*Lehrveranstaltung*

# Informationstheorie

— *Sommersemester 2023* —

Martin Mittelbach (Vorlesung, Tutorium), Anne Wolf (Übung, Tutorium)
{martin.mittelbach, anne.wolf}@tu-dresden.de

Professur für Informationstheorie und maschinelles Lernen, TU Dresden

*Vorlesung 3*
*19. April 2023*

# Inhalt der letzten Vorlesung

- Fortsetzung Grundbegriffe diskrete $\mathbb{W}$-Theorie
- Intuitive Einführung zur Datenkompression

---

- 1. Verlustlose Datenkompression mit Codes variabler Länge
- (1.1) Einführendes Beispiel, Modellbildung, Problemstellung
- (1.2) Quellen als Datenmodell

---

# Wiederholung

- **Mathematisches Datenmodell:** Folge diskreter Zufallsgrößen = Quelle

$$X_1,\ X_2,\ X_3,\ X_4,\ \ldots$$

- **Spezielle Quellen:**
  - stationäre Quellen
  - (stationäre) gedächtnislose Quellen
  - (stationäre) Markow-Quellen

| Informationstheorie | | Wahrscheinlichkeitstheorie |
|---|---|---|
| • (diskrete) Quelle | = | Folge diskreter Zufallsgrößen |
| • (diskrete) stationäre Quelle | = | stationäre Folge diskreter Zufallsgrößen |
| • (diskrete) gedächtnislose Quelle | = | unabhängige Folge diskreter Zufallsgrößen |
| • (diskrete) stationäre gedächtnislose Quelle | = | i.i.d.-Folge diskreter Zufallsgrößen |
| • (diskrete) Markow-Quelle | = | Markowkette diskreter Zufallsgrößen |
| • (diskrete) stationäre Markow-Quelle | = | stationäre Markowkette diskreter Zufallsgrößen |

# Inhalt Vorlesung 3

- 1. Verlustlose Datenkompression mit Codes variabler Länge
- (1.3) Codes variabler Länge
- (1.4) Huffman-Codes

---

- 2. Informationsmaße für diskrete Zufallsgrößen
- (2.1) Definition der Shannonschen Informationsmaße

---

# (1.3) Codes variabler Länge

- **Festlegungen:**
  - Im Folgenden werden wir nur

    stationäre Quellen

    betrachten und uns zunächst auf die

    Codierung der Werte einzelner Zufallsgrößen

    beschränken, die bei stationären Quellen alle identisch verteilt sind.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Quelle (Folge von Zufallsgrößen): | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | ... |
| mögliche Werte der Zufallsgrößen: | 1 | 2 | 1 | 1 | 3 | 2 | 1 | 4 | ... |
| | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | ... |
| Codierung: | 0 | 10 | 0 | 0 | 110 | 10 | 0 | 111 | ... |

  - Mathematisches Modell: Wir betrachten die

    (diskrete) stationäre Quelle $X = (X_k)_{k \in \mathbb{N}}$

    mit dem Alphabet $\mathcal{X} = \{1, 2, \ldots, M\}$ und der für alle Folgeglieder $X_k$ identischen $\mathbb{W}$-Funktion $p$.

# (1.3) Codes variabler Länge

- **$D$-wertiger Code:**
  - Ein $D$-wertiger (Quellen-) Code $\mathcal{C}$ für das Alphabet $\mathcal{X}$ ist eine Menge $\mathcal{C} = \{c(i) : i \in \mathcal{X}\}$, wobei jedes Codewort $c(i)$ eine endliche Folge von Symbolen eines $D$-wertigen Codealphabets $\{a_1, a_2, \ldots, a_D\}$ ist.
  - Die Codierung eines Wertes $i \in \mathcal{X}$ mit dem Code $\mathcal{C}$ erfolgt durch die Zuordnung

$$i \mapsto c(i).$$

Beispiel: Im einführenden Beispiel in Teilabschnitt (1.1) haben wir für das Alphabet $\mathcal{X} = \{1, 2, 3, 4\}$ folgende Quellencodes mit dem $2$-wertigen Codealphabet $\{0, 1\}$ betrachtet.

Variante 1: $\mathcal{C} = \{ \underset{c(1)}{\underset{||}{00}}, \underset{c(2)}{\underset{||}{01}}, \underset{c(3)}{\underset{||}{10}}, \underset{c(4)}{\underset{||}{11}} \}$ Variante 2: $\mathcal{C} = \{ \underset{c(1)}{\underset{||}{0}}, \underset{c(2)}{\underset{||}{10}}, \underset{c(3)}{\underset{||}{110}}, \underset{c(4)}{\underset{||}{111}} \}$

- **Mittlere Codewortlänge:** Ist $\ell(i)$ die Länge des Codewortes $c(i) \in \mathcal{C}$, so heißt

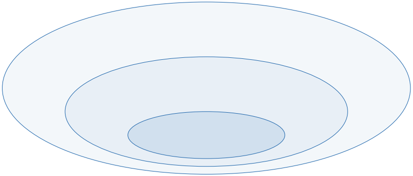
$$\bar{\ell} = \bar{\ell}(\mathcal{C}, p) = \sum_{i=1}^{M} p(i)\, \ell(i)$$

mittlere Codewortlänge des Codes $\mathcal{C}$ für die $\mathbb{W}$-Funktion $p$.

# (1.3) Codes variabler Länge

- **Codeklassen:** Ein Code $\mathcal{C} = \{c(i) : i \in \mathcal{X}\}$ heißt
  - nicht-singulär, falls die Zuordnung $c(i) \mapsto i$ für alle $i \in \mathcal{X}$ eindeutig ist, d. h. die (Codier-) Funktion $c : \mathcal{X} \to \mathcal{C}$ bijektiv ist.
  - eindeutig decodierbar, falls jede endliche Aneinanderreihung von Codewörtern ohne Trennzeichen rekonstruiert werden kann.
  - präfixfrei, falls kein Codewort Anfang eines anderen Codewortes ist.

- **Relationen zwischen Codeklassen:**
  - Präfixfreie Codes sind stets eindeutig decodierbar und
  - eindeutig decodierbare Codes sind stets nicht-singulär.



Wir sind ausschließlich an eindeutig decodierbaren Codes interessiert und innerhalb dieser Codeklasse insbesondere an den präfixfreien Codes.

# (1.3) Codes variabler Länge

- **Minimale mittlere Codewortlänge:** Die minimale mittlere Codewortlänge
  - eindeutig decodierbarer Codes für die $\mathbb{W}$-Funktion $p$ bezeichnen wir mit

$$\bar{\ell}^*_{\mathrm{ud}} = \bar{\ell}^*_{\mathrm{ud}}(p) = \min\left\{\bar{\ell}(\mathcal{C}, p) : \mathcal{C} \text{ eindeutig decodierbar}\right\}.$$

  - präfixfreier Codes für die $\mathbb{W}$-Funktion $p$ bezeichnen wir mit

$$\bar{\ell}^*_{\mathrm{pre}} = \bar{\ell}^*_{\mathrm{pre}}(p) = \min\left\{\bar{\ell}(\mathcal{C}, p) : \mathcal{C} \text{ präfixfrei}\right\}.$$

  - Offenbar gilt:

$$\bar{\ell}^*_{\mathrm{ud}}(p) \;\; ? \leq ? \geq ? \;\; \bar{\ell}^*_{\mathrm{pre}}(p).$$

- **Optimale Codes:**
  - Ein eindeutig decodierbarer Code $\mathcal{C}$ ist optimal für die $\mathbb{W}$-Funktion $p$, falls gilt:

$$\bar{\ell}(\mathcal{C}, p) = \bar{\ell}^*_{\mathrm{ud}}(p).$$

  - Ein präfixfreier Code $\mathcal{C}$ ist optimal für die $\mathbb{W}$-Funktion $p$, falls gilt:

$$\bar{\ell}(\mathcal{C}, p) = \bar{\ell}^*_{\mathrm{pre}}(p).$$

Achtung: Im Skript werden optimale Codes nur für die Klasse der präfixfreien Codes betrachtet.

# (1.3) Codes variabler Länge

- **Beispiel:** Wir betrachten das einführende Beispiel aus Teilabschnitt (1.1) für verschiedene 2-wertige (binäre) Codierungen.
  - Mittlere Codewortlängen:

| | | Code $\mathcal{C}_1$ | | Code $\mathcal{C}_2$ | | Code $\mathcal{C}_3$ | |
|---|---|---|---|---|---|---|---|
| $i$ | $p(i)$ | $c(i)$ | $\ell(i)$ | $c(i)$ | $\ell(i)$ | $c(i)$ | $\ell(i)$ |
| 1 | $\frac{1}{2}$ | 0 | 1 | 10 | 2 | 0 | 1 |
| 2 | $\frac{1}{4}$ | 01 | 2 | 00 | 2 | 10 | 2 |
| 3 | $\frac{1}{8}$ | 010 | 3 | 11 | 2 | 110 | 3 |
| 4 | $\frac{1}{8}$ | 10 | 2 | 110 | 3 | 111 | 3 |
| $\bar{\ell}$ | | 1.625 Bit / Symbol | | 2.125 Bit / Symbol | | 1.75 Bit / Symbol | |

  - Für die Zeichenfolge aus Beispiel (1.1) ergeben sich folgende Codierungen.

| Zeichenfolge: | 1 | 2 | 1 | 1 | 3 | 2 | 1 | 4 | ... |
|---|---|---|---|---|---|---|---|---|---|
| Codierung mit Code $\mathcal{C}_1$: | 0 | 01 | 0 | 0 | 010 | 01 | 0 | 10 | ... |
| Codierung mit Code $\mathcal{C}_2$: | 10 | 00 | 10 | 10 | 11 | 00 | 10 | 110 | ... |
| Codierung mit Code $\mathcal{C}_3$: | 0 | 10 | 0 | 0 | 110 | 10 | 0 | 111 | ... |

# (1.3) Codes variabler Länge

- **Beispiel:** Wir betrachten das einführende Beispiel aus Teilabschnitt (1.1) für verschiedene 2-wertige (binäre) Codierungen.
  - Für die Zeichenfolge aus Beispiel (1.1) ergeben sich folgende Codierungen.

| Zeichenfolge: | 1 | 2 | 1 | 1 | 3 | 2 | 1 | 4 | ... |
|---|---|---|---|---|---|---|---|---|---|
| Codierung mit Code $\mathcal{C}_1$: | 0 | 01 | 0 | 0 | 010 | 01 | 0 | 10 | ... |
| Codierung mit Code $\mathcal{C}_2$: | 10 | 00 | 10 | 10 | 11 | 00 | 10 | 110 | ... |
| ■ Codierung mit Code $\mathcal{C}_3$: | 0 | 10 | 0 | 0 | 110 | 10 | 0 | 111 | ... |

  - Code $\mathcal{C}_1$ ist nicht eindeutig decodierbar, da die Bitfolge beispielsweise auch in die Zeichenfolge 1 3 1 1 4 2 1 4 . . . decodiert werden kann.
  - Code $\mathcal{C}_2$ ist eindeutig decodierbar aber nicht präfixfrei, denn $c(3)$ ist Präfix von $c(4)$. Die mit $\uparrow$ markierte Stelle lässt sich dadurch erst durch Einbeziehen zukünftiger Bits eindeutig decodieren.
  - Code $\mathcal{C}_3$ ist präfixfrei und das Ende jedes Codewortes ist unmittelbar erkennbar.
  - Es gilt $\bar{\ell}(\mathcal{C}_1, p) \leq \bar{\ell}(\mathcal{C}_3, p)$. Für die in diesem Beispiel gegebene $\mathbb{W}$-Funktion $p$ werden wir in Kürze jedoch $\bar{\ell}^*_{\text{ud}}(p) \geq 1.75$ Bit zeigen.
  - Damit wäre Code $\mathcal{C}_3$ für die $\mathbb{W}$-Funktion $p$ ein optimaler eindeutig decodierbarer Code und, weil er präfixfrei ist, auch ein optimaler präfixfreier Code.

# (1.3) Codes variabler Länge

- **Kriterium für eindeutig decodierbare Codes:**
  - Definition: Sei $\mathcal{C}$ ein $D$-wertiger Code. Eine endliche Folge $s$ von Symbolen des $D$-wertigen Codealphabets heißt Suffix in $\mathcal{C}$, falls:

(i) $\exists c(i), c(j) \in \mathcal{C} : c(i) = c(j)s$ oder

(ii) $\exists c(i) \in \mathcal{C}$ und ein Suffix $\hat{s}$ in $\mathcal{C}$ : $\hat{s} = c(i)s$ oder

(iii) $\exists c(i) \in \mathcal{C}$ und ein Suffix $\hat{s}$ in $\mathcal{C}$ : $c(i) = \hat{s}s$.

  - Kriterium: (ohne Beweis)

$\mathcal{C}$ ist ein eindeutig decodierbarer Code $\Longleftrightarrow$ Kein Suffix ist Codewort in $\mathcal{C}$.

# (1.3) Codes variabler Länge

- **Kriterium für eindeutig decodierbare Codes:**
  - Definition: ... $s$ ... Suffix in $\mathcal{C}$ ...

(i) $\exists c(i), c(j) \in \mathcal{C} : c(i) = c(j)s$ oder

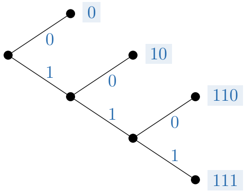
(ii) $\exists c(i) \in \mathcal{C}$ und ein Suffix $\hat{s}$ in $\mathcal{C}$ : $\hat{s} = c(i)s$ oder

(iii) $\exists c(i) \in \mathcal{C}$ und ein Suffix $\hat{s}$ in $\mathcal{C}$ : $c(i) = \hat{s}s$.

- Beispiele 2-wertiger Codierungen:
  - Code $\mathcal{C}_1 = \{0, 01, 010, 10\}$
  - Code $\mathcal{C}_2 = \{10, 00, 11, 110\}$
  - Code $\mathcal{C}_3 = \{0, 10, 110, 111\}$
  - Code $\mathcal{C}_4 = \{11, 11010, 01\}$

# (1.3) Codes variabler Länge

- **Codebaum eines präfixfreien Codes:**
  - Codebäume sind nützliche grafische Hilfsmittel zur Veranschaulichung von präfixfreien Codes.
  - Durch einen $D$-wertigen Codebaum lässt sich ein $D$-wertiger präfixfreier Code repräsentieren.
  - In diesem Codebaum hat jeder Knoten höchstens $D$ direkte Nachfolgeknoten und die Endknoten repräsentieren die Codewörter.
  - Die nachfolgende Grafik zeigt den Codebaum für Code $\mathcal{C}_3 = \{0, 10, 110, 111\}$ aus dem vorhergehenen Beispiel.

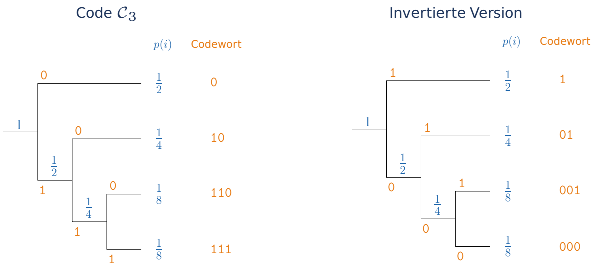


# (1.4) Huffman-Codes

- Huffman-Codes sind spezielle präfixfreie Codes.
- Die Konstruktion eines $D$-wertigen Huffman-Codes für das Alphabet $\mathcal{X}$ basiert auf der $\mathbb{W}$-Funktion $p$ und folgt einer einfachen Regel:

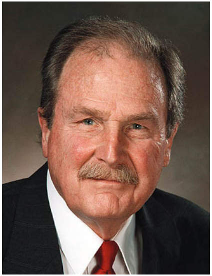
> Fasse stufenweise solange die $D$ kleinsten $\mathbb{W}$en zusammen,
> bis nur noch eine $\mathbb{W}$ (= 1) übrig bleibt.

- Durch diesen Konstruktionsmechanismus wird ein Codebaum erzeugt.
- Für die $\mathbb{W}$-Funktion $p$ für das Alphabet $\mathcal{X}$ aus dem einführenden Beispiel (1.1) lässt sich die Konstruktion eines binären Huffman-Codes wie folgt veranschaulichen.



# (1.4) Huffman-Codes

- **Optimalität:** Wir werden in Kürze u. a. zeigen, dass Huffman-Codes optimale präfixfreie Codes für diejenige $\mathbb{W}$-Funktion sind, für die sie jeweils konstruiert wurden.
- **Historisches:**
  - Der Huffman-Code ist nach David Albert Huffman benannt.
  - Diese Codierung wurde von Huffman 1951 im Rahmen einer Seminararbeit (Massachusetts Institute of Technology) entwickelt.



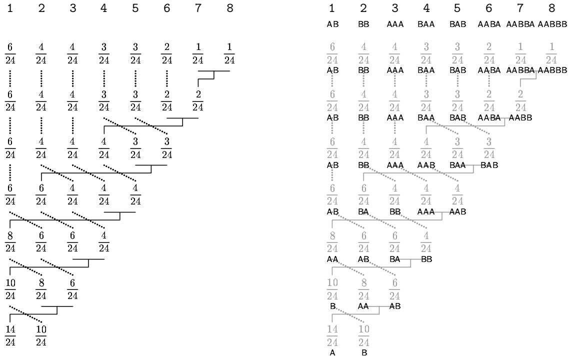
David A. Huffman (1925 – 1999)

# (1.4) Huffman-Codes

- **Beispiel: 2-wertiger Huffman-Code**
  - Gegeben ist das Alphabet $\mathcal{X} = \{1, 2, \ldots, 8\}$ und die $\mathbb{W}$-Funktion $p$ mit

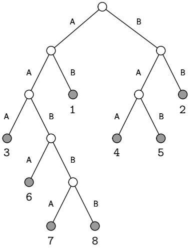
| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $p(i)$ | $\frac{1}{4}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{12}$ | $\frac{1}{24}$ | $\frac{1}{24}$ |

  - Zur Codierung soll das 2-wertige Alphabet $\{A, B\}$ verwendet werden.
  - Codebaumerzeugung:



# (1.4) Huffman-Codes

- **Beispiel: 2-wertiger Huffman-Code**
  - Codebaum für den konstruierten Huffman-Code $\mathcal{C} = \{$AB, BB, AAA, BAA, BAB, AABA, AABBA, AABBB$\}$:




  - (Mittlere) Codewortlänge(n):

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $p(i)$ | $\frac{1}{4}$ | $\frac{1}{6}$ | $\frac{1}{6}$ | $\frac{1}{8}$ | $\frac{1}{8}$ | $\frac{1}{12}$ | $\frac{1}{24}$ | $\frac{1}{24}$ |
| $c(i)$ | AB | BB | AAA | BAA | BAB | AABA | AABBA | AABBB |
| $\ell(i)$ | 2 | 2 | 3 | 3 | 3 | 4 | 5 | 5 |

$$\bar{\ell} = \bar{\ell}(\mathcal{C}, p) = \left(\left(\frac{1}{4} + \frac{1}{6}\right) \cdot 2 + \left(\frac{1}{6} + \frac{1}{8} + \frac{1}{8}\right) \cdot 3 + \frac{1}{12} \cdot 4 + \left(\frac{1}{24} + \frac{1}{24}\right) \cdot 5\right)$$

$= 2.833$ Bit / Quellensymbol

# Vorgehensweise / Themenübersicht

(1) Praktische Problemstellung

systematischer Entwurf
effizienter Datenkompression

systematischer Entwurf
effizienter zuverlässiger Datenübertragung

(2) Abstraktion / Mathematische Modellbildung

stochastisches Datenmodell
Quelle = Folge von Zufallsgrößen
deterministisches Kompressionsmodell
Quellen-Codier-Decodier-Schemata
operationelle Bewertungskenngrößen
(minimale) mittlere Codewortlänge / Coderate, Distortionmaß, Rate-Distortion-Funktion

stochastisches Übertragungsmodell
Kanal = bedingte Wahrscheinlichkeitsverteilung
deterministisches Kommunikationsmodell
Kanal-Codier-Decodier-Schemata
operationelle Bewertungskenngrößen
(maximale) Coderate, Decodierfehlerwahrscheinlichkeit, Codierkapazität

(3) Entwicklung mathematischer Werkzeuge für Modellanalyse

Informationsmaße
Entropie (bedingte, relative, differentielle, empirische ∼), Transinformation (bedingte, empirische ∼)
Eigenschaften, Rechenregeln
(Symmetrie, Nichtnegativität), Kettenregeln, Konvexität / Konkavität
Unabhängigkeit / (Gleichverteilung (Normalverteilung)) maximiert Entropie (differentielle ∼)
Ungleichungen, grundlegende Theoreme
Kraft-McMillan-Ungleichung, fundamentale Ungleichung, Datenverarbeitungsungleichung, Fano-Ungleichung
Shannon-McMillan-Theorem, Feinstein-Lemma
Informationsmaßebasierte Kenngrößen
Redundanz, Entropierate, Informations-Rate-Distortion-Funktion, Informationskapazität

(4) Herleitung fundamentaler Resultate für abstrahiertes Modell mit math. Werkzeugen

Quellencodierungstheoreme
Aussagen über theoretische Grenzen effizienter Datenkompression (bei definierter Güte)
Sinngemäß: Verlustlose (asymptotisch ∼) Datenkompression ist begrenzt durch Entropie(-rate)
(Entropie(-rate) $\leq$ min. mittl. Cwl (pro Sym.) / Coderate)
Sinngemäß: Verlustbehaftete Datenkompression ist begrenzt durch Informations-Rate-Distortion-Funktion
(Rate-Distortion-Funktion = Informations-∼)

Kanalcodierungstheoreme
Aussagen über theoretische Grenzen effizienter Datenübertragung bei definierter Zuverlässigkeit
Sinngemäß: Maximale Coderate bei zuverlässiger Datenübertragung ist begrenzt durch Informationskapazität
(Codierkapazität = Informationskapazität)

(5) Anwendung der Resultate auf praktische Problemstellung

Entwurfskriterien für optimale Datenkompression / -übertragung, Systembewertung

# 2. Informationsmaße für diskrete Zufallsgrößen

# Häufigkeitsverteilung im Skript

Für uns relevant sind statistische Kenngrößen von Daten. Verwendet man das Skript als Menge von Textdaten und ermittelt die häufigsten Wörter und mathematischen Symbole, so ergibt sich folgendes Bild. Die Schriftgröße innerhalb der "Wortwolke" ist proportional zur Häufigkeit.


Diese Wortwolke enthält die 75 häufigsten Wörter / Symbole im Skript ohne Kapitel 0, ausgenommen sind Füllwörter, Artikel usw. Die Wortwolke wurde mit Wordle™ (www.wordle.net) erzeugt.

# Festlegungen und Notation

- In diesem Abschnitt betrachten wir Informationsmaße für folgende Zufallsgrößen:

| | |
|---|---|
| $X$ | diskrete Zufallsgröße mit endlichem Alphabet $\mathcal{X}$ und Träger $\mathcal{S}_X$ |
| $Y$ | diskrete Zufallsgröße mit endlichem Alphabet $\mathcal{Y}$ und Träger $\mathcal{S}_Y$ |
| $Z$ | diskrete Zufallsgröße mit endlichem Alphabet $\mathcal{Z}$ und Träger $\mathcal{S}_Z$ |

mit den (gemeinsamen / bedingten) $\mathbb{W}$-Funktionen

| | | | |
|---|---|---|---|
| $p_{X,Y,Z}$ | gemeinsame $\mathbb{W}$-Funktion von | $X, Y, Z$ | |
| $p_{X,Y}$ | gemeinsame $\mathbb{W}$-Funktion von | $X, Y$ | |
| $p_X$ | $\mathbb{W}$-Funktion von | $X$ | |
| $p_Y$ | $\mathbb{W}$-Funktion von | $Y$ | |
| $p_Z$ | $\mathbb{W}$-Funktion von | $Z$ | |
| $p_{Y\|X}$ | bedingte $\mathbb{W}$-Funktion von | $Y$ | unter der Bedingung $X$ |
| $p_{X,Y\|Z}$ | bedingte $\mathbb{W}$-Funktion von | $(X, Y)$ | unter der Bedingung $Z$ |
| $p_{X\|Z}$ | bedingte $\mathbb{W}$-Funktion von | $X$ | unter der Bedingung $Z$ |
| $p_{Y\|Z}$ | bedingte $\mathbb{W}$-Funktion von | $Y$ | unter der Bedingung $Z$ |

# Festlegungen und Notation

- Wir definieren sämtliche Informationsmaße für den Logarithmus $\log_2$ zur Basis $2$. Die damit verbundene Einheit ist $\mathrm{bit}$.
- Je nach Anwendungsfall kann auch eine andere Basis $D$ für den Logarithmus sinnvoll sein. Falls wir davon Gebrauch machen, weisen wir explizit darauf hin.
- Für sämtliche Definitionen gelten folgende Konventionen.

$$
\begin{aligned}
0\log_2\frac{0}{a} &= 0 && \text{falls} \quad a \geq 0\\
0\log_2\frac{a}{0} &= +\infty && \text{falls} \quad a > 0
\end{aligned}
$$

- Wir beschränken uns auf diskrete Zufallsgrößen mit endlichem Alphabet. Eine Verallgemeinerung auf ein abzählbar unendliches Alphabet ist fast immer möglich. Eine Einschränkung auf ein endliches Alphabet genügt für unsere Anwendungen jedoch und vereinfacht die mathematischen Betrachtungen.
- Für Implikationen und Äquivalenzen verwenden wir die Notation $\Longrightarrow$ und $\Longleftrightarrow$.

  A $\Longrightarrow$ B bedeutet, aus A folgt B.

  A $\Longleftrightarrow$ B bedeutet, aus A folgt B und aus B folgt A.

- Mit $|\mathcal{X}|$ bezeichnen wir die Anzahl der Elemente in der Menge $\mathcal{X}$.

# Übersicht Shannonsche Informationsmaße

- Entropie, gemeinsame Entropie

$$H(X), H(X,Y)$$

- bedingte Entropie

$$H(Y|X)$$

- Transinformation

$$I(X;Y)$$

- bedingte Transinformation

$$I(X;Y|Z)$$

- relative Entropie

$$D(X||Y)$$

# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.1) Entropie:** Die Größe

$$H(X) = -\sum_{x \in \mathcal{X}} p_X(x) \log_2 p_X(x)$$

heißt Entropie der Zufallsgröße $X$ (auch: Shannon-Entropie) und die Größe

$$H(X,Y) = -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x,y) \log_2 p_{X,Y}(x,y)$$

heißt gemeinsame Entropie der Zufallsgrößen $X$ und $Y$.

**Bemerkungen:**

- Alternative Notation: $H(p_X)$ statt $H(X)$ und $H(p_{X,Y})$ statt $H(X,Y)$, da die Entropie nur von der $\mathbb{W}$-Funktion abhängt.
- Die Größe $\log_2 \frac{1}{p_X(x)}$ wird auch als "Informationsgehalt" des Ereignisses $X = x$ bezeichnet.

kleine $\mathbb{W} \Longrightarrow$ großer Informationsgehalt

große $\mathbb{W} \Longrightarrow$ kleiner Informationsgehalt

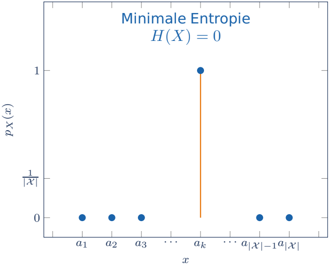
Damit entspricht die Entropie $H(X)$ auch dem mittleren Informationsgehalt der Zufallsgröße $X$.

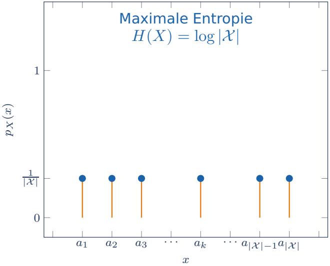
# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.1) Entropie:** $[\ldots]$

  **Bemerkungen:** $[\ldots]$

  - Die Entropie ist ein quantitatives Maß für den Grad der Zufälligkeit / Unbestimmtheit einer Zufallsgröße. Dabei ist die Zufälligkeit minimal für eine deterministische Größe (d. h. $H(X) = 0$) und maximal für eine Gleichverteilung (d. h. $H(X) = \log|\mathcal{X}|$). Siehe dazu (2.4.2) und (2.5.1).




  - Nachrichtentechnische Relevanz: Die Entropie ist u. a. eine untere Schranke für die (asymptotisch) verlustlose Datenkompression für gedächtnislose Modelle.

# (2.1) Definition der Shannonschen Informationsmaße

- **Zitate zur Entropie:**

***Shannon:*** *"My greatest concern was what to call it. I thought of calling it 'information', but the word was overly used, so I decided to call it 'uncertainty'. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, you should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."*

aus: M. Tribus, E. C. McIrvine: *Energy and information*, Scientic American, 224 (September 1971), 178-184.

***Jaglom:*** *"(...) Der wirkliche Wert des Begriffs Entropie wird in erster Linie dadurch bestimmt, daß der durch ihn ausgedrückte 'Grad der Unbestimmtheit' von Versuchen bei den mannigfaltigsten Prozessen in der Natur und Technik auftritt, die alle in irgendeiner Weise mit der Übertragung und Speicherung von Nachrichten zusammenhängen. (...)"*

A. M. Yaglom
(Russischer Mathematiker)

aus: A. M. Jaglom und I. M. Jaglom: *Wahrscheinlichkeit und Information*, Dt. Verl. d. Wiss., Berlin, 1988, S. 69.

# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.2) Bedingte Entropie:** Für alle $x \in \mathcal{X}$ heißt die Größe

$$H(Y|X = x) = H\Big(p_{Y|X}(\cdot|x)\Big) = -\sum_{y \in \mathcal{Y}} p_{Y|X}(y|x) \log_2 p_{Y|X}(y|x)$$

die bedingte Entropie von $Y$ unter der Bedingung $X = x$. Die Größe

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p_X(x) H(Y|X = x) \\ &= -\sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x, y) \log_2 p_{Y|X}(y|x) \end{aligned}$$

heißt die bedingte Entropie von $Y$ unter der Bedingung $X$.

**Bemerkungen:**

- Die Größe $H(Y|X = x)$ ist ein quantitatives Maß für den Grad der Unbestimmtheit der Zufallsgröße $Y$, wenn bereits feststeht, dass die Zufallsgröße $X$ den Wert $x$ angenommen hat.
- Dementsprechend quantifiziert $H(Y|X)$ die mittlere Unbestimmtheit der Zufallsgröße $Y$ bei Kenntnis der Werte der Zufallsgröße $X$.
- Nachrichtentechnische Relevanz: Die bedingte Entropie ist u. a. eine untere Schranke für die (asymptotisch) verlustlose Datenkompression für spezielle gedächtnisbehaftete Modelle.

# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.3) Transinformation:** Die Größe

$$
\begin{aligned}
I(X;Y) &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y)\log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \\
&= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_X(x)p_{Y|X}(y|x)\log_2 \frac{p_{Y|X}(y|x)}{\sum_{\tilde{x}\in\mathcal{X}} p_X(\tilde{x})p_{Y|X}(y|\tilde{x})}
\end{aligned}
$$

heißt Transinformation zwischen (den Zufallsgrößen) $X$ und $Y$.

**Bemerkungen:**

- Alternative Notation: $I(p_X, p_{Y|X})$ statt $I(X;Y)$ immer dann, wenn die Abhängigkeit von der $\mathbb{W}$-Funktion $p_X$ bzw. bedingten $\mathbb{W}$-Funktion $p_{Y|X}$ relevant ist.
- Man beachte die Verwendung des Semikolons bei $I(X;Y)$.
- Die Transinformation $I(X;Y)$ ist ein quantitatives Maß für den Grad der stochastischen Abhängigkeit der Zufallsgrößen $X$ und $Y$. Sie ist minimal, d. h. gleich $0$, genau dann, wenn $X$ und $Y$ unabhängig sind. Siehe dazu (2.4.4).
- Nachrichtentechnische Relevanz: Mit der Transinformation erhält man eine obere Schranke für die Datenrate für eine zuverlässige Datenübertragung bei gedächtnislosen Modellen.

# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.4) Bedingte Transinformation:** Für alle $z \in \mathcal{Z}$ heißt die Größe

$$I(X;Y|Z=z) = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y|Z}(x,y|z)\log_2\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}$$

die bedingte Transinformation zwischen (den Zufallsgrößen) $X$ und $Y$ unter der Bedingung $Z=z$. Die Größe

$$\begin{aligned} I(X;Y|Z) &= \sum_{z\in\mathcal{Z}} p_Z(z) I(X;Y|Z=z) \\ &= \sum_{(x,y,z)\in\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}} p_{X,Y,Z}(x,y,z)\log_2\frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)} \end{aligned}$$

heißt die bedingte Transinformation zwischen (den Zufallsgrößen) $X$ und $Y$ unter der Bedingung $Z$.

**Bemerkungen:**

- Die bedingte Transinformation $I(X;Y|Z)$ ist ein quantitatives Maß für den Grad der bedingten stochastischen Abhängigkeit der Zufallsgrößen $X$ und $Y$ gegeben $Z$. Sie ist minimal, d. h. gleich $0$, genau dann, wenn $X$ und $Y$ bedingt unabhängig gegeben $Z$ sind. Siehe dazu (2.4.5).
- Die bedingte Transinformation ist in dieser Lehrveranstaltung primär als mathematisches Hilfsmittel von Bedeutung.

# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.5) Darstellung als Erwartungswert transformierter Zufallsgrößen:**
  - Wir definieren folgende Funktionen.

$$
\begin{aligned}
g_1(x) &= -\log_2 p_X(x) & \qquad g_2(x,y) &= -\log_2 p_{X,Y}(x,y)\\
g_3(x,y) &= -\log_2 p_{Y|X}(y|x) & & \\
g_4(x,y) &= \log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} & \qquad g_5(x,y,z) &= \log_2 \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}\\
g_6(x) &= \log_2 \frac{p_X(x)}{p_Y(x)} & &
\end{aligned}
$$

  - Für die Informationsmaße gilt dann gemäß der Definitionen:

$$
\begin{aligned}
H(X) &= \mathbb{E}\Big(g_1(X)\Big) & \qquad H(X,Y) &= \mathbb{E}\Big(g_2(X,Y)\Big)\\
H(Y|X) &= \mathbb{E}\Big(g_3(X,Y)\Big) & &\\
I(X;Y) &= \mathbb{E}\Big(g_4(X,Y)\Big) & \qquad I(X;Y|Z) &= \mathbb{E}\Big(g_5(X,Y,Z)\Big)\\
D(X||Y) &= \mathbb{E}\Big(g_6(X)\Big) & &
\end{aligned}
$$

  - **Achtung:** $p_X(x)$ ist der Funktionswert der $\mathbb{W}$-Funktion $p_X$ für $x \in \mathcal{X}$. Aber $p_X(X)$ ist eine Zufallsgröße, die aus der Transformation der Zufallsgröße $X$ mit der Funktion $p_X$ resultiert.

# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.6) Darstellung mit Hilfe der relativen Entropie:**
  - (i) Relative Entropie: Gilt $\mathcal{X} \subset \mathcal{Y}$, dann heißt die Größe

$$D(X||Y) = \begin{cases} \sum_{x \in \mathcal{X}} p_X(x) \log_2 \frac{p_X(x)}{p_Y(x)} & \text{falls } \mathcal{S}_X \subseteq \mathcal{S}_Y \\ \infty & \text{sonst} \end{cases}$$

relative Entropie zwischen (den Zufallsgrößen) $X$ und $Y$ (auch: Kullback-Leibler-Abstand).

**Bemerkungen:**

- Alternative Notation: $D(p_X||p_Y)$ statt $D(X||Y)$.
- Die relative Entropie $D(X||Y)$ ist ein quantitatives Maß für den Unterschied der $\mathbb{W}$-Verteilungen der Zufallsgrößen $X$ und $Y$. Sie ist minimal, d. h. gleich 0, genau dann, wenn $X$ und $Y$ identisch verteilt sind. Siehe dazu (2.4.1).
- Die bisher eingeführten Informationsmaße lassen sich mit Hilfe der relativen Entropie darstellen. Dadurch ist die relative Entropie sehr hilfreich bei der Herleitung informationstheoretischer Eigenschaften.
- Nachrichtentechnische Relevanz: Die relative Entropie quantifiziert u. a. den zusätzlich erforderlichen Speicherbedarf bei suboptimaler Quellencodierung.

- **(2.1.6) Darstellung mit Hilfe der relativen Entropie:**
  - (ii) Entropie: ($p_U$: $\mathbb{W}$-Funktion einer Gleichverteilung auf $\mathcal{X}$)

$$\boxed{H(X) = H(p_X) = \log_2 |\mathcal{X}| - D(p_X || p_U)}$$

  - (iii) Bedingte Entropie: ($p_{\tilde{U}}$: $\mathbb{W}$-Funktion einer Gleichverteilung auf $\mathcal{Y}$)

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p_X(x) H\Big(p_{Y|X}(\cdot|x)\Big) \\ &= \log_2 |\mathcal{Y}| - \sum_{x \in \mathcal{X}} p_X(x) D\Big(p_{Y|X}(\cdot, x) || p_{\tilde{U}}\Big) \end{aligned}$$

  - (iv) Transinformation:

$$\boxed{I(X;Y) = D(p_{X,Y} || p_X \cdot p_Y)}$$

  - (v) Bedingte Transinformation:

$$\begin{aligned} I(X;Y|Z) &= \sum_{z \in \mathcal{Z}} p_Z(z) I(X;Y|Z=z) \\ &= \sum_{z \in \mathcal{Z}} p_Z(z) D\Big(p_{X,Y|Z}(\cdot,\cdot|z) || p_{X|Z}(\cdot|z) p_{Y|Z}(\cdot|z)\Big) \end{aligned}$$

# (2.1) Definition der Shannonschen Informationsmaße

- **Herleitungen zu (2.1.6):**

# (2.1) Definition der Shannonschen Informationsmaße

- **(2.1.7) Informationsmaße als Spezialfälle:**
  - (i) Transinformation als Spezialfall der bedingten Transinformation:

$$I(X;Y) = I(X;Y|Z) \qquad \text{falls } Z \text{ deterministisch}$$

  - (ii) Entropie als Spezialfall der bedingten Entropie:

$$H(X) = H(X|Z) \qquad \text{falls } Z \text{ deterministisch}$$

  - (iii) Bedingte Entropie als Spezialfall der bedingten Transinformation:

$$H(X|Z) = I(X;X|Z)$$

  - (iv) Entropie als Spezialfall der Transinformation:

$$H(X) = I(X;X)$$

Eine diskrete Zufallsgröße $X$ mit Alphabet $\mathcal{X}$ heißt deterministisch, falls sie mit $\mathbb{W}$ eins konstant ist, d. h. falls ein $x \in \mathcal{X}$ existiert mit $\mathbb{P}(X = x) = 1$.

- **(2.1.8) Diskrete Zufallsvektoren**
  - Die betrachteten diskreten Zufallsgrößen $X$, $Y$ und $Z$ in diesem Abschnitt kann man auch durch diskrete Zufallsvektoren ersetzen, da diskrete Zufallsvektoren auch als diskrete Zufallsgrößen mit kartesischem Produkt als Alphabet aufgefasst werden können (siehe Bemerkung Vorlesung 1, Folie 33).

    Beispielsweise:

    diskrete Zufallsgröße $X_1$ mit Alphabet $\mathcal{X}_1 = \{a_1, a_2, \ldots, a_m\}$

    diskrete Zufallsgröße $X_2$ mit Alphabet $\mathcal{X}_2 = \{b_1, b_2, \ldots, b_n\}$

    Der Zufallsvektor $(X_1, X_2)$ lässt sich auffassen als

    diskrete Zufallsgröße $X$ mit Alphabet $\mathcal{X} = \{(a_1, b_1), (a_1, b_2), \ldots, (a_m, b_n)\}$

  - Für einen $n$-dimensionalen diskreten Zufallsvektor $X = (X_1, X_2, \ldots, X_n)$ ergibt sich beispielsweise

$$H(X) = H\Big((X_1, X_2, \ldots, X_n)\Big)$$
$$H(Y|X) = H\Big(Y|(X_1, X_2, \ldots, X_n)\Big)$$
$$I(X;Y) = I\Big((X_1, X_2, \ldots, X_n); Y\Big).$$

  - Die violettmarkierten Klammern werden nachfolgend weggelassen.