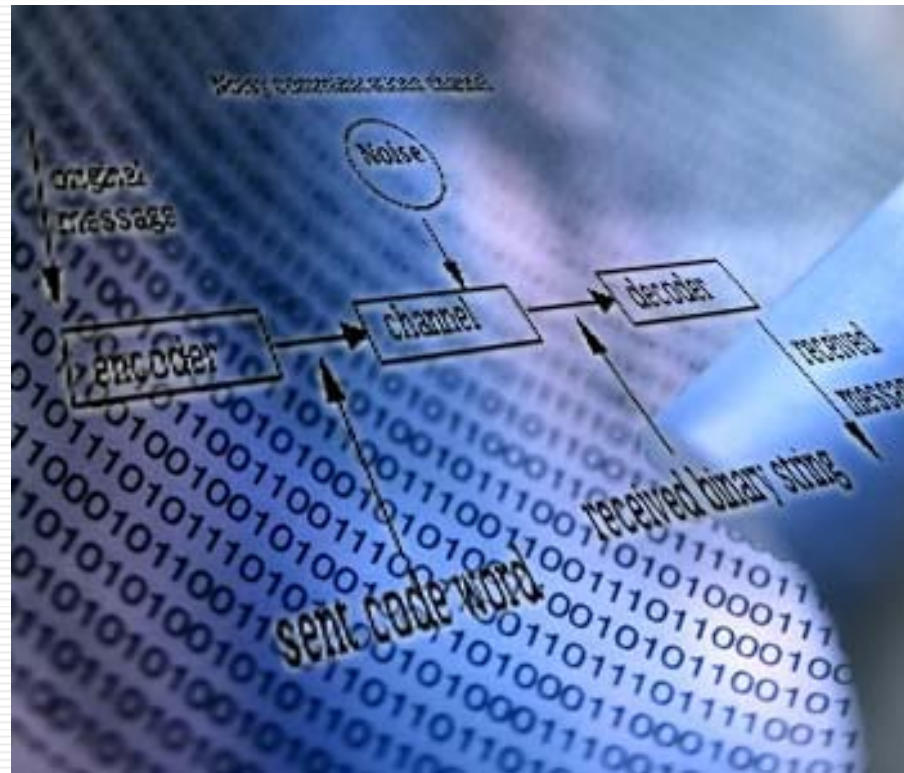


Kapitel 3: Entropie



- Wir erinnern uns: Um eine Zufallsvariable mit N verschiedenen, gleichwahrscheinlichen Zuständen binär zu codieren, benötigen wir

$$\lceil \log_2 N \rceil \text{ Bits} \qquad \lceil -\log_2 p_N \rceil \text{ Bits}$$

- Die Information steht in direktem Zusammenhang mit der **Unsicherheit (Entropie)** über den Ausgang eines Zufallsexperimentes
- Wie kann diese Unsicherheit quantitativ erfasst werden?
- Statt einer direkten Definition stellen wir eine Reihe von Anforderungen auf

- Die Unsicherheit über ein Experiment soll unabhängig von der Nomenklatur sein und nur von den Wahrscheinlichkeiten der Elementarereignisse abhängen

Beispiel: Beim fairen Münzwurf soll die Unsicherheit gleich gross sein, egal ob wir die Ereignisse „Kopf“ und „Zahl“ oder „0“ und „1“ nennen

- Die Unsicherheit über ein Experiment soll unabhängig von der Nomenklatur sein und nur von den Wahrscheinlichkeiten der Elementarereignisse abhängen
- Die Unsicherheit ist eine Funktion H , welche jeder Wahrscheinlichkeitsverteilung eine **reelle** Zahl zuordnet, im Falle endlicher Zufallsexperimente also jeder Liste $[p_1, \dots, p_L]$ sich zu 1 summierender Zahlen
- Ohne Beschränkung der Allgemeinheit können wir daher eine solche Liste mit $L \geq 1$ Elementen als **geordnet** auffassen, also

$$p_1 \geq p_2 \geq \dots \geq p_L$$

- Ereignisse mit der Wahrscheinlichkeit 0 sollen keinen Einfluss haben:

$$H([p_1, \dots, p_L]) = H([p_1, \dots, p_L, 0])$$

- Allgemein soll gelten, dass für Experimente mit gleichwahrscheinlichen Ereignissen die Entropie mit der Anzahl der möglichen Ereignisse zunimmt:

$$H\left(\left[\frac{1}{L}, \dots, \frac{1}{L}\right]\right) < H\left(\left[\frac{1}{L+1}, \dots, \frac{1}{L+1}\right]\right)$$

- Die Entropie eines Münzwurfes soll umso kleiner sein, je unfairer oder asymmetrischer die Münze ist, und ist maximal für einen fairen Münzwurf

$$p < q \leq 1/2 \Rightarrow H([p, 1-p]) < H([q, 1-q])$$

- $H([p_1, \dots, p_L])$ ist maximal, wenn $p_1 = \dots = p_L = 1/L$

- Die Entropie eines Experimentes, welches aus zwei unabhängigen Einzelexperimenten besteht, soll gleich der Summe der Einzelentropien sein

$$H\left(\left[\frac{1}{LM}, \dots, \frac{1}{LM}\right]\right) = H\left(\left[\frac{1}{L}, \dots, \frac{1}{L}\right]\right) + H\left(\left[\frac{1}{M}, \dots, \frac{1}{M}\right]\right)$$

- Wenn $L=M=1$, dann gilt $H(1)=0$



Die Entropie eines Experimentes mit nur einem einzigen möglichen Ausgang ist also 0. Sie enthält keine Information

- **Normierung:** Es ist sinnvoll, die Entropie eines fairen Münzwurfs auf 1 zu normieren, da man ein Bit benötigt, um das Resultat darzustellen

$$H\left(\left[\frac{1}{2}, \frac{1}{2}\right]\right) = 1$$

- **Glattheit:** Kleine Änderungen in der Wahrscheinlichkeitsverteilung sollen nur kleine Änderungen in der Entropie bewirken

- Man kann zeigen, dass die einzige Funktion, welche alle diese Forderungen erfüllt, wie folgt definiert sein muss
- **Definition 1:** Die Entropie einer diskreten Wahrscheinlichkeitsverteilung $[p_1, \dots, p_L]$ ist:

$$H([p_1, \dots, p_L]) = -\sum_{i=1}^L p_i \log_2 p_i$$

Beispiel: Die Entropie der dreiwertigen Verteilung $[0.7, 0.27655, 0.02345]$ ist 1 bit, wie beim fairen Münzwurf

- Besonders einfach ist die Entropieberechnung, wenn alle Wahrscheinlichkeiten negative Zweierpotenzen sind!

- Die Zufallsvariable X nehme die Zustände $\{x_1, \dots, x_L\}$ mit den Wahrscheinlichkeiten $p_X(x_i)$ an
- **Definition 2:** Die Entropie einer diskreten Zufallsvariablen X ist:

$$H(X) = - \sum_{i=1}^L p_X(x_i) \log_2 p_X(x_i)$$

- Anmerkung: Auch für $L=\infty$ kann die obige Summe einen endlichen Wert annehmen
- Die Menge $\{x_1, \dots, x_L\}$ aller Zustände von X nennen wir auch das **Alphabet** von X

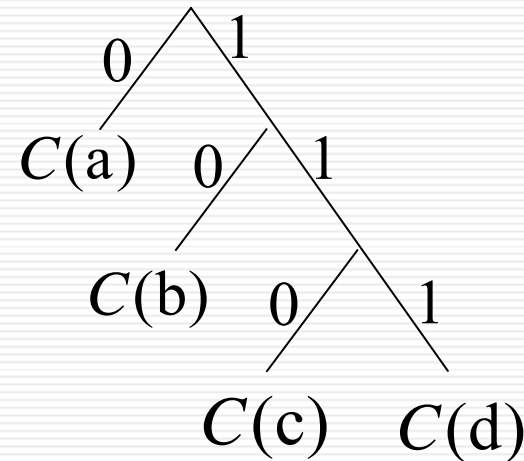
Informationsquelle 1

- Gegeben: Informationsquelle mit Alphabet $\{a,b,c,d\}$

$$p(a) = \frac{1}{2} \quad p(b) = \frac{1}{4} \quad p(c) = \frac{1}{8} \quad p(d) = \frac{1}{8}$$

- Ziel: möglichst optimale Codierung (Huffman)

$$\begin{aligned} C(a) &= 0 \\ C(b) &= 10 \\ C(c) &= 110 \\ C(d) &= 111 \end{aligned}$$

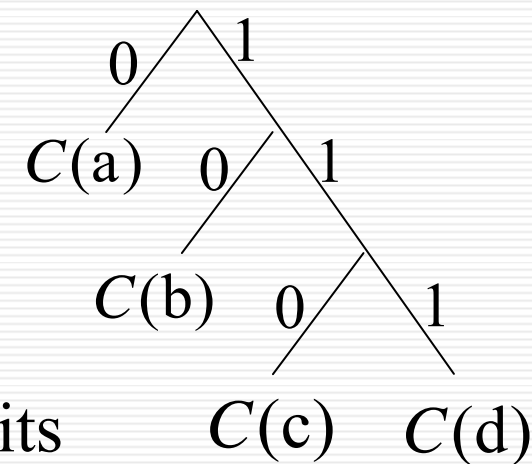


- Entropie:

$$H(X) = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 \cdot 2 = 1.75 \text{ Bits}$$

- Mittlere Codelänge:

$$\begin{aligned} L(C) &= E[l(C)] \\ &= \sum_{i=a}^d l(C(i)) \cdot p(C(i)) \\ &= \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 \cdot 2 = 1.75 \text{ Bits} \end{aligned}$$



Informationsquelle 2

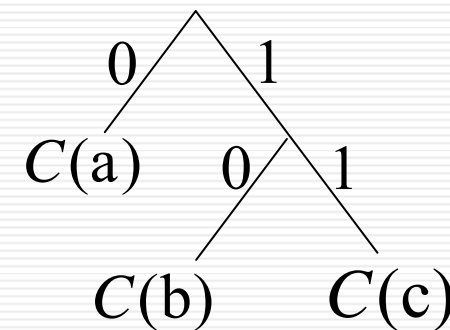
- Alphabet $\{a,b,c\}$

$$p(a) = \frac{1}{3} \quad p(b) = \frac{1}{3} \quad p(c) = \frac{1}{3}$$

- $C(a) = 0, C(b) = 10, C(c) = 11$

- $H(X) = \sum_{i=1}^3 \frac{1}{3} \log_2 3 = \log_2 3 = 1.58 \text{ Bits}$

- $$L(C) = \sum_{i=a}^c l(C(i)) \cdot p(C(i))$$
$$= \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 2 \cdot 2 = 1.66 \text{ Bits}$$



- Beides sind präfixfreie Codes

acdaab = 01101110010

ist eindeutig dekodierbar

- Minimale Anzahl Fragen zur Bestimmung von X
 - Ist $X=a$?
 - Ist $X=b$?

- Erwartungswert

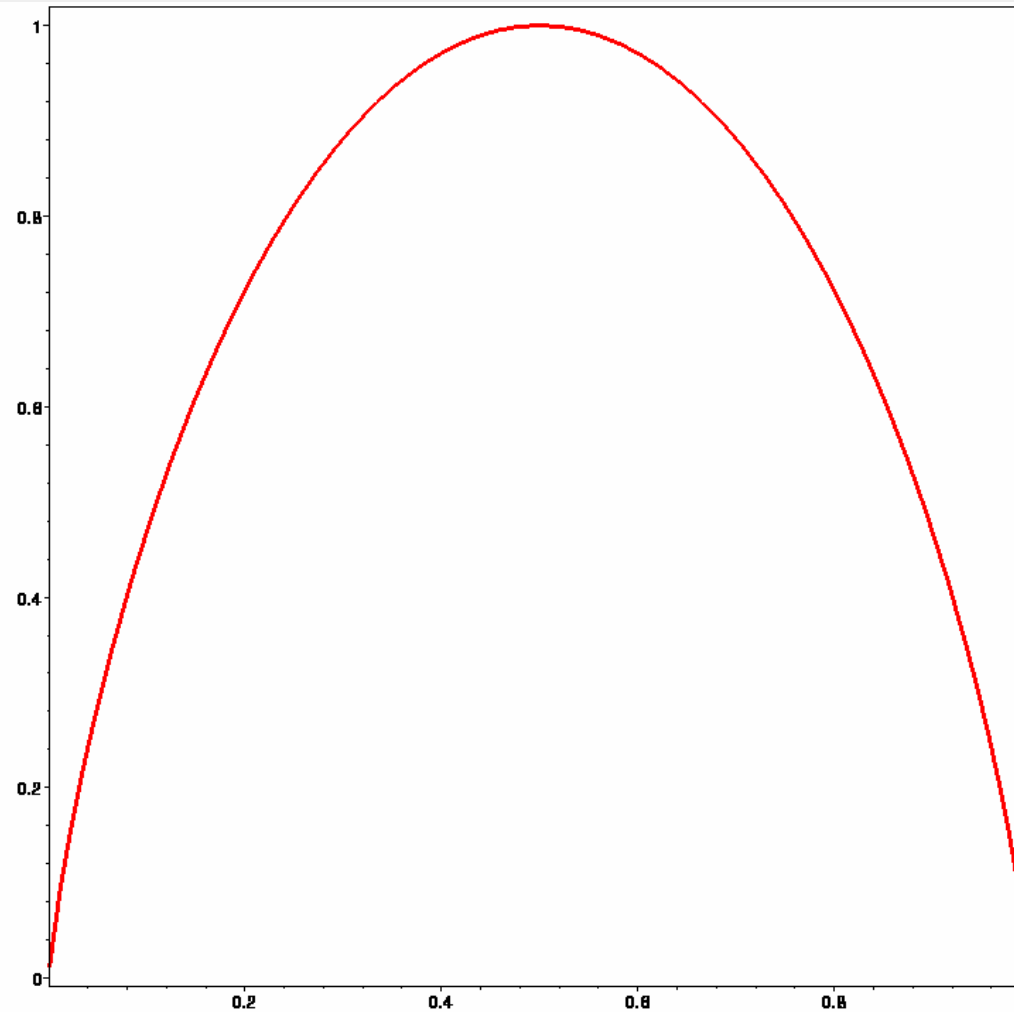
$$H(X) < E_{F_{\min}}[X] < H(X) + 1$$

- Die Wahrscheinlichkeitsverteilung einer binären Zufallsvariablen X ist durch $p_X(0) = p$ vollständig beschrieben, da gilt: $p_X(1) = 1 - p$
- Die Entropie kann also als Funktion von p aufgefasst werden

$$h(p) = -p \log_2 p - (1 - p) \log_2 (1 - p)$$

- Sie ist für $0 < p < 1$ sowie $h(0) = h(1) = 0$ definiert
- $h(p)$ ist strikt konkav und besitzt ein Maximum bei $p = 1/2$ (fairer Münzwurf)

Binäre Entropiefunktion



Entropie als Erwartungswert **ETH**

- Anmerkung: Wir verwenden die Konvention

$$0 \cdot \log_2(0) = 0$$

- Dennoch gilt, dass Werte mit der Wahrscheinlichkeit 0 von der Betrachtung ausgeschlossen werden
- Dann kann die Entropie auch als Erwartungswert einer reellwertigen Funktion aufgefasst werden:

$$H(X) = E[-\log_2 P_X(X)]$$

- **Theorem:** Es sei $\chi = \{x_1, \dots, x_L\}$ die Menge der möglichen Zustände der Zufallsvariablen X .

Dann gilt:

$$0 \leq H(X) \leq \log_2 |\chi|$$

oder auch

$$0 \leq H([p_1, \dots, p_L]) \leq \log_2 L$$

- Gleichheit gilt auf der linken Seite, wenn $p_X(x) = 1$ für genau ein x
- Gleichheit gilt auf der rechten Seite, wenn $p_1 = \dots = p_L = 1/L$

- **Beweis:** Der **linke** Teil der Ungleichung folgt direkt aus der Tatsache, dass die Funktion

$$f(x) = -x \log_2 x$$

für $0 < x < 1$ streng positiv ist und nur für $x = 1$ gemäss Konvention gleich 0 ist

- Die rechte Ungleichung folgt aus der **Jensen-Ungleichung** und der Tatsache, dass die Funktion

$$f(x) = \log_2 x$$

konkav ist

- Es gilt:

$$H(x) = E \left[\log_2 \frac{1}{P_X(X)} \right] \leq \log_2 \left(E \left[\frac{1}{P_X(X)} \right] \right) = \log_2 |\mathcal{X}|$$

- Der letzte Schritt folgt aus

$$E \left[\frac{1}{P_X(X)} \right] = \sum_{i=1}^L P_X(X) \frac{1}{P_X(X)} = |\mathcal{X}|$$

- **Konvention:** Im Folgenden wird die Basis 2 für den Logarithmus angenommen und nicht mehr explizit geschrieben

- **Definition:** Eine Funktion $f(x)$ heisst **konvex** auf einem Intervall $[a, b]$, wenn für alle

$x_1, x_2 \in [a, b], x_1 \neq x_2$ und $\lambda \in [0, 1]$ gilt:

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

- Eine Funktion ist strikt konvex, wenn Gleichheit nur für $\lambda=0$ und $\lambda=1$ gilt
- Eine Funktion f ist **konkav**, wenn $-f$ konvex ist
- Der Graph einer differenzierbaren, konvexen Funktion liegt immer oberhalb jeder Tangente

- Existiert $f''(x)$ einer Funktion f auf einem offenen oder geschlossenen Intervall $[a, b]$ und gilt $f''(x) > 0$, dann ist f konvex

- Taylorreihe von f um x_0

$$f(x) = f(x_0) + f'(x_0) \cdot (x - x_0) + f''(x^*) \cdot \frac{(x - x_0)^2}{2}$$

- für $x^* \in [x_0 \dots x]$ gilt $f''(x^*) \geq 0$

- sei $x_0 = \lambda x_1 + (1 - \lambda)x_2$

- $x = x_1$

$$\rightarrow f(x_1) \geq f(x_0) + f'(x_0) \cdot (1 - \lambda) \cdot (x_1 - x_2) \mid \cdot \lambda$$

- $x = x_2$

$$\rightarrow f(x_2) \geq f(x_0) + f'(x_0) \cdot \lambda \cdot (x_2 - x_1) \mid \cdot (1 - \lambda)$$

$$\oplus \rightarrow f(\lambda \cdot x_1 + (1 - \lambda) \cdot x_2) \leq \lambda \cdot f(x_1) + (1 - \lambda) \cdot f(x_2)$$

- **Theorem:** Für eine konvexe Funktion f und eine Zufallsvariable X gilt:

$$E[f(X)] \geq f(E[X])$$

- Entsprechend gilt für eine konkave Funktion g und eine Zufallsvariable X

$$E[g(X)] \leq g(E[X])$$

Beweis Jensen-Ungleichung **ETH**

- Wir gehen davon aus, dass f mindestens einmal differenzierbar ist
- Sei $ax+b$ die Tangente an f im Punkt $x=E[X]$
- Wir ersetzen die Funktion in x durch Ihre Tangente
- Dann gilt aufgrund der Linearität von E

$$f[E(X)] = aE[X] + b = E[aX + b] \leq E[f(X)]$$

- Dies gilt aufgrund der Tatsache, dass der Graph der konvexen Funktion immer oberhalb der Tangente liegt

- Es sei (X, Y) ein Paar von Zufallsvariablen
- Die gemeinsame Entropie (**Verbundentropie**) zweier Zufallsvariablen X und Y ist gegeben durch

$$H(XY) = - \sum_{(x,y)} p_{XY}(x, y) \log p_{XY}(x, y) = E[-\log p_{XY}(X, Y)]$$

- **Theorem:** Es gilt:

$$H(X) \leq H(XY)$$

- Gleichheit gilt genau dann, wenn Y durch Kenntnis von X eindeutig bestimmt ist, Y also keine neue Information enthält, also für ein y gilt

$$p_{YX}(x, y) = 1$$

- $H(X)$ und $H(XY)$ sind Erwartungswerte:

$$H(XY) = E[-\log p_{XY}(X, Y)]$$

- Im Rahmen der Verbundwahrscheinlichkeiten zweier Zufallsvariablen haben wir folgende Gesetzmässigkeit kennengelernt:

$$p_{XY}(x, y) \leq p_X(x)$$

- Dies gilt für alle möglichen Zustandspaare (x, y)

- Folgendes Theorem ist von zentraler Bedeutung:

- **Theorem:** Es gilt

$$H(XY) \leq H(X) + H(Y)$$

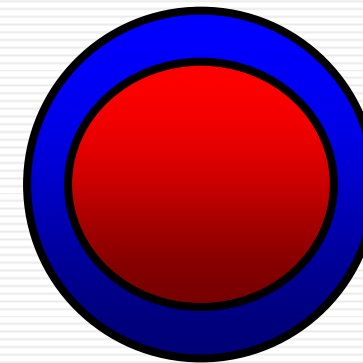
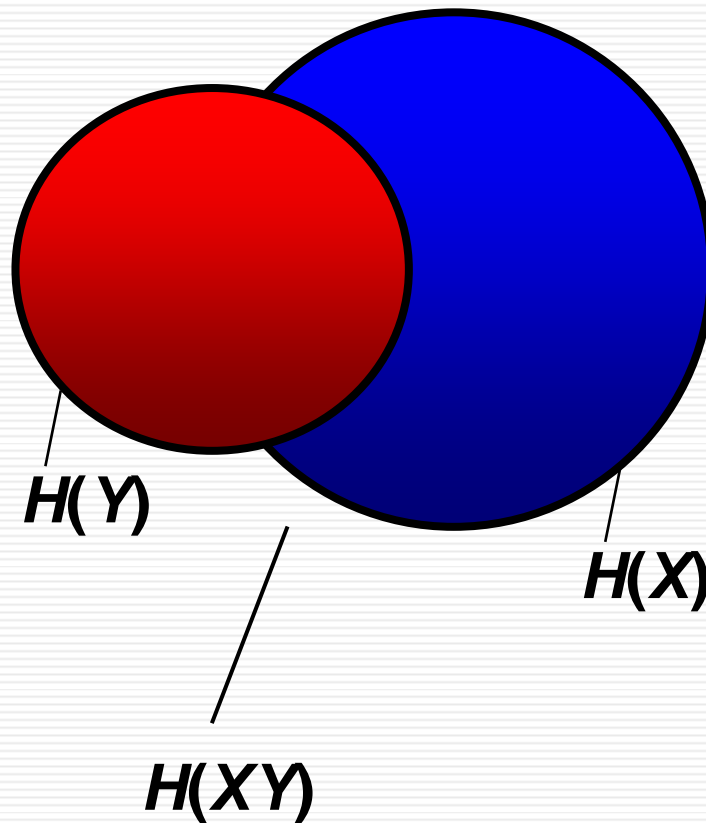
- Gleichheit gilt genau dann, wenn Y und X statistisch unabhängig sind
- Der Beweis erfolgt durch Einsetzen sowie Anwendung der Jensen-Ungleichung:

$$H(X) + H(Y) - H(XY) =$$
$$E[-\log p_X(X) - \log p_Y(Y) + \log p_{XY}(X, Y)]$$

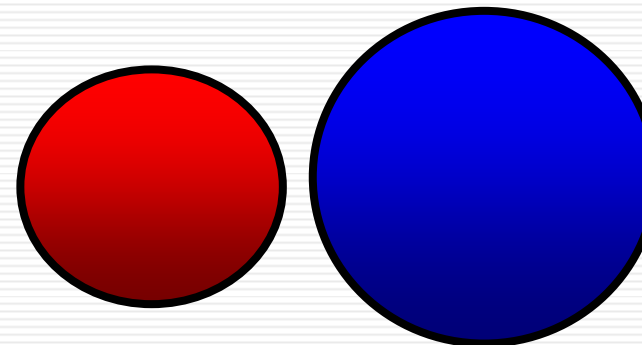


Die Entropie des Verbundereignisses XY ist höchstens so gross, wie die Summe der Entropien der Einzelereignisse.

Bild dazu



$$H(XY) = H(X)$$



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