1. Woche
Einführung in die Codierungstheorie, Definition Codes, Präfixcode, kompakte Codes

Unser Modell

- Shannon 1948: Informationstheorie und Mathematik der Kommunikation
- Hamming 1950: Erste Arbeit über fehlerkorrigierende Codes

Modell:

Sender $\rightarrow$ Codierer $\rightarrow$ Kanal $\rightarrow$ Decodierer $\rightarrow$ Empfänger

- Kanal ist bandbreitenbeschränkt (Kompression)
- Kanal ist fehleranfällig (Fehlerkorrektur)
  - Bits können ausfallen: $0 \mapsto \epsilon$, $1 \mapsto \epsilon$
  - Bits können kippen: $0 \mapsto 1$, $1 \mapsto 0$

Motivierendes Bsp: Datenkompression

Szenario:

- Kanal ist **fehlerfrei**.
- Übertragen gescannte Nachricht:
  Wahrscheinlichkeiten: 99% weißer, 1% schwarzer Punkt.
- Weiße Punkte erhalten Wert 0, schwarze Wert 1.

Codierer:

- Splitten Nachricht in Blocks der Größe 10.
- Wenn Block x=0000000000, codiere mit 0, sonst mit 1x.
- 1 dient als Trennzeichen beim Decodieren.

Decodierer:

- Lese den Code von links nach rechts.
- Falls 0, decodiere 0000000000.
- Falls 1, übernehme die folgenden 10 Symbole.

Erwartete Codelänge

Sei $q := \mathrm{Ws}[\text{Block ist } 0000000000] = (0.99)^{10} \geq 0.9$.
Sei Y Zufallsvariable für die Codewortlänge eines 10-Bit Blocks:

$$E[Y] = \sum_{y \in \{0,1x\}} |y| \cdot \mathrm{Ws}(Y = y) = 1 \cdot q + 11 \cdot (1 - q) = 11 - 10q.$$

- D.h. erwartete Länge der Codierung eines 10-Bit Blocks ist
$$11 - 10q \leq 2 \text{ Bit}.$$
- Datenkompression der Nachricht auf 20%.
- Können wir noch stärker komprimieren?
- Entropie wird uns Schranke für Komprimierbarkeit liefern.

Exkurs: Weitere Motivation

Skalarmultiplikation, i.e. für $g \in G$ Gruppe und $n \in \mathbb{Z}$, rechne das Vielfache $n \cdot g$ von g.

Interessant für Kryptographie (RSA, Elliptische Kurven).

Es gibt Methoden, die basieren auf Datenkompression.

Idee: Betrachte n als Bitfolge (binäre Darstellung), und komprimiere diese – interpretiere diese Kompression als Folge von Operationen, deren End-Resultat $n \cdot g$ ist.

Bocharova-Kudryasoul, und Yacobi – leider werden wir dies nicht in der VL betrachten.

Ausblick: fehlerkorrigierende Codes

Szenario: Binärer symmetrischer Kanal

- Bits 0,1 kippen mit Ws $p, p < \frac{1}{2}$ zu 1,0. (Warum $< \frac{1}{2}$?)
- Korrekte Übertragung $0 \mapsto 0$, $1 \mapsto 1$ mit Ws $1 - p$.
- In unserem Beispiel $p = 0.1$.

Codierer:

- Verdreifache jedes Symbol, d.h. $0 \mapsto 000$, $1 \mapsto 111$
- Repetitionscode der Länge 3.

Decodierer:

- Lese den Code in 3er-Blöcken.
- Falls mindestens zwei Symbole 0 sind, decodiere zu 0.
- Sonst decodiere zu 1.

Ws Decodierfehler

Symbol wird falsch decodiert, falls mind. zwei der drei Bits kippen.

$$\begin{aligned} & \text{Ws}(\text{Bit wird falsch decodiert}) \\ = \; & \text{Ws}(\text{genau 2 Bits kippen}) + \text{Ws}(\text{genau 3 Bits kippen}) \\ = \; & 3 * p^2 * (1-p) + p^3 = 3 * 10^{-2} * (1 - 10^{-1}) + 10^{-3} \end{aligned}$$

- Ohne Codierung Fehlerws von 0.1.
- Mit Repetitionscode Fehlerws von ≈ 0.03.
- Nachteil: Codierung ist dreimal so lang wie Nachricht.
- **Ziel:**
  Finde guten Tradeoff zwischen Fehlerws und Codewortlänge.

Ausblick: fehlertolerante Codes

Szenario: Binärer Ausfallkanal

- Bits 0,1 gehen mit Ws p, $p < \frac{1}{2}$ verloren,
  d.h. $0 \mapsto \epsilon$ bzw. $1 \mapsto \epsilon$.
- Korrekte Übertragung $0 \mapsto 0$, $1 \mapsto 1$ mit Ws $1 - p$.
- In unserem Beispiel $p = 0.1$.

Codierer: Repetitionscode der Länge 3.

Decodierer:

- Lese den Code in 3er-Blöcken xyz
- Ausgabe: x.
  Falls $y \neq x$, sei yz Anfang vom nächsten Block
  also lese nur ein extra Zeichen, um nächsten 3er-Block zu bilden
  Falls $y = x$ und $z \neq x$, sei z Anfang vom nächsten Block
  also lese zwei extra Zeichen, um nächsten 3er-Block zu bilden
  Falls $x = y = z$, bilde nächsten 3er-Block aus drei frischen Codezeichen

Fehler beim Decodieren: Alle drei Symbole gehen verloren.

- $\mathrm{Ws}(\text{Bit kann nicht decodiert werden}) = p^3 = 0.001$.
- Fehlerws kleiner beim Ausfallkanal als beim sym. Kanal.

Definition Code

- Alphabet $A = \{a_1, \ldots, a_n\}$, Menge von Symbolen a_i
- Nachricht $m \in A^*$

Definition Code

Sei A ein Alphabet. Eine (binäre) *Codierung C des Alphabets A* ist eine injektive Abbildung

$$\begin{aligned} C: \quad & A \to \{0,1\}^* \\ & a_i \mapsto C(a_i). \end{aligned}$$

Die *Codierung einer Nachricht* $m = a_{i_1} \ldots a_{i_\ell} \in A^*$ definieren wir als

$$C(m) = C(a_{i_1}) \ldots C(a_{i_\ell}) \quad \text{(Erweiterung von } C \text{ auf } A^*).$$

Die Abbildung C heißt *Code*.

Bezeichnungen Code

- Die Elemente $c_i := C(a_i)$ bezeichnen wir als *Codeworte*.
- Wir bezeichnen sowohl die Abbildung von Nachrichten auf Codeworte als auch die *Menge der Codeworte* mit dem Buchstaben C.
- Falls $C \subseteq \{0,1\}^n$ spricht man von einem *Blockcode* der Länge n. In einem Blockcode haben alle Codeworte die gleiche Länge.

Entschlüsselbarkeit von Codes

Szenario: Datenkompression in fehlerfreiem Kanal

Definition eindeutig entschlüsselbar

Ein Code heißt eindeutig entschlüsselbar, falls jedes Element aus $\{0,1\}^*$ Bild höchstens einer Nachricht ist. D.h. die Erweiterung der Abbildung C auf A^* muss injektiv sein.
Mit anderen Worten, ein Code C heißt *eindeutig entschlüsselbar* wenn füer jede Zeichenkette $x \in \{0,1\}^*$ höchstens eine Folge von Codeworten $c_1, c_2, \ldots, c_r$ existiert, derart dass $x = c_1 c_2 \ldots c_r$

Definition Präfixcode (eigentlich: präfixfreier Code)

Ein Code $C = \{c_1, \ldots, c_n\}$ heißt Präfixcode, falls es keine zwei Codeworte $c_i \neq c_j$ gibt mit

$$c_i \text{ ist Präfix (Wortanfang) von } c_j$$

i.e.

$$c_j = c_i s \quad \text{mit} \quad s \in \{0,1\}^* \ .$$

Sofortige Entschlüsselbarkeit

Definition Sofort entschlüsselbar

Ein Code C heißt *sofort entschlüsselbar* wenn ein Codewort lässt sich sofort dekodieren, sobald seine Zeichen bekannt sind, i.e. falls

$$x = x_1 x_2 \dots x_t x_{t+1} \dots$$

und t der kleinste Index ist, derart dass $c = x_1 x_2 \dots x_t$ ein Codewort ist, dann ist c die einzige Mögliche Dekodiertung des Anfangs von x und das nächste Codewort fängt mit x_{t+1} an.

Beobachtung

Es ist klar, dass

Präfixcode $\Leftrightarrow$ Sofort entschlüsselbar $\Rightarrow$ Eindeutig entschlüsselbar

Beispiel

	a_1	a_2	a_3
C_1	0	0	1
C_2	0	1	00
C_3	0	01	011
C_4	0	10	11

- C_1 ist kein Code, da $C_1 : A \to \{0,1\}^*$ nicht injektiv.
- C_2 ist nicht eindeutig entschlüsselbar, da $C_2 : A^* \to \{0,1\}^*$ nicht injektiv.
- C_3 ist eindeutig entschlüsselbar, aber kein Präfixcode.
- C_4 ist ein Präfixcode.

Weitere Beispiele (nicht unbedingt binär)

- Vollständige internationale Rufnummer: Präfixcode
- Morse Code ist kein Präfixcode, da
  - $A \mapsto \cdot -$
  - $L \mapsto \cdot - \cdot -$

  Aber: mit Pausen wird eindeutig entschlüsselbar, da Pausen wirken als Terminatoren
- UTF-8 ist Präfix code
- Die *Secondary Synchronization Codes* im UMTS Standard

Präfixcodes sind eindeutig entschlüsselbar.

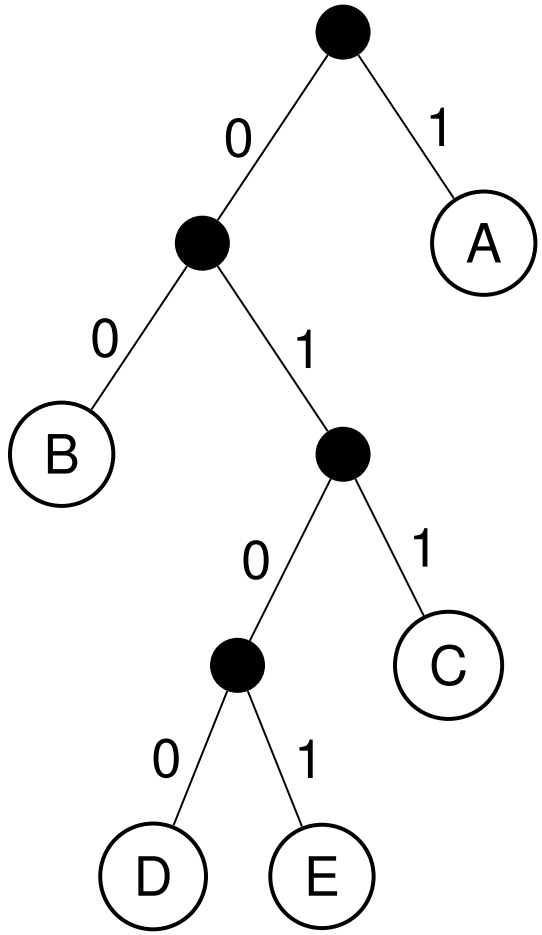
Satz Präfixcode eindeutig entschlüsselbar

Sei $C = \{c_1, \ldots, c_n\}$ ein Präfixcode. Dann kann jede codierte Nachricht $C(m)$ in Zeit $\mathcal{O}(|C(m)|)$ eindeutig zu m decodiert werden.

- Zeichne binären Baum
  - Kanten erhalten Label 0 für linkes Kind, 1 für rechtes Kind.
  - Codewort $c_i = c_{i_1} \ldots c_{i_k}$ ist Label des Endknoten eines Pfads von der Wurzel mit den Kantenlabeln $i_1, \ldots, i_n$
- **Präfixeigenschaft:** Kein einfacher Pfad von der Wurzel enthält zwei Knoten, die mit Codeworten gelabelt sind.
- Codewort c_i, oder das Ursprüngliche Alphabetsymbol, ist Blatt in Tiefe c_i

Beispiel: Code und entsprechender binärer Baum

$$C : \begin{cases} A \mapsto 1 \\ B \mapsto 00 \\ C \mapsto 011 \\ D \mapsto 0100 \\ E \mapsto 0101 \end{cases} \Rightarrow$$



Algorithmus Decodierung Präfix

Algorithmus Decodierung Präfix

1. Lese $C(m)$ von links nach rechts.
2. Starte bei der Wurzel. Falls 0, gehe nach links. Falls 1, gehe nach rechts.
3. Falls Blatt mit Codewort $c_i = C(a_i)$ erreicht, gib a_i aus und iteriere.

Laufzeit: $\mathcal{O}(|C(m)|)$

Woher kommen die Nachrichtensymbole?

Modell

- *Quelle Q* liefert Strom von Symbolen aus A.
- Quellwahrscheinlichkeit: $\mathrm{Ws}(\text{ Quelle liefert } a_i) = p_i$
- Ws p_i ist unabhängig von der Zeit und vom bisher produzierten Strom (gedächtnislose Quelle)
- X_i: Zufallsvariable für das Quellsymbol an der i-ten Position im Strom, d.h.

$$\mathrm{Ws}(X_i = a_j) = p_j \quad \text{für } j = 1, \ldots, n \text{ und alle } i.$$

Ziel: Codiere Elemente a_j mit großer Ws p_j mit kleiner Codewortlänge.

Kompakte Codes

Definition Erwartete Codewortlänge

Sei Q eine Quelle mit Alphabet $A = a_1, \ldots, a_n$ und Quellwahrscheinlichkeiten $p_1, \ldots, p_n$. Die Größe

$$E(C) := \sum_{i=1}^{n} p_i |C(a_i)|$$

bezeichne die erwartete Codewortlänge.

Definition Kompakter Code

Ein Code C heißt kompakt bezüglich einer Quelle Q, falls er *minimale erwartete Codewortlänge* besitzt.