3. Bayes' Theorem und Kanalkapazität

Konrad Voelkel

24. August 2010

Outline

Informationstheorie

Begriffe aus der Informatik

Nachrichten

- Gerhard Goos definiert in *Vorlesungen über Informatik, Band 1*:
  - "Die Darstellung einer Mitteilung durch die zeitliche Veränderung einer physikalischen Größe heißt ein *Signal*."
  - "Die dauerhafte Darstellung einer Mitteilung auf einem physikalischen Medium heißt *Inschrift*."
  - "Wenn wir bei der Darstellung und Weitergabe einer Mitteilung vom verwandten Medium und den Einzelheiten der Signale und Signalparameter abstrahieren, heißt die Mitteilung eine *Nachricht*."

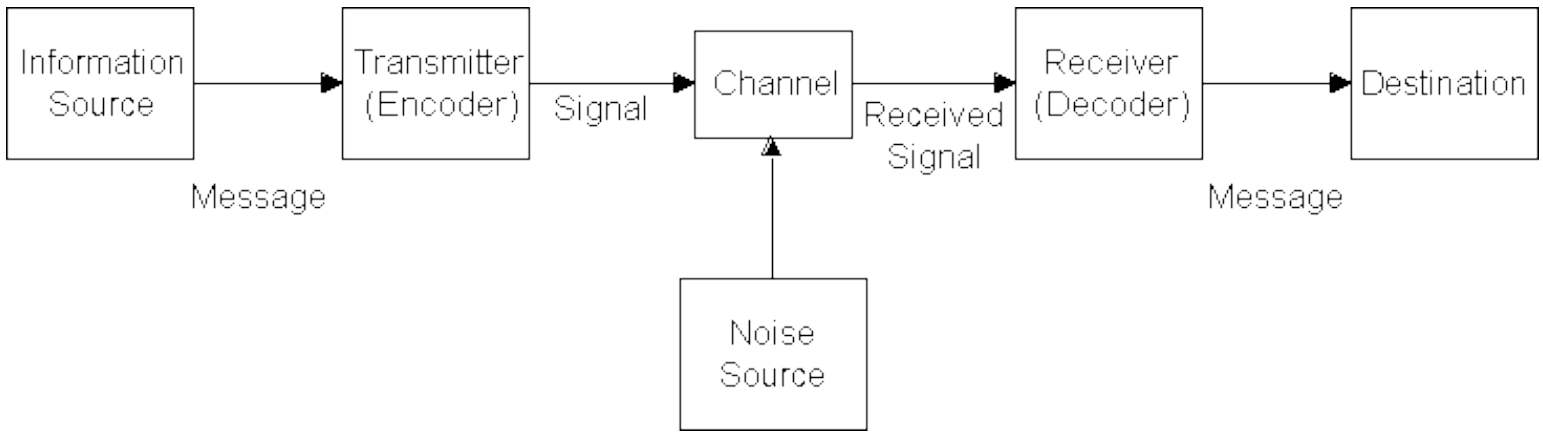
Begriffe aus der Informatik

Interpretation

- Gerhard Goos definiert in *Vorlesungen über Informatik, Band 1*:
  - "Die Kenntnisse, die man benötigt, um einer Nachricht Bedeutung zuzuordnen, nennen wir einen *Kontext* oder ein *Bezugssystem*."
  - "Die zugeordnete Bedeutung heißt eine *Information*."
  - "Man gewinnt sie durch *Interpretation* von Nachrichten auf der Grundlage eines Bezugssystems."
  - Eine *Interpretationsvorschrift*, angewendet auf eine Nachricht, liefert die zugeordnete Information.
  - Das Paar (Nachricht, zugeordnete Information) heißt *Datum*.
- Hier trägt der Begriff *Information* also eine andere, weniger präzise Bedeutung als in der Informationstheorie!

Setting von Shannon-Weaver

- Modell der Kommunikation nach Shannon und Weaver:



Setting von Shannon-Weaver

Mathematisches Modell

- Die Informationsquelle wird als *stochastischer Prozess* modelliert. Etwas einfacher wird die Theorie, wenn man sich auf *stationäre Prozesse* oder, meistens physikalisch gerechtfertigt, *ergodische Prozesse* beschränkt. Im diskreten Fall ist so ein Prozess eine Folge von *Zufallsvariablen* X_n.
- Encoder und Decoder sind Spezialfälle eines *Transcoders*, der kodierte Nachrichten in kodierte Nachrichten (mit ggf. anderem Code) umwandelt. So gesehen gibt es eigentlich nur kodierte Signale, während Nachrichten ein abstraktes Konzept sind, deren physikalische Entsprechung sich nur in kodierter Form (in Signalen) befindet.
- Wesentlich für die Theorie (und historisch Ausgangspunkt) ist die Fragestellung, wie sich eine Nachricht *digital* kodieren lässt, d.h. als Zeichenkette über einem endlichen Zeichenvorrat $\Sigma = \{z_1, ..., z_n\}$, z.B. dem Zeichenvorrat $\{0, 1\}$.

Setting von Shannon-Weaver

Rauschen

- Eine zentrale Frage ist, ob und wie sich Nachrichten digital so kodieren lassen, dass trotz Rauschen (also zufällig verändertem Signal) die Nachricht fehlerfrei dekodiert werden kann.
- Nebenbei ist es interessant, wie viele Nachrichten pro Sekunde über einen gegebenen Kanal übertragen werden können.
- Dafür ist die Frage nach dem Informationsgehalt der Nachricht und die Nutzbarmachung von Redundanz der Schlüssel.
- Im physikalischen Modell sind Kanal und Rauschquelle oft ein Bauteil. Es ist trotzdem sinnvoll, davon zu abstrahieren und eine separate Rauschquelle zu betrachten.
- In erster Näherung lässt sich rauschfreie Kommunikation untersuchen, in zweiter Näherung lässt sich Kommunikation unter *weissem Rauschen* untersuchen.

Grundlagen der diskreten Stochastik

- Wir wiederholen kurz einige Grundbegriffe der Stochastik für den diskreten Fall, insbesondere für Anwendungen zufälliger Zeichenfolgen, wie sie in der Shannonschen Darstellung der Informationstheorie genutzt werden, bevor wir die weiteren Begriffe der Stochastik parallel mit den Begriffen der Informationstheorie entwickeln.

Wahrscheinlichkeitsräume

Ereignisse

- Man betrachtet eine Menge Ω, genannt *Grundgesamtheit* und eine σ-*Algebra* von Teilmengen von Ω, die man *Ereignisse* nennt. Ein Ereignis modelliert einen möglichen Ausgang eines Experiments oder einer Messung bzw. Beobachtung. Insbesondere können Ereignisse einander enthalten, einander ausschliessen, viel oder wenig miteinander zu tun haben. Die Grundgesamtheit modelliert alle möglichen Ereignisse. Eine Menge von Ereignissen bildet eine σ-Algebra, wenn
  - es mindestens ein Ereignis gibt,
  - es zu jedem Ereignis $A \subset \Omega$ das Ereignis "A findet nicht statt", also $\Omega \setminus A$ gibt,
  - es für eine Folge von Ereignissen $A_n \subset \Omega$ auch das Ereignis "eines der A_n tritt ein", also $\bigcup_{n \geq 0} A_n$ gibt.
- Daraus folgt, dass es immer das "Kein-Ereignis" $\emptyset$ (die leere Menge) sowie ihr Komplement, dass "Irgendein-Ereignis" Ω gibt.

Wahrscheinlichkeitsräume

Wahrscheinlichkeitsmaße

- Ein *Wahrscheinlichkeitsmaß* auf Ω ist eine Abbildung P von der σ-Algebra in die reellen Zahlen zwischen 0 und 1 einschliesslich, die ein *Maß* auf der σ-Algebra definiert. Das bedeutet, jedem Ereignis wird eine Wahrscheinlichkeit zugeordnet, derart dass
  - für eine Folge sich gegenseitig ausschließender Ereignisse $A_n \subset \Omega$ die Wahrscheinlichkeit, dass irgendeines der Ereignisse eintritt, genau so groß ist wie die Summe der Wahrscheinlichkeiten aller Ereignisse, also $P(\bigcup_{n\geq 0} A_n) = \sum_{n\geq 0} P(A_n)$,
  - das Kein-Ereignis die Wahrscheinlichkeit 0 hat, also $P(\emptyset) = 0$.
- Daraus folgt, dass das Irgendein-Ereignis die Wahrscheinlichkeit 1 hat.

Zufallsvariable

- Eine *Zufallsvariable* modelliert Eigenschaften zufälliger Geschehnisse.
- Z.B. kann man den Wahrscheinlichkeitsraum Ω aller Konfigurationen von Teilchen im Universum betrachten (darauf gibt es ein Wahrscheinlichkeitsmaß P, das angibt, wie wahrscheinlich bestimmte Arten von Konfigurationen sind). Wie weit ein bestimmtes Teilchen von uns entfernt ist, ist nun eine Zufallsvariable mit Werten in den positiven reellen Zahlen.
- Formal sind Zufallsvariable sog. "meßbare" Abbildungen von Wahrscheinlichkeitsräumen in andere Wahrscheinlichkeitsräume, also $X : \Omega_1 \to \Omega_2$. In der Praxis ist dabei oft $\Omega_2 = \mathbb{R}$, die reellen Zahlen oder sogar, für uns relevant, $\Omega_2 = \mathbb{N}$, die natürlichen Zahlen.
- Eine Zufallsvariable hat eine *Verteilung*, d.h. man kann sich zu jedem Ereignis in Ω_2 fragen, wie wahrscheinlich es ist, dass X einen Wert daraus annimmt.

Diskrete Wahrscheinlichkeitsräume

- Ein Wahrscheinlichkeitsraum Ω heißt *diskret*, wenn er nur endlich viele oder abzählbar viele Elemente hat (also so viele, dass man sie mit natürlichen Zahlen durchnummerieren kann) und wenn für jedes Element $\omega \in \Omega$ die einelementige Teilmenge $\{\omega\} \subset \Omega$ ein Ereignis ist.
- Gegeben eine endliche Menge $\{z_1, ..., z_n\}$ lässt sich darauf die Struktur eines diskreten Wahrscheinlichkeitsraums definieren, indem man jede Teilmenge als mögliches Ereignis definiert und dann eine *Gleichverteilung* festlegt, d.h. jeder Menge $\{z_i\}$ die Wahrscheinlichkeit $P(\{z_i\}) := \frac{1}{n}$ zuweist. Damit hat eine k-elementige Teilmenge dann automatisch die Wahrscheinlichkeit $\frac{k}{n}$.

Diskrete Stochastik ohne Maßtheorie

- Wir betrachten von nun an nur noch diskrete Wahrscheinlichkeitsräume.
- Die Funktion $p : \Omega \to [0, 1]$, die jedem $\omega \in \Omega$ die Wahrscheinlichkeit $P(\{\omega\})$ zuordnet, heißt auch *Wahrscheinlichkeitsfunktion* vom Wahrscheinlichkeitsraum Ω.
- Aus der Wahrscheinlichkeitsfunktion lässt sich das Wahrscheinlichkeitsmaß P herleiten, d.h. wenn man nur mit diskreten Wahrscheinlichkeitsräumen arbeitet, benötigt man keine Begriffe der Maßtheorie.
- In diesem Kontext heißt für eine Zufallsvariable $X : \Omega_1 \to \Omega_2$ die Funktion $p^X : \Omega_2 \to [0, 1]$, die jedem $\omega \in \Omega_2$ die Wahrscheinlichkeit $P(X = \omega)$ zuorndet, die *Verteilungsfunktion* von X. Sie bestimmt eine diskrete Zufallsvariable eindeutig.

Diskrete Zufallsvariable

Ein Modell für zufällige Zeichenketten

- Eine in der Shannonschen Informationstheorie wichtige Situation ist die eines endlichen *Zeichenvorrats* $\Sigma = \{z_1, ..., z_n\}$ (z.B. das lateinische Alphabet zusammen mit dem Leerzeichen, also $n = 27$) und zufälliger Zeichenketten $w \in \Sigma^*$, wobei $\Sigma^* := \bigcup_{k \geq 0} \Sigma^k$ definiert ist und Σ^k die Menge aller Zeichenketten der Länge k aus dem Zeichenvorrat Σ bezeichnet.
- Die einzelnen Zeichen aus einer Zeichenkette werden als Zufallsvariable $X_j : \Omega \to \Sigma$ modelliert, eine Zeichenkette der Länge k ist damit eine Zufallsvariable $X : \Omega^k \to \Sigma^k$.
- Der Zufall steckt hier im Wahrscheinlichkeitsraum Ω, der in der Regel nicht näher spezifiziert wird. Stattdessen interessiert man sich nur für die Verteilung der Zufallsvariable, also für die Wahrscheinlichkeiten $P(X_j = z_i)$ oder auch $P(X = w)$, wenn $w \in \Sigma^k$.

Markov-Ketten

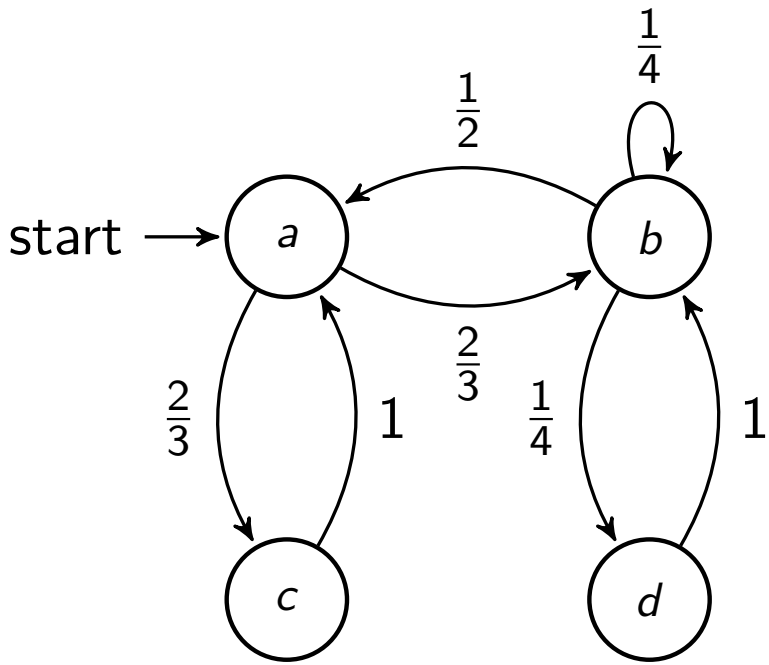
Ein besseres Modell

- Eine Nachricht in englischer Sprache die mit “th” anfängt, liefert schon einen Hinweis darauf, dass es vermutlich mit “e” weitergeht, jedoch nicht mit “q”. Das bedeutet, die Wahrscheinlichkeit von X_3 hängt ab vom Ausgang der Zufallsexperimente X_1 und X_2. Ein spezieller Fall so einer Art von Abhängigkeit ist die Markov-Kette, bei der man die Modell-Annahme trifft, ein Zeichen hänge nur vom vorhergehenden Zeichen ab.
- Markov-Ketten sind somit ein besseres Modell für Zeichenfolgen, die in der Natur vorkommen, als komplett unabhängige Zeichenfolgen. Auf der anderen Seite beinhalten sie den komplett unabhängigen Fall, sind also strikt allgemeiner.

Markov-Ketten

Graphische Darstellung

- Um eine Markov-Kette anzugeben, gibt man nicht Wahrscheinlichkeiten für die Zeichen an, sondern *Übergangswahrscheinlichkeiten*, für die Wahrscheinlichkeit, dass auf z_i ein z_j folgt. Man kann eine Markov-Kette graphisch darstellen:



- Typische Ausgabe: *acababbdbdbacabd*...

Markov-Ketten

Verallgemeinerungen

- Stochastiker fassen eine Markov-Kette gern in einer *stochastischen Matrix* zusammen, das ist die Adjazenzmatrix der graphischen Darstellung (und sie erlaubt es, mittels *Markovkernen* die Theorie auf stetige Zufallsvariable zu verallgemeinern, das heisst dann *stochastischer Prozess*).
- Informatiker fassen eine Markov-Kette gern als Verallgemeinerung von *endlichen Automaten* auf.
- Es gibt auch Markov-n-Ketten, die dann einen stochastischen n-Tensor als Datum haben. Hier gibt man an, wie hoch die Wahrscheinlichkeit ist, nach einer bestimmten Zeichenkette der Länge n ein bestimmtes Zeichen zu erhalten (und das für jede Zeichenkette der Länge n und jedes Zeichen, das folgen könnte).

Bayessche Netzwerke

Kurzer Ausblick

- Eine mögliche Verallgemeinerung von Markov-Ketten sind sogenannte *Bayessche Netze*.
- Hier betrachtet man n Zufallsvariable $X_1, ..., X_n$, die man als Knoten in einem Graphen repräsentiert.
- Die Kanten im (gerichteten, azyklischen) Graphen modellieren Abhängigkeit, d.h. eine Zufallsvariable hängt nur von genau den Zufallsvariablen ab, die im Graphen einen Verbindungspfeil auf sie haben (man nennt dies die *lokale Markov-Eigenschaft*. Eine Zufallsvariable, auf dessen Knoten im Graph kein Pfeil zeigt, ist von allen anderen unabhängig.
- Markov-Ketten sind der Spezialfall, in dem der Graph nur eine lineare Kette darstellt, d.h. jeder Knoten X_i höchstens auf den nächsten Knoten X_{i+1} zeigt, mehr nicht.

Grenzwertsätze

- Grenzwertsätze dienen dazu, das Vorgehen zu legitimieren, von einer kleinen Zahl von Experimenten auf langfristiges Verhalten zu schliessen.

Gesetze großer Zahlen

Das schwache Gesetz der großen Zahlen

- Gesetze großer Zahlen sagen im Wesentlichen aus, dass eine Folge von Zufallsvariablen "wahrscheinlich" auf lange Sicht keine große mittlere Abweichung von den Mittelwerten von ihren Erwartungswerten hat.
- Präzise, in Formeln, sagt man, dass eine Folge X_n dem *schwachen Gesetz der großen Zahlen* genügt, wenn

$$P - \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}(X_k)) = 0.$$

- Dabei erfüllt nicht jede Folge von Zufallsvariablen ein solches Gesetz! (Auflösung kommt gleich)

Gesetze großer Zahlen

P-fast-sichere Konvergenz

- $P - \lim_{n\to\infty}$ ist ein *stochastischer Limes*, es ist für eine Folge von Zufallsvariablen X_n mit Werten in $\mathbb{R}$ definiert:

$$P - \lim_{n\to\infty} X_n = 0 :\Leftrightarrow \forall \epsilon > 0 : \lim_{n\to\infty} P(|X_n| > \epsilon) = 0.$$

- Die Namensgebung kommt daher, dass die Ereignisse $|X_n| < \epsilon$ mit einer Wahrscheinlichkeit eintreten, die sich für $n \to \infty$ der 1 nähern. Ereignisse mit Wahrscheinlichkeit 1 nennt man *fast sicher*.

Gesetze großer Zahlen

Das starke Gesetz

- Eine Folge X_n von Zufallsvariablen genügt dem *starken Gesetz der großen Zahlen*, wenn

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (X_k - \mathbb{E}(X_k)) = 0 \quad P\text{-fast-sicher.}$$

- Erfüllt eine Folge von Zufallsvariablen das starke Gesetz, so auch das schwache.

Zentraler Grenzwertsatz

- Eine Folge von Zufallsvariablen genügt dem *zentralen Grenzwertsatz*, wenn

$$P_{S_n^*} \xrightarrow{schwach} \mathcal{N}(0,1)$$

- Dabei bezeichnet $P_{S_n^*}$ die Verteilung der auf Erwartungswert 0 und Varianz 1 normierten Summe der Zufallsvariablen X_i von $i = 1$ bis n und $\mathcal{N}(0,1)$ die *Standard-Normalverteilung*.
- Schwache Konvergenz nennt man auch *Konvergenz in Verteilung*, definiert ist die schwache Konvergenz von Verteilungen P_i gegen eine Verteilung P_∞ so, dass für jede Zufallsvariable Y der Grenzwert der Erwartungswerte von Y bezüglich der Verteilungen P_i gegen den Erwartungswert von Y bezüglich der Verteilung P_∞. In Formeln:

$$\forall Y : \lim_{i \to \infty} \mathbb{E}_{P_i}(Y) = \mathbb{E}_{P_\infty}(Y).$$

Überblick über einige stochastische Konvergenzbegriffe

- Sei X_n eine Folge von Zufallsvariablen und X eine Zufallsvariable.
- P-fast-sichere Konvergenz:

$$P(\lim_{n\to\infty} X_n = X) = 1$$

- impliziert stochastische Konvergenz:

$$\forall \epsilon > 0 : \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

- impliziert Konvergenz in Verteilung:

$$\forall Y : \lim_{n\to\infty} \mathbb{E}_{PX_n}(Y) = \mathbb{E}_{PX}(Y).$$

Normalverteilung

Definition

- Eine Zufallsvariable $X : \Omega \to \mathbb{R}$ heißt *normalverteilt* ($X \sim \mathcal{N}(\mu, \sigma)$), wenn $P(X \leq k) = \int_0^k \phi(x)dx$, wobei ϕ eine Funktion ist, die von den *Parametern* der Normalverteilung abhängt. Diese Parameter sind Erwartungswert μ und Standardabweichung σ.

$$\phi(x) = \phi_{\mu,\sigma^2}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right).$$

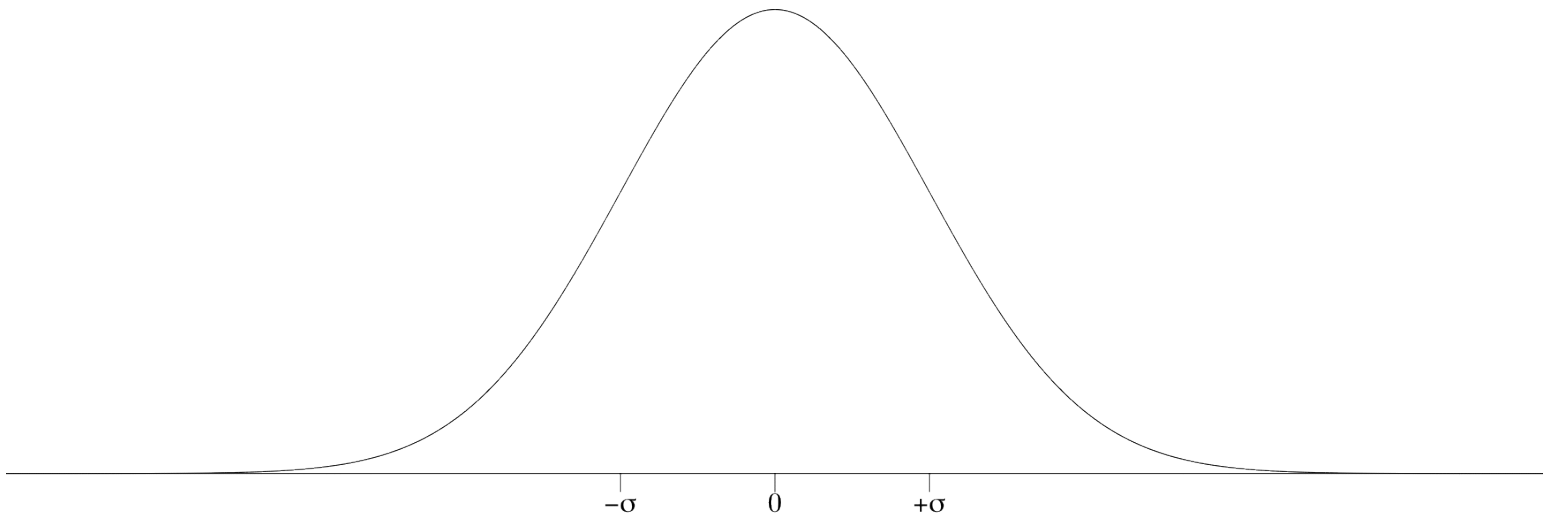
- Im Falle der *Standard-Normalverteilung* $\mathcal{N}(0, 1)$ ist $\mu = 0$ und $\sigma = 1$, dann ist

$$\phi(x) = \phi_{0,1}(x) = \frac{1}{2\pi} e^{-x^2}$$

Normalverteilung

Gausssche Glockenkurve

- Der Graph der Funktion ϕ ist die bekannte *Gausssche Glockenkurve*, die Funktion $P(X \leq k)$ ist der Flächeninhalt unter dem Graphen bis zur Stelle k.



- Man spricht im Kontext der Rauschanalyse nicht von der Normalverteilung sondern von *weissem Gaussschen Rauschen*.

Grenzwertsätze

Sätze

- Sei X_n eine Folge von Zufallsvariablen (mit wohldefiniertem Erwartungswert und Varianz).
- Wenn die X_n unabhängig voneinander und gleich verteilt sind, so erfüllen sie das starke Gesetz großer Zahlen.
- Sind die X_n bloß paarweise unkorreliert und gleich verteilt, so erfüllen sie das schwache Gesetz großer Zahlen.
- Sind die X_n unabhängig voneinander und gleich verteilt, so erfüllen sie den zentralen Grenzwertsatz.
- Moralisch gesehen heißt das: Führt man ein Experiment unabhängig voneinander wiederholt durch, so nähern sich die Daten einer Normalverteilung mit Parametern Erwartungswert und Varianz der ursprünglichen Verteilung (selbst wenn diese keine Normalverteilung war).
- Das erklärt auch die zentrale Rolle der Normalverteilung für die Biologie.

Bedingte Wahrscheinlichkeiten und Information

- Nun werden wir untersuchen, wie Ereignisse miteinander zusammen hängen können.
- Ereignisse können zusammengesetzt und zerlegt werden: Verbundwahrscheinlichkeit.
- Ereignisse können betrachtet werden unter der Prämisse, dass andere Ereignisse eingetreten sind: bedingte Wahrscheinlichkeit.
- Dabei betrachten wir, wie sich die Entropie verhält.

Entropie von Zufallsvariablen

Entropie

Definition

- Entropie ist ein Maß für die *Unsicherheit*, die mit einer Zufallsvariable verbunden wird.
- Ihre Maßeinheit in der Shannonschen Theorie ist das *bit*, abgekürzt für *binary digit*.
- Sei X eine diskrete Zufallsvariable mit Werten in einem endlichen Zeichenvorrat $\Sigma = \{z_1, .., z_n\}$ und Verteilungsfunktion $p : \Sigma \to [0, 1]$. Dann ist die *Entropie* von X definiert als

$$H(X) := \mathbb{E}(I_X) = -\sum_{i=1}^{n} p(z_i) \log_2 p(z_i).$$

Entropie

Erklärung der Definition

- Die Notation I_X bezeichnet die *Selbstinformation* von X (auch: *Überraschung*), das ist eine Zufallsvariable, die jedem Zeichen z_i seine Entropie zuordnet, d.h. die die Verteilungsfunktion hat:

$$I_X(z_i) := -\log_2 p(z_i).$$

- Die Notation $\mathbb{E}$ bezeichnet den *Erwartungswert*. Der Erwartungswert einer Zufallsvariable $I : \{z_1, ..., z_n\} \to [0, 1]$ bezüglich der Verteilung p von X ist gegeben durch

$$\mathbb{E}(I) := \sum_{i=1}^{n} p(z_i) I(z_i).$$

Entropie

Erläuterung anhand einer Gleichverteilung

- Wir betrachten den einfachen Fall, dass die Zufallsvariable X gleichverteilt ist, also dass $p(z_i) = \frac{1}{n}$ ist für alle $z_i \in \{z_1, ..., z_n\}$.
- Weiterhin betrachten wir den Fall, dass es genau $n = 2^m$ viele Zeichen gibt, dann ist $p(z_i) = \frac{1}{2}^m = 2^{-m}$.
- Die Selbstinformation ist damit $I_X(z_i) = -\log_2 p(z_i) = -\log_2 2^{-m} = -(-m) = m$.
- Der Erwartungswert der Selbstinformation ist $\mathbb{E}(I_X) = \sum_{i=1}^{n} p(z_i) I_X(z_i) = \sum_{i=1}^{n} \frac{1}{n} m = n\frac{1}{n} m = m$.
- Damit sehen wir: die Selbstinformation ist genau die Entropie einer gleichverteilten Zufallsvariablen.

Entropie

Konzeptuelle Erläuterung

- Um zwischen 2^m verschiedenen Zeichen eines auszuwählen, müssen m Ja/Nein-Fragen beantwortet werden (etwa: “liegt das gesuchte Zeichen in der ersten Hälfte aller Zeichen?” usw.).
- Sind manche Zeichen wahrscheinlicher als andere (so wie bei Sätzen in natürlicher Sprache), so ist eine Gleichverteilung ein schlechtes Modell. Ist aber ein Zeichen, z.B. “e”, deutlich wahrscheinlicher, so ist auch die Unsicherheit, ein “e” zu erhalten, nicht so groß. Die Selbstinformation des Zeichens “e” ist um so geringer, je wahrscheinlicher es ist.

$$p(z_1) > p(z_2) \implies \log_2 p(z_1) > \log_2 p(z_2)$$

$$\implies -\log_2 p(z_1) < -\log_2 p(z_2) \implies I(z_1) < I(z_2).$$

- Die Gleichverteilung ist die Verteilung mit der größten Entropie.

Tukey vs. Shannon

Definition

- Shannon erwähnt als erster schriftlich den Begriff “bit”, gibt aber an, dass er von Tukey stammt.
- Tukey nutze aber den Begriff “Bit” ausschließlich im Sinne des Speicherns von Daten. Die Zeichenkette 000000 bestand für Tukey aus 6 Bits, selbst wenn von vornherein klar wäre, dass sie nur aus Nullen bestünde.
- In Shannons Sinne kommt es auf die Verteilung an. Sind die Zeichenketten, die von einer Quelle stammen etwa so verteilt, dass nur entweder Nullen oder ausschliesslich Einsen erwartet werden können, so enthält die Zeichenkette 000000 genau ein bit an Information (denn sie teilt uns mit, dass eben nicht 111111 gesendet wurde).
- Einige Autoren unterscheiden daher zwischen *Bit* (Tukey) und *bit* (Shannon), so etwa Gerhard Goos, Vorlesungen über Informatik, 1995 Springer-Verlag Berlin Heidelberg.

bit vs. nat vs. hartley

Basiswechsel bei Logarithmen

- Anstelle des *logarithmus dualis*, dem Logarithmus zur Basis 2, lässt sich auch eine andere Basis verwenden, z.B. 10, dann spricht man von der Einheit hartley.
- Die Basis e (eulersche Zahl) entspricht dem *natürlichen Logarithmus*, der wohl in der reinen Mathematik am geläufigsten ist. Die entsprechende Einheit heißt nat.
- Für drei reelle Zahlen a, b und x gilt $a^{\log_a(x)} = x = b^{\log_b(x)}$. Ausserdem ist $b = a^{\log_a(b)}$, also

$$a^{\log_a(x)} = x = b^{\log_b(x)} = \left(a^{\log_a(b)}\right)^{\log_b(x)} = a^{\log_a(b)\log_b(x)}$$

Damit ist $\log_a(x) = \log_a(b)\log_b(x)$ und die verschiedenen Einheiten unterscheiden sich jeweils nur um Konstanten.

Zusammengesetze Systeme

- Ein Wort besteht aus mehreren Zeichen. Eine zufällige Zeichenkette lässt sich in mehrere (i.A. nicht unabhängige) Zufallsvariable zerlegen.
- Mehrere Zeichen bilden eine Zeichenkette. Zufallsvariable lassen sich zusammensetzen zu “größeren” Zufallsvariablen.
- Man muss aufpassen: Setzt man eine zufällige Zeichenkette und eines der darin enthaltenen zufälligen Zeichen zusammen, so erhält man keine neue Information!

Verbundwahrscheinlichkeiten

Definition

- Gegeben zwei diskrete Zufallsvariable X und Y, bezeichnet man die Wahrscheinlichkeit $P(X = x \text{ und } Y = y)$ als *Verbundwahrscheinlichkeit* und notiert auch $P(X = x, Y = y)$. Mengentheoretisch entspricht dem Wort "und" der Schnitt der Ereignisse $\{X = x\} \cap \{Y = y\}$ innerhalb einer Menge Ω auf der X und Y definiert sind.
- Wegen $\sum_y P(Y = y) = 1$ folgt daraus

$$P(X = x) = \sum_y P(X = x, Y = y).$$

- Sind beispielsweise zwei Zufallsvariable X_1, X_2 mit Werten in einem endlichen Zeichenvorrat $\Sigma = \{z_1, ..., z_n\}$ gegeben, die ein Wort der Länge 2 modellieren, also $X := (X_1, X_2)$ mit Werten in Σ^2, so gibt die Verbundwahrscheinlichkeit $P(X_1 = z_i, X_2 = z_j)$ die Wahrscheinlichkeit $P(X = z_i z_j)$ an.

Verbundentropie

Entropie eines zusammengesetzten Systems

- Gegeben zwei diskrete Zufallsvariable X und Y mit Werten in einem endlichen Zeichenvorrat $\{z_1, ..., z_n\}$, bezeichnet man die Größe $H(X, Y) := \mathbb{E}(I_{X,Y})$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{n} P(X = z_i, Y = z_j) \log_2 (P(X = z_i, Y = z_j))$$

  als *Verbundentropie* von X und Y.

- Dies ist genau das selbe wie die Entropie der Zufallsvariable (X, Y) mit Werten in Σ^2.
- Verbundentropie im Verhältnis zur Entropie der Teilsysteme:

$$H(X) + H(Y) \;\geq\; H(X, Y) \;\geq\; \max(H(X), H(Y))$$

Bedingte Wahrscheinlichkeiten

Intuitive Bedeutung

- Man kann fragen, wie wahrscheinlich ein Wort w der Länge 2 ist, wenn das erste Zeichen bereits auf ein $z_i \in \Sigma$ festgelegt ist. Dies bezeichnet man mit $P(X = w|X_1 = z_i)$, es handelt sich um eine *bedingte Wahrscheinlichkeit* mit Bedingung $X_1 = z_i$.
- Wir können in diesem Fall (zusammengesetztes System) auch schreiben

$$P(X = w|X_1 = z_i) = P((X_1, X_2) = (w_1, w_2)|X_1 = z_i)$$

$$= P(X_1 = w_1, X_2 = w_2|X_1 = z_i)$$

  und es ist klar, dass diese bedingte Wahrscheinlichkeit 0 ist, wenn $w_1 \neq z_i$.
- Im Falle einer Markovkette hängt X_2 von X_1 ab, also ist die *Übergangswahrscheinlichkeit* $P(X_2 = w_2|X_1 = w_1)$ im Allgemeinen nicht die selbe wie $P(X_2 = w_2)$.

Bedingte Wahrscheinlichkeiten

Definition

- Mathematisch definiert man für zwei Zufallsvariable X und Y die bedingte Wahrscheinlichkeit

$$P(X = x|Y = y) := \frac{P(X = x, Y = y)}{P(Y = y)}.$$

- Man rechnet leicht nach, dass stets $P(X = x, Y = y) \leq P(Y = y)$ und $\sum_x P(X = x, Y = y) = 1$ gilt, somit $x \mapsto P(X = x|Y = y)$ eine Wahrscheinlichkeitsfunktion in x liefert, sofern $Y = y$ nicht unmöglich ist.
- Die Definitionen sind so gegeben, dass

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y).$$

Unabhängige Ereignisse

Definition

- Zwei Ereignisse heißen *unabhängig*, wenn die Wahrscheinlichkeit, dass beide eintreten, genau so groß ist wie das Produkt der Wahrscheinlichkeiten der einzelnen Ereignisse.
- In Formeln: $\{X = x\}$ und $\{Y = y\}$ heißen *unabhängig*, wenn $P(X = x, Y = y) = P(X = x)P(Y = y)$.
- In diesem Fall ist

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

$$= \frac{P(X = x)P(Y = y)}{P(Y = y)} = P(X = x),$$

also hängt die Wahrscheinlichkeit für $\{X = x\}$ nicht davon ab, ob das Ereignis $\{Y = y\}$ eintritt.

Verbundentropie

Für unabhängige Zufallsvariablen

- Betrachten wir nun Verbundentropie im Fall unabhängiger Zufallsvariablen:
- Der Logarithmus wandelt Produkte in Summen um, also ist

$$\log_2 \left(P(X = z_i)P(Y = z_j)\right) = \log_2 P(X = z_i) + \log_2 P(Y = z_j).$$

und damit $H(X, Y)$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{n} P(X = z_i, Y = z_j)\left(\log_2 P(X = z_i) + \log_2 P(Y = z_j)\right)$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{n}\left(P(X = z_i, Y = z_j)\log_2 P(X = z_i) + P(X = z_i, Y = z_j)\log_2 P(Y = z_j)\right)$$

$$= -\sum_{i=1}^{n}\left(P(X = z_i)\log_2 P(X = z_i) + \sum_{j=1}^{n} P(Y = z_j)\log_2 P(Y = z_j)\right)$$

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

Gemeinsame Information

- Die Größe

$$I(X;Y) := \sum_{i=1}^{n}\sum_{j=1}^{n} P(X = z_i, Y = z_j) \log_2 \left(\frac{P(X = z_i, Y = z_j)}{P(X = z_i)P(Y = z_j)} \right)$$

  heißt *gemeinsame Information* von X und Y, sie misst die Menge an Information, die X und Y gemeinsam haben.

- Entsprechend lässt sich leicht beweisen:

$$I(X;Y) = H(X) + H(Y) - H(X,Y).$$

- Unabhängige Zufallsvariablen haben daher keine gemeinsame Information, $I(X;Y) = 0$.
- $I(X;X)$ ist die *Selbstinformation* von X. Es ist

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

Bedingte Entropie

Auch: Equivokation (nach Shannon)

- Gegeben zwei diskrete Zufallsvariable X und Y mit Werten in einem endlichen Zeichenvorrat $\{z_1, ..., z_n\}$, bezeichnet man die Größe

$$H(X|Y) := \sum_{i=1}^{n} \sum_{j=1}^{n} P(X = z_i, Y = z_j) \log_2 \left(\frac{P(Y = z_j)}{P(X = z_i, Y = z_j)} \right)$$

  als *bedingte Entropie* von X unter der Bekanntheit von Y.
- Der Logarithmus wandelt Brüche in Differenzen um, also ist

$$\log_2 \left(\frac{P(Y = z_j)}{P(X = z_i, Y = z_j)} \right) = \log_2 P(Y = z_j) - \log_2 P(X = z_i, Y = z_j)$$

  und damit $H(X|Y) = H(X, Y) - H(Y)$.
- Im Falle unabhängiger Zufallsvariablen ist also

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

Beispiel

Bedingte Wahrscheinlichkeiten von Zeichenfolgen

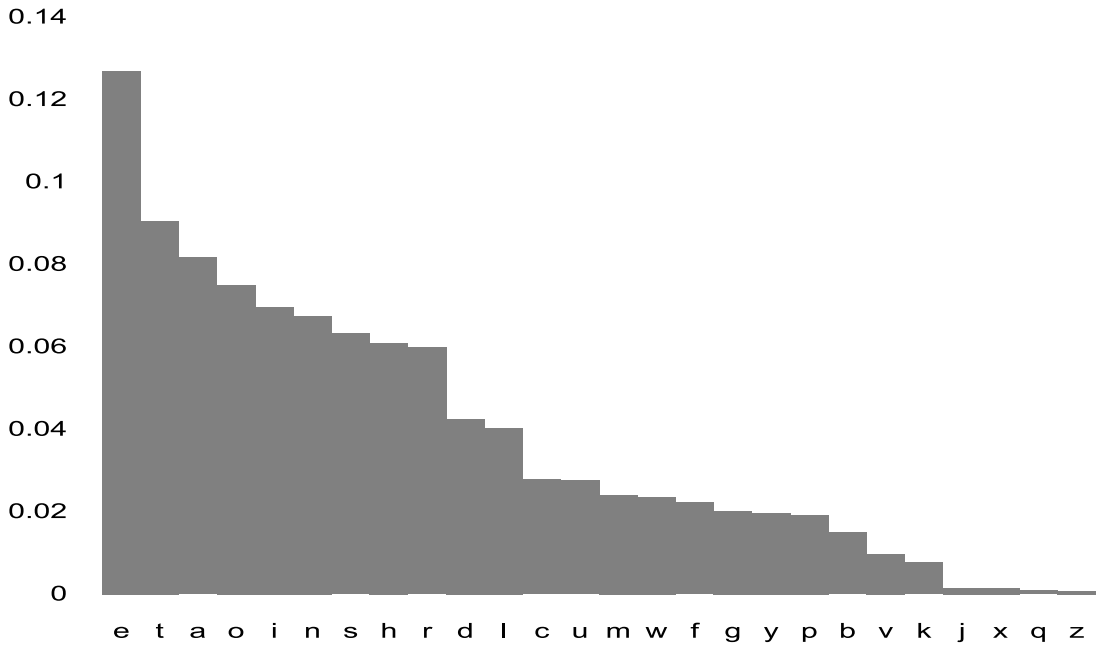
- Bereits Shannon hat in seinem Artikel von 1977 als Beispiel versucht, die Entropie der englischen Sprache zu bestimmen.
- In erster Näherung wollen wir eine Nachricht in englischer Sprache modellieren durch eine Folge unabhängiger Zufallsvariablen X_i mit Werten im Alphabet $\Sigma := \{A, B, C, ..., Y, Z, \_\}$ und Gleichverteilung, also $P(X_i = z) = \frac{1}{27}$ für alle $z \in \Sigma$. Dabei ist _ das Leerzeichen.
- Die Selbstinformation eines Zeichens $z \in \Sigma$ ist somit $I_i(z) = -\log_2 P(X_i = z) = -\log_2 \frac{1}{27} = \log_2 27 \approx 4,75\text{bit}$.
- Die Entropie der Zufallsvariablen X_i ist damit auch etwa $4,75\text{bit}$ (wegen Gleichverteilung).
- Da die Zufallsvariablen nicht voneinander abhängen ist damit die Entropie im Grenzwert

$$H(X) = \lim_{n\to\infty} \frac{H(X_1, ..., X_n)}{n} = \lim_{n\to\infty} n\frac{4,75\text{bit}}{n} = 4,75\text{bit}.$$

Beispiel

Eine realistischere Verteilung der Zeichen

- Die Häufigkeit der Buchstaben in der englischen Sprache ist nicht gleichverteilt:



- Die Selbstinformation eines Zeichens $z \in \Sigma$ ist somit z.B. $I_i(e) = -\log_2 P(X_i = e) = 3\text{bit}$ oder aber $I_i(v) = 6,64\text{bit}$.
- Die Entropie der Zufallsvariablen X_i und damit $H(X)$ lässt sich anhand dieser Tabelle berechnen: $\approx 4\text{bit}$.

Beispiel

Weitere Methoden

- Mit vorliegenden Daten über die Häufigkeit von 3-Zeichen-Kombinationen in der englischen Sprache rechnet Shannon mittels Markov-3-Ketten noch einen genaueren Wert für die Entropie eines Zeichens aus, nämlich etwas mehr als 2 bit.
- Der letzte Stand der Forschung vermutet für die mittlere Entropie eines Zeichens in englischer Sprache einen Wert zwischen $1,1$ und $1,9$, sodass 2 bit im Mittel genügen, um ein Zeichen in englischer Sprache zu kodieren. Das bedeutet aber nicht, dass man eine Tabelle anlegen kann, die jedem Zeichen ein-eindeutig einen 2-bit-Code zuweist!
- Der PPM (prediction by partial matching) Algorithmus erreicht eine mittlere Effizienz von $1,5$ bit pro Zeichen für Texte in englischer Sprache.
- Später mehr zu Redundanz und Kodierung.

Kullback-Leibler-Divergenz

Ein geringfügig allgemeinerer Begriff als *bedingte Entropie*

- Für zwei Wahrscheinlichkeitsmaße (z.B. Verteilungen von diskreten Zufallsvariablen) P und Q auf einem Raum (genau genommen einer σ-Algebra) Ω sodass Q *absolutstetig* bezüglich P ist (im Falle diskreter Verteilungen genügt $P(i) > 0 \implies Q(i) > 0$), definiert man die *Kullback-Leibler-Divergenz* (auch: *Informationsgewinn*) von P bezüglich Q als

$$D_{KL}(P||Q) := -\int_{\Omega} \log \frac{dQ}{dP} dP,$$

  wobei $\frac{dQ}{dP}$ die *Radon-Nikodym-Ableitung* von Q nach P ist.
- Sind P und Q durch *Dichten* gegeben (z.B. im diskreten Fall durch Wahrscheinlichkeitsfunktionen p und q), so vereinfacht sich die Definition zu

$$D_{KL}(P||Q) = \int_{\Omega} p(x) \log \frac{p(x)}{q(x)} dx.$$

Kullback-Leibler-Divergenz

Spezialfälle der Kullback-Leibler-Divergenz: diskrete Verteilungen

- Im diskreten Fall ist das Integral eine Summe, also für Zufallsvariable X und Y mit Werten in einem Zeichenvorrat $\{z_1, ..., z_n\}$:

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(X = z_i) \log \frac{P(X = z_i)}{P(Y = z_i)}.$$

- Wir werden nun sehen, dass viele zuvor eingeführten Größen als Kullback-Leibler-Divergenzen definiert werden können. Das erweist sich oft als der Schlüssel, um die diskrete Definition auf den stetigen Fall zu verallgemeinern.
- Die *Selbstinformation* eines Zeichens z_i ist $I(z_i) = D_{KL}(\delta_{ij}||P_i)$, wobei P_i die Gleichverteilung mit Wahrscheinlichkeit $P(X = z_i)$ ist und δ_{ij} das *Kronecker-Delta*.
- Die *gemeinsame Information* von X und Y ist $I(X; Y) = D_{KL}(P(X, Y)||P(X)P(Y))$.

Kullback-Leibler-Divergenz

Spezialfälle der Kullback-Leibler-Divergenz: Entropie und bedingte Entropie

- Die *Entropie* von X ist
  $H(X) = \mathbb{E}(I) = \log N - D_{KL}(P(X)||P_U(X))$, wobei P_U die Gleichverteilung ist und N die Entropie von X, wenn Gleichverteilung vorläge. Die hier auftretende KL-Divergenz misst also, wie viele bits der Verteilung noch fehlen bis zur Entropie einer Gleichverteilung.
- Die *bedingte Entropie* von X unter Bekanntheit von Y ist

  $$H(X|Y) = H(X) - I(X;Y) = \log N - D_{KL}(P(X,Y)||P_U(X)P(Y)),$$

  wobei wieder P_U die Gleichverteilung ist und N die Entropie, die bei Gleichverteilung von X vorläge.

Bayessche Induktion

- Nun wagen wir einen kurzen Ausflug in die Anfänge der statistische Inferenztheorie.

Satz von Bayes

- Der *Satz von Bayes* besagt, dass für zwei Zufallsvariable X und Y gilt:

$$P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X = x)}{P(Y = y)}$$

- Man beweist dies leicht, indem man mit $P(Y = y)$ durchmultipliziert und die Definition der bedingten Wahrscheinlichkeit einsetzt:

$$\frac{P(X = x, Y = y)}{P(Y = y)} P(Y = y) = \frac{P(Y = y, X = x)}{P(X = x)} P(X = x)$$

Satz von Bayes

Bedeutung des Satzes I

- Mit dem Satz von Bayes lässt sich eine der Größen $P(X = x),\ P(Y = y),\ P(X = x|Y = y),\ P(Y = y|X = x)$ berechnen, wenn die anderen 3 bekannt sind.
- Modelliert man mit $Y = y$ die vorliegenden Daten eines Experiments und mit $X = x$ eine zu testende Hypothese, so fragt $P(X = x|Y = y)$ nach der Wahrscheinlichkeit der Hypothese unter den Daten, hingegen $P(Y = y|X = x)$ nach der Wahrscheinlichkeit der Daten unter der Hypothese.
- Die Wahrscheinlichkeit der Hypothese $P(X = x)$ lässt sich nicht durch Experimente feststellen (da Hypothesen reine Gedankenkonstruktionen sind).
- Die Wahrscheinlichkeit der Daten $P(Y = y)$ lässt sich durch Experimente feststellen (wiederholte Messung und Häufigkeitszählung).

Satz von Bayes

Bedeutung des Satzes II

- $P(Y = y|X = x)$ lässt sich oft aus einem Modell berechnen. Nimmt man also die Hypothese als besonders wahrscheinlich an mit $P(X = x) = 0.95$, so kann ein Experiment über die Häufigkeit der Daten $Y = y$ uns verraten, wie groß $P(Y = y)$ ist und mit Bayes' Formel schliesslich $P(X = x|Y = y)$. Wenn dieser Wert klein ist, so ist die Hypothese unwahrscheinlich unter den gegebenen Daten, selbst wenn man die Hypothese als wahrscheinlich annimmt!
- Man muss folglich die Hypothese verwerfen, wenn $P(X = x|Y = y)$ klein ist. Ist $P(X = x|Y = y)$ hingegen groß, so hat man nichts gelernt, geht man doch von vornherein davon aus, dass die Hypothese gilt. In diesem Fall aber lässt sich probehalber die Gegenhypothese $X \neq x$ annehmen.

Satz von Bayes

Bayessche Induktion

- In diesem Zusammenhang ist es eigentlich wichtig, welche möglichen Verteilungen für die Daten man überhaupt in Betracht zieht. Wenn man hier rigoros arbeitet (und das muss man), so spricht man von *Parameterschätzern* in der Statistik.
- Wir sehen, dass sich mit dem Satz von Bayes keine absoluten Aussagen über Hypothesen treffen lassen, einzig relative Wahrscheinlichkeiten können mit Sicherheit betrachtet werden.
- Diese Schlussweise "angenommen A gilt, aber B ist unwahrscheinlich, wenn A gilt, obwohl B wahrscheinlich ist - also müssen wir A verwerfen" nennt man *Bayessche Induktion*.
- Dies ist die einzige gültige Schlussweise um wissenschaftliche Hypothesen am Experiment zu falsifizieren.

Induktion vs. Deduktion

Begrifflichkeit

- Seien A und B Aussagen. Wir wollen verschiedene Schlussweisen begrifflich voneinander trennen.
- *Deduktion* ist der Schluss von A auf B, wenn B eine logische Konsequenz aus A ist. Eine gültige Deduktion ist rein formal und innerhalb eines Kalküls sicher (der Kalkül kann allerdings durchaus nicht vollständig oder nicht widerspruchsfrei sein, aber das ist eine andere Geschichte). Ein gutes Beispiel ist der Syllogismus "Aristoteles ist ein Mensch, Menschen sind sterblich, also ist Aristoteles sterblich".
- *Induktion* ist der Schluss von A auf B ohne dass B aus A folgt. Das kann z.B. erfolgen, wenn A sehr viele Indizien für B liefert, ein Gegenbeispiel für B jedoch nicht ausgeschlossen ist (nur noch nicht gefunden wurde). Ein gutes Beispiel ist der Schluss "Alle Schwäne, die ich bis jetzt gesehen habe sind weiß, also sind alle Schwäne weiß".

Induktion vs. Deduktion vs. Abduktion

Wissenschaftstheorie

- *Abduktion* ist der Schluss von B auf A indem man B durch A erklärt. Auch dieser Schluss muss im Gegensatz zur Deduktion nicht immer sicher korrekte Antworten liefern, da es mehrere Erklärungen für eine Aussage geben kann.
- In der Wissenschaft (und strenggenommen Erkenntnistheorie) sind wir im Bezug auf eine äußere Realität (jenseits der Mathematik) auf Induktion und Abduktion angewiesen.
- Der wissenschaftliche Prozess, Phänomene durch Theorien zu "erklären" und zu testen ist Abduktion.
- Bayes kommt hier ins Spiel, da man eine Erklärung (Hypothese) A nicht direkt testet, sondern als Erklärung der Daten B überprüft. Wenn dies konsistent ist, lässt sich unter den verschiedenen Erklärungen (Hypothesen) noch diejenige mit den besten Eigenschaften (Einfachheit, Eleganz, Vorhersagekraft, Signifikanz der Übereinstimmung mit Daten, usw.) auswählen.

Bayes für bedingte Entropie

Korollar

- Der Satz von Bayes
  $P(X = x|Y = y) = P(Y = y|X = x)\frac{P(X=x)}{P(Y=y)}$ impliziert für gleichverteilte X, Y durch Logarithmierung:

  $$\log_2 P(X = x|Y = y) = \log_2 P(Y = y|X = x) + \log_2 P(X = x) - \log_2$$

  und Durchmultiplizieren mit -1 und Erwartungswertbildung liefert die Definition der Entropie bei Gleichverteilung, also

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

- Dieser Zusammenhang gilt auch ohne Annahme der Gleichverteilung, wie sich daraus oder auch direkter mit den vorherigen Resultaten beweisen lässt (Übungsaufgabe!)
- Es folgt durch weitere Anwendung dieses Satzes auf $-H(Z|X)$ und Addition:

  $$H(X|Y) - H(Z|X) = H(Y|X) - H(X|Z) + H(Z) - H(Y).$$

Korrelation

- Korrelation bezeichnet einen sichtbaren Zusammenhang zwischen Größen (Zufallsvariablen).

Kovarianz und Korrelation

Definition

- Für zwei Zufallsvariable X und Y heißt

$$\mathrm{Cov}(X, Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

  die *Kovarianz* von X und Y.
- Es ist $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
- X und Y heißen *unkorreliert*, wenn $\mathrm{Cov}(X, Y) = 0$ ist.
- Für unabhängige Zufallsvariable ist $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, also sind unabhängige Zufallsvariable stets unkorreliert.
- Es bezeichnet $\sigma(X) := \sqrt{\mathrm{Cov}(X, X)}$ die *Standardabweichung* und

$$\rho(X, Y) := \frac{\mathrm{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

  den *Pearson-Korrelationskoeffizienten* von X und Y.

Korrelation und Information

Ein anderes Korrelationsmaß

- Die *totale Korrelation* von Zufallsvariablen $X_1, ..., X_n$ ist definiert als

$$C(X_1, ..., X_n) := \sum_{i=1}^{n} H(X_i) - H(X_1, ..., X_n).$$

- Es handelt sich hierbei um die Kullback-Leibler-Divergenz zwischen der gemeinsamen Verteilung und der Produktverteilung. Bei Unabhängigkeit sind diese Verteilungen identisch, dann ist die totale Korrelation 0.

Grenzfälle

Gänzlich unkorrelierte Zufallsvariablen

- Eine Folge von Experimenten mit gleichen Versuchsaufbau modelliert man als Folge von Zufallsvariablen, die identisch verteilt sind. Sind diese Zufallsvariablen paarweise unkorreliert (im Sinne von Pearson, z.B. unabhängig), so genügen sie dem schwachen Gesetz der großen Zahlen, was für die Praxis bedeutet, dass ihre Mittelwerte $\frac{1}{n}\sum_{k=1}^{n} X_k$ immer bessere Näherungen an den Erwartungswert liefern.
- Kennt man z.B. die Verteilung nicht (etwa in einem physikalischen Experiment), interessiert sich aber für den Erwartungswert (weil er erste Informationen über die Verteilung liefert), so ist der Mittelwert nach n Experimenten eine gute Näherung (und eine um so bessere, je öfter man das Experiment wiederholt).

Grenzfälle

Abhängige Zufallsvariablen

- Ist eine Zufallsvariable X vollständig von Y abhängig, so ist die Information von X vollständig in der von Y enthalten.
- In Formeln: $P(X, Y) = P(Y)$ oder auch $P(X|Y) = \frac{P(X)}{P(Y)}$.
- Für die Entropie heisst das: $H(X, Y) = H(Y)$ oder auch $H(X|Y) = 0$.

Korrelation und Kausalität oder Erklärung

- Korrelation muss nichts mit Erklärung oder Kausalität zu tun haben!
- Wenn das Ereignis A stets mit B einher geht, kann das eine gemeinsame Ursache haben,
- kann A die Ursache für B oder umgekehrt sein,
- es kann sich aber auch um einen zufälligen Zusammenhang handeln!

Zusammenfassung

- Wir wollen noch einmal zusammenfassen, was wir uns bis hier hin merken wollen.

Zusammenfassung der wichtigsten Größen

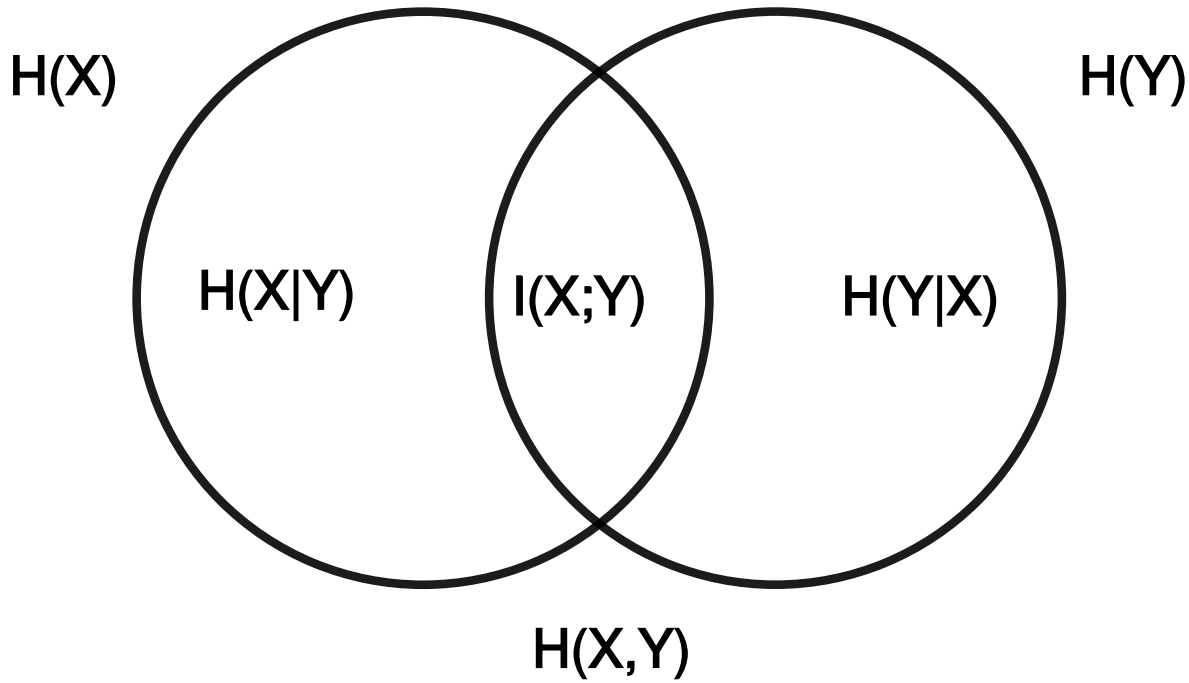
Mathematisch

- Eine Zufallsvariable $X : \Omega \to \Sigma = \{z_1, ..., z_n\}$ ist bestimmt durch ihre *Verteilungsfunktion* $z \mapsto P(X = z)$.
- Die *Selbstinformation* von X ist die Funktion $I_X : z \mapsto -\log_2 P(X = z)$.
- Die *Entropie* von X ist der Erwartungswert der Selbstinformation: $H(X) := \mathbb{E}(I_X)$.
- Die *Verbundentropie* von X und Y ist die Entropie der komponierten Zufallsvariablen: $H(X, Y) := H((X, Y))$.
- Die *gemeinsame Information* von X und Y ist $I(X; Y) = H(X) + H(Y) - H(X, Y)$.
- Die *bedingte Entropie* von X bezüglich Y ist $H(X|Y) = H(X, Y) - H(Y)$.
- Alle Definitionen lassen sich auf den Begriff der bedingten Entropie zurückführen (und dies ist ein Spezialfall der Kullback-Leibler-Divergenz, und lässt sich somit auf stetige Zufallsvariablen übertragen).

Zusammenfassung der wichtigsten Größen

Diagrammatisch

- Für zwei Zufallsvariablen X und Y lassen sich die grundlegenden informationstheoretischen Größen in einem Diagramm darstellen:



Redundanz und Kodierung

- Die optimale Kodierung im rauschfreien Kanal enthält keine Redundanz (minimale Kodierungslänge).
- Bei Rauschen ist Redundanz unbedingt notwendig, um fehlerfrei eine Nachricht zu rekonstruieren (zu dekodieren).

Rauschen

- Ein Signal, das verrauscht wird, modelliert man durch zwei reelle Zufallsvariablen: X ist das ursprüngliche Signal, Y ein Störsignal, $X + Y$ schließlich das verrauschte Signal.
- Rauschen folgt also auch einer bestimmten Verteilung.
- Im diskreten Fall modelliert man Rauschen z.B. durch eine Zufallsvariable Y, die von X abhängt (etwa, indem der Buchstabe “e” häufiger zu einem “i” verändert wird als zu einem “x” - so verhält sich z.B. das Rauschen in der natürlichen Sprache bei Lärm).
- Im Anwendungsfall der digitalen Kommunikation über einen anlogen Kanal ist es ausreichend, Rauschen in Normalverteilung zu betrachten. Wenn man den Fall der Kommunikation über Photonen ohne absorbierendes oder streuendes Material betrachtet, so ist diese Näherung sogar sehr präzise, da Wärmestrahlung (die dann einzige Störquelle) normalverteilt ist.

Redundanz und Fehlerkorrektur

- Die *(absolute) Redundanz* einer Zeichenkette ω ist die Differenz aus Kodierungslänge und Entropie. Die *relative Redundanz* ist das Verhältnis.
- Prinzipiell sind zum Dekodieren einer Zeichenkette nur so viele bits (Ja/Nein-Fragen) notwendig, wie die Entropie angibt. Jedes zusätzliche bit ist überflüssig.
- Verfälscht man jedoch eine Nachricht, so lässt sich Redundanz ausnutzen, um Fehler zu korrigieren.
- Sendet man eine Nachricht (z.B. ein einzelnes Bit) drei Mal, so lässt sich bereits ein einmal auftretender Fehler sicher korrigieren: $000, 001, 010, 100$ werden als 0 dekodiert, während $111, 110, 101, 011$ als 1 dekodiert werden. Hier beträgt die Entropie 1 bit, die Kodierungslänge 3 Bit, also beträgt die Redundanz 2 Bit, die zur Fehlerkorrektur offensichtlich notwendig sind.

Redundanz in der englischen Sprache

- Die Redundanz der englischen Sprache lässt sich aus der mittleren Kodierungslänge und der Entropie bestimmen.
- Legen wir als Entropie eines lateinischen Buchstabens den Wert $1,5$ bit fest, so ergibt sich zusammen mit der mittleren Wortlänge von $5,5$ Buchstaben eine Entropie von $8,25$ bit pro Wort (die wir nun als elementare Zeichen betrachten werden).
- Die typische Wortlänge von $5,5$ Buchstaben zusammen mit der Kodierungslänge von lateinischen Buchstaben in bit, nämlich $\log_2 27 \approx 4,75$ (wegen 26 Buchstaben und einem Leerzeichen), liefert die mittlere Kodierungslänge eines Wortes, nämlich $26,15$ bit.
- Als Differenz ergibt sich eine (absolute) Redundanz von $17,9$ bit, anders ausgedrückt, eine relative Redundanz von $68,45\%$.
- Man kann also einen englischen Text noch fehlerfrei rekonstruieren, wenn man ein Drittel der Zeichen weglässt - und einen geeigneten Code wählt!

Beispiele für Codes

Fehlerkorrigierende Codes

- Weit verbreitete Beispiele für fehlerkorrigierende Codes sind *Paritätscodes*, die alle k bit ein zusätzliches Bit hinzufügen, nämlich die Parität der Summe der vorangegangenen k Bits. Ist etwa $k = 2$, so würde der Code die Nachricht 00 mit 000 kodieren, 01 mit 011 und 11 mit 110. Die Nachrichten 111 und 001 sind unmöglich, also kann beim dekodieren von einem Übertragungsfehler ausgegangen werden.
- Solche Codes erfordern, dass die Nachricht bei erkanntem Fehler erneut gesendet wird. Spezialfälle solcher Situationen (und Codes) sind ISBN-Codes und die meisten Matrikelnummern an Universitäten.
- Elaboriertere Versionen dieser Codes enthalten genug Redundanz um z.B. einen Zahlendreher tatsächlich zu korrigieren, ohne ein erneutes Senden der Information zu erfordern.

Beispiele für Codes

Effiziente Codes

- Neben Fehlerfreiheit möchte man auch möglichst große Übertragungseffizienz, d.h. kleine Kodierungslänge, möglichst nah am Limit, der Entropie.
- Ein klassisches Beispiel für effiziente Codes, das noch vor Shannons Theorie erfunden wurde, ist der Morse-Code.
- Im Morse-Code sind häufig auftretende Signale durch kurze Zeichenketten repräsentiert, selten auftretende Signale durch lange Zeichenketten. Das ist effizienter als ein Code, der jedes Signal durch ein gleich langes Signal repräsentiert.
- Der Grund dafür liegt darin, dass die Entropie einer Gleichverteilung maximal ist. Wenn ein System also nicht gleichverteilt ist, so ist die optimale Kodierung auch nicht gleichverteilt.

Beispiele für Codes

Verlustbehaftete Codes

- Natürliche Sprache ist auch ein fehlerkorrigierender Code: Bei Lärm versteht man zwar meist nur die Hälfte, aber doch oft die Nachricht. Dabei betrachtet man also nicht den exakten Klang als Nachricht, sondern die abstrakte Aussage (wie sie auch geschrieben kodiert sein könnte).
- Viele moderne Codes, wie z.B. MP3 für englische Sprache, sind nicht vollständig fehlerkorrigierend. Stattdessen arbeiten sie mit wissenschaftlichen Erkenntnissen darüber, was Menschen hörphysiologisch und hörpsychologisch nicht wahrnehmen können, und verzichten nur auf diese Information.
- Vom Standpunkt der Informationstheorie hängt es nun davon ab, was man als "Nachricht" bezeichnet, ob man solch einen Code (unabhängig vom Rauschen) als verlustbehaftet ansieht, oder nicht.

Die Bedeutung des Satzes von Bayes

In der Kodierungstheorie

- Das Dekodieren einer verrauschten Nachricht lässt sich stets mathematisch beschreiben:
- Gegeben (unbekannte) gesendete Daten x, modelliert durch eine Zufallsvariable X mit bekannter Verteilung (gegeben durch Bekanntheit des verwendeten Codes) und gestörte (bekannte) Daten y, modelliert durch eine Zufallsvariable Y mit bekannter Verteilung (im Zweifelsfall nimmt man hier $Y = X + \mathcal{N}$ an, ein weisses Rauschen), so ist das gesuchte, aber unbekannte x dasjenige, welches die Größe

$$P(X = x|Y = y) = P(Y = y|X = x)\frac{P(X = x)}{P(Y = y)}$$

maximiert (denn sonst wäre der Code ungeeignet gewählt für das gegebene Rauschniveau).

Kanalkapazität

- Die *Kanalkapazität* ist definiert als die stärkste obere Schranke für die Menge an Information, die über einen Kommunikationskanal übertragen werden kann.
- Formal definiert man

$$C := \sup_X(I(X;Y)),$$

  X modelliert die gesendete Nachricht, Y die empfangene (verrauschte).
- Die Information des Rauschens steckt vollständig in der Wahrscheinlichkeitsverteilung $z \mapsto P(Y = z|X = x)$, wobei x die tatsächlich gesendete (unbekannte) Nachricht ist.

Shannons Quellenkodierungstheorem

- Shannons Theorem zur Quellenkodierung, oder auch Shannons Theorem zur unverrauschten Kommunikation, besagt, dass N unabhängige, identisch verteilte Zufallsvariablen (z.B. eine Nachricht aus Zeichen über einem endlichen Zeichenvorrat mit bekannter Verteilung) $X_1, ..., X_N$ mit jeweiliger Entropie $H(X_i) = H(X_1)$ komprimiert werden können in (unwesentlich) mehr als $NH(X_1)$ bits mit beliebig kleinem Fehler.
- Komprimiert man sie jedoch in weniger als $NH(X_1)$ bits, so sind nicht-korrigierbare Fehler P-fast-sicher.
- Mit "auf n bits komprimiert" ist hier gemeint, dass es einen Code gibt, der für jeden möglichen Ausgang der Zufallsvariablen die resultierende Zeichenkette $x_1...x_N$ in eine Zeichenkette der Länge k mit $k \leq n$ kodiert.
- Das bedeutet entsprechend, dass die mittlere Kodierungslänge eines Zeichens maximal der Entropie der Zeichenverteilung entsprechen kann, ansonsten sind Fehler nicht zu korrigieren.

Shannons Quellenkodierungstheorem

Die Rolle der Redundanz

- Ein Code, der im Mittel mehr bits pro Zeichen verwendet als die Entropie der Zeichenverteilung, ist redundant.
- Das Quellenkodierungstheorem sagt nun, dass man bei Abwesenheit von Rauschen nicht auf Redundanz verzichten kann, diese jedoch beliebig klein halten kann.

Shannons Kanalkodierungstheorem

Kodierungstheorem für verrauschte Kanäle

- Für jedes $\epsilon > 0$ und jede Übertragungsrate (von Zeichen pro Sekunde) R gilt:
  ist $R < C$, die Kanalkapazität, so gibt es einen Code, deren Fehlerrate pro Zeichen für hinreichend lange Zeichenketten kleiner als ϵ wird.
- Man kann also durch Wahl eines geeigneten Codes beliebig nah mit beliebig kleinem Fehler an die obere Schranke der Übertragungsrate $R = C$ gelangen. In der Tat ist diese Schranke in einigen Fällen von Nachrichten (Text, Bild, Ton) in den letzten 10 Jahren bereits zu über 90% erreicht worden.

Shannons Theorem für verrauschte Kanäle

Die Rolle der Redundanz

- Im Gegensatz zum unverrauschten Kanal lässt sich bei Anwesenheit von Rauschen keine Kodierung finden, die beliebig nah an das Limit der Entropie eines Zeichens herankommt.
- Es ist notwendig, zur Fehlerkorrektur einen Code mit längerer Kodierungslänge zu wählen. Man fügt also Redundanz hinzu.
- In der gesprochenen Sprache als Kanal für geschriebene Sätze wird die Redundanz der natürlichen Sprache zur Fehlerkorrektur eingesetzt.

Shannon-Hartley-Theorem

Bandbreitenbeschränkter Kanal mit weissem Rauschen

- Angenommen, das Rauschen eines Kanals ist jederzeit Datenunabhängig Standard-Normalverteilt, man spricht dann auch von additiven weissen Gaussschen Rauschen. “Weiss” heisst hierbei gleiche Menge an Rauschen auf allen Frequenzen innerhalb der Kanalbandbreite, während “farbig” gesagt wird, wenn das Rauschen von der Frequenz (also der Wellenlänge bzw. Farbe des Lichts) abhängt.
- Bezeichne mit S/N das Signal-Rausch-Verhältnis (signal-noise-ratio) und mit B die Bandbreite (in Hertz). Die Bandbreite ist gegeben durch die Anzahl an Zeichen, die pro Sekunde übertragen werden können.
- Das Shannon-Hartley-Theorem sagt nun, dass

$$C = B \log_2 \left(1 + \frac{S}{N}\right).$$

WLAN-Signale

Weisses Rauschen im stetigen Signal

- Bezeichne mit P die empfangene *Leistung* (in Watt) und N_0 die Rauschleistungsspektraldichte (in Wattsekunden).
- Dann ist die Kanalkapazität

$$C = W \log_2 \left(1 + \frac{P}{N_0 W}\right) \text{ in bits/Hertz.}$$

- Die Größe $\frac{P}{N_0 W}$ bezeichnet man als SNR (signal-noise-range).
- Ist die SNR groß, so ist die Kapazität linear in der Bandbreite und logarithmisch in der Leistung. Dies nennt man *Bandbreiten-beschränktes Regime*.
- Ist die SNR klein, so hängt die Kapazität linear von der Leistung aber unwesentlich von der Bandreite ab. Dies nennt man *Leistungs-beschränktes Regime*.

Satz von Wolfowitz

Klippen-Effekt

- Der Satz von Wolfowitz (1957) besagt, dass es eine endliche positive Konstante A gibt, sodass

$$P_{error} \geq 1 - \frac{4A}{n(R-C)^2} e^{-n(R-C)}.$$

- Daraus folgt, dass für Kommunikation bei Raten oberhalb der Kanalkapazität die Fehlerrate exponentiell gegen 1 geht.
- Graphisch ist die Fehlerrate in Abhängigkeit der Differenz aus Übertragungsrate und Kanalkapazität also eine sehr steil ansteigende Funktion, zentriert um die Kanalkapazität, beinahe 0 vor Erreichen der Kapazität, fast sofort 1 nach überschreiten der Kanalkapazität.
- Darum brechen Handy-Gespräche bei schlechtem Empfang oft auch abrupt vollständig ab, anstatt zunächst mit schlechterer Übertragungsqualität weiter zu arbeiten.

Weisses Rauschen als Modell

- Weisses Rauschen ist nicht nur dann ein gutes Modell, wenn die Kommunikation nur durch Thermalstrahlung gestört wird.
- Die Summe (und damit das arithmetische Mittel) unabhängiger, normalverteilter Zufallsvariablen ist wieder Normalverteilt.
- Nach dem zentralen Grenzwertsatz konvergiert die Verteilung von n unabhängigen identisch (nicht notwendig normal) verteilten Zufallsvariablen für großes n gegen eine Normalverteilung.
- Damit lässt sich oft rechtfertigen, die empfangenen Zeichen als normalverteilt zu modellieren (da in diesem Zustand die Übertragungsrate sehr nah an der Kanalkapazität wäre).
- Ebenso lassen sich andere Störsignale mit einer Normalverteilung modellieren. Die resultierende Verteilung des empfangenen Signals ist dann automatisch auch normalverteilt.

Beweisidee des Kanalkodierungssatzes

- Der Beweis ist nichtkonstruktiv, d.h. es werden keine Codes angegeben, die die gewünschten Eigenschaften haben.
- Dass man Raten $R < C$ mit Fehler $< \epsilon$ erreichen kann, beweist man durch einen zufällig gewählten Code (repräsentiert durch eine Zufallsvariable) und Analyse des Fehlers bei diesem Code. Dieser Fehler wird über alle Symbole des Codes gemittelt und diese Größe wird über alle möglichen (zufällig gewählten) Codes gemittelt.

Shannon-Entropie

- Shannon-Entropie ist der Erwartungswert der Selbstinformation.
- Relative Shannon-Entropie ist Kullback-Leibler-Divergenz.

Fisher-Information

- Fisher-Information ist die *Krümmung* der Kullback-Leibler-Divergenz.

Rényi-Entropie

- Rényi-Divergenz verallgemeinert die Kullback-Leibler-Divergenz und damit (unter dem Namen Rényi-Entropie) die Shannon-Entropie, mit einem Parameter α, sodass dem Fall $\alpha = 1$ die Kullback-Leibler-Divergenz bzw. die Shannon-Entropie entspricht.
- Nur in diesem Fall ($\alpha = 1$) kann man sinnvoll die Begriffe der bedingten und gemeinsamen Information entwickeln.

Kolmogorov-Komplexität

- Die *Kolmogorov-Komplexität* einer Zeichenkette ist die minimale Länge eines Algorithmus in einer festen Beschreibungssprache.
- Kolmogorov konnte zeigen, dass sich die Kolmogorov-Komplexität von Zeichenketten bezüglich zwei verschiedener Beschreibungssprachen nur um eine Konstante unterscheiden.
- Während die Shannon-Entropie einer Zeichenkette 000000 davon abhängt, welche Nachrichten prinzipiell erwartet werden, ist die Kolmogorov-Komplexität unabhängig davon.