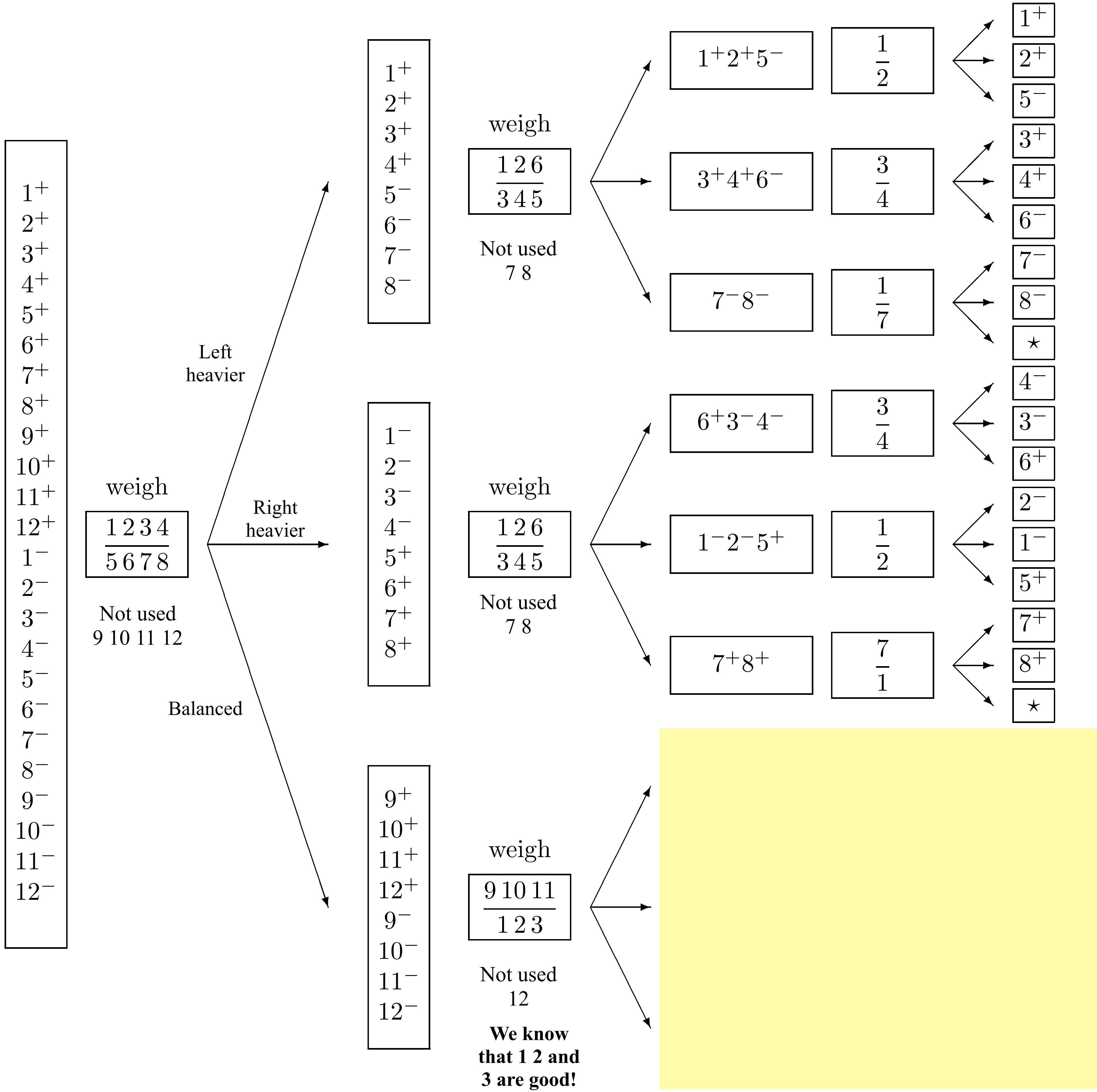

1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4 / 5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−
1− 2− 3− 4− 5+ 6+ 7+ 8+
9+ 10+ 11+ 12+ 9− 10− 11− 12−
weigh
1 2 6 / 3 4 5
Not used
7 8
weigh
1 2 6 / 3 4 5
Not used
7 8



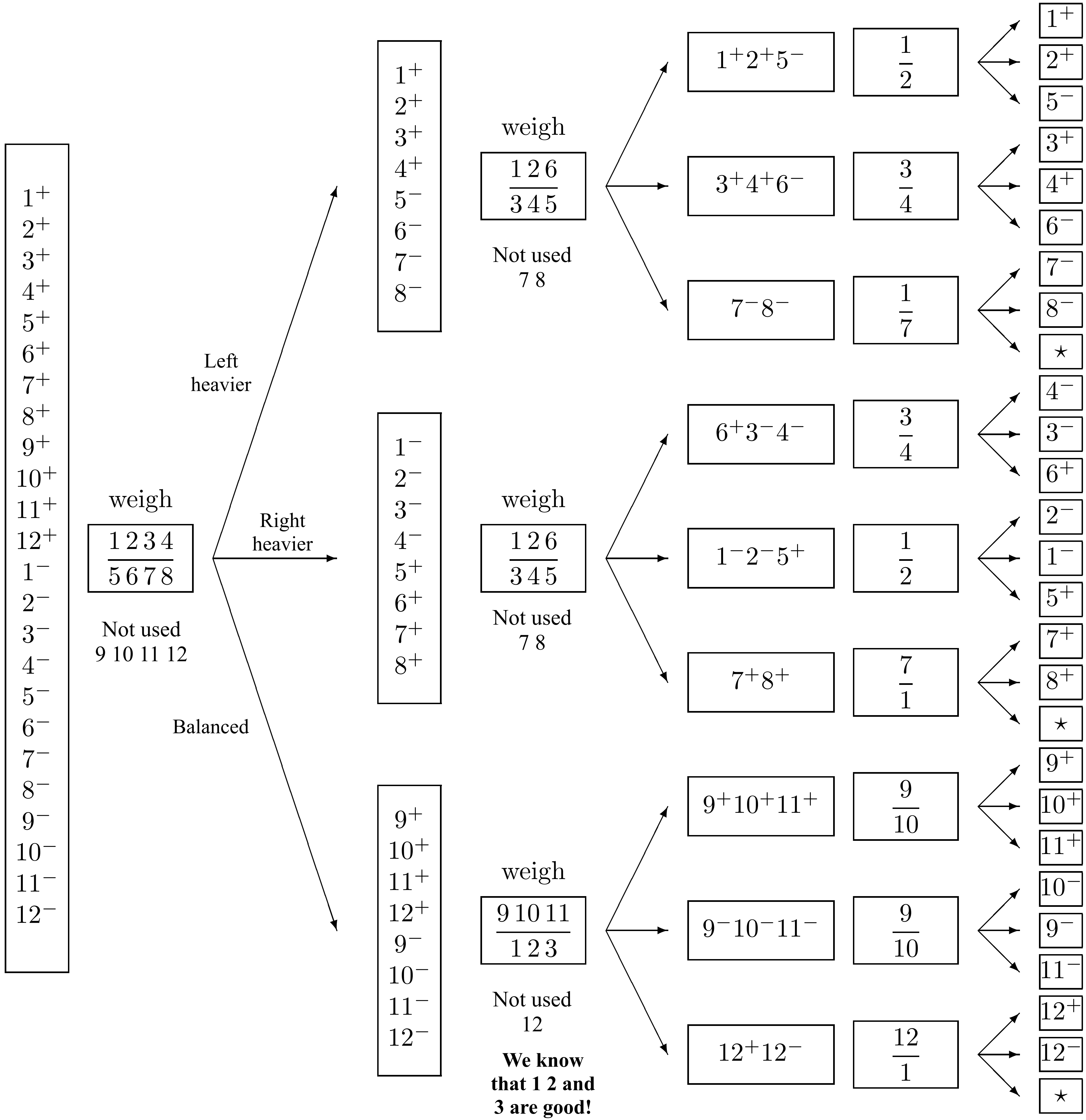
1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4
5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−
weigh
1 2 6
3 4 5
Not used
7 8
1− 2− 3− 4− 5+ 6+ 7+ 8+
weigh
1 2 6
3 4 5
Not used
7 8
9+ 10+ 11+ 12+ 9− 10− 11− 12−
weigh
9 10 11
1 2 3
Not used
12
We know that 1, 2 and 3 are good!



1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4 / 5 6 7 8
Not used 9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−
1− 2− 3− 4− 5+ 6+ 7+ 8+
9+ 10+ 11+ 12+ 9− 10− 11− 12−
weigh
1 2 6 / 3 4 5
Not used 7 8
weigh
1 2 6 / 3 4 5
Not used 7 8
weigh
9 10 11 / 1 2 3
Not used 12
We know that 1 2 and 3 are good!
1+2+5−
3+4+6−
7−8−
1 / 2
3 / 4
1 / 7
1+ 2+ 5− 3+ 4+ 6− 7− 8− ⋆



1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4
5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−
weigh
1 2 6
3 4 5
Not used
7 8
1+2+5−
3+4+6−
7−8−
1
2
3
4
1
7
1+
2+
5−
3+
4+
6−
7−
8−
⋆
1− 2− 3− 4− 5+ 6+ 7+ 8+
weigh
1 2 6
3 4 5
Not used
7 8
6+3−4−
1−2−5+
7+8+
3
4
1
2
7
1
4−
3−
6+
2−
1−
5+
7+
8+
⋆
9+ 10+ 11+ 12+ 9− 10− 11− 12−
weigh
9 10 11
1 2 3
Not used
12
We know that 1 2 and 3 are good!



1+ 2+ 3+ 4+ 5+ 6+ 7+ 8+ 9+ 10+ 11+ 12+ 1− 2− 3− 4− 5− 6− 7− 8− 9− 10− 11− 12−
weigh
1 2 3 4 / 5 6 7 8
Not used
9 10 11 12
Left heavier
Right heavier
Balanced
1+ 2+ 3+ 4+ 5− 6− 7− 8−
1− 2− 3− 4− 5+ 6+ 7+ 8+
9+ 10+ 11+ 12+ 9− 10− 11− 12−
weigh
1 2 6 / 3 4 5
Not used
7 8
weigh
1 2 6 / 3 4 5
Not used
7 8
weigh
9 10 11 / 1 2 3
Not used
12
We know that 1 2 and 3 are good!
1+2+5−
3+4+6−
7−8−
6+3−4−
1−2−5+
7+8+
9+10+11+
9−10−11−
12+12−
1 / 2
3 / 4
1 / 7
3 / 4
1 / 2
7 / 1
9 / 10
9 / 10
12 / 1
1+
2+
5−
3+
4+
6−
7−
8−
⋆
4−
3−
6+
2−
1−
5+
7+
8+
⋆
9+
10+
11+
10−
9−
11−
12+
12−
⋆

The weighting problem: some maths

- In the three uses of the balance – which reads either 'left heavier', 'right heavier', or 'balanced' – the **number of conceivable outcomes is $3^3 = 27$**,
- The **number of possible states of the world is 24**: the odd ball could be any of twelve balls, and it could be heavy or light
- So in principle, the **problem might be solvable in three weighings**
  - but not in two, since $3^2 < 24$.
- Why the strategy was optimal? What is it about your series of weighings that allows useful information to be gained as quickly as possible?
  - **At each step of an optimal procedure, the three outcomes** ('left heavier', 'right heavier', and 'balance') **are as close as possible to equiprobable**.

The weighting problem: some maths

- In the three uses of the balance – which reads either 'left heavier', 'right heavier', or 'balanced' – the **number of conceivable outcomes is $3^3 = 27$**,
- The **number of possible states of the world is 24**:
- **At each step of an optimal procedure, the three outcomes** ('left heavier', 'right heavier', and 'balance') **are as close as possible to equiprobable**.

- Strategies, such as weighing balls 1–6 against 7–12 on the first step, do not achieve all outcomes with equal probability: these two sets of balls can never balance, so the only possible outcomes are 'left heavy' and 'right heavy'.
  - Such a binary outcome rules out only half of the possible hypotheses, so a strategy that uses such outcomes must sometimes take longer to find the right answer.

The weighting problem: some maths

- In the three uses of the balance – which reads either ‘left heavier’, ‘right heavier’, or ‘balanced’ – the **number of conceivable outcomes is $3^3 = 27$**,
- The **number of possible states of the world is 24**:
- **At each step of an optimal procedure, the three outcomes** (‘left heavier’, ‘right heavier’, and ‘balance’) **are as close as possible to equiprobable**.

- An optimal strategy:
  - The first weighing must divide the 24 possible hypotheses into three groups of eight.
  - Then the second weighing must be chosen so that there is a 3:3:2 split of the hypotheses.

> the outcome of a random experiment is guaranteed to be most informative if the probability distribution over outcomes is uniform.

The Shannon information content of an outcome

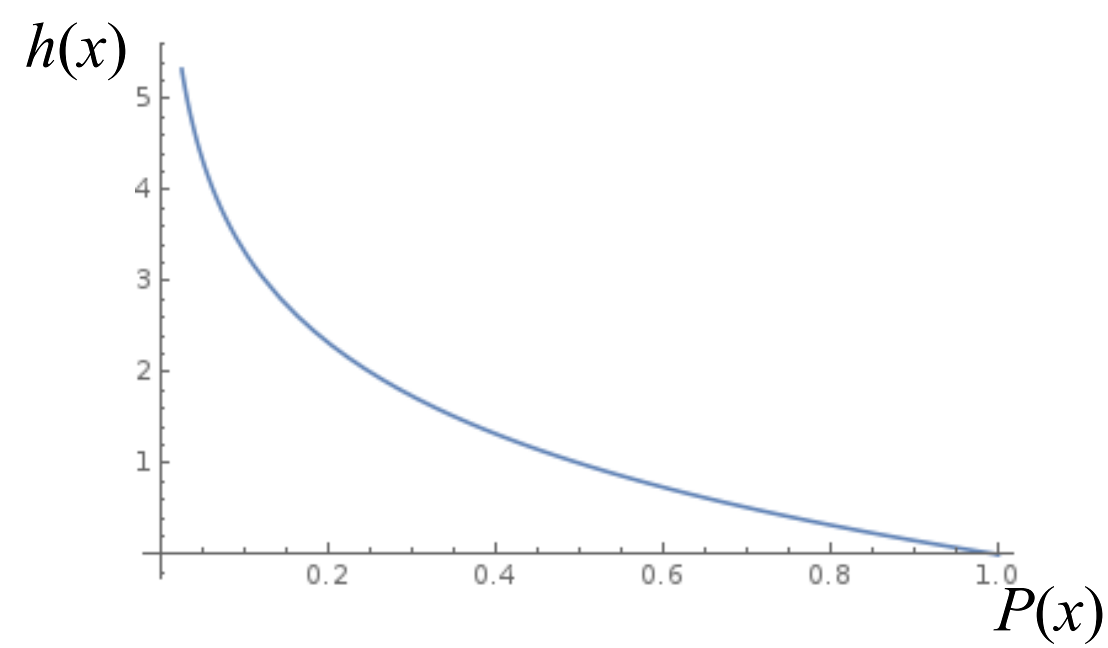
- The **Shannon information content** of an outcome x is defined to be

$$h(x) = \log_2 \frac{1}{P(x)} = -\log_2 P(x)$$

- It is measured in **bits**
  - The word bit is is also used to denote a variable whose value is 0 or 1 (**b**inary dig**it**)
- $h(a_i)$ is indeed a natural **measure of the information content** of the event $x = a_i$.
  - When a_i is almost certain ($P(ai)$ near to 1)

    the occurrence of a has a small information content
  - When a_i is very unlikely ($P(ai)$ near to 0)

    the occurrence of a has a large information content

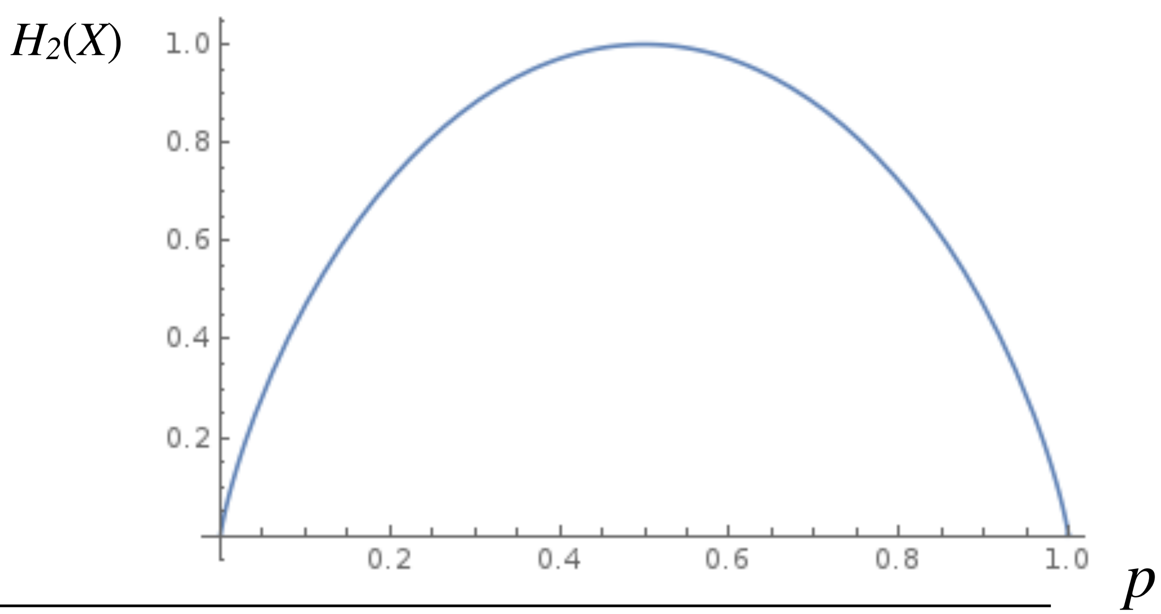


Entropy of an ensemble X

- The entropy of an ensemble X is defined to be the average Shannon information content of an outcome:

$$H(x) = \sum_{x \in A_X} P(x) \log_2 \frac{1}{P(x)} = -\sum_{x \in A_X} P(x) \log_2 P(x)$$

- $H(X) \geq 0$
  - $H(X) = 0$ if and only if $p_i = 1$ for one i.
- Entropy is **maximized** if ***p*** is uniform $H(X) \leq \log(|A_X|)$
  - $H(X) = \log(|A_X|)$ if and only if $p_i = \frac{1}{|A_X|}$ for all i
- Binary case, $H_2(X)$



Guessing Games

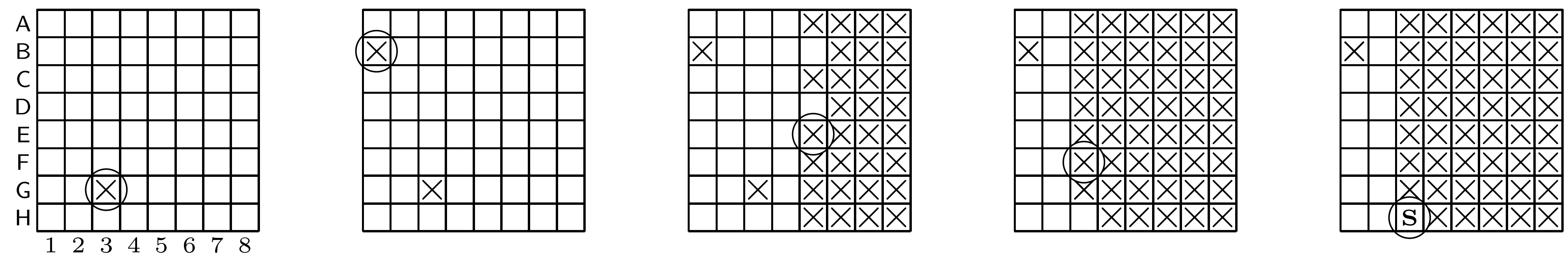
- Guess a **hidden number between 0 and 63** with a serie of questions that have an answer **yes/no**. How many questions are necessary to ensure that we discover the number?
- Intuitively, the best questions successively divide the 64 possibilities into equal sized sets.
- Six questions suffice: $2^6 = 64$

$$
\begin{aligned}
&1\text{: is } x \geq 32?\\
&2\text{: is } x \bmod 32 \geq 16?\\
&3\text{: is } x \bmod 16 \geq 8?\\
&4\text{: is } x \bmod 8 \geq 4?\\
&5\text{: is } x \bmod 4 \geq 2?\\
&6\text{: is } x \bmod 2 = 1?
\end{aligned}
$$

- Assuming that all values of x are equally likely, then the answers to the questions are independent and each has Shannon information content $\log_2(1/0.5) = 1$bit

The game of submarine: how many bits can one bit convey?

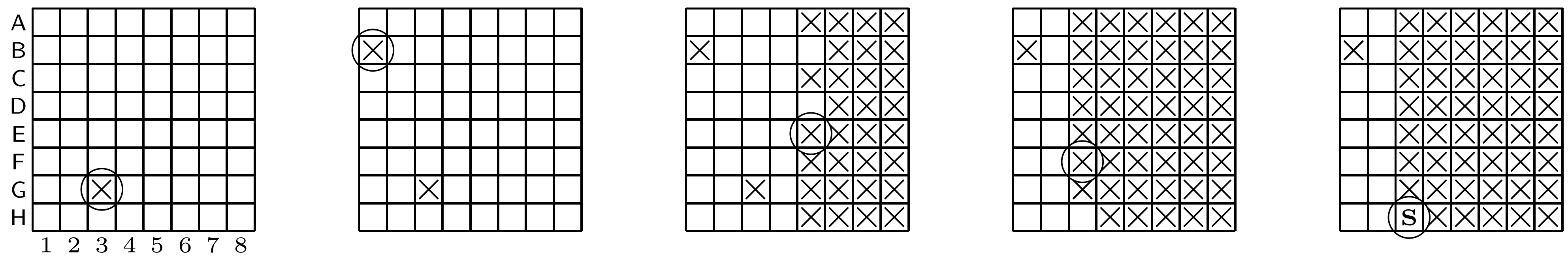
- In a simplified version of battleships called **submarine**, each player **hides just one submarine** in one square of an eight-by-eight grid.



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$

- The circle represents the square that is being fired at,
  - The X show the squares in which the outcome was a miss, $x = n$;
  - The submarine is hit (outcome $x =$ y shown by the symbol s)

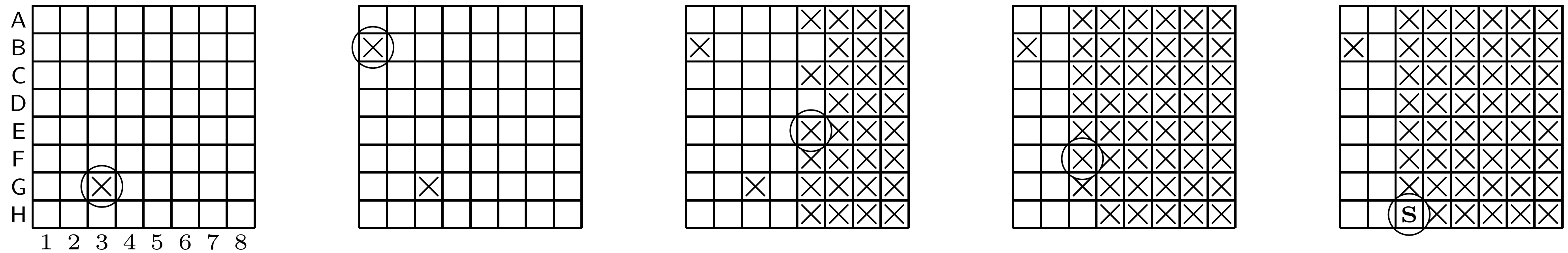
The game of submarine: how many bits can one bit convey?



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$

- Each shot made by a player defines an ensemble.
  - The two possible outcomes are $\{y, n\}$.
  - Their probabilities depend on the state of the board.

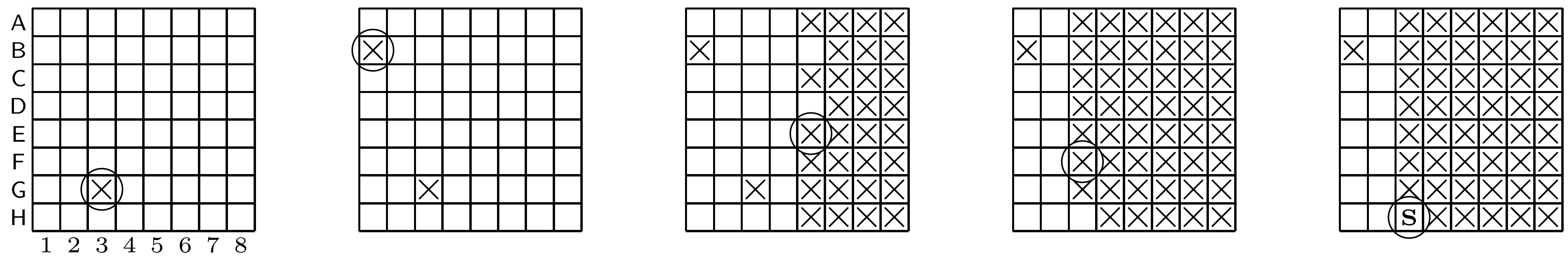
The game of submarine: how many bits can one bit convey?



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$

- Each shot made by a player defines an ensemble.
  - The two possible outcomes are $\{y, n\}$.
  - Their probabilities depend on the state of the board.
- At the beginning, $P(y) = 1/64$ and $P(n) = 63/64$.
- At the second shot, <u>if the first shot missed</u>, $P(y) = 1/63$ and $P(n) = 62/63$.

The game of submarine: how many bits can one bit convey?



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$

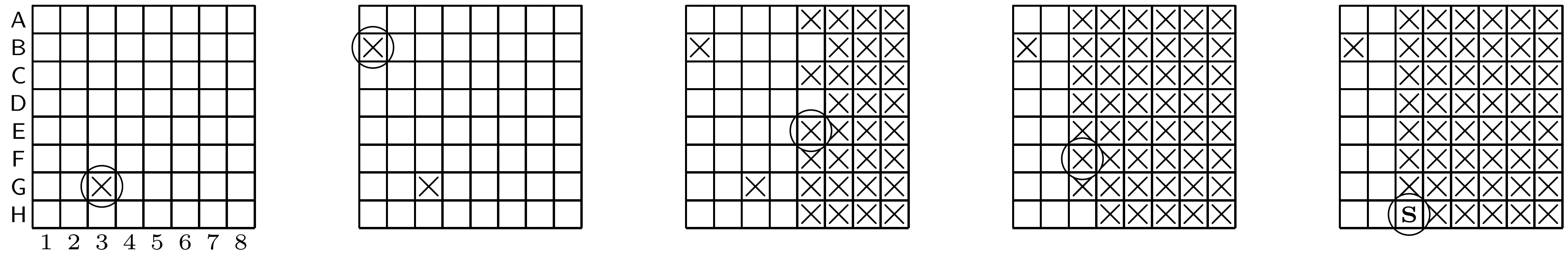
- The Shannon information gained from an outcome x is $h(x) = \log(1/P(x))$.
  - If we are lucky, and hit the submarine on the first shot, then

$$h(x) = h_{(1)}(\mathtt{y}) = \log_2 64 = 6 \text{ bits.} \qquad !!!$$

  - If we miss the shot, then

$$h(x) = h_{(1)}(\mathtt{n}) = \log_2 \frac{64}{63} = 0.0227 \text{ bits.}$$

The game of submarine: how many bits can one bit convey?

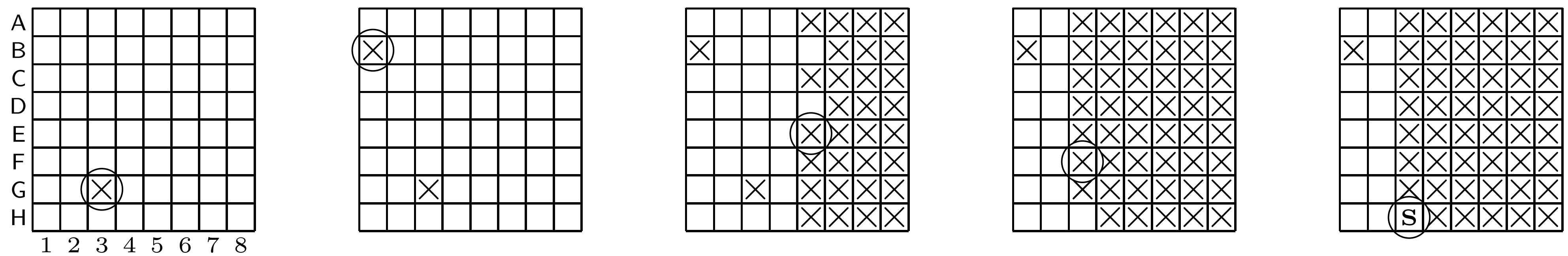


move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

- If we miss thirty-two times (firing at a new square each time), the total Shannon information gained is

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \cdots + \log_2 \frac{33}{32}$$
$$= 0.0227 + 0.0230 + \cdots + 0.0430 \;=\; 1.0\,\text{bits.} \qquad \text{Why?}$$

The game of submarine: how many bits can one bit convey?



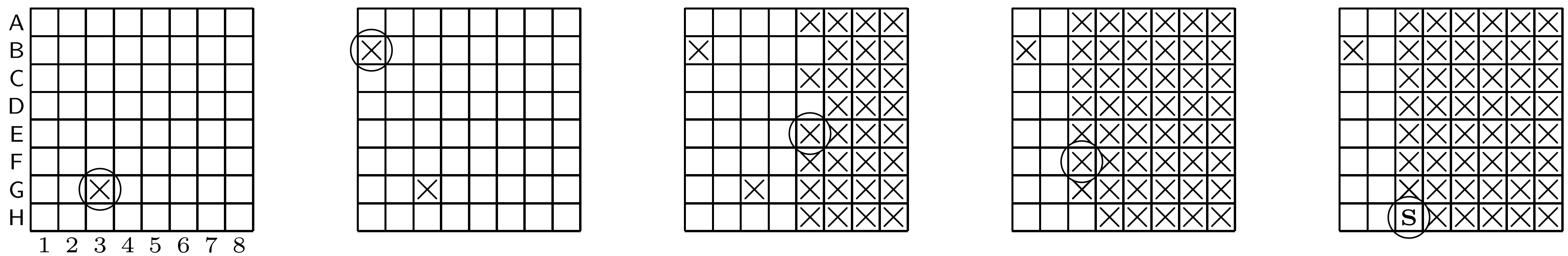
move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0 Why?	6.0

- If we miss thirty-two times (firing at a new square each time), the total Shannon information gained is

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \cdots + \log_2 \frac{33}{32}$$
$$= 0.0227 + 0.0230 + \cdots + 0.0430 \quad = \quad 1.0\,\text{bits}.$$

FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

The game of submarine: how many bits can one bit convey?

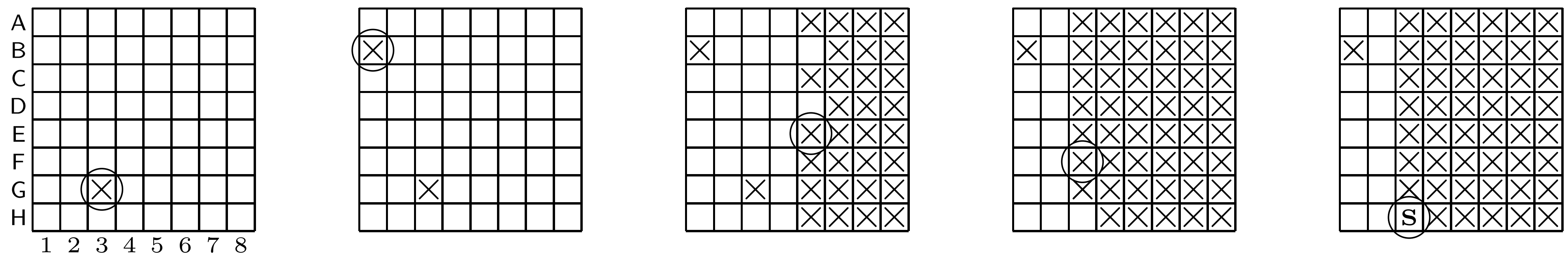


move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

- What if we hit the submarine on the 49th shot, when there were **16 squares left**? The Shannon information content of this outcome is

$$h_{(49)}(\mathtt{y}) = \log_2 16 = 4.0 \text{ bits}.$$

The game of submarine: how many bits can one bit convey?



move #	1	2	32	48	49
question	G3	B1	E5	F3	H3
outcome	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{n}$	$x = \mathtt{y}$
$P(x)$	$\frac{63}{64}$	$\frac{62}{63}$	$\frac{32}{33}$	$\frac{16}{17}$	$\frac{1}{16}$
$h(x)$	0.0227	0.0230	0.0443	0.0874	4.0
Total info.	0.0227	0.0458	1.0	2.0	6.0

- The total Shannon information content of all the outcomes is

$$\log_2 \frac{64}{63} + \log_2 \frac{63}{62} + \cdots + \log_2 \frac{17}{16} + \log_2 \frac{16}{1}$$
$$= \quad 0.0227 + 0.0230 + \cdots + 0.0874 + 4.0 \quad = \quad 6.0\,\text{bits}.$$

Further Reading and Summary

Q&A

Further Reading

- **Recommend Readings**
  - Information Theory, Inference, and Learning Algorithms from David MacKay, 2015, pages 32 - 36.

What you should know

- The definition and the meaning of Shannon information content
- The diference between Binary Digit and Bit as unit of Shannon information content
- The definition and the meaning of Entropy
- Understand the equation 0 ≤ Entropy ≤ log cardinality. In which conditions the equalities arise.
- The joint entropy of two independent ensembles
- Decomposability of the entropy. How to use
- The relative Entropy (or Kullback–Leibler divergence)
- Gibbs' inequality
- Jensen's inequality for convex functions. How to use
- How to think to Design informative experiments.

Further Reading and Summary

Q&A