Institut für Nachrichtentechnik

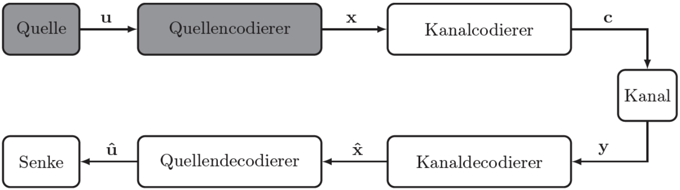
Grundlagen der Informationstechnik - Nachrichtentechnik

Vorlesung: Eduard A. Jorswieck
Übung: Dr. Bile Peng

Wintersemester 2023-2024, 2. November 2023

Quellen und Quellencodierung I

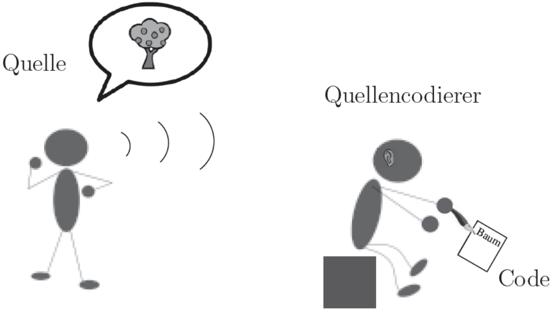
Kapitel 4 in M. Bossert 'Einführung in die Nachrichtentechnik'



Quelle und Quellencodierung im Modell der Informationstheorie

- Als *Quelle* bezeichnen wir alles, was diskrete Daten liefert - erinnere *Digitalisierung*. Quellen werden *statistisch* beschrieben.
- Die Quelle produziert Wörter $\boldsymbol{u}$, welche der *Quellencodierer* in *Codewörter* $\boldsymbol{x}$ umwandelt.

Quellen und Quellencodierung II



Schrift als Beispiel für Quellencodierung

Quellen und Quellencodierung III

- Schrift ist eine sehr effiziente Quellencodierung, die akustische Signale als geschriebene Text beschreibt.
- Die Wörter bezeichnen wir als *Codewörter*, die der Codierer liefert.
- Diese besitzen variable Länge und bestehen aus *Symbolen* aus einem diskreten endlichen *Alphabet*.
- Die Symbole haben unterschiedliche relative Häufigkeiten (erinnere Morsecode).
- Allerdings bestehen Abhängigkeiten zwischen Buchstabenfolgen, z.B. nach `sc` folgt sehr wahrscheinlich ein `h`.
- Bei einer Quelle nennt man so etwas *Gedächtnis*.

Quellen und Quellencodierung IV

- Die Quelle liefert in diesem Beispiel ein akustisches Signal, das der Quellencodierer mit Hilfe seiner Ohren empfängt und in Codewörter umsetzt, die auf ein Blatt Papier geschrieben werden.
- Es handelt sich zunächst um *verlustbehaftete Codierung*, da Information verloren geht (z.B. über die Lautstärke des Sprechers).
- Der Code enthält jedoch genug Information, um den Sinn zu verstehen. (Das schafft der Quellencodierer sogar ohne den Inahlt (Semantik) und die Regeln der Sprache (Grammatik) zu verstehen.
- Offensichtlich enthält die Schrift *Redundanz*, weil bestimmte Wörter nicht existieren. $\Longleftrightarrow$ Wir können die Sprache mit weniger Buchstaben erneut quellencodieren.

Quellen und Quellencodierung V

- Wir benötigen ein Maß für die Redundanz, das Sematik und Grammatik ignoriert und führen dafür eine *diskrete, gedächtnislose Quelle* ein, die zufällig Buchstaben aus einem endlichen Alphabet erzeugt.

$\Longrightarrow$ **Quellencodierungstheorem**

Quellencodierungstheorem I

Beispiel: Quellencodierung

Gegeben sei eine gedächtnislose Quelle, die zufällig einen der Buchstaben $\{a, b, c, d\}$ erzeugt. Die Buchstaben treten mit den folgenden Wahrscheinlichkeiten auf: $P(a) = 1/2$, $P(b) = 1/4$ und $P(c) = P(d) = 1/8$.

Verschiedene Codierungsansätze:

Institut für Nachrichtentechnik

Quellencodierungstheorem II

1. Eine einfache Möglichkeit der Codierung besteht nun darin, jedem Symbol ein Codewort bestehend aus $\log_2 4 = 2$ Bits zuzuordnen. Eine mögliche, eindeutig umkehrbare Zuordnung ist

   $$a \to 00, \quad b \to 01, \quad c \to 10, \quad d \to 11.$$

   Bei dieser Codierung werden $2n$ Bits benötigt, um eine Zeichenkette der Länge n zu codieren.

2. Effizienter wäre es, wenn wir den Buchstaben unterschiedlich lange Codewörter zuordnen:

   $$a \to 0, \quad b \to 10, \quad c \to 110, \quad d \to 111.$$

Quellencodierungstheorem III

Definition: Präfix

Gegeben sei ein Codewort $\boldsymbol{x}_i$ der Länge w_i mit $\boldsymbol{x}_i = [x_{i1}, x_{i2}, ..., x_{iw_i}]$.
Als *Präfix* bezeichnet man alle Vektoren
$[x_{i1}], [x_{i1}, x_{i2}], ..., [x_{i1}, x_{i2}, ..., x_{iw_{i-1}}]$.

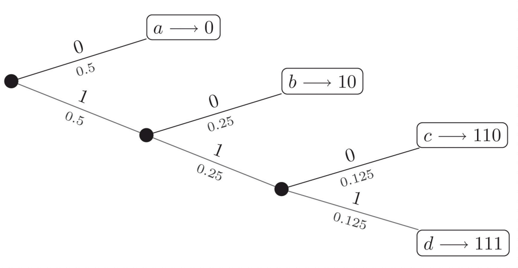
Satz: Umkehrbare Codierung

Eine Codierung durch Codewörter mit variabler Länge ist umkehrbar eindeutig, wenn der Code präfixfrei ist, d.h., kein Codewort Präfix eines anderen Codewortes ist.

Quellencodierungstheorem IV

Beispiel: Codebaum

In der Abbildung ist ein Codebaum für den Beispielcode variabler Länge von oben dargestellt.



Der Code ist präfixfrei, weil genau ein Pfad von der Wurzel zu jedem Enknoten (Codewort) führt.

Quellencodierungstheorem V

Satz: Kraft-Ungleichung

Ein D-närer präfixfreier Code mit den Codewortlängen $w_1, \ldots, w_L$ existiert genau dann, wenn gilt:

$$\sum_{i=1}^{L} D^{-w_i} \leqslant 1.$$

- Das Quellencodierungstheorem gibt an, wie viele Bits man mindestens benötigt, um die Symbole einer bestimmten Quelle zu codieren.
- Dafür führen wir im folgenden die Unsicherheit ein, die auch als *Entropie* bezeichnet wird.

Quellencodierungstheorem VI

Beispiel: Bar-Kochba-Spiel - Umfrage: `https://www.vote.ac`

Wir spielen auf einer Party mit 33 Personen das folgende Spiel: Eine Person, der Rater, verlässt den Raum, und die anderen 32 Personen einigen sich auf eine bestimmte Person, den Erwählten. Der Rater soll nun den Erwählten mit möglichst wenigen Fragen finden. Wie viele Fragen, auf die nur mit ja oder nein geantwortet werden darf, muss der Rater mindestens stellen, um den Erwählten sicher zu finden?

- Mit Wahrscheinlichkeit $1/32$ ist eine zufällig gewählte Person der Erwählte. Wenn der Rater eine Person nach der anderen zufällig auswählt, so kann es vorkommen, dass er bis zu 31 Fragen benötigt.

Quellencodierungstheorem VII

- Mit folgender Methode benötigt der Rater jedoch nur genau 5 Fragen:
  1. Er teilt die Personen in zwei gleich große Gruppen ein und stellt einer Gruppe die Frage: Ist der Erwählte in dieser Gruppe?
  2. Lautet die Antwort ja, wird diese Gruppe, andernfalls die andere Gruppe in zwei weitere gleich große Gruppen aufgeteilt.
  3. Die Aufteilung in Gruppen wird solange wiederholt bis nur noch eine Person in der entsprechenden Gruppe ist.
- Im Allgemeinen beträgt die Anzahl an notwendigen Fragen bei L Personen (Hartley 1933)

$$H(L) = \log_2(L).$$

Quellencodierungstheorem VIII

Nun formen wir um: $H(L) = -\sum_{i=1}^{L} \frac{1}{L} \log_2 \left(\frac{1}{L}\right)$.

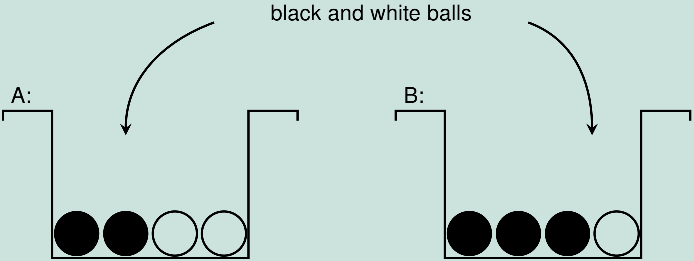
Definition: Shannons Unsicherheit, Entropie

Gegeben sei eine gedächtnislose Quelle mit dem Symbolalphabet $\mathcal{U} = \{u_1, \ldots, u_L\}$ und den Auftrittswahrscheinlichkeiten $P(u_i) = \mathbb{P}(U = u_i) = p_i$, $i = 1, \ldots, L$ mit $\sum_{i=1}^{L} p_i = 1$. Dann ist die *Unsicherheit* definiert durch

$$H(U) = -\sum_{i=1}^{L} p_i \log p_i.$$

Quellencodierungstheorem IX

Beispiel zu Entropie



Wir ziehen zufällig Bälle aus Urne A oder B in der Abbildung.
Berechne die Unsicherheit (Entropie)!

Quellencodierungstheorem X

Definition: Binäre Unsicherheit, binäre Entropie

Gegeben sei eine binäre Zufallsvariable X aus dem Alphabet $\{0, 1\}$ mit $P(0) = 1 - p$ und $P(1) = p$. Die binäre Unsicherheit berechnet sich zu

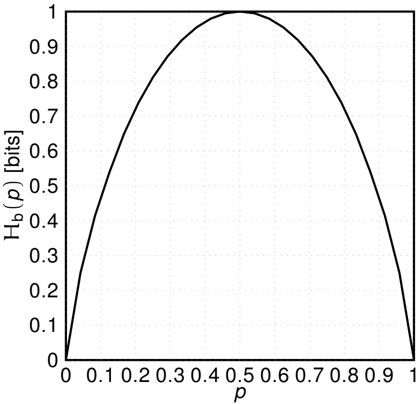
$$H(X) = -p \log_2(p) - (1 - p) \log_2(1 - p) = H_B(p).$$

$\Longrightarrow$ Diskutiere die Spezialfälle:

$p = 0, p = 1$ Kein Zufall, deterministisch, Unsicherheit ist gleich Null

$p = 1/2 = 1 - p$ Gleichverteilt, sehr unsicher, Unsicherheit ist gleich Eins

Quellencodierungstheorem XI



Quellencodierungstheorem XII

Satz: Schranken der Unsicherheit

Ist X eine Zufallsvariable mit dem Alphabet $\mathcal{U} = \{u_1, ..., u_K\}$ und den Auftrittswahrscheinlichkeiten $P(u_i) = p_i$, $i = 1, 2, ..., L$ mit $\sum_{i=1}^{L} p_i = 1$, so gilt für die Unsicherheit

$$0 \leqslant H(X) = -\sum_{i=1}^{L} p_i \log_2(p_i) \leqslant \log_2(L).$$

Quellencodierungstheorem XIII

Satz: Unsicherheit von Vektoren

Seien X und Y statistisch unabhängige Zufallsvariablen mit den Alphabeten $\mathcal{U} = \{u_1, ..., u_L\}$ und $\mathcal{V} = \{v_1, ..., v_L\}$ und den Auftrittswahrscheinlichkeiten $P(u_i)$ und $P(v_j)$. Dann errechnet sich die Unsicherheit des Vektors $\boldsymbol{Z} = [X, Y]$ zu

$$H(\boldsymbol{Z}) = H(X) + H(Y).$$

Quellencodierungstheorem XIV

Satz: Shannons Quellencodierungstheorem

Gegeben sei eine gedächtnislose Quelle, beschrieben durch die Zufallsvariable X mit dem Alphabet $\mathcal{U} = \{u_1, u_2, ..., u_L\}$ und den Auftrittswahrscheinlichkeiten $P(u_i) = p_i$ und der Unsicherheit $H(X)$. Die Quelle liefert für die Sequenz $\boldsymbol{X} = [X_1, X_2, ..., X_n]$ mit statistisch unabhängigen $X_j \sim X$.

Für $n \to \infty$ kann die Quelle im Mittel mit $H(X)$ Bits pro Symbol verlustlos codiert werden. Man benötigt also für die Sequenz der Länge n näherungsweise $H(\boldsymbol{X}) = nH(X)$ Bits.