Week 3: Conditional Entropy (Book Chapter 4)

Prof. Michael Gastpar

Slides by Prof. M. Gastpar and Prof. em. B. Rimoldi

EPFL

Spring Semester 2025

LAST WEEK

MAITRE CORBEAU SUR UN

A	
A	1100
B	⋮
C	
D	
E	101
⋮	

$$nH(X)$$

$$nH(X_1, X_2)$$

Letter	Prob.
A	0.0811
B	0.0081
C	0.0338
D	0.0428
E	0.1769
F	0.0113
G	0.0119
H	0.0074
I	0.0724
J	0.0018
K	0.0002
L	0.0599
M	0.0229
N	0.0768
O	0.0520
P	0.0292
Q	0.0083
R	0.0643
S	0.0887
T	0.0744
U	0.0523
V	0.0128
W	0.0006
X	0.0053
Y	0.0026
Z	0.0012

$$H(X) \leq L(X, \Gamma) < H(X) + 1$$

Letter	Prob.
A	0.0811
B	0.0081
C	0.0338
D	0.0428
E	0.1769
F	0.0113
G	0.0119
H	0.0074
I	0.0724
J	0.0018
K	0.0002
L	0.0599
M	0.0229
N	0.0768
O	0.0520
P	0.0292
Q	0.0083
R	0.0643
S	0.0887
T	0.0744
U	0.0523
V	0.0128
W	0.0006
X	0.0053
Y	0.0026
Z	0.0012

LAST WEEK

MAITRE CORBEAU SUR UN ...

$A \times A$	
AA	$\frac{1}{2} H(X_1, X_2)$
AB	$\leq$
AC	$\frac{1}{2} L((X_1 X_2), 17)$
⋮	$\leq$
AZ	$\frac{1}{2} H(X_1, X_2) + 1/2$
Au	
A!	
⋮	
BA	

Outline

Introduction and Organization

Entropy and Data Compression

Probability Review

Sources and Entropy

The Fundamental Compression Theorem: The IID Case

Conditional Entropy

Entropy and Algorithms

Prediction, Learning, and Cross-Entropy Loss

Summary of Chapter 1

Cryptography

Channel Coding

Key Idea

- **Pack multiple symbols into "supersymbols"!**
- $(S_1, S_2, S_3, \ldots, S_n)$
- Now, apply our Main Result to such supersymbols:

Theorem (Textbook Thm 3.3)

The average codeword-length of a uniquely decodable code Γ for S must satisfy

$$H_D(S_1, S_2, \ldots, S_n) \leq L((S_1, S_2, \ldots, S_n), \Gamma)$$

and there exists a uniquely decodable code Γ_{SF} satisfying

$$L((S_1, S_2, \ldots, S_n), \Gamma_{SF}) < H_D(S_1, S_2, \ldots, S_n) + 1.$$

- Why is this clever?
- Let us study the entropy of the supersymbol $H_D(S_1, S_2, \ldots, S_n)$ next.

Our Next Nugget

- Understand the behavior of

$$H_D(S_1, S_2, \ldots, S_n)$$

when $S_1, S_2, \ldots, S_n$ are not independent random variables following the same distribution.

Key steps to get there:

- Understand **conditional entropy**
- Understand how to model "many" random variables (a.k.a. **random processes**)

Our Next Nugget

Example: Standard text.

- After a letter “q”, we have a letter “u” with very high probability (probability 1 in some languages).
- After a letter “c”, we have a letter “h” with higher probability than many other letters.
- After a letter “i”, it is extremely unlikely to have yet another letter “i”. And so on.

Our Next Nugget

Example: Audio recoding.

- Why?

Example: Image.

Example: Video recording.

KEY (SIMPLE) EXAMPLE 1 : INDEPENDENT

DEFINITION (COIN-FLIP SOURCE)

The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

KEY (SIMPLE) EXAMPLE 1 : INDEPENDENT

DEFINITION (COIN-FLIP SOURCE)

The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

KEY (SIMPLE) EXAMPLE 1 : INDEPENDENT

DEFINITION (COIN-FLIP SOURCE)

The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

Key (Simple) Example 1 : Independent

Definition (Coin-Flip Source)

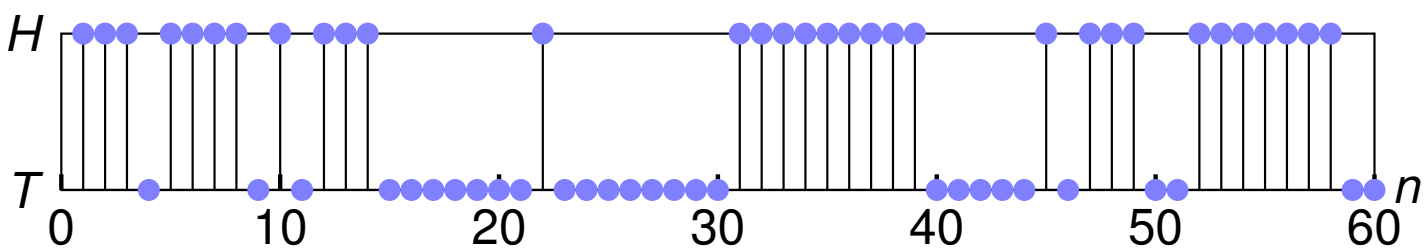
The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$



Key (Simple) Example 2 : Not Independent

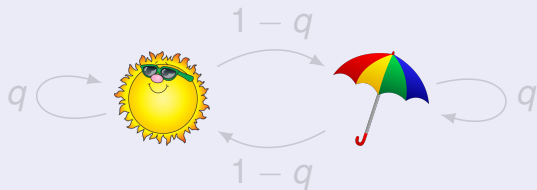
Definition (Sunny-Rainy Source)

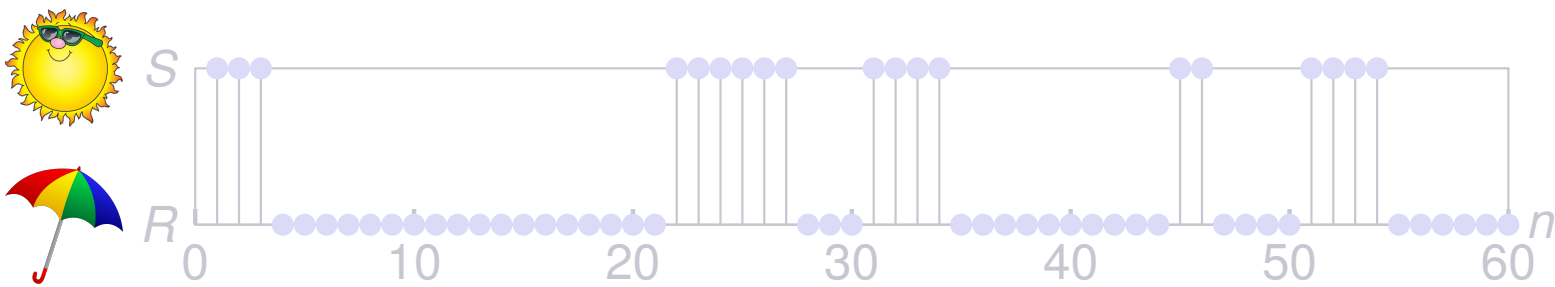
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.

For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.





Key (Simple) Example 2 : Not Independent

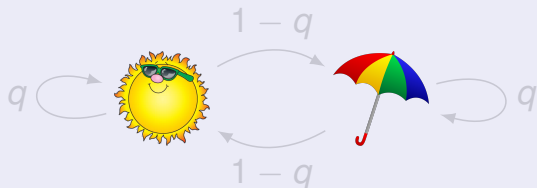
Definition (Sunny-Rainy Source)

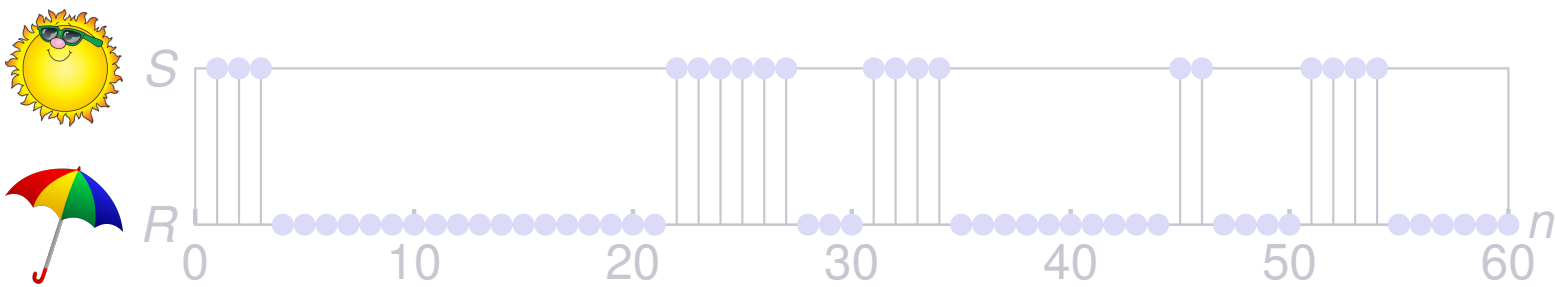
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.

For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.





Key (Simple) Example 2 : Not Independent

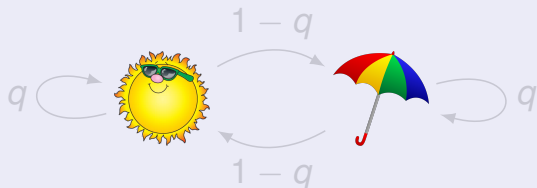
Definition (Sunny-Rainy Source)

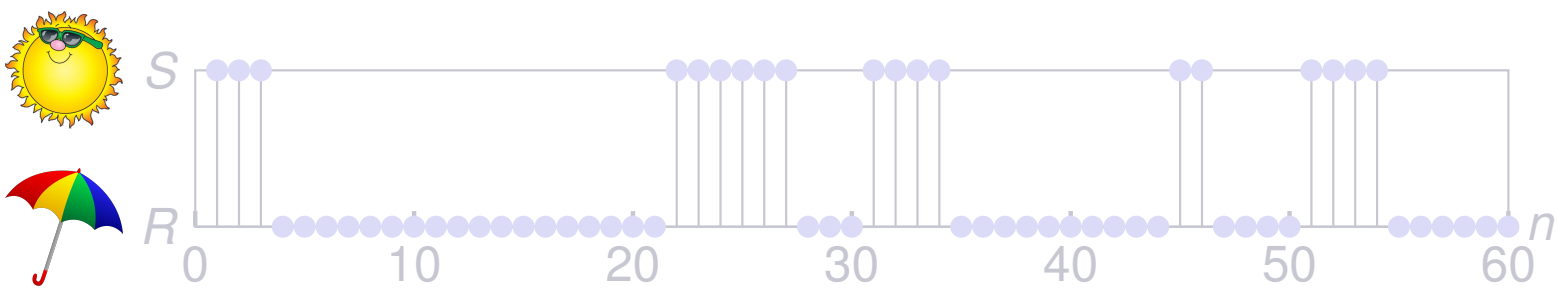
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.

For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.





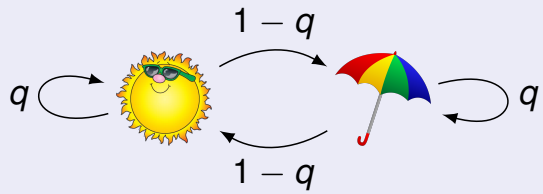
Key (Simple) Example 2 : Not Independent

Definition (Sunny-Rainy Source)

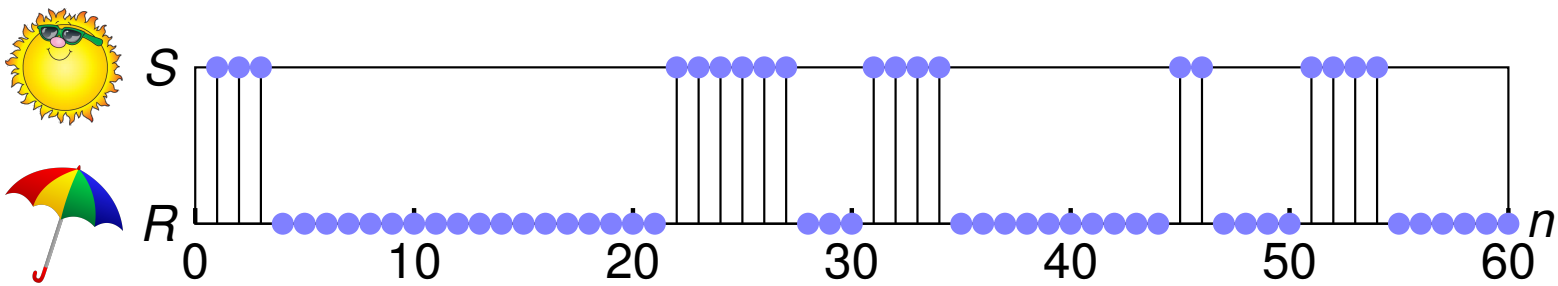
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.



For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.



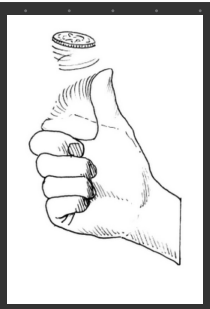
Recall how to determine the conditional probability:

$$p_{X|Y}(x|y) \stackrel{def}{=} \frac{p_{X,Y}(x,y)}{p_Y(y)}.$$

It gives the probability of the event $X = x$, given that the event $Y = y$ has occurred.

It is defined for all y for which $p_Y(y) > 0$.

TWO CARD DECKS



Y: FAIR COIN TOSS

IF HEAD

IF TAIL

DECK A

18 RED
18 BLACK

DECK B

36 BLACK

DRAW A SINGLE CARD

$$X = \begin{cases} 0, & \text{IF CARD IS BLACK} \\ 1, & \text{IF CARD IS RED} \end{cases}$$

$$P(X=0 \mid Y \text{ is HEAD}) = 1/2$$

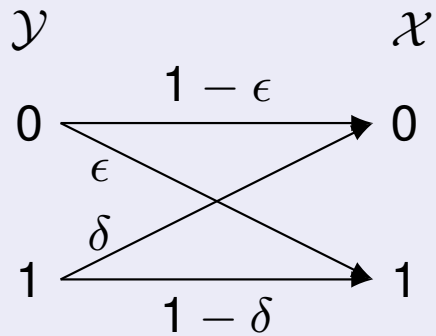
$$\Rightarrow P(X=1 \mid Y \text{ is HEAD}) = 1/2$$

$$P(X=0 \mid Y \text{ is TAIL}) = 1$$

$$\Rightarrow P(X=1 \mid Y \text{ is TAIL}) = 0.$$

$$P(X=0) = 3/4$$

EXAMPLE ("BIT FLIPPER CHANNEL")



$\Longleftrightarrow$

$p(x\|y)$	$y=0$	$y=1$
$x=0$	$1-\epsilon$	δ
$x=1$	ϵ	$1-\delta$

Suppose Y is uniformly distributed. Then, the joint distribution of X, Y is

$p(x,y)$	$y=0$	$y=1$
$x=0$	$\frac{1}{2}(1-\epsilon)$	$\frac{1}{2}\delta$
$x=1$	$\frac{1}{2}\epsilon$	$\frac{1}{2}(1-\delta)$

EXERCISE ("BIT FLIPPER CHANNEL")

As we have seen, for the bit flipper channel with uniform input Y, the joint distribution of X, Y is

$p(x,y)$	$y=0$	$y=1$
$x=0$	$\frac{1}{2}(1-\epsilon)$	$\frac{1}{2}\delta$
$x=1$	$\frac{1}{2}\epsilon$	$\frac{1}{2}(1-\delta)$

- Find the conditional distribution $p(y|x)$ (input given the output).

SOLUTION ("BIT FLIPPER CHANNEL")

The general formula is

$$p(y|x) = \frac{p(x,y)}{p(x)}.$$

Hence, we need the marginal distribution of X :

$p(x,y)$	$y=0$	$y=1$	Marginal distribution $p(x)$
$x=0$	$\frac{1}{2}(1-\epsilon)$	$\frac{1}{2}\delta$	$\frac{1}{2}(1-\epsilon)+\frac{1}{2}\delta$
$x=1$	$\frac{1}{2}\epsilon$	$\frac{1}{2}(1-\delta)$	$\frac{1}{2}\epsilon+\frac{1}{2}(1-\delta)$

Hence, we find the desired object:

$p(y\|x)$	$y=0$	$y=1$
$x=0$	$\frac{1-\epsilon}{1-\epsilon+\delta}$	$\frac{\delta}{1-\epsilon+\delta}$
$x=1$	$\frac{\epsilon}{1-\delta+\epsilon}$	$\frac{1-\delta}{1-\delta+\epsilon}$

- Convince yourself that indeed, $p(y|x)$ is a valid probability distribution for each fixed value of x.

Conditional expectation of X given $Y = y$

$p_{X|Y}(\cdot|y)$ is a probability distribution on the alphabet of X, just like $p_X(\cdot)$

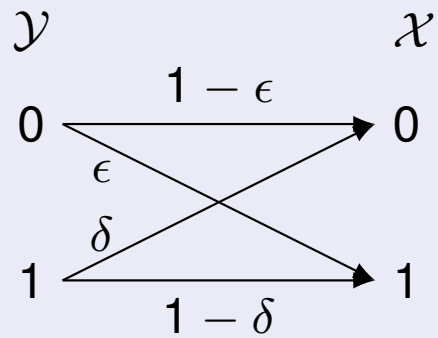
Definition

The conditional expectation of X given $Y = y$ is defined as

$$\mathbb{E}[X|Y = y] \stackrel{def}{=} \sum_{x \in \mathcal{X}} x p_{X|Y}(x|y).$$

Conditional expectation of X given $Y = y$

Exercise ("Bit flipper channel")



$\Longleftrightarrow$

$p(x\|y)$	$y=0$	$y=1$
$x=0$	$1-\epsilon$	δ
$x=1$	ϵ	$1-\delta$

- Find the conditional expectations $\mathbb{E}[X|Y=y]$ for $y=0$ and for $y=1$.

$$\mathbb{E}[X|Y=0] = 0 \cdot p(x=0|y=0) + 1\, p(x=1|y=0)$$
$$= \epsilon$$
$$\mathbb{E}[X|Y=1] = 1-\delta.$$

Conditional entropy of X given $Y = y$

$p_{X|Y}(\cdot|y)$ is a probability distribution on the alphabet of X, just like $p_X(\cdot)$

Every probability distribution has an entropy associated to it:

- $p_X(\cdot) \longrightarrow H(X)$
- $p_{X|Y}(\cdot|y) \longrightarrow H(X|Y = y)$

Definition

The conditional entropy of X given $Y = y$ is defined as

$$H_D(X|Y = y) \stackrel{def}{=} -\sum_{x \in \mathcal{X}} p_{X|Y}(x|y) \log_D p_{X|Y}(x|y).$$

Conditional entropy of X given $Y = y$

Exercise ("Bit flipper channel")

For the Bit flipper channel with uniform input, calculate:

- $H(X|Y = y)$ for each fixed y,
- $H(Y|X = x)$ for each fixed x.

Solution

TWO CARD DECKS

$H(X|Y \text{ IS HEAD}) = 1$ BIT.

$H(X|Y \text{ IS TAIL}) = 0$

$H(X) = h(1/4) = h(3/4)$

NOTE : HERE (SEE WEEK 1)

$h(p)$ IS THE BINARY ENTROPY FUNCTION

$$h(p) = -p \log p - (1-p) \log (1-p)$$

TWO CARD DECKS

$H(X \mid Y \text{ IS HEAD}) = 1$ BIT.

$H(X \mid Y \text{ IS TAIL}) = 0$

$$H(X) = h(1/4) = h(3/4)$$

$$H(X|Y) = p(Y \text{ IS HEAD}) H(X|Y \text{ IS HEAD}) + p(Y \text{ IS TAIL}) H(X|Y \text{ IS TAIL})$$

$$= 1/2 \cdot 1 + 1/2 \cdot 0 = \underline{1/2}$$

THEOREM (BOUNDS ON CONDITIONAL ENTROPY OF X GIVEN $Y = y$)

The conditional entropy of a discrete random variable $X \in \mathcal{X}$ conditioned on $Y = y$ satisfies

$$0 \leq H_D(X|Y = y) \leq \log_D |\mathcal{X}|,$$

with equality on the left iff $p_{X|Y}(x|y) = 1$ for some x, and with equality on the right iff $p_{X|Y}(x|y) = \frac{1}{|\mathcal{X}|}$ for all x.

The proof is identical to our proof of the basic entropy bounds.

EXAMPLE ("BIT FLIPPER CHANNEL")

For the Bit flipper channel, verify the entropy bounds.

Question: Do we also have the following entropy bound:

$$H_D(X|Y=y) \overset{???}{\leq} H_D(X)?$$

Answer: No!

EXAMPLE (BIT FLIPPER WITH UNIFORM INPUT Y)

(Or "counterexample," if you prefer). Just for ease of calculation, let us set $\delta = 0$ (but this is not necessary for the example to work!). Then, we have:

$$H_D(X|Y=0) = h_D(\epsilon) \quad \text{and} \quad H_D(X|Y=1) = 0.$$

where $h_D(\cdot)$ is the binary entropy function (with $\log_D(\cdot)$). But we have

$$H_D(X) = h_D\left(\frac{1-\epsilon}{2}\right).$$

(Set, for example, $\epsilon = 3/8$, thus $\frac{1-\epsilon}{2} = 5/16$.)

Conditional entropy of X given Y

The most useful and impactful definition is the *average* conditional entropy of X given $Y = y$, averaged over all values of y under the marginal distribution $p_Y(y)$. Formally, we thus define:

Definition

The conditional entropy of X given Y is defined as

$$H_D(X|Y) \stackrel{def}{=} \sum_{y \in \mathcal{Y}} p_Y(y) \left(- \sum_{x \in \mathcal{X}} p_{X|Y}(x|y) \log_D p_{X|Y}(x|y) \right).$$

EXAMPLE ("BIT FLIPPER CHANNEL")

For the Bit flipper channel, we have

$$H_D(X|Y) = p(Y=0)H_D(X|Y=0) + p(Y=1)H_D(X|Y=1).$$

We have already calculated

$$H_D(X|Y=0) = h_D(\epsilon) \quad \text{and} \quad H_D(X|Y=1) = h_D(\delta).$$

For example, when Y is uniform, we have

$$H_D(X|Y) = \frac{h_D(\epsilon) + h_D(\delta)}{2}.$$

THEOREM (BOUNDS ON CONDITIONAL ENTROPY OF X GIVEN Y)

The conditional entropy of a discrete random variable $X \in \mathcal{X}$ conditioned on Y satisfies

$$0 \leq H_D(X|Y) \leq \log_D |\mathcal{X}|,$$

with equality on the left iff for every y there exists and x such that $p_{X|Y}(x|y) = 1$, and with equality on the right iff $p_{X|Y}(x|y) = \frac{1}{|\mathcal{X}|}$ for all x and all y.

This follows directly from our bounds on $H_D(X|Y = y)$.

Note: Having $p_{X|Y}(x|y) = \frac{1}{|\mathcal{X}|}$ for all x and all y implies that X and Y are independent random variables.

SUPPOSE WE HAVE:

$$p(x|y) = \frac{1}{|\mathcal{X}|} \quad \text{for all } y$$

THEN:

$$p(x) = \sum_{y \in \mathcal{Y}} p(y)\, p(x|y)$$

$$= \sum_{y} p(y)\, \frac{1}{|\mathcal{X}|}$$

$$= \frac{1}{|\mathcal{X}|} \sum_{y} p(y)$$

$$= 1$$

HENCE, IN THIS EXAMPLE:

WE HAVE THAT X IS INDEPENDENT OF Y.

EXERCISE ("BIT FLIPPER CHANNEL")

Verify the bounds for the bit-flipper channel.

The following bound is important and impactful (and also intuitively pleasing!):

THEOREM (CONDITIONING REDUCES ENTROPY)

For any two discrete random variables X and Y,

$$H_D(X|Y) \leq H_D(X)$$

with equality iff X and Y are independent random variables.

In words: **On average,** the uncertainty about X can only become smaller if we know Y.

Note Bene: As we have seen, this is *not true point-wise:* We may have $H_D(X|Y = y) > H_D(X)$ for some values of y.

$$H(X|Y) - H(X)$$

$$= \sum_y p(y) \left[- \sum_x p(x|y) \log p(x|y) \right] + \sum_x p(x) \log p(x)$$

✓ (*)

$$= \sum_{x,y} p(y) p(x|y) \log \frac{1}{p(x|y)} + \sum_{x,y} p(y|x) p(x) \log p(x)$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x)}{p(x|y)}$$

$$\leq \sum_{x,y} p(x,y) \left(\frac{p(x)p(y)}{p(x|y)p(y)} - 1 \right) \log(e)$$

$$= \sum_{x,y} p(x,y) \left(\frac{p(x)p(y)}{p(x,y)} - 1 \right) \log(e)$$

$$= \sum_{x,y} \left(p(x)p(y) - p(x,y) \right) \log(e)$$

$$= \left\{ \underbrace{\left(\sum_{x,y} p(x)p(y) \right)}_{=1} - \underbrace{\left(\sum_{x,y} p(x,y) \right)}_{=1} \right\} \log(e)$$

$$= 0.$$

JUSTIFICATION OF (*)

$$\sum_x \sum_y p(y|x)\, p(x) \log p(x)$$

$$= \sum_x \left\{ p(x) \log p(x) \underbrace{\left(\sum_y p(y|x) \right)}_{=1} \right\}$$

$$= \sum_x p(x) \log p(x).$$

Proof [Conditioning reduces entropy]:

$$H_D(X|Y) - H_D(X) = \mathbb{E}\left[\log_D \frac{1}{p_{X|Y}(X|Y)}\right] + \mathbb{E}[\log_D p_X(X)]$$

$$= \mathbb{E}\left[\log_D \frac{p_X(X)}{p_{X|Y}(X|Y)}\right]$$

$$= \mathbb{E}\left[\log_D \frac{p_X(X)p_Y(Y)}{p_{X|Y}(X|Y)p_Y(Y)}\right] = \mathbb{E}\left[\log_D \frac{p_X(X)p_Y(Y)}{p_{X,Y}(X,Y)}\right]$$

$$\overset{\text{(IT-Inequality)}}{\leq} \mathbb{E}\left[\frac{p_X(X)p_Y(Y)}{p_{X,Y}(X,Y)} - 1\right]\log_D(e)$$

$$= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \left[p_X(x)p_Y(y) - p_{X,Y}(x,y)\right]\log_D(e)$$

$$= [1-1]\log_D(e) = 0.$$

The condition for equality is $\frac{p_X(x)p_Y(y)}{p_{X,Y}(x,y)} = 1$ for all x and y, i.e., equality holds iff X and Y are independent random variables.

CONDITIONAL ENTROPY OF $f(X)$.

Let X be an arbitrary random variable.

Let $f(X)$ be a (deterministic) function of X.

$$H(f(X)|X) = 0$$

let: $Y = f(x)$

$$p(y|x) = \begin{cases} 1, & y = f(x) \\ 0, & y \neq f(x) \end{cases}$$

$$\Rightarrow H(Y|X) = H(f(x)|X) = 0$$

A generalization of the previous bound is also of interest to us:

THEOREM (CONDITIONING REDUCES ENTROPY)

For any three discrete random variables X, Y and Z,

$$H_D(X|Y,Z) \leq H_D(X|Z)$$

with equality iff X and Y are conditionally independent random variables given Z (that is, if and only if $p(x,y|z) = p(x|z)p(y|z)$ for all x, y, z).

Proof [Conditioning reduces entropy, generalized version]:

$$H_D(X|Y,Z) - H_D(X|Z) = \mathbb{E}\left[\log_D \frac{1}{p_{X|Y,Z}(X|Y,Z)}\right] + \mathbb{E}[\log_D p_{X|Z}(X|Z)]$$

$$= \mathbb{E}\left[\log_D \frac{p_{X|Z}(X|Z)}{p_{X|Y,Z}(X|Y,Z)}\right]$$

$$= \mathbb{E}\left[\log_D \frac{p_{X|Z}(X|Z)p_{Y|Z}(Y|Z)p_Z(Z)}{p_{X|Y,Z}(X|Y,Z)p_{Y|Z}(Y|Z)p_Z(Z)}\right]$$

$$= \mathbb{E}\left[\log_D \frac{p_{X|Z}(X|Z)p_{Y|Z}(Y|Z)p_Z(Z)}{p_{X,Y,Z}(X,Y,Z)}\right]$$

$$\overset{\text{(IT-Inequality)}}{\leq} \mathbb{E}\left[\frac{p_{X|Z}(X|Z)p_{Y|Z}(Y|Z)p_Z(Z)}{p_{X,Y,Z}(X,Y,Z)} - 1\right]\log_D(e)$$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}}\left[p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z) - p_{X,Y,Z}(x,y,z)\right]\log_D(e)$$

$$= [1-1]\log_D(e) = 0.$$

The condition for equality is $\frac{p_{X|Z}(X|Z)p_{Y|Z}(Y|Z)p_Z(Z)}{p_{X,Y,Z}(X,Y,Z)} = 1$ for all x, y, z, i.e., equality holds iff X and Y are conditionally independent random variables given Z.

PROOF THAT:

$$p_{X,Y,Z}(x,y,z)$$

$$= p_{X|Y,Z}(x|y,z)\, p_{Y|Z}(y|z)\, p_Z(z)$$

STEP 1: OBSERVE THAT:

$$p_{Y|Z}(y|z)\, p_Z(z) = p_{Y,Z}(y,z)$$

STEP 2: NEXT PAGE:

$$P_{Y|Z}(y|z)\,p_Z(z) = P_{Y,Z}(y,z)$$

$W := (Y, Z)$

WITH THIS, REWRITE:

$$P_{X|Y,Z}(x|y,z) = p_{X|W}(x|w)$$

$$P_{X|Y,Z}(x|y,z)\,P_{Y|Z}(y|z)\,p_Z(z)$$

$$= p_{X|W}(x|w)\,p_W(w) = P_{X,W}(x,w)$$

$$= P_{X,Y,Z}(x,y,z)$$

FIRST QUIZ

- OPENS TODAY AT 17:00
- CLOSES MONDAY (MARCH 10) AT 23:59

- FORMULA COLLECTION.

YESTERDAY

CONDITIONAL ENTROPY

$$H(X|Y) = \sum_y p(y)\ H(X|Y=y)$$

$$= -\sum_y p(y) \sum_x p(x|y) \log p(x|y)$$

$$= -\sum_y \sum_x p(y)p(x|y) \log p(x|y)$$

$$= -\sum_{y}\sum_{x} p(x,y) \log p(x|y)$$

$$0 \le H_D(X|Y) \le H_D(X) \le \log_D |X|$$

X, Y, Z:

$$0 \le H(X|Y,Z) \le H(X|Z) \le H(X) \le \log_D |X|$$

Recall: When we simply write $H(X)$, suppressing the subscript D, then we mean $D = 2$.

EXAMPLE

Let $X \in \{0, 1\}$ be uniformly distributed and let $Y = X$. Then

$$H(X|Y) = 0 \text{ and } H(X) = 1.$$

EXAMPLE

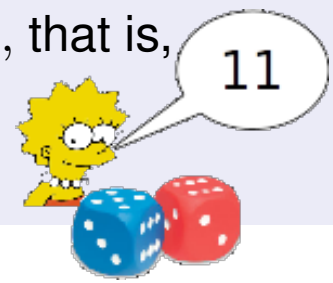
Let $X \in \{0, 1\}$ and $Y \in \{0, 1\}$ be uniformly distributed and independent. Then

$$H(X|Y) = 1 \text{ and } H(X) = 1.$$

EXERCISE (LISA ROLLS TWO DICE)

- Lisa rolls two dice and announces the sum L written as a two digit number.
- The alphabet of $L = L_1 L_2$ is $\{02, 03, 04, 05, 06, 07, 08, 09, 10, 11, 12\}$.
  - The alphabet of L_1 is $\{0, 1\}$.
  - The alphabet of L_2 is $\{0, 1, \ldots, 9\}$.
- Determine the probability that $L_2 = 2$, knowing that $L_1 = 1$, that is,

$$p_{L_2|L_1}(2|1).$$



LISA ROLLS TWO DICE

SOLUTION

Using the definition (and calculations from Lecture 1),

$$p_{L_2|L_1}(2|1) = \frac{p_{L_1,L_2}(1,2)}{p_{L_1}(1)} = \frac{1/36}{1/6} = \frac{1}{6}.$$

After running over all possible values for (i, j), we obtain

$L_2 = j$	$L_1 = i$ $p_{L_1,L_2}(i,j)$	0	1	$p_{L_2}(j)$
0		0	3/36	3/36
1		0	2/36	2/36
2		1/36	1/36	2/36
3		2/36	0	2/36
4		3/36	0	3/36
5		4/36	0	4/36
6		5/36	0	5/36
7		6/36	0	6/36
8		5/36	0	5/36
9		4/36	0	4/36
	$p_{L_1}(i)$	5/6	1/6	

$L_2 = j$	$p_{L_2\|L_1}(j\|0)$	$p_{L_2\|L_1}(j\|1)$
0	0	3/6
1	0	2/6
2	1/30	1/6
3	2/30	0
4	3/30	0
5	4/30	0
6	5/30	0
7	6/30	0
8	5/30	0
9	4/30	0

Lisa rolls two dice

Example

$$H(L_2|L_1=1) = \frac{3}{6}\log\frac{6}{3} + \frac{2}{6}\log\frac{6}{2} + \frac{1}{6}\log 6$$
$$= 1.459 \text{ bits}$$
$$H(L_2|L_1=0) = \cdots = 2.857 \text{ bits}$$

$L_2 = j$	$p_{L_2\|L_1}(j\|0)$	$p_{L_2\|L_1}(j\|1)$
0	0	3/6
1	0	2/6
2	1/30	1/6
3	2/30	0
4	3/30	0
5	4/30	0
6	5/30	0
7	6/30	0
8	5/30	0
9	4/30	0

EXAMPLE

$$H(L_2|L_1) = p_{L_1}(0)H(L_2|L_1 = 0) + p_{L_1}(1)H(L_2|L_1 = 1)$$
$$= \frac{5}{6} \times 2.857 + \frac{1}{6} \times 1.459 = 2.624 \text{ bits}$$

Now, we can observe that

$$2.624 = H(L_2|L_1) \leq H(L_2) = 3.22,$$

exactly like it has to be according to our theorems.

Recall that the joint entropy of two random variables X, Y is completely naturally defined as

$$H_D(X, Y) = -\sum_x \sum_y p_{X,Y}(x, y) \log_D p_{X,Y}(x, y).$$

Using the fact that $p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x)$, we can write this as

$$\begin{aligned} H_D(X, Y) &= -\sum_x p_X(x) \left(\sum_y p_{Y|X}(y|x) \log_D \left(p_X(x)p_{Y|X}(y|x) \right) \right) \\ &= -\sum_x p_X(x) \left(\sum_y p_{Y|X}(y|x) \left(\log_D p_X(x) + \log_D p_{Y|X}(y|x) \right) \right) \\ &= -\sum_x p_X(x) \left\{ \left(\sum_y p_{Y|X}(y|x) \log_D p_X(x) \right) \right. \\ &\qquad \left. + \left(\sum_y p_{Y|X}(y|x) \log_D p_{Y|X}(y|x) \right) \right\} \end{aligned}$$

$$H(X,Y)$$

$$= -\sum_{x,y} p(x,y) \log p(x,y)$$

$$= -\sum_{x,y} p(x,y) \log \left[p(x)\, p(y|x) \right]$$

$$= -\sum_{x,y} p(x,y) \left\{ \log p(x) + \log p(y|x) \right\}$$

$$= \left(- \sum_{x,y} p(x,y) \log p(x) \right) + \left(- \sum_{x,y} p(x,y) \log p(y|x) \right)$$

$$= \left(- \sum_{x,y} p(x) p(y|x) \log p(x) \right) + \left(- \sum_{x,y} p(x,y) \log p(y|x) \right)$$

$$= \underbrace{\left(- \sum_{x} p(x) \log p(x) \left(\underbrace{\sum_{y} p(y|x)}_{=1} \right) \right)}_{H(X)} + \text{---''---}$$

$$= H(X) + H(Y|X)$$

HENCE:

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

$$\| \qquad = H(Y) + H(X|Y)$$

$$H(Y,X)$$

$$p(X=x, Y=y) = p(Y=y, X=x)$$

But now, we observe:

$$
\begin{aligned}
H_D(X,Y) &= -\sum_x p_X(x) \left\{ \left(\sum_y p_{Y|X}(y|x) \log_D p_X(x) \right) \right. \\
&\qquad \left. + \left(\sum_y p_{Y|X}(y|x) \log_D p_{Y|X}(y|x) \right) \right\} \\
&= \underbrace{-\sum_x p_X(x) \left(\sum_y p_{Y|X}(y|x) \log_D p_X(x) \right)}_{H_D(X)} \\
&\qquad + \underbrace{\sum_x p_X(x) \left(-\sum_y p_{Y|X}(y|x) \log_D p_{Y|X}(y|x) \right)}_{H_D(Y|X)} \\
&= H_D(X) + H_D(Y|X).
\end{aligned}
$$

Let us write this once more and enjoy it properly:

$$H_D(X, Y) = H_D(X) + H_D(Y|X).$$

In words: To find the joint entropy of two random variables, we can first calculate the entropy of one of the two, and then add to it the conditional entropy of the second, given the first.

Of course, what we could do once, we can do again!

THEOREM (CHAIN RULE FOR ENTROPIES)

Let $S_1, \ldots, S_n$ be discrete random variables. Then

$$H_D(S_1, S_2, \ldots, S_n) = H_D(S_1) + H_D(S_2|S_1) + \cdots + H_D(S_n|S_1, \ldots, S_{n-1}).$$

The above result says that the uncertainty of a collection of random variables (in any order) is the uncertainty of the first, plus the uncertainty of the second when the first is known, plus the uncertainty of the third when the first two are known, etc.

Proof [Chain rule for entropy]:

$$p_{S_1,S_2,\ldots,S_n}(s_1,\ldots,s_n) = p_{S_1}(s_1)\prod_{i=2}^{n} p_{S_i|S_1,\ldots,S_{i-1}}(s_i|s_1,\ldots,s_{i-1})$$

$$-\log_D\left(p_{S_1,S_2,\ldots,S_n}(s_1,\ldots,s_n)\right) = -\log p_{S_1}(s_1) - \sum_{i=2}^{n} \log_D\left(p_{S_i|S_1,\ldots,S_{i-1}}(s_i|s_1,\ldots,s_{i-1})\right)$$

The expected value of the LHS is $H_D(S_1, S_2, \ldots, S_n)$.

The expected value of the RHS is
$H_D(S_1) + H_D(S_2|S_1) + \cdots + H_D(S_n|S_1,\ldots,S_{n-1})$. □

EXAMPLE

Let X, Y, Z be discrete random variables. We have:

$$\begin{aligned} H(X,Y,Z) &= H(X) + H(Y|X) + H(Z|X,Y) \\ &= H(X) + H(Z|X) + H(Y|X,Z) \\ &= H(Y) + H(X|Y) + H(Z|X,Y) \\ &= H(Y) + H(Z|Y) + H(X|Y,Z) \\ &= H(Z) + H(X|Z) + H(Y|X,Z) \\ &= H(Z) + H(Y|Z) + H(X|Y,Z), \end{aligned}$$

where we omitted the subscript D for compact notation, but these relationships hold for all integers $D \geq 2$.

The chain rule for entropy and the fact that conditioning reduces entropy, proves the following theorem which was stated last week without proof:

THEOREM

Let $S_1, \ldots, S_n$ be discrete random variables. Then

$$H(S_1, S_2, \ldots, S_n) \leq H(S_1) + H(S_2) + \cdots + H(S_n),$$

with equality iff $S_1, \ldots, S_n$ are independent.

$$H(S_1, S_2, S_3) = H(S_1) + H(S_2|S_1) + H(S_3|S_1, S_2)$$
$$\leq H(S_1) + H(S_2) + H(S_3)$$

Sometimes it is convenient to compute the conditional entropy using the chain rule for entropies. For instance:

$$H(X|Y) = H(X, Y) - H(Y).$$

$$H(X,Y) = H(X) + H(Y|X) \geq H(X)$$

COROLLARY

$$H(X, Y) \geq H(X);$$
$$H(X, Y) \geq H(Y).$$

The above inequalities follow from the chain rule for entropies and the fact that entropy (conditional or not) is nonnegative.

EXAMPLE (LISA ROLLS TWO DICE)

From

$$H(L_1, L_2) = 3.2744 \text{ bits}$$
$$H(L_1) = 0.6500 \text{ bits}$$
$$H(L_2) = 3.2188 \text{ bits},$$

we compute

$$H(L_2|L_1) = H(L_1, L_2) - H(L_1) = 3.2744 - 0.6500 = 2.624 \text{ bits}$$
$$H(L_1|L_2) = H(L_1, L_2) - H(L_2) = 3.2744 - 3.2188 = 0.056 \text{ bits},$$

and verify that indeed

$$H(L_1|L_2) \leq H(L_1) \leq H(L_1, L_2)$$
$$H(L_2|L_1) \leq H(L_2) \leq H(L_1, L_2).$$

EXERCISE

Determine $H(L_1, L_2 | S_1, S_2)$.

WHERE:

S_1: OUTCOME OF BLUE DIE

S_2: OUTCOME OF RED DIE

EXERCISE

Determine $H(L_1, L_2|S_1, S_2)$.

SOLUTION

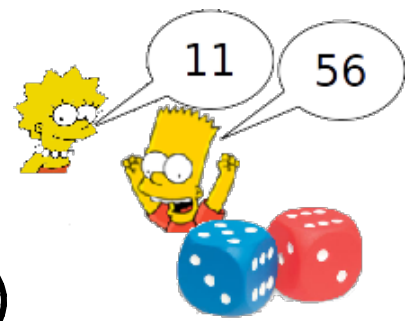
L_1 and L_2 are deterministic functions of S_1 and S_2.
Hence $H(L_1, L_2|S_1, S_2) = 0$.

EXAMPLE

Determine $H(S_1, S_2|L_1, L_2)$ knowing that $H(S_1, S_2) = 5.1699$ bits and $H(L_1, L_2) = 3.2744$ bits.

$$H(S_1, S_2, L_1, L_2)$$
$$= H(S_1, S_2) + H(L_1, L_2 | S_1, S_2)$$
$$= H(L_1, L_2) + H(S_1, S_2 | L_1, L_2)$$

$H(L_1, L_2 | S_1, S_2) = 0$



LISA ROLLS TWO DICE

EXAMPLE

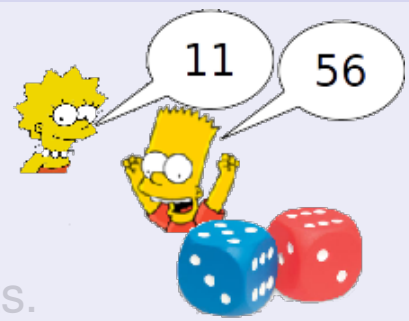
Determine $H(S_1, S_2|L_1, L_2)$ knowing that $H(S_1, S_2) = 5.1699$ bits and $H(L_1, L_2) = 3.2744$ bits.

SOLUTION

$H(S_1, S_2|L_1, L_2) = H(S_1, S_2, L_1, L_2) - H(L_1, L_2)$.

But $H(S_1, S_2, L_1, L_2) = H(S_1, S_2)$. (Can you say why?)

Hence $H(S_1, S_2|L_1, L_2) = H(S_1, S_2) - H(L_1, L_2) = 1.896$ bits.



LISA ROLLS TWO DICE

EXAMPLE

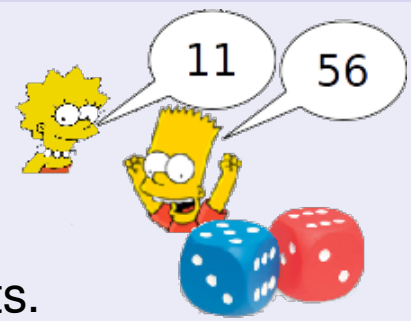
Determine $H(S_1, S_2|L_1, L_2)$ knowing that $H(S_1, S_2) = 5.1699$ bits and $H(L_1, L_2) = 3.2744$ bits.

SOLUTION

$H(S_1, S_2|L_1, L_2) = H(S_1, S_2, L_1, L_2) - H(L_1, L_2)$.

But $H(S_1, S_2, L_1, L_2) = H(S_1, S_2)$. (Can you say why?)

Hence $H(S_1, S_2|L_1, L_2) = H(S_1, S_2) - H(L_1, L_2) = 1.896$ bits.



Definition (Coin-Flip Source)

The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

Definition (Coin-Flip Source)

The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

Definition (Coin-Flip Source)

The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$

Definition (Coin-Flip Source)

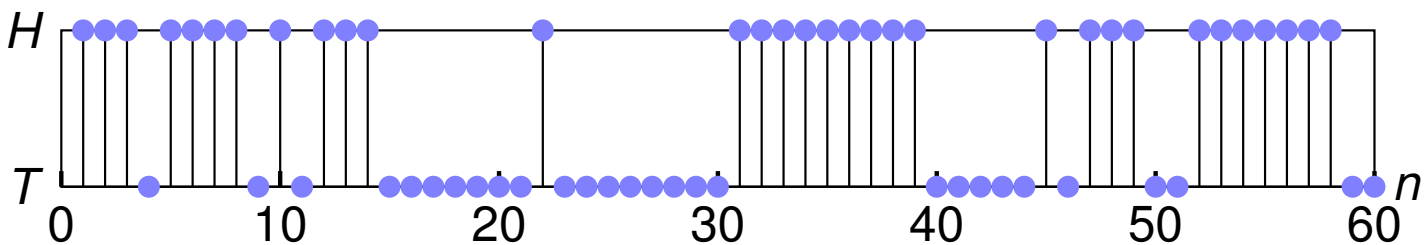
The source models a sequence $S_1, S_2, \ldots, S_n$ of n coin flips.

So $S_i \in \mathcal{A} = \{H, T\}$, where H stands for heads, T for tails, $i = 1, 2, \ldots, n$.

$p_{S_i}(H) = p_{S_i}(T) = \frac{1}{2}$ for all i, and coin flips are independent.

Hence,

$$p_{S_1,S_2,\ldots,S_n}(s_1, s_2, \ldots, s_n) = \frac{1}{2^n} \quad \text{for all } (s_1, s_2, \ldots, s_n) \in \mathcal{A}^n$$



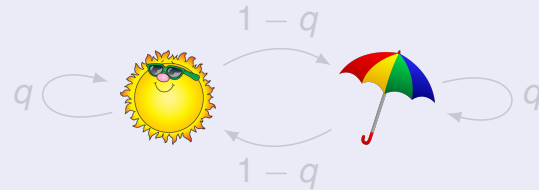
Definition (Sunny-Rainy Source)

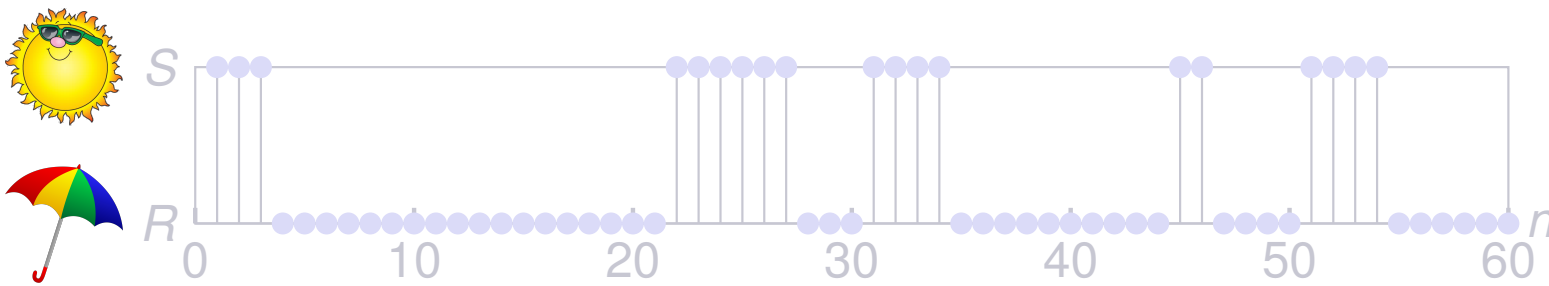
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.

For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.



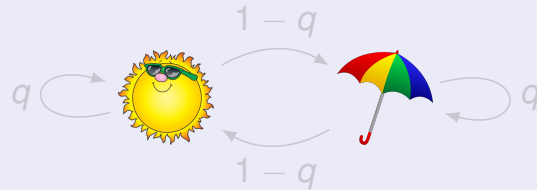


Definition (Sunny-Rainy Source)

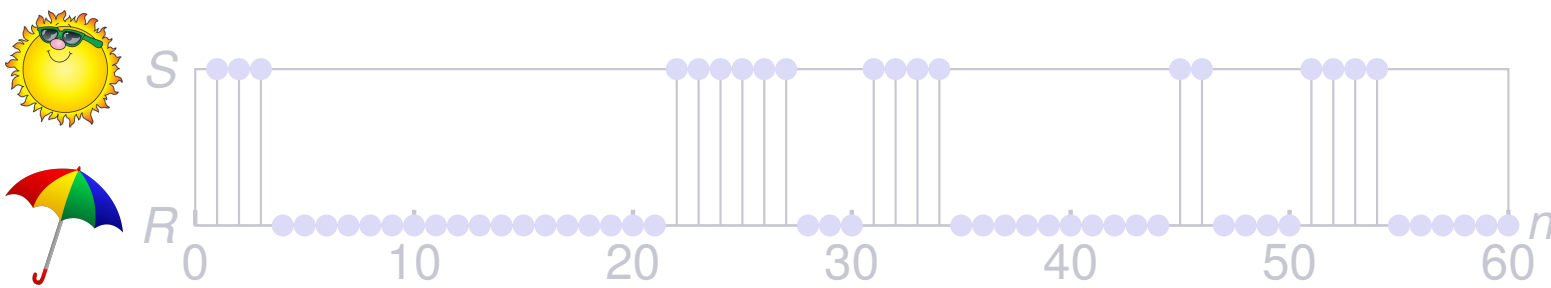
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.



For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.



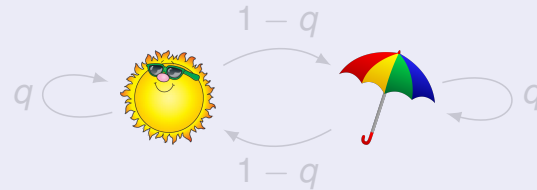
Definition (Sunny-Rainy Source)

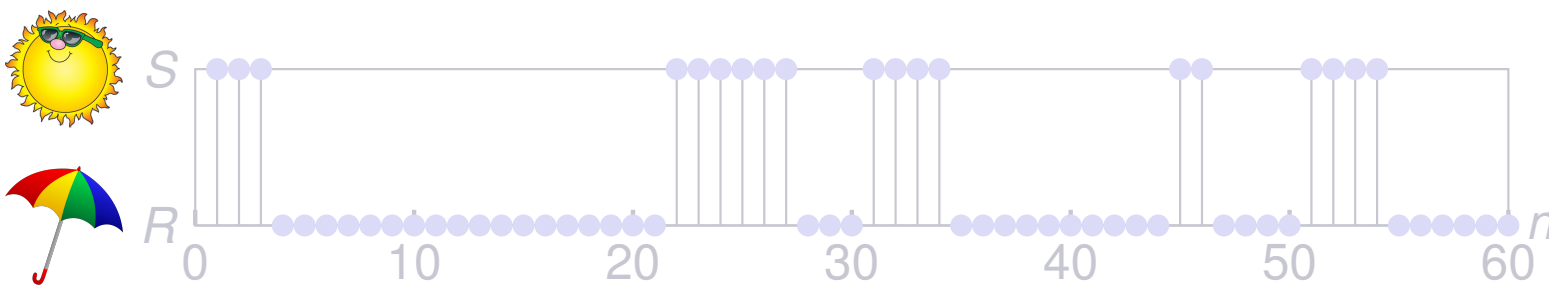
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.

For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.



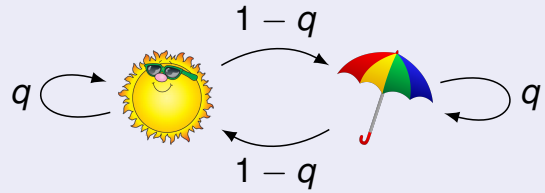


Definition (Sunny-Rainy Source)

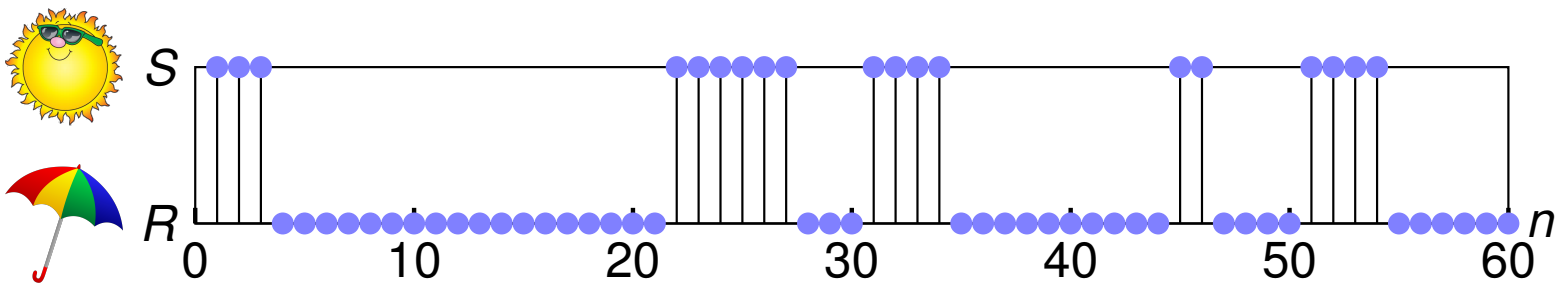
The source models a sequence $S_1, S_2, \ldots, S_n$ of weather conditions.

So $S_i \in \mathcal{A} = \{S, R\}$, where S stands for sunny, R for rainy, $i = 1, 2, \ldots, n$.

The weather on the first day is uniformly distributed in $\mathcal{A}$.



For all other days, with probability $q = \frac{6}{7}$ the weather is as for the day before.



$$p(S_n = R \mid S_1 = S, S_2 = S, \ldots, S_{n-1} = R) = p(S_n = R \mid S_{n-1} = R)$$

"Markov"

Example

For the Sunny-Rainy source:

- $p_{S_1}(S) = \frac{1}{2}$
- $p_{S_1,S_2}(R,R) = p_{S_1}(R)p_{S_2|S_1}(R|R) = \frac{1}{2}q$
- $p_{S_1,S_2}(R,S) = p_{S_1}(R)p_{S_2|S_1}(S|R) = \frac{1}{2}(1-q)$
- $p_{S_1,S_2,S_3,S_4}(R,S,S,R) = \frac{1}{2}(1-q)q(1-q) = \frac{1}{2}q(1-q)^2$

In general, if c is the number of weather changes ($0 \leq c \leq n-1$), then

$$p_{S_1,S_2,\ldots,S_n}(s_1,s_2,\ldots,s_n) = \frac{1}{2}q^{n-1-c}(1-q)^c.$$

Example

For the Sunny-Rainy source:

- $p_{S_1}(S) = \frac{1}{2}$
- $p_{S_1,S_2}(R,R) = p_{S_1}(R)p_{S_2|S_1}(R|R) = \frac{1}{2}q$
- $p_{S_1,S_2}(R,S) = p_{S_1}(R)p_{S_2|S_1}(S|R) = \frac{1}{2}(1-q)$
- $p_{S_1,S_2,S_3,S_4}(R,S,S,R) = \frac{1}{2}(1-q)q(1-q) = \frac{1}{2}q(1-q)^2$

In general, if c is the number of weather changes ($0 \leq c \leq n-1$), then

$$p_{S_1,S_2,\ldots,S_n}(s_1,s_2,\ldots,s_n) = \frac{1}{2}q^{n-1-c}(1-q)^c.$$

EXAMPLE

For the Sunny-Rainy source:

- $p_{S_1}(S) = \frac{1}{2}$
- $p_{S_1,S_2}(R,R) = p_{S_1}(R)p_{S_2|S_1}(R|R) = \frac{1}{2}q$
- $p_{S_1,S_2}(R,S) = p_{S_1}(R)p_{S_2|S_1}(S|R) = \frac{1}{2}(1-q)$
- $p_{S_1,S_2,S_3,S_4}(R,S,S,R) = \frac{1}{2}(1-q)q(1-q) = \frac{1}{2}q(1-q)^2$

In general, if c is the number of weather changes ($0 \leq c \leq n-1$), then

$$p_{S_1,S_2,\ldots,S_n}(s_1,s_2,\ldots,s_n) = \frac{1}{2}q^{n-1-c}(1-q)^c.$$

Example

For the Sunny-Rainy source:

- $p_{S_1}(S) = \frac{1}{2}$
- $p_{S_1,S_2}(R,R) = p_{S_1}(R)p_{S_2|S_1}(R|R) = \frac{1}{2}q$
- $p_{S_1,S_2}(R,S) = p_{S_1}(R)p_{S_2|S_1}(S|R) = \frac{1}{2}(1-q)$
- $p_{S_1,S_2,S_3,S_4}(R,S,S,R) = \frac{1}{2}(1-q)q(1-q) = \frac{1}{2}q(1-q)^2$

In general, if c is the number of weather changes ($0 \leq c \leq n-1$), then

$$p_{S_1,S_2,\ldots,S_n}(s_1,s_2,\ldots,s_n) = \frac{1}{2}q^{n-1-c}(1-q)^c.$$

Example

For the Sunny-Rainy source:

- $p_{S_1}(S) = \frac{1}{2}$
- $p_{S_1,S_2}(R,R) = p_{S_1}(R)p_{S_2|S_1}(R|R) = \frac{1}{2}q$
- $p_{S_1,S_2}(R,S) = p_{S_1}(R)p_{S_2|S_1}(S|R) = \frac{1}{2}(1-q)$
- $p_{S_1,S_2,S_3,S_4}(R,S,S,R) = \frac{1}{2}(1-q)q(1-q) = \frac{1}{2}q(1-q)^2$

In general, if c is the number of weather changes ($0 \leq c \leq n-1$), then

$$p_{S_1,S_2,\ldots,S_n}(s_1,s_2,\ldots,s_n) = \frac{1}{2}q^{n-1-c}(1-q)^c.$$

EXERCISE

Let $i = 2, 3, \ldots$

For the Sunny-Rainy source:

- Find $p_{S_i}(s_i)$
- Find $p_{S_i|S_{i-1}}(s_i|s_{i-1})$
- Are S_i and S_{i-1} independent?

Solution (Sunny-Rainy source)

By definition, $p_{S_i|S_{i-1}}(j|k) = q$ if $j = k$ and $(1-q)$ otherwise.

Hence S_{i-1} and S_i are not independent.

To determine the statistic of the marginals, we use the law of total probability and induction to show that p_{S_i} is uniform.

It is true by definition for $i = 1$.

Suppose that p_{S_i} is uniform for $i = 1, \ldots, n-1$. We show that it is uniform also for $i = n$:

$$p_{S_n}(j) = \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k) p_{S_{n-1}}(k) = \frac{1}{2} \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k)$$
$$= \frac{1}{2}(q + (1-q)) = \frac{1}{2}.$$

Hence the marginals are uniformly distributed (like for the Coin-Flip source).

Solution (Sunny-Rainy source)

By definition, $p_{S_i|S_{i-1}}(j|k) = q$ if $j = k$ and $(1-q)$ otherwise.

Hence S_{i-1} and S_i are not independent.

To determine the statistic of the marginals, we use the law of total probability and induction to show that p_{S_i} is uniform.

It is true by definition for $i = 1$.

Suppose that p_{S_i} is uniform for $i = 1, \ldots, n-1$. We show that it is uniform also for $i = n$:

$$p_{S_n}(j) = \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k) p_{S_{n-1}}(k) = \frac{1}{2} \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k)$$
$$= \frac{1}{2}(q + (1-q)) = \frac{1}{2}.$$

Hence the marginals are uniformly distributed (like for the Coin-Flip source).

Solution (Sunny-Rainy source)

By definition, $p_{S_i|S_{i-1}}(j|k) = q$ if $j = k$ and $(1-q)$ otherwise.

Hence S_{i-1} and S_i are not independent.

To determine the statistic of the marginals, we use the law of total probability and induction to show that p_{S_i} is uniform.

It is true by definition for $i = 1$.

Suppose that p_{S_i} is uniform for $i = 1, \ldots, n-1$. We show that it is uniform also for $i = n$:

$$p_{S_n}(j) = \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k) p_{S_{n-1}}(k) = \frac{1}{2} \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k)$$
$$= \frac{1}{2}(q + (1-q)) = \frac{1}{2}.$$

Hence the marginals are uniformly distributed (like for the Coin-Flip source).

Solution (Sunny-Rainy source)

By definition, $p_{S_i|S_{i-1}}(j|k) = q$ if $j = k$ and $(1-q)$ otherwise.

Hence S_{i-1} and S_i are not independent.

To determine the statistic of the marginals, we use the law of total probability and induction to show that p_{S_i} is uniform.

It is true by definition for $i = 1$.

Suppose that p_{S_i} is uniform for $i = 1, \ldots, n-1$. We show that it is uniform also for $i = n$:

$$p_{S_n}(j) = \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k) p_{S_{n-1}}(k) = \frac{1}{2} \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k)$$
$$= \frac{1}{2}(q + (1-q)) = \frac{1}{2}.$$

Hence the marginals are uniformly distributed (like for the Coin-Flip source).

Solution (Sunny-Rainy source)

By definition, $p_{S_i|S_{i-1}}(j|k) = q$ if $j = k$ and $(1-q)$ otherwise.

Hence S_{i-1} and S_i are not independent.

To determine the statistic of the marginals, we use the law of total probability and induction to show that p_{S_i} is uniform.

It is true by definition for $i = 1$.

Suppose that p_{S_i} is uniform for $i = 1, \ldots, n-1$. We show that it is uniform also for $i = n$:

$$p_{S_n}(j) = \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k) p_{S_{n-1}}(k) = \frac{1}{2} \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k)$$
$$= \frac{1}{2}(q + (1-q)) = \frac{1}{2}.$$

Hence the marginals are uniformly distributed (like for the Coin-Flip source).

Solution (Sunny-Rainy source)

By definition, $p_{S_i|S_{i-1}}(j|k) = q$ if $j = k$ and $(1-q)$ otherwise.

Hence S_{i-1} and S_i are not independent.

To determine the statistic of the marginals, we use the law of total probability and induction to show that p_{S_i} is uniform.

It is true by definition for $i = 1$.

Suppose that p_{S_i} is uniform for $i = 1, \ldots, n-1$. We show that it is uniform also for $i = n$:

$$p_{S_n}(j) = \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k) p_{S_{n-1}}(k) = \frac{1}{2} \sum_{k \in \{S,R\}} p_{S_n|S_{n-1}}(j|k)$$
$$= \frac{1}{2}(q + (1-q)) = \frac{1}{2}.$$

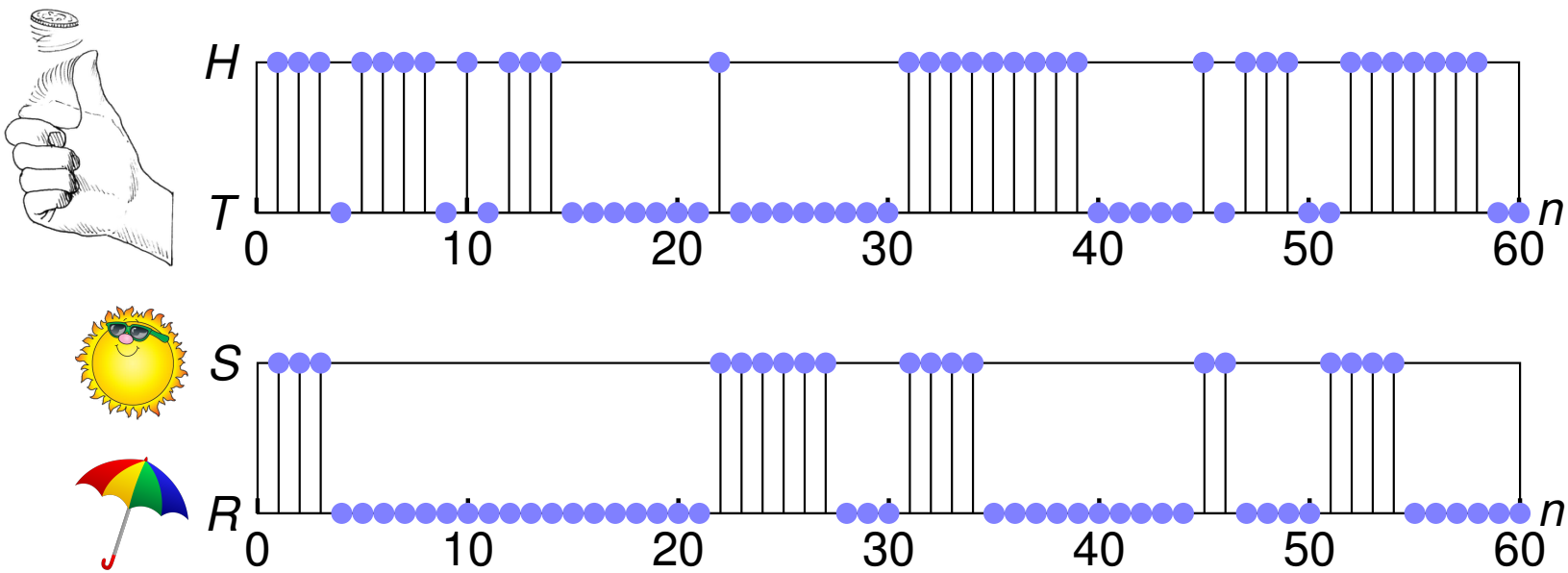
Hence the marginals are uniformly distributed (like for the Coin-Flip source).

EXERCISE

Let $i = 2, 3, \ldots$

For the Coin-Flip (CF) and Sunny-Rainy (SR) sources:

- Compute $H(S_i)$
- Compute $H(S_i|S_1, \ldots, S_{i-1})$



Solution ($H(S_i)$)

The entropy depends only on the distribution, and for a uniform distribution, it is the log of the alphabet's cardinality. Hence

$$H_{CF}(S_i) = H_{SR}(S_i) = \log 2 = 1$$

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Coin-Flip source)

S_i is independent of $S_1, \ldots, S_{i-1}$

Hence, $H(S_i|S_1, \ldots, S_{i-1}) = H(S_i)$.

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Coin-Flip source)

S_i is independent of $S_1, \ldots, S_{i-1}$

Hence, $H(S_i|S_1, \ldots, S_{i-1}) = H(S_i)$.

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Sunny-Rainy source)

S_i depends only on S_{i-1}. Hence

$$H_{SR}(S_i|S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = H_{SR}(S_i|S_{i-1} = s_{i-1}).$$

When $S_{i-1} = k \in \{S, R\}$, the probabilities for S_i are q and $(1-q)$. Hence

$$H_{SR}(S_i|S_{i-1} = s_{i-1}) = -q \log q - (1-q) \log(1-q).$$

Taking the average on both sides yields

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1-q) \log(1-q).$$

For $q = \frac{6}{7}$, we have

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1-q) \log(1-q) = 0.592.$$

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Sunny-Rainy source)

S_i depends only on S_{i-1}. Hence

$$H_{SR}(S_i|S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = H_{SR}(S_i|S_{i-1} = s_{i-1}).$$

When $S_{i-1} = k \in \{S, R\}$, the probabilities for S_i are q and $(1 - q)$. Hence

$$H_{SR}(S_i|S_{i-1} = s_{i-1}) = -q \log q - (1 - q) \log(1 - q).$$

Taking the average on both sides yields

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1 - q) \log(1 - q).$$

For $q = \frac{6}{7}$, we have

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1 - q) \log(1 - q) = 0.592.$$

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Sunny-Rainy source)

S_i depends only on S_{i-1}. Hence

$$H_{SR}(S_i|S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = H_{SR}(S_i|S_{i-1} = s_{i-1}).$$

When $S_{i-1} = k \in \{S, R\}$, the probabilities for S_i are q and $(1 - q)$. Hence

$$H_{SR}(S_i|S_{i-1} = s_{i-1}) = -q \log q - (1 - q) \log(1 - q).$$

Taking the average on both sides yields

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1 - q) \log(1 - q).$$

For $q = \frac{6}{7}$, we have

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1 - q) \log(1 - q) = 0.592.$$

Solution ($H(S_i|S_1, \ldots, S_{i-1})$ for the Sunny-Rainy source)

S_i depends only on S_{i-1}. Hence

$$H_{SR}(S_i|S_1 = s_1, \ldots, S_{i-1} = s_{i-1}) = H_{SR}(S_i|S_{i-1} = s_{i-1}).$$

When $S_{i-1} = k \in \{S, R\}$, the probabilities for S_i are q and $(1 - q)$. Hence

$$H_{SR}(S_i|S_{i-1} = s_{i-1}) = -q \log q - (1 - q) \log(1 - q).$$

Taking the average on both sides yields

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1 - q) \log(1 - q).$$

For $q = \frac{6}{7}$, we have

$$H_{SR}(S_i|S_{i-1}) = -q \log q - (1 - q) \log(1 - q) = 0.592.$$

EXERCISE

Determine $H(S_1, S_2, \ldots, S_n)$ for the Coin-Flip source.

Exercise

Determine $H(S_1, S_2, \ldots, S_n)$ for the Coin-Flip source.

Solution

The source produces independent and identically distributed symbols. Hence

$$H(S_1, S_2, \ldots, S_n) \overset{\text{(indep.)}}{=} H(S_1) + H(S_2) + \cdots + H(S_n) \overset{\text{(identically distributed)}}{=} nH(S_1)$$

Moreover, the distribution is uniform, therefore $H(S_1) = 1$ bit. Putting things together,

$$H(S_1, S_2, \ldots, S_n) = n \text{ bits}$$

EXERCISE

Determine $H(S_1, S_2, \ldots, S_n)$ for the Sunny-Rainy source with $q = \frac{6}{7}$.

For $i = 2, 3, \ldots, n$, the statistic of S_i depends only on S_{i-1}. Hence

$$H(S_i|S_1, S_2, \ldots, S_{i-1}) = H(S_i|S_{i-1})$$

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \cdots + H(S_n|S_{n-1})$$

We have already determined that $H(S_1) = 1$ bit and $H(S_i|S_{i-1}) = 0.592$ bits. Therefore

$$H(S_1, S_2, \ldots, S_n) = 1 + 0.592(n-1) \text{ bits}$$

EXERCISE

Determine $H(S_1, S_2, \ldots, S_n)$ for the Sunny-Rainy source with $q = \frac{6}{7}$.

SOLUTION

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \cdots + H(S_n|S_1, \ldots, S_{n-1})$$

For $i = 2, 3, \ldots, n$, the statistic of S_i depends only on S_{i-1}. Hence

$$H(S_i|S_1, S_2, \ldots, S_{i-1}) = H(S_i|S_{i-1})$$

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \cdots + H(S_n|S_{n-1})$$

We have already determined that $H(S_1) = 1$ bit and $H(S_i|S_{i-1}) = 0.592$ bits. Therefore

$$H(S_1, S_2, \ldots, S_n) = 1 + 0.592(n-1) \text{ bits}$$

Exercise

Determine $H(S_1, S_2, \ldots, S_n)$ for the Sunny-Rainy source with $q = \frac{6}{7}$.

Solution

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \cdots + H(S_n|S_1, \ldots, S_{n-1})$$

For $i = 2, 3, \ldots, n$, the statistic of S_i depends only on S_{i-1}. Hence

$$H(S_i|S_1, S_2, \ldots, S_{i-1}) = H(S_i|S_{i-1})$$

$$H(S_1, S_2, \ldots, S_n) = H(S_1) + H(S_2|S_1) + \cdots + H(S_n|S_{n-1})$$

We have already determined that $H(S_1) = 1$ bit and $H(S_i|S_{i-1}) = 0.592$ bits. Therefore

$$H(S_1, S_2, \ldots, S_n) = 1 + 0.592(n-1) \text{ bits}$$

Summary : The Main Result of Source Coding / Data Compression

Theorem (Textbook Thm 3.3)

The average codeword-length of a uniquely decodable code Γ for S must satisfy

$$H_D(S_1, S_2, \ldots, S_n) \leq L((S_1, S_2, \ldots, S_n), \Gamma)$$

and there exists a uniquely decodable code Γ_{SF} satisfying

$$L((S_1, S_2, \ldots, S_n), \Gamma_{SF}) < H_D(S_1, S_2, \ldots, S_n) + 1.$$

- And in many cases, as n becomes large, the upper and the lower bound are arbitrarily close!

Additional Questions of interest include:

- What if the source alphabet is not finite?
- What if we do not know the source distribution $p_X(x)$? (Universal source coding)

What if the source alphabet is infinite?

- In all of our previous discussion on actual codes, we have assumed that the source alphabet is discrete and finite.
- What if it is discrete but infinite?
- ... is this just an academic endeavour?
- In this class, we only touch the top of this iceberg...

BINARY PREFIX-FREE CODE FOR POSITIVE INTEGERS

The set of positive integers is infinite and no probability is assigned to its elements. Hence we cannot use Huffman's construction to encode integers.

First Attempt to Encode Positive Integers: "Standard Method"

n	$c(n)$
1	1
2	10
3	11
4	100
5	101
$\vdots$	$\vdots$

The code is not prefix-free.

The length of $c(n)$ is $l(n) = \lfloor \log_2 n \rfloor + 1$.

Note: The first digit is always 1.

Second Attempt: “Elias Code 1"

We prefix code $c(n)$ with $l(n) - 1$ zeros.

n	$c_1(n)$
1	1
2	010
3	011
4	00100
5	00101
$\vdots$	$\vdots$

The code is prefix-free. (Codewords of different length cannot have the same number of leading zeros.)

The length of $c_1(n)$ is

$$l_1(n) = l(n) - 1 + l(n) = 2\lfloor \log_2 n \rfloor + 1.$$

Note: we are essentially doubling the length to make the code prefix-free.

Third Attempt: “Elias Code 2"

Instead of $l(n) - 1$ zeros followed by a 1, we prefix with $c_1\left(l(n)\right)$, which is also prefix-free (hence can be identified). Like the zeros, it tells the length of the codeword.

Notation: $\tilde{c}(n)$ is $c(n)$ without the leading 1.

n	$c(n)$	$l(n)$	$c_1(n)$	$c_1\left(l(n)\right)\tilde{c}(n)$
1	1	1	1	$c_1(1) = 1$
2	10	2	010	$c_1(2)\,0 = 010\,0$
3	11	2	011	$c_1(2)\,1 = 010\,1$
4	100	3	00100	$c_1(3)\,00 = 011\,00$
5	101	3	00101	$c_1(3)\,01 = 011\,01$
⋮	⋮			

The code is prefix-free.

The codeword length is
$l_2(n) = l_1\left(l(n)\right) + l(n) - 1 = 2\lfloor \log_2(\lfloor \log_2 n \rfloor + 1) \rfloor + 1 + \lfloor \log_2 n \rfloor$.

What if the source distribution is not known?

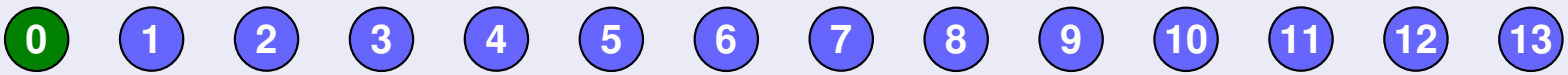
- Universal source coding.

LZW

- Practically important algorithms: “Lempel-Ziv” (LZ77, LZ78). Time permitting, we briefly discuss how they work. An analysis is beyond the scope of AICC-2.

EXERCISE

There are 14 billiard balls numbered as shown:



Among balls 1 - 13, at most one **could** be heavier/lighter than the others.

What is the minimum number of weightings to simultaneously determine:

- if one ball is different . . .
- if there is such a ball, which one, . . .
- and whether the different ball is heavier/lighter.