

Dr. Jürg M. Stettbacher

Neugutstrasse 54

CH-8600 Dübendorf

Telefon: +41 43 299 57 23

E-Mail: dsp@stettbacher.ch

Informationstheorie

Grundlagen der Quellen- und Kanalcodierung

Version 2.00
2017-10-08

Zusammenfassung: Es werden die mathematischen Grundlagen der Informationstheorie eingeführt und Anwendungen in der Quellen- und Kanalcodierung aufgezeigt. Das grosse Thema der Kryptografie wird jedoch nur in der Einleitung angeschnitten.

Inhaltsverzeichnis

1 Zweck	4
2 Einleitung	4
2.1 Kurze Geschichte der Kryptografie	4
2.2 Geburt der Informationstheorie	8
2.3 Wesen der Informationstheorie	10
2.4 Anwendung der Informationstheorie	10
3 Information	11
3.1 Begriff	11
3.2 Mass	12
3.3 Informationsgehalt von Ereignissen	13
3.3.1 Quellen-Modell	13
3.3.2 Wahrscheinlichkeiten	14
3.3.3 Informationsgehalt	16
3.3.4 Eigenschaften	17
3.3.5 Mehrfach-Ereignisse	18
3.3.6 Statistisch unabhängige Mehrfach-Ereignisse	19
3.3.7 Masseinheit	21
3.4 Zusammenfassung	21
4 Entropie	21
4.1 Begriff	21
4.2 Entropie von Quellen	22
4.2.1 Eigenschaften	25
4.2.2 Mehrfach-Quellen	27
4.2.3 Statistisch unabhängige Mehrfach-Quellen	28
4.2.4 Masseinheit	29
4.3 Zusammenfassung	30

5 Anhang	31
5.1 Wahrscheinlichkeitstheorie	31
5.1.1 Zufallsvariablen	31
5.1.2 Statistische Abhangigkeit und Unabhangigkeit	32
5.1.3 Wahrscheinlichkeit	33
5.1.4 Verbund-Wahrscheinlichkeit	35
5.1.5 Totale Wahrscheinlichkeit	37
5.1.6 Bedingte Wahrscheinlichkeit	38
5.2 Logarithmen	40
5.3 Information von statistisch abhangigen Ereignissen	41
5.3.1 Verbund-Informationsgehalt	41
5.3.2 Bedingter Informationsgehalt	41
5.4 Entropie von statistisch abhangigen Quellen	43
5.4.1 Verbund-Entropie	44
5.4.2 Bedingte Entropie	45

Information is not knowledge
knowledge is not wisdom
wisdom is not truth
truth is not beauty
beauty is not love
love is not music
music is the best

Frank Zappa

1 Zweck

Ziel dieses Dokuments ist es,

- den Geltungsbereich der *Informationstheorie* abzustecken,
- zu zeigen, wie Information in diesem Kontext *gemessen* wird,
- die Begriffe *Entropie* und *Redundanz* einzuführen,
- sowie Anwendungen und Beispiele¹ davon zu zeigen.

2 Einleitung

Die Informationstheorie ist eine recht junge Wissenschaft. Ihre Quelle war die Kryptografie², also die uralte Absicht der Menschen, Information so festzuhalten oder zu übertragen, dass sie vor unbefugten Augen und Ohren verborgen bleibt.

Die folgenden Abschnitte beleuchten die Entstehungsgeschichte der Informationstheorie. Wer sich nur für harte Fakten interessiert kann diese Einleitung getrost überspringen.

2.1 Kurze Geschichte der Kryptografie

Vor 2000 Jahren Der Überlieferung nach, soll schon Julius Cäsar³ geheime Meldungen verschlüsselt haben, indem er die Nachricht aufschrieb und dabei die Buchstaben des Textes mit Hilfe einer Tabelle durch andere ersetzte⁴ (siehe Abbildung 1). Wer im Besitz derselben Tabelle war, konnte den verschlüsselten Text wieder entschlüsseln und lesbar machen.

¹ In diesem Dokument sind Beispiele zur Kennzeichnung von den Symbolen ▼ und ▲ umschlossen.

² Kryptografie stammt aus dem Griechischen und heisst ungefähr *verborgenes Schreiben* oder *geheimes Schreiben*.

³ Gaius Julius Cäsar, 100 bis 44 v. Chr., römischer Feldherr und Staatsmann.

⁴ Man nennt diese Verfahren auch *monoalphabetische Substitution*.

Cäsars Methode war allerdings leicht zu knacken, denn in jeder Sprache treten die einzelnen Buchstaben mit einer spezifischen Häufigkeit auf. In der deutschen Sprache trifft man zum Beispiel *x* und *y* sehr viel seltener an als die Buchstaben *a* oder *e*. Hat man also einen genügend langen, verschlüsselten Text vorliegen, so lässt sich mit Hilfe einer Häufigkeitsanalyse die Verschlüsselungstabelle erraten.

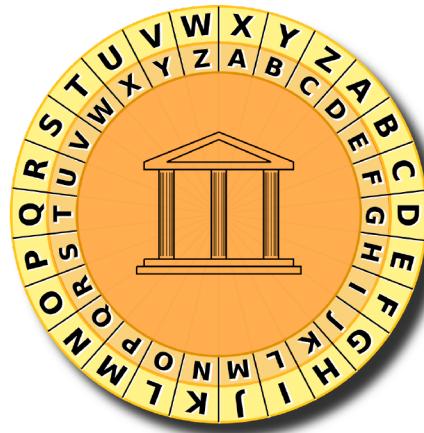


Abbildung 1: Cäsar wählte eine Tabelle, in der die Buchstaben des verschlüsselten Textes im Alphabet um drei Stellen nach rechts verschoben waren. Beispiel: Aus dem Originaltext KRYPTOGRAFIE (äusserer Ring) wird der verschlüsselte Text NUBSWRJUDILH (innerer Ring).

Vor 450 Jahren Trotz dieser Schwäche wurde Cäsars Verfahren über Jahrhunderte hinweg mit leichten Abwandlungen verwendet. Erst zu Beginn der Neuzeit publizierte Blaise de Vigenère⁵ ein Verfahren, bei welchem Cäsars Methode durch einen Tabellenwechsel erweitert wurde. Aus einer Sammlung von verfügbaren Verschlüsselungstabellen wurde ein fixes Set mit definierter Reihenfolge ausgewählt. Bei der Chiffrierung wurde der Reihe nach je ein Buchstabe des Originaltextes mit Hilfe einer Tabelle aus dem Set vertauscht, wobei das Set der Tabellen zyklisch verwendet wurde. Damit wurde die Häufigkeit der einzelnen Buchstaben verwischt und der Code galt für lange Zeit als unknackbar.

⁵ Blaise de Vigenère, 1523 bis 1596, französischer Diplomat und Autor mit Interesse an Kryptografie.

Vor 250 Jahren 1854 jedoch beschrieb Charles Babbage⁶ eine Methode, mit der sich Vigenères Code auch ohne Kenntnis des Tabellensets entschlüsseln liess. Es reicht nämlich, wenn man die Anzahl Tabellen im Set erraten kann. Damit ist auf Grund der zyklischen Anwendung der Tabellen sofort klar, welche Buchstaben im Text jeweils mit der selben Tabelle ersetzt wurden. Diese Buchstaben wählt man aus und kann anschliessend, mit Hilfe der Buchstabenhäufigkeiten wie bei Cäsars Verfahren, die betreffende Tabelle rekonstruieren.

Vor 100 Jahren Im ersten Weltkrieg von 1914 bis 1918 spielte die von Fritz Nebel⁷ entwickelte ADFGX⁸ Verschlüsselung des deutschen Heeres in Frankreich eine wichtige Rolle. Bei diesem Verfahren wird zuerst jeder Buchstabe der Originalnachricht mit Hilfe einer Tabelle auf eine Zweiergruppe von Buchstaben aus der Menge $\{A, D, F, G, X\}$ abgebildet. Diese fünf Buchstaben sind im Morsecode, welcher bei der Funkübertragung von Meldungen und Befehlen verwendet wurde, besonders deutlich unterscheidbar. Im zweiten Schritt war die resultierende Buchstabenreihe zeilenweise in eine weitere Tabelle einzutragen und danach wieder spaltenweise auszulesen. Diese Transposition vermischt die Buchstaben und löste die ursprünglichen Buchstabenpaare auf, was dann die verschlüsselte Nachricht ergab (siehe Abbildung 2).

Es war klar, dass der Feind die per Funk übermittelten ADFGX-Botschaften mithören würde. Aber war es menschenmöglich, die Mitteilungen zu entziffern? Garantierte ADFGX totale Sicherheit?

Nein, es sollte sich bald zeigen, dass auch dieses Verfahren nicht standhalten konnte. Durch die Anwendung von ausgedehnten statistischen Analysen und viel Knochenarbeit, die damals selbstverständlich von Hand geleistet wurde, gelang es dem französischen Nachrichtenoffizier Georges Painvin⁹ im Juni 1918, das Verfahren zu knacken¹⁰. Es zeigte sich, dass analog zu Vigenères Technik nur der zweite Schritt, die Transposition, wirklich kritisch und aufwändig zu entschlüsseln war.

Nach dem Krieg war klar, dass Kryptografie in Zukunft - im zivilen wie im militärischen Bereich - eine ganz wichtige Rolle spielen würde. Und es musste mehr in die Forschung investiert werden, denn praktisch alle bis dahin bekannten Verschlüsselungsverfahren hatten letztlich versagt. Zudem kamen mit der sich rasch entwickelnden Telekommunikation¹¹ neue Anwendungsfelder hinzu. Die zentrale Frage war: Gibt es absolute Sicherheit in der Kryptografie?

⁶ Charles Babbage, 1791 bis 1871, englischer Mathematiker und Erfinder einer mechanischen Rechenmaschine, die heute als Meilenstein auf dem Weg zum modernen Computer angesehen wird.

⁷ Fritz Nebel, 1891 bis 1967, deutscher Nachrichtenoffizier im ersten Weltkrieg.

⁸ Tatsächlich gab es zwei ähnliche Verfahren, nämlich ADFGX und ADFGVX, die beide im Jahr 1918 eingeführt wurden. Im Gegensatz zu ADFGX konnte ADFGVX auch die Ziffern 0 bis 9 verschlüsseln.

⁹ Georges Jean Painvin, 1886 bis 1980, französischer Industrieller. Während dem 1. Weltkrieg war er als Kryptoanalytiker tätig, allerdings ohne spezielle Vorbildung.

¹⁰ Painvins Durchbruch bei der Entschlüsselung von ADFGX im April 1918 gilt heute als massgeblicher Faktor für die Wende im Krieg zugunsten der Entente-Mächte an der deutschen Westfront. Insbesondere konnte das deutsche Heer vor Paris erfolgreich gestoppt werden.

¹¹ Die Telekommunikation umfasste zuerst Telegrafie und Telefonie via Kabel oder kabellos. Später kamen Radio und Fernsehen, sowie die allgemeine Datenkommunikation dazu.

1. Tabelle

		2				
		A	D	F	G	X
1	A	D	R	E	O	Z
	D	Y	U	K	A	S
	F	Q	H	N	W	B
	G	I	C	V	F	G
	X	T	M	L	X	P

2. Tabelle

2	3	5	1	7	4	6

Beispiel

Originaltext: KRYPTOGRAFIE

- (a) Wähle ein Tabellenset aus (siehe oben).
Beachte: Die zweite Tabelle ist noch leer.

- (b) Übersetze jeden Buchstaben des Originaltextes mit Hilfe der ersten Tabelle in ein Buchstabenpaar:
DF AD DA XX XA AG GX AD DG GG GA AF

- (c) Setzen die Buchstabenpaare zeilenweise in die zweite Tabelle ein:

- (d) Lese die Buchstaben spaltenweise aus der zweiten Tabelle aus:
aus: DAGDX AAFXD AAGGA ADFXX GDGG

2	3	5	1	7	4	6
D	F	A	D	D	A	X
X	X	A	A	G	G	X
A	D	D	G	G	G	G
A	A	F				

Abbildung 2: Die ADFGX-Verschlüsselung verwendet zwei Tabellen, wovon die zweite am Anfang leer ist. Im Beispiel wird die erste Tabelle verwendet um den Originaltext KRYPTOGRAFIE buchstabenweise in Doppelbuchstaben umzuwandeln. Aus K wird DF, aus R wird AD, usw. Beachte, dass die erste Tabelle nur 25 helle Felder hat und daher ein Buchstabe des Alphabets fehlt. Es wird darum J = I gesetzt. Anschliessend wird der neue Text zeilenweise in die zweite Tabelle eingetragen. Die Nummern im Kopf der zweiten Tabelle geben an, in welcher Reihenfolge die Spalten auszulesen sind. Das Resultat des Auslesens ist die verschlüsselte Nachricht.

Ja, und diese Frage war sicher brennend. Aber nach der Lektüre dieses Abschnitts mag man sich zu Recht fragen, wo denn die Informationstheorie bleibt. Hier kommt sie ...

2.2 Geburt der Informationstheorie

1920 Kurz nach dem ersten Weltkrieg tauchte eine weitere kryptografische Variante auf. Und wiederum war es nicht wirklich eine revolutionäre Idee, die Vernam¹² und Mauborgne¹³ um 1920 präsentierten. Neu war der Vorschlag auch nicht, denn Frank Miller¹⁴ hatte ihn bereits 1882 skizziert, allerdings ohne viel Beachtung zu finden. Ähnlich wie bei Vigenère wurden wieder die Zeichen des Originaltextes substituiert. Aber statt der wiederholten Anwendung eines Sets von Übersetzungstabellen wurde sozusagen für jeden Buchstaben der Nachricht eine eigene Tabelle benutzt. In der Praxis handelte es sich nicht mehr um echte Tabellen, sondern es wurde für jeden Buchstaben nur die Verschiebung im Alphabet als Zahl gespeichert. Die Liste all dieser Zahlen bildete einen sogenannten *Schlüssel* für die Chiffrierung. Um einem allfälligen Angreifer das Leben noch schwerer zu machen, wurde empfohlen, die Elemente des Schlüssels zufällig zu wählen und jeden Schlüssel nur einmal zu verwenden und anschliessend zu vernichten. Daher nennt man derartige Verfahren *One-Time Pad*¹⁵ (OTP). Technisch gesehen handelt es sich um einen sogenannten *Stream Cipher*¹⁶, der jeden vorbei kommenden Buchstaben einer Nachricht einzeln verschlüsselt und sogleich weiter schickt.

Man war sich weitgehend einig, dass der neue OTP Cipher sehr viel schwieriger zu knacken war, als alle bisherigen Verfahren. Aber *wie* schwierig war er zu knacken? War es überhaupt möglich, ihn zu knacken? Diese Fragen konnte vorderhand niemand beantworten. Trotzdem wurde das OTP Verfahren eingesetzt, zum Beispiel in diplomatischen Diensten und in der Spionage. Günstig war, dass sich der Algorithmus sowohl für die manuelle, wie für die maschinelle Verarbeitung eignete.

1945 Beim Ausbruch des zweiten Weltkriegs verstärken die involvierten Nationen ihre kryptologische Forschung. So betrieben die USA zahlreiche geheime, militärische Forschungsprogramme mit Wissenschaftern verschiedener Universitäten und privater Organisationen. Zu letzteren zählten die Bell Labs¹⁷, die weltweit zu den grössten und innovativsten privaten Forschungs- und Entwicklungseinrichtungen zählen. Dort arbeitete zu der Zeit Claude Shannon¹⁸ nebst zahlreichen weiteren

¹² Gilbert Sandford Vernam, 1890 bis 1960, arbeitete als US-amerikanischer Ingenieur bei AT&T (später Bell Labs). Er erfand und patentierte 1917 eine Maschine für die fortlaufende Chiffrierung von Text.

¹³ Joseph Oswald Mauborgne, 1881 bis 1971, hatte als Kryptologe eine führende Funktion in der US-amerikanischen Armee.

¹⁴ Frank Miller, 1842 bis 1920, US-amerikanischer Bankier, mit Interesse an Kryptografie.

¹⁵ Mit Pad bezeichnete man einen Block mit kleinen Abreisszetteln, welche die Schlüssel enthielten. Die Zettel wurden für die manuelle Ver- und Entschlüsselung benutzt und anschliessend verbrannt. Jeder Zettel enthielt nebst dem Schlüssel eine Nummer, welche der Codierten Nachricht voran gestellt wurde, so dass der Empfänger wusste, mit welchem Schlüssel die Nachricht zu dechiffrieren war.

¹⁶ Im Gegensatz dazu verarbeitet ein *Block Cipher* nur Gruppen einer bestimmten Anzahl von Buchstaben.

¹⁷ Die Bell Laboratories wurden 1925 von Western Electric und AT&T gegründet. Beide Firmen waren führend in der Telekommunikation. Nebst eigenen Projekten übernahmen die Bell Labs auch staatliche Aufträge, insbesondere in den Bereichen Verteidigung und Grundlagenforschung. Seit 1996 gehören die Bell Labs zu Lucent Technologies. Zu den wichtigsten Erfolgen der Bell Labs zählen die Entwicklung des ersten Bipolartransistors (1947), der ersten Silizium-Solarzelle (1954), des ersten Gaslasers (1960), des ersten CCD-Bildsensors (1969), des Unix-Betriebssystems (1969), sowie der Programmiersprachen C (1972) und C++ (1985).

Exponenten und Pionieren der (späteren) Mathematik, Kommunikationstechnik, Signalverarbeitung, Informationstheorie, Kryptologie, Informatik, usw.

Shannon gelang 1945¹⁹ ein formaler Beweis dafür, dass der OTP Cipher unknackbar ist, sofern die folgenden Bedingungen erfüllt sind:

- Der Schlüssel muss die selbe Länge haben wie die Nachricht.
- Die unterschiedlichen Verschiebungszahlen im Schlüssels müssen gleich häufig vorkommen und zufällig angeordnet sein.
- Jeder Schlüssel darf nur einmal verwendet werden.
- Die Schlüssel müssen geheim bleiben.

Für den Beweis hatte Shannon eine eigene Theorie der technischen Kommunikation entwickelt. Aus Gründen der Geheimhaltung durfte er sie während dem Krieg jedoch nicht publizieren. Dies wurde ihm erst danach erlaubt. Sodann publizierte er die drei wegweisenden Papers:

1. *A Mathematical Theory of Communication* (1948).

Der Artikel konstituierte sozusagen die moderne Informationstheorie, indem Shannon drei Dinge zeigte, nämlich

- wie Information gemessen werden kann,
- wie stark sich Nachrichten komprimieren lassen ohne Information zu verlieren, und
- wie viel Information sich fehlerfrei über einen fehlerbehafteten Kommunikationskanal übertragen lässt.

2. *Communication in the Presence of Noise* (1949).

In diesem Text bewies Shannon das Abtasttheorem²⁰. Dieses zeigt, unter welchen Bedingungen man aus einem abgetasteten Signal das ursprüngliche kontinuierliche Signal verlustfrei wieder herstellen kann. Das Abtasttheorem ist von zentraler Bedeutung bei abtastenden Systemen und in der digitalen Signalverarbeitung.

¹⁸ Claude Elwood Shannon, 1916 bis 2001, studierte Elektrotechnik und Mathematik. Nach seinem Doktorat forschte er am Institute for Advanced Study in Princeton an einer mathematischen Behandlung der technischen Kommunikation. Ab 1941 arbeitete er bei den Bell Labs in den Bereichen Signalverarbeitung und Kryptografie. 1956 wurde Shannon Professor am Massachusetts Institute of Technology (MIT), wo er bis zu seiner Pensionierung 1978 blieb. Seine wichtigsten Leistungen waren die Formulierung des Abtasttheorems (1949), sowie seine grundlegenden Beiträge zur mathematischen Informationstheorie.

¹⁹ Unabhängig davon hatte der russische Wissenschaftler Vladimir Aleksandrovich Kotelnikov (1908 bis 2005) den selben Beweis bereits 1941 erbracht.

²⁰ Das Abtasttheorem wurde in ähnlicher Form unabhängig von Shannon bereits formuliert durch E. T. Whittaker (1915), H. Nyquist (1928), V. A. Kotelnikov (1933) und weiteren.

3. *Communication Theory of Secrecy Systems* (1949).

In dieser Schrift behandelt Shannen die Kryptografie mit mathematischen Methoden. Unter anderem erbringt er den Beweis, dass das OTP-Verfahren unknackbar ist, wenn es richtig angewendet wird.

Wie die Titel seiner Publikationen verraten, ging es Shannon immer um technische Kommunikation. Dank seinem mathematischen Ansatz gelangen ihm zahlreiche wichtige Beweise. Und wie schon erwähnt legten seine Ausführungen das Fundament für die heutige Informationstheorie.

2.3 Wesen der Informationstheorie

Entsprechend Shannons Vorlage ist die Informationstheorie eine technische Wissenschaft, die sich gerne mathematischer Methoden bedient, insbesondere der Wahrscheinlichkeitstheorie. Entsprechend wird der Begriff der Information sehr nüchtern betrachtet. Er ist bar jeder emotionalen oder persönlichen Komponente. So freuen wir uns beispielsweise, wenn wir die Nachricht bekommen, wir hätten im Lotto gewonnen. Für uns impliziert das, dass wir reich geworden sind, wir endlich Häuser und italienische Sportwagen kaufen können, nicht mehr arbeiten müssen, usw. Der technische Informati onsbe griff kümmert sich um all das nicht. Statt dessen weist er dem Lotto-Gewinn bei 6 aus 42 einen Informationsgehalt von 22.3 Bit zu. Der Nachricht, dass wir im Lotto keinen Sechser erzielt haben, gibt er dagegen einen Informationsgehalt von ungefähr 0.0 Bit. Warum das so ist, werden wir in Kürze sehen.

2.4 Anwendung der Informationstheorie

Die moderne Informationstheorie umfasst die folgenden drei hauptsächlichen Anwendungsbereiche. Wie schon erwähnt, werden wir uns hier nur mit den ersten beiden davon beschäftigen.

1. *Datenkompression*.

Wie stark kann man Daten komprimieren ohne Information zu verlieren?

2. *Datenübertragung*.

Wieviel Information kann man pro Zeiteinheit über einen nicht idealen Kanal übertragen?

3. *Datenverschlüsselung*.

Wie kann man Information unkenntlich machen ohne sie zu verlieren?

Die drei typischen Anwendungsbereiche lassen sich in einem allgemeinen Kommunikationssystemen identifizieren (siehe Abbildung 3). Wir werden uns im Folgenden auf die Übertragung oder Speicherung von digitalen Daten beschränken. Dabei werden die Daten einer Quelle nacheinander komprimiert, verschlüsselt und bei der Übertragung oder Speicherung mit einem Fehlerschutz versehen. Auf dem Kanal oder Speichermedium können die Daten verfälscht werden. Der Empfänger versucht nun aufgrund des Fehlerschutzes allfällige Fehler zu korrigieren, dann wird die Verschlüsselung und

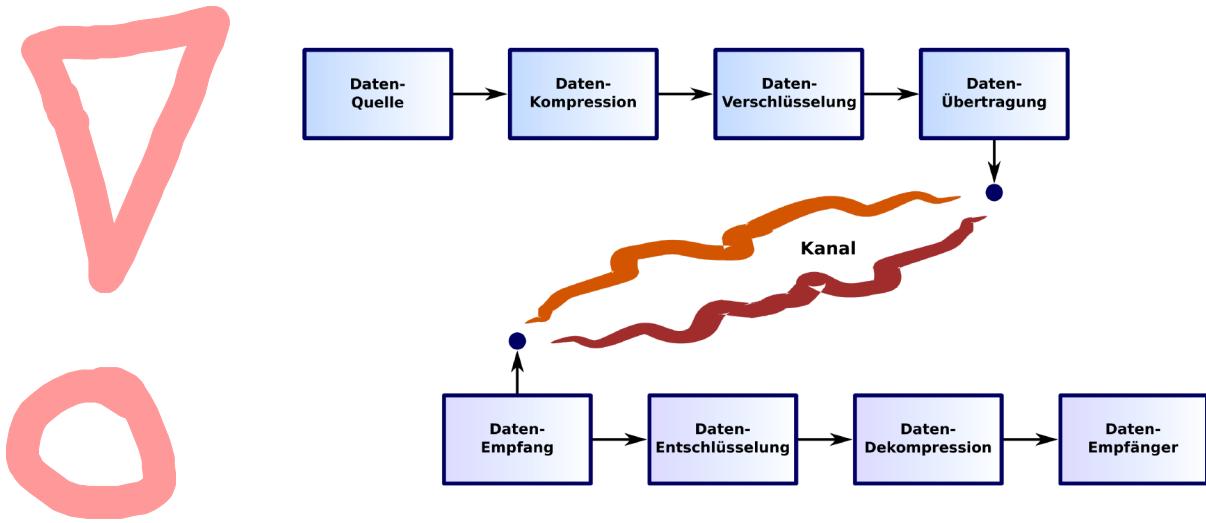


Abbildung 3: Schema eines allgemeinen Kommunikationssystems.

die Kompression rückgängig gemacht, bevor die Daten in der (hoffentlich²¹) ursprünglichen Form dem Empfänger übergeben werden. Beispiele für derartige Systeme sind alle Einrichtungen für die Telekommunikation, Datenübertragung, Datenspeicherung, usw.

3 Information

Information ist allgegenwärtig und wichtig, denn sie kann unser Verhalten und unser Zusammenleben beeinflussen. Aber versucht man, das Wesen der Information zu erklären oder zu definieren, so gerät man rasch in Schwierigkeiten.

3.1 Begriff

Auch das Lexikon tut sich schwer mit dem Begriff der Information und seinen zahlreichen Facetten. Man findet die folgenden Grundzüge:

²¹ Wir werden sehen, dass es den perfekten Fehlerschutz nicht gibt. Daher besteht immer eine Wahrscheinlichkeit für einen Restfehler bei der Übertragung.

- Information ist irgendwie mitgeteiltes, erlangtes oder gespeichertes Wissen.
- Information erweitert das Wissen, resp. hebt Unwissen auf.
- Information hat einen Nutzen und einen Wert.

Information kommt oft in der Gestalt von Nachrichten oder Mitteilungen daher. Wir lesen, sehen oder hören Information. Im Grunde können wir sie mit all unseren Sinnen wahrnehmen. Wenn wir Information empfangen, so erkennen wir sie als solche. Wir erhalten damit Wissen, das wir vorher noch nicht hatten. Wenn wir dieselbe Mitteilung erneut erhalten, so gibt sie uns keine Information mehr. Nur neue Information ist Information.

Die Informationstheorie liefert uns mathematische Werkzeuge, mit denen wir Information messen können. Objektive Aussagen zur Grösse von Information sind jedoch nur denkbar, wenn jegliche subjektiven Aspekte ausgeblendet werden. Wie aber geht das, wenn ich beispielsweise von einem Lotto-Gewinn spreche?

3.2 Mass

Problem Wenn wir die Länge einer Holzlatte messen wollen, oder den elektrischen Strom durch einen Draht, so ist schnell klar, wie wir vorzugehen haben. Für die Latte nehmen wir ein Referenzmass, zum Beispiel ein Messband. Beim Strom ist es etwas schwieriger, denn wir haben keinen direkten Zugriff auf die bewegten Ladungsträger, die in der Summe den Strom ausmachen. Ein klassisches Ampèremeter misst daher einen sekundären Effekt des Stroms, zum Beispiel den Spannungsabfall über einem Referenzwiderstand oder das magnetische Feld, das vom Strom verursacht wird. Im Gegensatz zu diesen physikalisch greifbaren Grössen ist Information ein abstrakter Begriff.

▼ Nehmen wir zum Beispiel die Mitteilung: *Es ist schlechtes Wetter*. Die Information bewirkt vielleicht, dass wir einen geplanten Ausflug verschieben, oder dass wir erleichtert sind, dass die Blumen im Garten endlich Wasser kriegen. Dies sind aber eben subjektive Implikationen, die wir gerne ausblenden möchten. Ferner stellen wir fest, dass die Nachricht je nach Kontext von ganz unterschiedlicher Qualität ist. In den Tropen, wo es fast täglich regnet, dürfte die Mitteilung bestenfalls müdes Achselzucken erregen. Der Informationsgehalt der Nachricht ist hier gering. Für die Bewohner einer Wüstengegend sieht das ganz anders aus. Wir stellen fest: Der Informationsgehalt einer Mitteilung hängt weniger mit dem Inhalt der Nachricht zusammen, als vielmehr mit dem Überraschungseffekt, den die Nachricht bewirkt. ▲

▼ Oder nehmen wir das Lottospiel²² *6 aus 42*: Alle spielen mit, im Wissen, dass sie mit grösster Wahrscheinlichkeit nie gewinnen werden. Man kann sogar leicht ausrechnen wie gross die Gewinnwahrscheinlichkeit in einem Durchgang ist: Wenn die erste der sechs Kugeln aus einem Pool von 42

Kugeln gezogen wird, so ist die Wahrscheinlichkeit²³, dass sie eine meiner sechs Zahlen trägt genau $6/42$. Angenommen sie trägt tatsächlich eine meiner Zahlen, so bleiben auf meiner Liste fünf übrig, die noch zu ziehen wären. Und im Pool sind noch 41 Kugeln. Die Wahrscheinlichkeit, dass eine von meinen verbleibenden Zahlen gezogen wird ist demnach $5/41$. Entsprechend geht es weiter, bis sechs Kugeln gezogen sind. Die gesamte Wahrscheinlichkeit P für sechs richtige Zahlen ist also²⁴:

$$P = \frac{6}{42} \cdot \frac{5}{41} \cdot \frac{4}{40} \cdot \frac{3}{39} \cdot \frac{2}{38} \cdot \frac{1}{37} \approx 1.91 \cdot 10^{-7}$$

Anders ausgedrückt ist die Gewinnwahrscheinlichkeit P nur gerade eins zu 5.2 Mio. Und die Wahrscheinlichkeit für das Gegenereignis, nämlich dass ich nicht gewinne, resp. nicht sechs richtige Zahlen getippt habe, ist $1 - P$, also sozusagen eins, und damit so gut wie sicher. ▲

Wahrscheinlichkeit Aus derartigen Überlegungen heraus stiess Shannon auf ein Mass der Information, das allein vom Überraschungseffekt einer Nachricht abhängt. Technisch gesprochen meinte er den Kehrwert der Auftretenswahrscheinlichkeit einer Nachricht. Wahrscheinlichkeiten werden also im Folgenden eine wichtige Rolle spielen. Dem Leser, der sich diesbezüglich unsicher fühlt, wird der Anhang über Wahrscheinlichkeit im Kapitel 5.1 zur vorgängigen Lektüre empfohlen.

3.3 Informationsgehalt von Ereignissen

Im Folgenden geht es um Nachrichtenquellen, Wahrscheinlichkeiten, Zufallsvariablen, Ereignisse, Symbole, usw. Diese Begriffe haben in der Informationstheorie sehr scharf umrissene Bedeutungen, siehe Anhang 5.1. Anfangen wollen wir mit einem Modell für Informations- oder Datenquellen.

3.3.1 Quellen-Modell

Wir gehen aus von einer Datenquelle, welche periodisch, zu jedem Zeitpunkt $k \geq 0$ eine zufällige Nachricht X_k mit $k = 0, 1, 2, \dots$ abgibt²⁵. X_0 ist die nullte Nachricht, welche die Quelle nach dem

²² Es geht darum, dass man aus den Zahlen 1 bis 42 sechs Stück auswählt. Anschliessend werden aus 42 entsprechend beschrifteten Kugeln nach einem Zufallsprinzip sechs gezogen. Die betreffenden sechs Zahlen sind die Gewinnzahlen.

²³ Wir verwenden hier das allgemein übliche Wahrscheinlichkeitsmass, das aus dem Quotienten aller *günstigen Fälle* durch alle *möglichen Fälle* gebildet wird. Die *günstigen Fälle* sind die, die zum gesuchten Ereignis führen.

²⁴ Sind für ein Ereignis mehrere unabhängige Teilereignisse notwendig, so werden die Wahrscheinlichkeiten der unabhängigen Teilereignisse multipliziert um die Wahrscheinlichkeit für das Gesamt Ereignis zu erhalten. Was es mit dem Attribut *unabhängig* auf sich hat, wird im Verlauf dieses Skripts noch speziell erläutert.

²⁵ Die Datenquelle betrachten wir als diskreten stochastischen Prozess im Sinn der Wahrscheinlichkeitstheorie.

Einschalten ausspuckt, X_1 die erste, usw. Die Menge der Nachrichten können wir beispielsweise in einem Vektor anordnen, wie in Abbildung 4 dargestellt. Im Allgemeinen fassen wir diesen Vektor als unendlich lang auf. Beachte, dass wir bisher mit X_k eine Nachricht bezeichnet haben. Wir wissen aber noch nicht, von welchem Ereignis diese Nachricht berichtet.

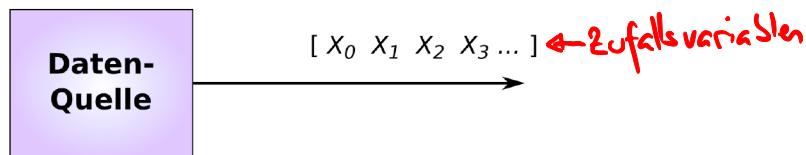


Abbildung 4: Modell einer Datenquelle.

Wortvorrat: Symbole, Ereignisse

x₀ x₁ x₂ ...

Wahrscheinlichkeit: P(x_n)

Realisierung: X_n = x_n

▼ Beispiel für eine derartige Quelle könnte die wöchentliche Ziehung der Lottozahlen sein oder der tägliche Wetterbericht im Radio. Im ersten Beispiel nummeriert k die Wochen und jede Nachricht X_k enthält die sechs Gewinnzahlen der betreffenden Woche. Im zweiten Fall zählt k die Tage und die Nachricht W_k ist die betreffende Wettermeldung für einen Ort. Beachte, dass wir die Nachrichten der beiden Beispiele unterschiedlich bezeichnen, damit wir sie unterscheiden können. Wir sind hier ganz frei in der Namensgebung. ▲

3.3.2 Wahrscheinlichkeiten

Da die Datenquelle stochastischen Charakter hat und nicht vorhersagbar ist, was der Inhalt der nächsten Nachricht sein wird, verwenden wir die Methoden der Wahrscheinlichkeitstheorie, um sie zu beschreiben. Aus diesem Grund definieren wir die Nachrichten X_k als Zufallsvariablen. Der Wertevorrat der Quelle seien N möglichen Ereignissen²⁶ x_n mit $n = 0 \dots N - 1$, so dass jede Zufallsvariable X_k mit einer gewissen Wahrscheinlichkeit²⁷ $P(x_n)$ von einem dieser Ereignisse x_n berichtet. Vereinfacht sagen wir, die Nachricht oder Zufallsvariable X_k enthält, resp. liefert das Ereignis x_n . Die x_n sind im Grunde nummerierte Konstanten oder Symbole²⁸, die je für ein bestimmtes Ereignis steht.

²⁶ Wir gehen hier davon aus, dass es sich um diskrete Ereignisse handelt. Also ist $N \in \mathbb{N}$ und typischerweise gilt $N \geq 1$.

²⁷ Im Folgenden werden wir uns auf stationäre Quelle beschränken. Das heisst, dass sich die Wahrscheinlichkeiten $P(x_n)$ der Zufallsvariablen X_k in Abhängigkeit von k nicht ändern. Oder anders ausgedrückt: Alle X_k haben identische Wahrscheinlichkeiten $P(x_n)$.

Beachte, dass wir die Zufallsvariable X_k mit Grossbuchstaben bezeichnen und die möglichen Ereignisse x_n mit Kleinbuchstaben. Das ist zwar nicht zwingend, aber in der Literatur verbreitet. Die Ausgabe einer Quelle könnte also so aussehen:

$$[X_0 = x_4 \quad X_1 = x_3 \quad X_2 = x_0 \quad X_3 = x_2 \quad \dots]$$

▼ Ein bestimmtes Lotto-Ereignis x_η könnte diese Form haben:

$$x_\eta = (4, 17, 19, 28, 31, 36)$$

Das Ereignis besteht also aus einem Set von sechs Zahlen. Beim Lotto spielt die Reihenfolge der Zahlen keine Rolle. In der Regel werden sie aber in aufsteigender Folge sortiert. Wie wir früher schon gezeigt haben gibt es sehr viele derartige Ereignisse, nämlich $N \approx 5.2$ Mio. ▲

▼ Beim Wetterbericht könnten wir etwa die folgenden $M = 3$ Fälle (Ereignisse) unterscheiden:

w_0 = “Es ist schlechtes Wetter.”

w_1 = “Es ist gutes Wetter.”

w_2 = “Es ist wechselhaft.”

Beachte, dass wir die Wetter-Ereignisse mit w_m ($m = 0 \dots M - 1$) bezeichnet haben, damit wir sie nicht mit den Lotto-Ereignissen verwechseln. Wir können nun aufeinander folgende Wetterberichte als eine Sequenz von Zufallsvariablen W_k auffassen. Jede Zufallsvariable ist sozusagen ein Platzhalter, der mit den entsprechenden Auftretenswahrscheinlichkeiten $P(w_0)$, $P(w_1)$, $P(w_2)$ die Werte w_0 , w_1 , w_2 annehmen kann. Auf Grund langjähriger (subjektiver) Beobachtungen gilt beispielsweise für Zürich:

$$P(w_0) = \frac{101}{365} = 0.28 \quad (\text{schlechtes Wetter})$$

$$P(w_1) = \frac{97}{365} = 0.26 \quad (\text{gutes Wetter})$$

$$P(w_2) = \frac{167}{365} = 0.46 \quad (\text{wechselhaft})$$

Beachte, dass die Summe dieser Wahrscheinlichkeiten eins sein muss, denn eines von den drei Ereignissen tritt in jedem Fall auf. ▲

²⁸ Es gibt auch Quellen, deren Ereignisse stetig (zum Beispiel reellwertig) sind und daher nicht abzählbar. Auf solche Ereignisse wollen wir aber nicht eingehen.

3.3.3 Informationsgehalt

Wenn nun das Ereignis $X_k = x_n$ auftritt, so können wir dessen **Informationsgehalt** $I(x_n)$ angeben²⁹:

$$I(x_n) = \frac{1}{P(x_n)} \quad (1)$$

Der Informationsgehalt hängt also nur vom Kehrwert der Auftretenswahrscheinlichkeit des betreffenden Ereignisses ab. Für eine kleine Wahrscheinlichkeit $P(x_n) \approx 0$ wird der Kehrwert - und damit der Informationsgehalt - gross, für eine grosse Wahrscheinlichkeit $P(x_n) \approx 1$ wird er ungefähr eins. Man könnte daher den Informationsgehalt eines Ereignisses als ein Mass für den weiter oben erwähnten Überraschungseffekt gelten lassen.

Wir werden weiter unten sehen, dass der Informationsgehalt von grosser Bedeutung für die Codierung von Ereignissen ist. Da wir dabei in erster Linie an die binäre Codierung denken, geben wir den Informationsgehalt im Folgenden nur noch in der **Einheit Bit** an und verwenden daher diese Definition:

$$I(x_n) = \log_2 \frac{1}{P(x_n)} \quad (\text{Bit}) \quad (2)$$

Durch Verwendung eines anderen Logarithmus könnte der Informationsgehalt auch für Codes angegeben werden, die auf einem anderen als dem binären System beruhen. Das werden wir aber nicht weiter verfolgen. Wer nicht weiss, wie er einen Zweier-Logarithmus berechnen soll, sei auf den Anhang 5.2 verwiesen.

▼ Nehmen wir als Quelle wieder das Lottospiel und definieren zwei neue Ereignisse:

$$\begin{aligned} \lambda_0 &= \text{"verloren"} = P(\lambda_0) \approx 1 \\ \lambda_1 &= \text{"gewonnen"} = P(\lambda_1) \approx 1 \cdot 10^{-7} \end{aligned}$$

Mit den weiter oben gefundenen Wahrscheinlichkeiten $P(\lambda_0)$ und $P(\lambda_1)$ folgen die Werte für den Informationsgehalt:

$$I(\lambda_0) = \log_2 \frac{1}{P(\lambda_0)} \approx 2.8 \cdot 10^{-7} \text{ Bit}$$

$$I(\lambda_1) = \log_2 \frac{1}{P(\lambda_1)} \approx 22.3 \text{ Bit}$$

²⁹ Die korrekte Schreibweise für den Informationsgehalt wäre $I_X(x_n)$, um anzugeben, dass sich der Informationsgehalt auf die Zufallsvariable X bezieht. Aber solange aus dem Kontext hervor geht, was gemeint ist, bevorzugen wir die einfachere Form.

Wir sehen, dass das häufige Ereignis λ_0 praktisch keinen Informationsgehalt trägt, das sehr seltene Ereignis λ_1 jedoch viel. Dies deckt sich mit unserer Wahrnehmung von Information, und scheint daher zweckmäßig. ▲

Für verschiedene Wahrscheinlichkeiten $P(\cdot)$ können wir den dazu gehörigen Informationsgehalt $I(\cdot)$ grafisch darstellen, siehe Abbildung 5.

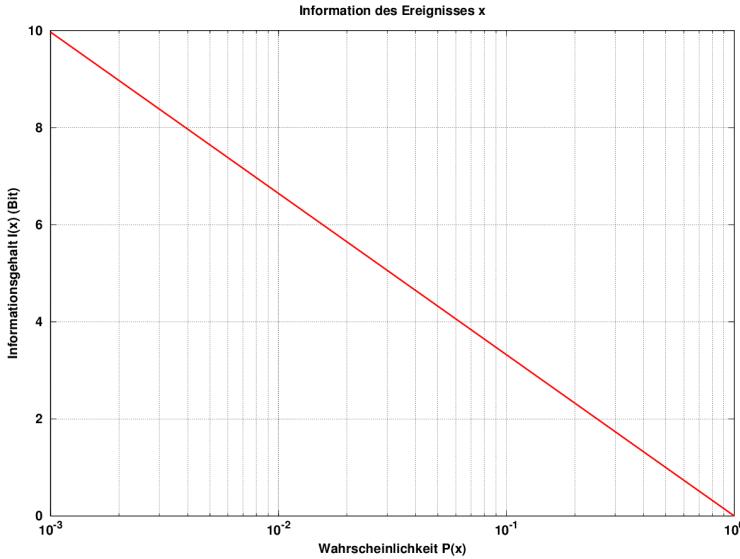


Abbildung 5: Informationsgehalt I für logarithmisch aufgezeichnete Wahrscheinlichkeiten P .

Würfel: $x_1 = 1 \dots x_6 = 6 \quad P(4) = \frac{1}{6}$

$I(x_4) = \log_2 \frac{1}{P(x_4)} = \log_2 \frac{1}{\frac{1}{6}} = 2.58 \text{ Bit}$

Es gilt als Mittelwert $\rightarrow 1x$ wird es als 3 Bit & $1x$ wird es als 2 Bit angezeigt usw...

3.3.4 Eigenschaften

Denken wir uns wieder eine Quelle, die N verschiedene Ereignisse, resp. Symbole x_n erzeugen kann.

- Haben alle Ereignisse x_n dieselbe Auftretenswahrscheinlichkeit $P(x_n) = 1/N$, dann haben auch alle Ereignisse denselben Informationsgehalt, nämlich:

$$I(x_n) = \log_2 N$$

Beachte, dass wir bei einer derartigen Quelle, und wenn N eine Potenz von 2 ist, genauso viele Bits brauchen, nämlich $\log_2 N$, um alle möglichen Ereignisse x_n binär zu nummerieren. Ist beispielsweise $N = 2^4 = 16$, so brauchen wir $\log_2 N = 4$ Bits, um die 16 Ereignisse (0000_b bis 1111_b) durchzuzählen. Wir können uns auch vorstellen, dass wir jedem dieser 16 Ereignisse eine dieser 4 Bit langen Nummern als Code zuordnen.

2. Der **minimale Wert** I_{\min} der Information tritt auf bei einem Ereignis x_n , das praktisch sicher auftritt, das also eine **Wahrscheinlichkeit** $P(x_n) \rightarrow 1$ hat.

$$I_{\min} = \lim_{P(x_n) \rightarrow 1} \log_2 \frac{1}{P(x_n)} = 0 \quad (\text{Bit})$$

3. Im Gegensatz dazu erhalten wir den **maximalen Wert** I_{\max} der Information für ein Ereignis x_n , das sozusagen nie auftritt, das also eine **Auftretenswahrscheinlichkeit** $P(x_n) \rightarrow 0$ hat.

$$I_{\max} = \lim_{P(x_n) \rightarrow 0} \log_2 \frac{1}{P(x_n)} \rightarrow \infty \quad (\text{Bit})$$

3.3.5 Mehrfach-Ereignisse

Analog zum Informationsgehalt für ein Ereignis können wir auch den Informationsgehalt eines **Doppel- oder Mehrfach-Ereignisses** angeben. Bei Mehrfach-Ereignissen spielt es keine Rolle, ob es sich um aufeinander folgende Ereignisse der selben Quelle handelt, oder um parallel erzeugte Ereignisse aus verschiedenen Quellen.

▲ Ein Mehrfach-Ereignis ist es beispielsweise, wenn wir vier mal hintereinander Lotto spielen und vier mal verlieren. Formal würden wir das Ereignis als $(\lambda_0, \lambda_0, \lambda_0, \lambda_0)$ notieren. Das Experiment (mehrfach Lotto spielen), welches derartige Mehrfach-Ereignis liefert, können wir beliebig wiederholen. Vielleicht haben wir dann Glück und es tritt einmal zum Beispiel das Ereignis $(\lambda_0, \lambda_1, \lambda_0, \lambda_0)$ oder das Ereignis $(\lambda_0, \lambda_0, \lambda_0, \lambda_1)$ auf, nämlich der Fall, dass wir in vier aufeinander folgenden Spielen genau einmal gewinnen. Noch besser wäre natürlich der Fall, dass wir in zwei, drei oder gar vier von vier aufeinander folgenden Lottospielen gewinnen. ▲

Stellvertretend für alle möglichen Fälle wollen wir zwei Quelle betrachten, deren Zufallsvariablen X_k und Y_k verschiedene Ereignisse x_n mit $n = 0 \dots N - 1$ und y_m mit $m = 0 \dots M - 1$ annehmen können. Wir fassen nun aber für jeden Index k die betreffenden Ereignisse x_n und y_m zusammen und definieren die Verbund-Ereignisse (x_n, y_m) . Für diese Doppel-Ereignisse wollen wir den Verbund-Informationsgehalt $I(x_n, x_m)$ angeben:

$$I(x_n, x_m) = \log_2 \frac{1}{P(x_n, x_m)} \quad (\text{Bit}) \quad (3)$$

Dabei ist $P(x_n, x_m)$ die Verbund-Wahrscheinlichkeit des Doppel-Ereignisses (x_n, x_m) .

▼ Ein Verbund-Ereignis können wir beispielsweise bilden aus dem Wetterbericht für einen Ort an zwei aufeinander folgenden Tagen. Die Ereignisse w_0 bis w_2 wurden schon eingeführt. Dies sind die

möglichen Verbund-Ereignisse v_σ mit $\sigma = 0 \dots \Sigma - 1$ und $\Sigma = N^2 = 9$:

$$\begin{array}{lll} v_0 = (w_0, w_0) & v_3 = (w_1, w_0) & v_6 = (w_2, w_0) \\ v_1 = (w_0, w_1) & v_4 = (w_1, w_1) & v_7 = (w_2, w_1) \\ v_2 = (w_0, w_2) & v_5 = (w_1, w_2) & v_8 = (w_2, w_2) \end{array}$$

Das Verbundereignis v_5 sagt zum Beispiel, dass es am ersten Tag gutes Wetter war und am zweiten Tag veränderlich. Auf Grund von Beobachtungen können wir die Wahrscheinlichkeiten der Verbund-Ereignisse angeben:

$$\begin{array}{lll} P(v_0) = P(w_0, w_0) = 0.18 & P(v_3) = P(w_1, w_0) = 0.04 & P(v_6) = P(w_2, w_0) = 0.06 \\ P(v_1) = P(w_0, w_1) = 0.03 & P(v_4) = P(w_1, w_1) = 0.14 & P(v_7) = P(w_2, w_1) = 0.09 \\ P(v_2) = P(w_0, w_2) = 0.07 & P(v_5) = P(w_1, w_2) = 0.08 & P(v_8) = P(w_2, w_2) = 0.31 \end{array}$$

Was sagen diese Zahlen aus? Vergleichen wir beispielsweise $P(w_0, w_0) = 0.18$ und $P(w_0, w_1) = 0.03$: Wir sehen, dass es wahrscheinlicher ist, dass auf einen Regentag wieder ein Regentag folgt (Ereignis v_0), als dass auf einen Regentag ein Sonnentag folgt (Ereignis v_1). Vergleichen wir dagegen $P(w_0, w_0)$ und $P(w_1, w_1)$, so zeigt sich, dass die Wahrscheinlichkeit grösser ist, dass es an aufeinander folgenden Tagen regnet, als dass an beiden Tagen die Sonne scheint. Die Summe aller 9 Verbund-Wahrscheinlichkeiten muss eins sein, denn eine von allen möglichen Abfolgen tritt sicher auf.

Nun können wir von jedem Doppel-Ereignis den Informationsgehalt angeben. Wir tun das hier für das Ereignis (w_1, w_2) :

$$I(w_1, w_2) = \log_2 \frac{1}{P(w_1, w_2)} \approx 3.64 \text{ Bit}$$

Wir wollen gleich noch zeigen, wie man die totale Wahrscheinlichkeit für Regen (unabhängig vom nachfolgenden oder voraus gehenden Tag) berechnen kann:

$$\begin{aligned} P(w_0) &= \sum_{n=0}^2 P(w_n, w_0) = P(w_0, w_0) + P(w_1, w_0) + P(w_2, w_0) \approx 0.28 \\ &= \sum_{n=0}^2 P(w_0, w_n) = P(w_0, w_0) + P(w_0, w_1) + P(w_0, w_2) \approx 0.28 \end{aligned}$$

Ebenso erhalten wir die Wahrscheinlichkeiten $P(w_1) = 0.26$ und $P(w_2) = 0.46$. Dieselben Wahrscheinlichkeiten haben wir weiter oben bereits gefunden. Wiederum muss die Summe aller Wahrscheinlichkeiten eins ergeben, was natürlich der Fall ist. ▲

3.3.6 Statistisch unabhängige Mehrfach-Ereignisse

An dieser Stelle gehen wir nur auf statistisch unabhängige Ereignisse ein. Der andere Fall wird im Anhang, Kapiel 5.3 behandelt. Sind die Ereignisse x_n und x_m statistisch unabhängig, dann (und nur

dann) gilt für die Verbund-Wahrscheinlichkeit:

$$P(x_n, x_m) = P(x_n) \cdot P(x_m)$$

Dies setzen wir in Formel 3 ein:

$$I(x_n, x_m) = \log_2 \frac{1}{P(x_n, x_m)} = \log_2 \frac{1}{P(x_n) \cdot P(x_m)} = \log_2 \frac{1}{P(x_n)} + \log_2 \frac{1}{P(x_m)}$$

Für statistisch unabhängige Ereignisse x_n und x_m folgt demnach:

$$I(x_n, x_m) = I(x_n) + I(x_m) \quad (4)$$

▼ Das Ziehen der Gewinnzahlen beim Lotto hat absolut keinen kausalen Zusammenhang mit den voraus gegangenen Ziehungen. Daher sind aufeinander folgende Ereignisse statistisch unabhängig von einander.

Wir definieren nun für die zwei aufeinander folgenden Ziehungen X_k und X_{k+1} die Doppel-Ereignisse $\varphi_r = (\lambda_n, \lambda_m)$ mit $r = 0 \dots 3$ und $n, m = 0 \dots 1$. Dabei ist k die Nummer der Woche und λ_n , resp. λ_m sind die möglichen Ereignisse.

φ_0	$= (\lambda_0, \lambda_0)$	(2 mal verloren)
φ_1	$= (\lambda_1, \lambda_0)$	(zuerst gewonnen, dann verloren)
φ_2	$= (\lambda_0, \lambda_1)$	(zuerst verloren, dann gewonnen)
φ_3	$= (\lambda_1, \lambda_1)$	(2 mal gewonnen)

Für jedes Verbund-Ereignis φ_r können wir die Wahrscheinlichkeit angeben:

$$\begin{aligned} P(\varphi_0) &= P(\lambda_0, \lambda_0) = P(\lambda_0) \cdot P(\lambda_0) \approx 1 \\ P(\varphi_1) &= P(\lambda_1, \lambda_0) = P(\lambda_1) \cdot P(\lambda_0) \approx 1.91 \cdot 10^{-7} \\ P(\varphi_2) &= P(\lambda_0, \lambda_1) = P(\lambda_0) \cdot P(\lambda_1) \approx 1.91 \cdot 10^{-7} \\ P(\varphi_3) &= P(\lambda_1, \lambda_1) = P(\lambda_1) \cdot P(\lambda_1) \approx 0 \end{aligned}$$

Der Informationsgehalt lässt sich aber direkt angeben, ohne Umweg über die Verbund-Wahrscheinlichkeit. Wir betrachten zuerst das Doppel-Ereignis φ_3 (ich gewinne diese Woche im Lotto und ich gewinne nächste Woche im Lotto):

$$I(\varphi_3) = I(\lambda_1, \lambda_1) = I(\lambda_1) + I(\lambda_1) = 2 \cdot I(\lambda_1) = 44.6 \text{ Bit}$$

Zusammengefasst erhalten wir:

$$\begin{aligned} I(\varphi_0) &= I(\lambda_0) + I(\lambda_0) \approx 0 \text{ Bit} \\ I(\varphi_1) &= I(\lambda_1) + I(\lambda_0) \approx 22.3 \text{ Bit} \\ I(\varphi_2) &= I(\lambda_0) + I(\lambda_1) \approx 22.3 \text{ Bit} \\ I(\varphi_3) &= I(\lambda_1) + I(\lambda_1) \approx 44.6 \text{ Bit} \end{aligned}$$



3.3.7 Masseinheit

Wie wir gesehen haben, messen wir die Information gemäss Formel 2 mit *Bit*. Was heisst das? Angenommen wir haben irgend ein Symbol x_n mit $P(x_n) = 0.25$, dann folgt $I(x_n) = 2$ Bit. Wenn wir nun weiter annehmen, dass x_2 eines von mehreren möglichen Symbolen ist, die alle die selbe Wahrscheinlichkeit $P(x_n)$ haben, so bezeichnet $I(x_n)$ die notwendige Anzahl Bits, mit denen sich alle x_n durchnummerieren liessen.

Mit den zuvor genannten Zahlen gäbe es also $N = 4$ Symbole x_0 bis x_3 , die alle die Wahrscheinlichkeit $P(x_2) = 0.25$ haben. Mit $I(x_n) = 2$ Bit liessen sich die x_n zum Beispiel so nummerieren, resp. codieren: $x_0 \hat{=} (00_b)$, $x_1 \hat{=} (01_b)$, $x_2 \hat{=} (10_b)$ und $x_3 \hat{=} (11_b)$. Beachte, dass wir Codeworte in runden Klammern schreiben und die Zuordnung des Symbols zum Codewort nicht mit einem Gleichzeichen angeben, sondern mit dem Zeichen $\hat{=}$ für die Entsprechung.

3.4 Zusammenfassung

Mit der Zufallsvariable X_k und den Ereignissen x_n und x_m gilt:

- Der Informationsgehalt beim Auftreten von $X_k = x_n$ ist:

$$I = \log_2 \frac{1}{P(x_n)} \quad (\text{Bit})$$

- Der Informationsgehalt beim Auftreten eines statistisch unabhängigen Wertepaares (x_n, x_m) ist:

$$I(x_n, x_m) = I(x_n) + I(x_m) \quad (\text{Bit})$$

- Die Masseinheit der Information gibt an, wie viele Bits für das Codieren des betreffenden Symbols notwendig wären, wenn alle Symbole dieselbe Auftretenswahrscheinlichkeit hätten.

4 Entropie

Nachdem zuvor der Begriff der Information, resp. des Informationsgehalts, sorgfältig eingeführt wurde, kann nun die Entropie etwas kürzer behandelt werden. Zudem soll der Begriff wieder anhand zahlreicher Beispiele erläutert werden.

4.1 Begriff

Der Begriff der Entropie stammt aus der Thermodynamik, wo er den Unordnungszustand eines Systems beschreibt. Eine ähnliche Funktion hat die Entropie in der Informationstheorie:

- Eine Datenquelle hat eine tiefe Entropie, wenn das nächste Ereignis (Symbol) mit hoher Wahrscheinlichkeit korrekt vorhersagbar ist.
- Und eine Quelle hat eine hohe Entropie, wenn die Wahrscheinlichkeit klein ist, dass das nächste Ereignis korrekt vorhergesagt werden kann.

Dabei gehen wir davon aus, dass eine Testperson die Datenquelle eine Weile beobachtet hat. Auf Grund der Beobachtung soll der Proband dann jeweils eine Prognose für das nächste Ereignis abgeben. Das wollen wir gleich an einem Beispiel erläutern.

▼ Wir beobachten die Ausgänge von drei binären Datenquellen mit den Zufallsvariablen A_k bis C_k mit $k = 0, 1, 2, \dots$, wobei wir annehmen, dass jede Beobachtung repräsentativ sei für die betreffende Quelle. Welche Prognosen für das jeweils nächste, noch nicht sichtbare Ereignis, würden wir abgeben?

$$\begin{aligned} A_k &= [\quad 1 \quad 0 \quad 1 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad \dots] \\ B_k &= [\quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 1 \quad \dots] \\ C_k &= [\quad 1 \quad 1 \quad 0 \quad 0 \quad \dots] \end{aligned}$$

Betrachten wir A_k , so stellen wir fest, dass von den 20 bereits vorliegenden Ereignissen nur gerade zwei Nullen sind. Folglich tritt in 90 % der Fälle eine Eins auf. Wir sind also gut beraten, für das nächste Ereignis auf eine Eins zu tippen. Wir werden dann in 90 % der Fälle richtig liegen, die Entropie dieser Quelle ist klein.

Wenn wir dagegen B_k betrachten, so stellen wir fest, dass Nullen und Einsen gleich häufig und in zufälliger Anordnung auftreten. Das nächste Ereignis kann demnach genauso gut eine Null wie eine Eins sein. Unsere Prognose - ob Null oder Eins - ist in jedem Fall unsicher. Wir werden nur in 50 % der Fälle richtig liegen, die Entropie der Quelle ist gross.

Wie bei B_k kommen bei C_k Nullen und Einsen gleich häufig vor. Es besteht aber eine feste Abfolge, denn die Periode 1100 wiederholt sich stets. Die Ereignisse sind statistisch voneinander abhängig. Damit können wir auf Grund der Beobachtung sicher vorhersagen, dass das nächste Ereignis eine Eins sein wird. Die Entropie dieser Quelle ist null.

4.2 Entropie von Quellen

An dieser Stelle betrachten wir nur Quellen (stochastische Prozesse), die statistisch unabhängige Zufallsvariablen, resp. Ereignisse erzeugen. Wer sich auch für den anderen Fall interessiert, sei auf den Anhang 5.4 verwiesen.

Informationstheorie

Die Entropie $H(X)$ einer Quelle³⁰ X , die statistisch unabhängige Symbole (Ereignisse) x_n liefert, ist definiert als der Erwartungswert der Information $I(x_n)$ dieser Symbole. Mit dem Erwartungswert meinen wir einen gewichteten Mittelwert. Gehen wir davon aus, dass X genau N verschiedene Symbole x_n mit $n = 0 \dots N-1$ liefert, so ist die Entropie:

Erwartungswert der Informations -> Entropie

$$H(X) = \sum_{n=0}^{N-1} P(x_n) \cdot I(x_n) \quad (5)$$

Die Einheit der Entropie ist dieselbe wie jene der Information. Um anzugeben, dass es sich bei der Entropie aber um einen Mittelwert der Information pro Symbol handelt, verwenden wir die Einheit *Bit/Symbol*.

Man gewichtet also den Informationsgehalt $I(x_n)$ von jedem Symbol mit der Auftretenswahrscheinlichkeit $P(x_n)$ des Symbols und zählt dann alle diese Produkte zusammen. Damit trägt man dem Umstand Rechnung, dass ein seltenes Symbol (kleine Wahrscheinlichkeit) zwar einen hohen Informationsgehalt hat, es aber wegen dem seltenen Auftreten im Mittelwert wenig gewichtet wird. Damit stimmt der Erwartungswert mit dem überein, was man erhält, wenn man die Zufallsvariable eine sehr grosse Zahl von Symbolen erzeugen lässt und dann davon den linearen Mittelwert bildet.

Durch Einsetzen der Gleichung 2 in 5 folgt, dass die Entropie nur von den Symbolwahrscheinlichkeiten $P(x_n)$ abhängt:

$$H(X) = \sum_{n=0}^{N-1} P(x_n) \cdot \log_2 \frac{1}{P(x_n)} \quad (\text{Bit/Symbol}) \quad (6)$$

In der Literatur findet man oft auch die folgenden Schreibweise, die mit $\log(\frac{1}{x}) = -\log(x)$ folgt:

$$H(X) = - \sum_{n=0}^{N-1} P(x_n) \cdot \log_2 P(x_n)$$

Es sei noch einmal betont, dass die Gleichungen 5 und 6 für stochastische Prozesse, resp. Quellen gelten, die statistisch unabhängige Symbole produzieren. Man nennt eine derartige Quelle *gedächtnislos*. Der Ausdruck kommt daher, dass die Quelle in dem Moment, wo sie ein Symbol produziert, sofort alles darüber vergisst, so dass jedes nachfolgende Symbol keinen Bezug zu irgend einem vergangenen Symbol haben kann.

Wir wollen nun einige Beispiele ansehen.

▼ Beim Lottospiel L haben wir zwei Ereignisse kennen gelernt, nämlich λ_0 (verloren) und λ_1 (gewonnen). Wir kennen auch bereits die Auftretenswahrscheinlichkeiten $P(\lambda_1) \approx 1.91 \cdot 10^{-7}$ und

³⁰ Wir verwenden von hier an für die Quelle dieselbe Bezeichnung, wie für die Zufallsvariablen, die von der Quelle produziert werden. Die Quelle X erzeugt also lauter identische Zufallsvariablen X_k .

Würfel: 6 Ereignisse mit $\frac{1}{6}$ Wahrscheinlichkeit
 $H = 2.58 \text{ Bit/Symbol}$ (Da alle gleiche Chancen \rightarrow Entropie = Informationsgehalt)

Informationstheorie $H = \log_2 N = \log_2 6 \Rightarrow H = \sum P(x_n) \cdot \log_2 \frac{1}{P(x_n)} = 6 \cdot \frac{1}{6} \cdot \log_2 6 = \log_2 6$

$P(\lambda_0) = 1 - P(\lambda_1)$. Folglich ist die Entropie $H(L)$ des Lottospiels:

$$H(L) = P(\lambda_0) \cdot \log_2 \frac{1}{P(\lambda_0)} + P(\lambda_1) \cdot \log_2 \frac{1}{P(\lambda_1)} \approx 4.54 \cdot 10^{-6} \text{ Bit/Symbol}$$

Obwohl - wie wir früher schon gezeigt haben - die Information eines Gewinns $I(\lambda_1) \approx 22.3$ Bit ist, resultiert für die Entropie $H(L)$ ein sehr kleiner Wert, weil das Gewinnereignis mit der sehr kleinen Auftretenswahrscheinlichkeit $P(\lambda_1)$ gewichtet wird. Man würde also erwarten, dass Lottospielen insgesamt eine ausgesprochen langweilige Sache ist, da im Durchschnitt der Informationsgehalt nahe bei null liegt. ▲

▼ Betrachtet man eine allgemeine binäre, gedächtnislose Quelle (BMS³¹), so kann man ihre Entropie H_b folgendermassen angeben:

$$H_b = p \cdot \log_2 \frac{1}{p} + (1-p) \cdot \log_2 \frac{1}{1-p} \quad [\text{Bit/Symbol}]$$

Dabei ist p die Auftretenswahrscheinlichkeit des einen Symbols und $1-p$ jene des anderen Symbols. Den Verlauf der Entropie H_b kann man in Funktion der Wahrscheinlichkeit p grafisch darstellen, siehe Abbildung 6. Beachte, dass gilt:

$$\lim_{p \rightarrow 0} p \cdot \log_2 \frac{1}{p} = 0$$

Die Funktion $H_b = H_b(p)$ wird auch *binäre Entropiefunktion* genannt. ▲

▼ Eine Quelle Ψ liefert die Buchstaben A, E, O, U und R. Die Auftretenswahrscheinlichkeiten der Symbole sind:

$P(A) = 0.70$	$I(A) = 0.51$	}
$P(E) = 0.10$		
$P(O) = 0.08$		
$P(U) = 0.07$		
$P(R) = 0.05$	$I(R) = 4.32$	

$H = 1.47 \text{ Bit/Symbol}$

Wir stellen fest, dass die Summe aller Symbolwahrscheinlichkeiten eins ist. Das heisst, dass wir kein Symbol vergessen haben. Damit können wir die Entropie der Buchstabenquelle berechnen:

$$H(\Psi) = \sum_{x=A}^R P(x) \cdot \log_2 \frac{1}{P(x)} \approx 1.47 \text{ Bit/Symbol}$$

³¹ Mit *binär* meint man, dass die Quelle nur zwei verschiedene Symbole erzeugt. Für eine gedächtnislose binäre Quelle verwendet man oft die Abkürzung BMS (Englisch für *Binary Memoryless Source*).

Binäre Entropiefunktion: $P(x_1)=p \quad P(x_2)=1-p$

$$H_b = p \cdot \log_2 \frac{1}{p} + (1-p) \cdot \log_2 \frac{1}{1-p}$$

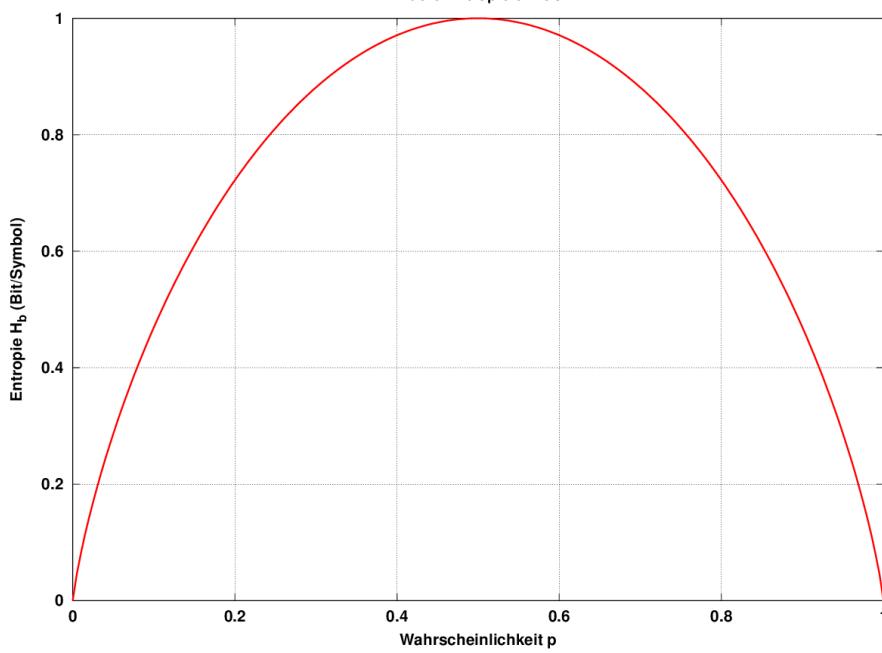


Abbildung 6: Binäre Entropiefunktion $H_b(p)$.

Am Ausgang der Quelle treten fünf verschiedene Symbole auf. Wollten wir sie binär durchnummerieren, resp. codieren, so würden dafür 3 Bit benötigen, zum Beispiel $A \hat{=} (000)_b$, $E \hat{=} (001)_b$, $O \hat{=} (010)_b$, $U \hat{=} (011)_b$ und $R \hat{=} (100)_b$. Im Durchschnitt beträgt die Information jedes Symbols jedoch nur 1.47 Bit. ▲

4.2.1 Eigenschaften

Wir wollen die Entropie noch weiter charakterisieren.

1. Eine Quelle X liefere N verschiedene Symbole x_n , die alle identische Wahrscheinlichkeiten $P(x_n) = \frac{1}{N}$ haben, dann gilt:

$$H(X) = \sum_{n=0}^{N-1} P(x_n) \cdot \log_2 \frac{1}{P(x_n)} = N \cdot \frac{1}{N} \cdot \log_2 N = \log_2 N$$

2. Minimale Entropie: Falls es am Ausgang einer Quelle X ein Symbol x_n gibt mit $P(x_n) = 1$, so ist $H(X) = 0$. Dies folgt direkt aus Abbildung 6. Es bedeutet, dass bei dieser Quelle immer nur das Symbol x_n auftritt. Es gibt daher keinen Überraschungseffekt. Die minimale Wert H_{min} der Entropie ist demnach:

$$H_{min} = 0 \quad (\text{Bit/Symbol}) \quad (7)$$

ZF:

- 1. ater Symbole mit gleicher Wahrscheinlichkeit
 $H = \log_2 N$

- $H_{\min} = 0$

- $H_{\max} = \log_2 N$

3. Maximale Entropie: Wenn N die Anzahl möglicher Symbole einer Quelle X ist, so gilt $0 \leq H(X) \leq \log_2 N$. Anders ausgedrückt: der maximale Wert H_{max} der Entropie bei N Symbolen ist:

$$H_{max} = \log_2 N \quad (\text{Bit/Symbol}) \quad (8)$$

Dieser Fall tritt genau dann auf, wenn jedes der N Symbole x_n die selbe Auftretenswahrscheinlichkeit $P(x_n) = \frac{1}{N}$ hat, siehe oben. Als anschauliche Begründung sei auf Abbildung 6 verwiesen.

- ▼ Die Quelle Ψ mit den Symbolen A, E, O, U und R erreicht die minimale Entropie $H_{min}(\Psi)$ beispielsweise dann, wenn $P(A) = 1$ und $P(E) = P(O) = P(U) = P(R) = 0$. Es folgt:

$$H_{min}(\Psi) = 1 \cdot \log_2 \frac{1}{1} + 4 \cdot 0 \cdot \log_2 \frac{1}{0} = 0 \quad \text{Bit/Symbol}$$

Die maximal mögliche Entropie H_{max} finden wir mit $P(A) = P(E) = P(O) = P(U) = P(R) = \frac{1}{5}$:

$$H_{max}(\Psi) = 5 \cdot \frac{1}{5} \cdot \log_2 5 = \log_2 5 \approx 2.32 \quad \text{Bit/Symbol}$$

Weiter oben haben wir gesehen, dass mit den dort gegebenen Wahrscheinlichkeiten die Entropie der Quelle Ψ den Wert $H(\Psi) \approx 1.47$ Bit/Symbol erreichte, also zwischen $H_{min}(\Psi)$ und $H_{max}(\Psi)$ lag. ▲

- ▼ Die Entropie eines fairen Würfels F ist:

$$H(F) = \log_2 6 \approx 2.58 \quad \text{Bit/Symbol}$$

Dabei haben wir Gebrauch gemacht vom Umstand, dass beim fairen Würfel jede Augenzahl mit der gleichen Wahrscheinlichkeit von $\frac{1}{6}$ auftritt. ▲

- ▼ Wir betrachten die Quellen X, Y und Z die je die vier Symbole 00, 01, 10 und 11 liefern. Die Wahrscheinlichkeiten sind jedoch bei jeder Quelle anders. Wir berechnen für jeden Fall die Entropie:

$X:$	$P(00) = 0.25$	$P(01) = 0.25$	$P(10) = 0.25$	$P(11) = 0.25$	$H(X) = 2.00$ Bit/Symbol
$Y:$	$P(00) = 0.70$	$P(01) = 0.10$	$P(10) = 0.10$	$P(11) = 0.10$	$H(Y) \approx 1.36$ Bit/Symbol
$Z:$	$P(00) = 0.94$	$P(01) = 0.02$	$P(10) = 0.02$	$P(11) = 0.02$	$H(Z) \approx 0.42$ Bit/Symbol

Je weiter die Verteilung der Wahrscheinlichkeiten vom ersten Fall X abweicht, umso kleiner wird die Entropie. ▲

4.2.2 Mehrfach-Quellen

Unter einer Mehrfach-Quelle verstehen wir eine Anordnung von Quellen, die Verbund-Ereignisse liefert. Die Einzel-Ereignisse können dabei aus verschiedenen Quellen stammen (zum Beispiel zwei Würfel) oder es sind aufeinander folgende Ereignisse aus derselben Quelle (zum Beispiel zweimal nacheinander würfeln).

Wir nennen die Mehrfach-Quelle nun (X, Y) , sie erzeugt Verbund-Ereignisse (x_n, y_m) . Die Verbund-Entropie $H(X, Y)$ ist dann:

$$H(X, Y) = \sum_{n,m} P(x_n, y_m) \cdot \log_2 \frac{1}{P(x_n, y_m)} \quad (\text{Bit/Symbol}) \quad (9)$$

Beachte, dass die Formel 9 für statistisch abhängige und unabhängige Zufallsvariablen gilt und dass sie analog auch für mehr als zwei Quellen geschrieben werden kann.

▼ Wir nehmen nochmals das Beispiel mit zwei Würfeln X und Y , definieren nun aber die Zufallsvariable Q :

$$Q = \begin{cases} q_0 & \text{falls } X < Y \\ q_1 & \text{falls } X \geq Y \end{cases}$$

Beachte, dass X und Q statistisch abhängig voneinander sind. Wir suchen nun die Verbund-Entropie $H(X, Q)$. Zu diesem Zweck müssen wir zuerst die Verbund-Wahrscheinlichkeiten $P(x_n, q_p)$ ermitteln. Wir tun das mit Hilfe einer Tabelle, wobei jede Zeile der Tabelle die Wahrscheinlichkeit $P(x_n, y_m) = \frac{1}{36}$ hat.

X	Y	Q	X	Y	Q	X	Y	Q
x_1	y_1	q_1	x_3	y_1	q_1	x_5	y_1	q_1
x_1	y_2	q_0	x_3	y_2	q_1	x_5	y_2	q_1
x_1	y_3	q_0	x_3	y_3	q_1	x_5	y_3	q_1
x_1	y_4	q_0	x_3	y_4	q_0	x_5	y_4	q_1
x_1	y_5	q_0	x_3	y_5	q_0	x_5	y_5	q_1
x_1	y_6	q_0	x_3	y_6	q_0	x_5	y_6	q_0
x_2	y_1	q_1	x_4	y_1	q_1	x_6	y_1	q_1
x_2	y_2	q_1	x_4	y_2	q_1	x_6	y_2	q_1
x_2	y_3	q_0	x_4	y_3	q_1	x_6	y_3	q_1
x_2	y_4	q_0	x_4	y_4	q_1	x_6	y_4	q_1
x_2	y_5	q_0	x_4	y_5	q_0	x_6	y_5	q_1
x_2	y_6	q_0	x_4	y_6	q_0	x_6	y_6	q_1

Nun können wir auszählen, wie oft jedes Verbund-Ereignis (x_n, q_p) auftritt und die entsprechende Wahrscheinlichkeit $P(x_n, q_p)$ angeben:

(x_n, q_p)	Anzahl	$P(x_n, q_p)$	(x_n, q_p)	Anzahl	$P(x_n, q_p)$
(x_1, q_1)	1	$\frac{1}{36}$	(x_4, q_1)	4	$\frac{4}{36}$
(x_1, q_0)	5	$\frac{5}{36}$	(x_4, q_0)	2	$\frac{2}{36}$
(x_2, q_1)	2	$\frac{2}{36}$	(x_4, q_1)	5	$\frac{5}{36}$
(x_2, q_0)	4	$\frac{4}{36}$	(x_5, q_0)	1	$\frac{1}{36}$
(x_3, q_1)	3	$\frac{3}{36}$	(x_5, q_1)	6	$\frac{6}{36}$
(x_3, q_0)	3	$\frac{3}{36}$	(x_6, q_0)	0	0

Mit den Wahrscheinlichkeitswerten $P(x_n, q_p)$ können wir nun die Verbund-Entropie $H(X, Q)$ ausrechnen. Man erhält:

$$H(X, Q) \approx 3.27 \text{ Bit/Symbol}$$

$H(X, Q)$ ist in diesem Fall nur unwesentlich kleiner als $H_{max} = 3.46$ Bit/Symbol bei 11 möglichen Verbund-Ereignissen. ▲

4.2.3 Statistisch unabhängige Mehrfach-Quellen

Falls die stochastischen Prozesse X und Y statistisch unabhängig von einander sind, können wir weiter schreiben:

$$\begin{aligned} H(X, Y) &= \sum_{n,m} P(x_n) P(y_m) \cdot \log_2 \frac{1}{P(x_n) P(y_m)} \\ &= \sum_{n,m} P(x_n) P(y_m) \cdot \left\{ \log_2 \frac{1}{P(x_n)} + \log_2 \frac{1}{P(y_m)} \right\} \\ &= \sum_{n,m} \left\{ P(x_n) P(y_m) \cdot \log_2 \frac{1}{P(x_n)} + P(x_n) P(y_m) \cdot \log_2 \frac{1}{P(y_m)} \right\} \\ &= \sum_n P(x_n) \cdot \log_2 \frac{1}{P(x_n)} + \sum_m P(y_m) \cdot \log_2 \frac{1}{P(y_m)} \end{aligned}$$

Schliesslich folgt für statistisch unabhängige Quellen X und Y :

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

(10)

Aus Gleichung 10 folgt schliesslich die Aussage, dass die Verbund-Entropie $H(X, Y)$ maximal wird, wenn erstens die Quellen X und Y von einander unabhängig sind, und zweitens, wenn sowohl alle Ereignisse x_n als auch alle y_m je untereinander gleich häufig auftreten.

▼ Es seien X und Y zwei unterscheidbare, faire Würfel. Damit erhalten wir Verbund-Ereignisse wie beispielsweise $(5, 1)$ oder $(3, 4)$. Aus einem früheren Beispiel kennen wir die Entropie $H(X) = H(Y) \approx 2.58$ Bit/Symbol. Damit folgt die Verbund-Entropie:

$$H(X, Y) = H(X) + H(Y) \approx 5.17 \text{ Bit/Symbol}$$

▲

▼ Als Fortsetzung betrachten wir nochmals die Würfel X und Y , sowie die Zufallsvariable Q mit:

$$Q = \begin{cases} q_0 & \text{falls } X < Y \\ q_1 & \text{falls } X \geq Y \end{cases}$$

Vom Würfel wissen wir, dass $P(x_n) = \frac{1}{6}$, für alle $n = 1 \dots 6$. Aus der Tabelle weiter oben lesen wir, dass $P(q_0) = \frac{15}{36}$ und $P(q_1) = \frac{21}{36}$. Damit folgt die schon bekannte Entropie des Würfels $H(X) \approx 2.58$ Bit/Symbol und die Entropie $H(Q) \approx 0.98$ Bit/Symbol. Im Vergleich mit der weiter oben berechneten Verbund-Entropie $H(X, Q) \approx 3.27$ Bit/Symbol folgt:

$$H(X, Q) \neq H(X) + H(Q)$$

Der Grund besteht darin, dass die Zufallsvariablen X und Q statistisch abhängig voneinander sind. ▲

4.2.4 Masseinheit

Wir erinnern uns: Der Informationswert $I(x_n)$ eines bestimmten Symbols x_n ist die Anzahl Bits, die notwendig wären um alle Symbole der Quelle X zu nummerieren, sofern alle Symbole dieselbe Wahrscheinlichkeit $P(x_n)$ hätten. Die Entropie $H(X)$ ist der Erwartungswert der Informationen aller Symbole der Quelle X . Das heisst, dass die Information der Quelle X im Durchschnitt mit $H(X)$ Bit pro erzeugtem Symbol darstellbar ist. In der praktischen Codierung der Symbole dürfen selbstverständlich mehr als $H(X)$ Bits verwendet werden. Hingegen ist es nicht möglich, die Information der Quelle mit weniger als $H(X)$ Bits pro Symbol darzustellen.

4.3 Zusammenfassung

- Die Entropie der Quelle X ist:

$$H(X) = \sum_{n=0}^{N-1} P(x_n) \cdot \log_2 \frac{1}{P(x_n)}$$

- Die Entropie von statistisch unabhängigen Quellen X und Y ist:

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

- Die Maßeinheit der Entropie ist *Bit/Symbol*. Sie gibt an, wie viele Bits im Durchschnitt pro Symbol notwendig sind, um die Information der Quelle darzustellen.

5 Anhang

5.1 Wahrscheinlichkeitstheorie

Die Wahrscheinlichkeitsrechnung ist eine zentrale **Grundlage der Informationstheorie**. Aus diesem Grund werden die wichtigsten Konzepte hier kurz zusammengefasst. Dabei erläutern wir nicht alle Facetten der Wahrscheinlichkeit, sondern legen den Fokus auf die (pragmatische) Anwendung in der Informationstheorie.

Je planmäßiger der Mensch vorgeht,
um so wirkungsvoller trifft ihn der Zufall.

Friedrich Dürrenmatt

5.1.1 Zufallsvariablen

Eine **Zufallsvariable³²** X_k ist eine Art Platzhalter für ein **zufälliges Ereignis** x_n , das in einem stochastischen Prozess auftreten kann. Einen **stochastischen Prozess** nennen wir auch ein **Zufallsexperiment**. Ein typisches Schulbeispiel ist das Würfeln³³. Das Werfen des Würfels ist der stochastische Prozess. Der Zufallsvariable X_k entspricht das Resultat des k-ten Wurfs des Würfels. Das Ereignis x_n steht für eine Augenzahl des Würfels, also für ein mögliches Resultat des Zufallsexperiments. Die Ereignisse x_n sind im Grunde nichts anderes als nummerierte Konstanten für mögliche Ergebnisse³⁴. Anstelle

³² Man könnte auch $X[k]$ schreiben, um den diskreten Zeit- oder Ortsaspekt der Nachricht X_k hervor zu heben. Wir wählen aber die einfachere Schreibweise mit Index, um die Lesbarkeit zu verbessern.

³³ Wir setzen dabei einen fairen Würfel voraus. Das heisst, dass der Würfel so gefertigt ist, dass jede der sechs möglichen Augenzahlen im Durchschnitt gleich häufig auftritt und dass die Resultate von mehreren Würfen unabhängig voneinander sind. Wir nehmen ausserdem an, dass der Würfel nie auf einer Kante stehen bleibt oder vom Tisch fällt, sondern bei jedem Wurf eine gültige Augenzahl zeigt. Der faire Würfel ist somit ein besonders gutmütiger Zufallsgenerator.

von Ereignissen sprechen wir manchmal auch von Symbolen. Beim Würfeln gibt es $N = 6$ mögliche Ereignisse x_n . Der Übersichtlichkeit halber wählen wir beim Würfel $n = 1 \dots N$ mit der folgenden (willkürlichen aber intuitiven) Zuordnung der Augenzahlen:

$$x_1 = 1 \quad x_2 = 2 \quad x_3 = 3 \quad x_4 = 4 \quad x_5 = 5 \quad x_6 = 6$$

Wenn wir mehrere Male hintereinander würfeln, zum Beispiel drei mal, oder drei Würfel gleichzeitig werfen³⁵, so erhalten wir ein Tripel von Zufallsvariablen (X_0, X_1, X_2) . Beim Wurf entsteht ein dreifaches Ereignis, zum Beispiel dieses:

$$(X_0 = x_5, \quad X_1 = x_2, \quad X_2 = x_2)$$

Oft schreiben wir das kompakter als (x_5, x_2, x_2) , wobei wir annehmen, dass die erste Position mit X_0 korrespondiert, die zweite mit X_1 und die dritte mit X_2 . Mit der Klammer deuten wir an, dass die drei Ereignisse zusammen gehören und gemeinsam das erwähnte Tripel bilden.

5.1.2 Statistische Abhängigkeit und Unabhängigkeit

Eine wichtige Eigenschaft von Zufallsvariablen leitet sich von der Frage ab, ob das Resultat eines Zufallsexperiments vom Resultat eines anderen Zufallsexperiment abhängt oder nicht. Haben Zufallsvariablen, resp. deren Ereignisse, keinen Einfluss aufeinander, so sagt man, sie seien statistisch unabhängig. Sonst sind sie statistisch abhängig.

▼ Wir werfen gleichzeitig zwei verschiedenfarbige Würfel X und Y . Das Ereignis y_n des Würfels Y ist offenkundig unabhängig vom Ereignis x_m des Würfels X , denn die beiden Würfel wissen nichts voneinander. ▲

▼ Wir werfen gleichzeitig zwei verschiedenfarbige Würfel, einen roten und einen blauen. Wir definieren die Zufallsvariable X als die Augenzahl des roten Würfels und Y als Augenzahl des blauen. Dann definieren wir eine neue Zufallsvariable $S = X + Y$. In diesem Fall besteht offensichtlich eine statistische Abhängigkeit zwischen den beiden Zufallsvariablen X und S , ebenso wie zwischen Y und S . Liefert beispielsweise die Zufallsvariable X das Ereignis $x_4 = 4$, so kann S zwischen $s_5 = 5$ und $s_{10} = 10$ liegen. Dagegen ist das Ereignis $s_4 = 4$ nicht möglich. Liefert X dagegen das Ereignis $x_3 = 3$, so ist das Ereignis $s_4 = 4$ durchaus möglich.▲

³⁴ Es gibt auch Zufallsvariablen, deren Ereignisse stetig, also zum Beispiel reellwertig, und daher nicht abzählbar sind. Auf solche Zufallsvariablen und Ereignisse wollen wir aber nicht eingehen.

³⁵ Beim gleichzeitigen Werfen von mehreren Würfeln ist es von Vorteil, wenn die Würfel unterscheidbar sind, zum Beispiel indem sie verschiedene Farben haben. Auf diese Weise lässt sich jeder Würfel fix einer Zufallsvariable zuordnen, etwa indem ich sage: X_0 ist der gelbe Würfel, X_1 ist der rote, usw.

5.1.3 Wahrscheinlichkeit

Die Wahrscheinlichkeit³⁶ $P(x_n)$ eines Ereignisses x_n ist seine relative Häufigkeit (Formel 11). Wir finden sie entweder durch Auszählen von vielen Experimenten, oder durch Anschauung. Mit Auszählern meinen wir, dass wir das Zufallsexperiment K mal durchführen und zählen, wie oft das gesuchte Ereignis x_n dabei auftritt. Das Resultat nennen wir absolute Häufigkeit κ_n . Wie wir unten noch zeigen werden, muss K eine grosse Zahl sein. Unter Anschauung verstehen wir, dass man versucht, die inneren Abläufe des Zufallsexperiments zu verstehen, um daraus die Häufigkeit κ_n bei K Versuchen zu erraten, schätzen oder berechnen. Die Wahrscheinlichkeit $P(x_n)$ ist dann:

$$P(x_n) = \frac{\kappa_n}{K} \quad (11)$$

Daraus folgt mit $0 \leq \kappa_n \leq K$:

$$0 \leq P(x_n) \leq 1 \quad (12)$$

Ein Ereignis x_n mit der Wahrscheinlichkeit $P(x_n) = 0$ tritt nie ein. Ein Ereignis x_n mit der Wahrscheinlichkeit $P(x_n) = 1$ tritt immer auf.

▼ Das sei wieder anhand des Würfels illustriert. Ein idealer Würfel ist ein homogenes, symmetrisches Gebilde, das bezüglich jeder Seite gleich aufgebaut ist. Folglich hat bei einem Wurf keine der sechs Seiten einen Vorzug, und wir erwarten, dass alle sechs Augenzahlen die gleiche Häufigkeit haben. Im Durchschnitt sollte demnach jeweils einer von sechs Würfen eine 1, 2, 3, bis 6 liefern. Wir beziffern also die Wahrscheinlichkeit $P(x_n) = \frac{1}{6}$ für alle $n = 1 \dots 6$. ▲

Es gilt für jeden stochastischen Prozess, der N unterschiedliche Ereignisse x_n mit $n = 0 \dots N - 1$ liefern kann, dass die Summe aller Wahrscheinlichkeiten $P(x_n)$ eins sein muss:

$$\sum_{n=0}^{N-1} P(x_n) = 1 \quad (13)$$

Angewendet auf das Würfel-Experiment heisst das, dass bei jedem Wurf (Zufallsexperiment) sicher eine der sechs Augenzahlen (Ereignisse) auftreten wird. Nehmen wir den Würfel gleich nochmals

³⁶ Eigentlich sollten wir statt $P(x_n)$ die Wahrscheinlichkeit mit $P_X(x_n)$ bezeichnen, um anzudeuten, dass die Wahrscheinlichkeit eine Eigenschaft der Zufallsvariable X , resp. des betreffenden stochastischen Prozesses ist. Wir wählen aber die einfachere Schreibweise, um die Lesbarkeit zu verbessern, solange aus dem Kontext hervor geht, was gemeint ist.

und werfen ihn sechs mal, so kommen beispielsweise diese Augenzahlen heraus:

$$(2, 5, 4, 2, 6, 1)$$

Stimmt nun unsere Annahme von oben? Fälschlicherweise könnten wir ausgehend vom Resultat annehmen, dass der Würfel das Ereignis 2 doppelt so oft erzeugt, wie die Ereignisse 1, 4, 5, 6, und dass das Ereignis 3 gar nie auftritt. Das ist aber falsch. Würden wir beispielsweise eine Million Würfe auszählen, so würden wir feststellen, dass die oben angegebene Wahrscheinlichkeit von $P(x_n) = \frac{1}{6}$ recht gut stimmt. Bei einer Milliarde mal würfeln würde es sogar fast perfekt stimmen. Man spricht in diesem Zusammenhang vom *Gesetz der grossen Zahl*. Es besagt, dass sich die relative Häufigkeit auf Grund von Experimenten der wahren Wahrscheinlichkeit annähert, wenn die Anzahl Wiederholungen K des Experiments gegen unendlich strebt.

▼ In typischer Irrtum: Jemand würfelt 50 mal und es resultiert nie eine 3. Wie gross ist die Wahrscheinlichkeit, dass beim 51. Wurf endlich eine 3 erscheint? Man könnte meinen, die Wahrscheinlichkeit für eine 3 sei jetzt höher, weil alle Symbole doch im Durchschnitt jedes sechste Mal auftreten sollte. Das ist aber nicht so, denn der Würfel weiss nichts von seiner Vergangenheit, wie wir oben schon gezeigt haben. Mehrere Würfe des Würfels sind statistisch unabhängig voneinander. Folglich wird die Wahrscheinlichkeit für eine 3 im 51. und in jedem folgenden Wurf immer genau $P(3) = \frac{1}{6}$ sein. ▲

▼ Wie gross ist die Wahrscheinlichkeit, dass der Würfel eine Primzahl liefert? Wir wissen schon, dass der Würfel die Augenzahlen $x_1 = 1$ bis $x_6 = 6$ mit je der selben Wahrscheinlichkeit $P(x_n) = \frac{1}{6}$, mit $n = 1 \dots 6$, erzeugt. Wir nennen diese nun Elementar-Ereignisse. Von den sechs möglichen Elementar-Ereignissen sind $x_2 = 2$, $x_3 = 3$ und $x_5 = 5$ Primzahlen. Wenn eines von diesen dreien auftrifft, so nennen wir es ein Primzahl-Ereignis und bezeichnen es mit ρ .

Da alle sechs Elementar-Ereignisse die selbe Wahrscheinlichkeit haben, können wir die Wahrscheinlichkeit für das Primzahl-Ereignis ρ ausrechnen als *günstige Fälle über mögliche Fälle*. Zu den *günstigen Fällen* zählen die drei Elementar-Ereignisse x_2 , x_3 , x_5 . Die *möglichen Fälle* sind alle sechs Elementar-Ereignisse x_1 bis x_6 . Folglich:

$$P(\rho) = \frac{3}{6} = 0.50$$

Im Durchschnitt ist jede zweite Augenzahl eine Primzahl.

Noch etwas kann man aus diesem Beispiel ableiten: Gibt es unabhängige alternative Elementar-Ereignisse, die zu einem Ereignis führen, so addieren sich die Wahrscheinlichkeiten der Alternativen zur Wahrscheinlichkeit des Ereignisses.

$$P(\rho) = P(x_2) + P(x_3) + P(x_5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 0.50$$

Im vorliegenden Fall sind die drei Elementar-Ereignisse x_2 , x_3 , x_5 die alternativen Möglichkeiten für das Primzahl-Ereignis ρ . ▲

▼ Wie gross ist die Wahrscheinlichkeit, dass der Würfel beim ersten Wurf die Augenzahl 1 und beim zweiten Wurf die Augenzahl 4 liefert? Wir suchen also ein Doppel-Ereignis (x_1, x_4) , das aus zwei Elementar-Ereignissen x_1 und x_4 besteht (Kapitel 5.1.4 behandelt derartige Mehrfach-Ereignisse). Wir stellen folgende Überlegungen an: Die Elementar-Ereignisse x_n sind statistisch unabhängig voneinander. Es gibt total $M = N \cdot N = 36$ Doppel-Ereignisse. Die Wahrscheinlichkeit für eines von diesen Doppel-Ereignissen ist demnach $P(x_n, x_m) = \frac{1}{N \cdot N} = \frac{1}{36}$, wobei $\frac{1}{N}$ die Wahrscheinlichkeit eines Elementar-Ereignisses ist. Folglich können wir für diesen Fall schreiben:

$$P(x_n, x_m) = P(x_n) \cdot P(x_m)$$

Wir finden: Die Wahrscheinlichkeit eines Mehrfach-Ereignisses aus statistisch unabhängigen und notwendigen Elementar-Ereignissen ist das Produkt der Elementar-Wahrscheinlichkeiten. ▲

Aus den Beispielen leiten wir die folgenden wichtigen Regeln ab:

1. Entsteht ein Ereignis y aus verschiedenen alternativen und statistisch unabhängigen Elementar-Ereignissen x_n mit $n = 0 \dots N - 1$, so gilt:

$$P(y) = \sum_{n=0}^{N-1} P(x_n) \quad (14)$$

2. Besteht ein Ereignis y aus verschiedenen notwendigen und statistisch unabhängigen Elementar-Ereignissen x_n mit $n = 0 \dots N - 1$, so gilt:

$$P(y) = \prod_{n=0}^{N-1} P(x_n) \quad (15)$$

Beachte, dass sich insbesondere Formel 15 auch verwenden lässt als Test dafür, ob Zufallsvariablen unabhängig voneinander sind oder nicht: Sind nämlich von einem Zufallsexperiment die Werte $P(y)$ und alle $P(x_n)$ gegeben und erfüllen diese Werte Formel 15, so sind die Zufallsvariablen x_n statistisch unabhängig voneinander, sonst nicht.

5.1.4 Verbund-Wahrscheinlichkeit

Analog zur Wahrscheinlichkeit für *ein* Ereignis können wir auch die Wahrscheinlichkeit für ein Doppel- oder Mehrfach-Ereignis angeben. $P(x_n, x_m)$ ist die Wahrscheinlichkeit des Doppel-Ereignisses (x_n, x_m) . Wir nennen (x_n, x_m) auch Verbund-Ereignis und $P(x_n, x_m)$ die *Verbund-Wahrscheinlichkeit*. Wir finden die Verbund-Wahrscheinlichkeit wiederum experimentell oder durch Anschauung.

Einige Beispiele dazu:

▼ Wir betrachten das gleichzeitige Werfen von zwei verschiedenfarbigen Würfeln X und Y und fragen uns, wie gross die Wahrscheinlichkeit dafür ist, dass beide Würfel eine Primzahl liefern. Gefragt ist also die Wahrscheinlichkeit des Doppel-Primzahlereignisses (ρ, ρ) . Der erste Eintrag im Tupel bezieht sich auf den Würfel X , der zweite Eintrag auf Y . Wir wissen schon, dass $P(\rho) = 0.5$ ist. Und da die Würfel-Ereignisse statistisch unabhängig voneinander sind gilt:

$$P(\rho, \rho) = P(\rho) \cdot P(\rho) = 0.25$$

Im Durchschnitt ist jeder vierte Wurf ein Doppel-Primzahlereignis. ▲

▼ Wir nehmen ein Beispiel von weiter oben nochmals auf: Man würfelt mit zwei Würfeln, einem roten und einem blauen. Dann definieren wir zwei Zufallsvariablen, nämlich X , die Augenzahl des roten Würfels, und S , die Summe der Augenzahlen beider Würfel. Wir haben schon gezeigt, dass die beiden Zufallsvariablen statistisch abhängig voneinander sind.

X kann die Ereignisse $x_1 = 1$ bis $x_6 = 6$ annehmen, S die Ereignisse $s_2 = 2$ bis $s_{12} = 12$. Die Wahrscheinlichkeiten für X kennen wir bereits. Mit Blick auf S wollen wir im Moment nur das Ereignis s_{11} näher betrachten. Es tritt ein, wenn die beiden Würfel beim gemeinsamen Wurf entweder $(5, 6)$ oder $(6, 5)$ zeigen. Jede dieser Kombinationen hat die Wahrscheinlichkeit von $\frac{1}{36}$. Da jede von beiden alternativ möglich ist, hat das Ereignis s_{11} die Wahrscheinlichkeit $P(s_{11}) = \frac{2}{36}$.

Nun gilt aber beispielsweise für das Doppelereignis (x_2, s_{11}) :

$$P(x_2, s_{11}) = 0$$

Wenn die Zufallsvariable X nicht ein Ereignis x_5 oder x_6 liefert, so ist das Ereignis s_{11} gar nicht möglich. Wir stellen fest, dass wegen der statistischen Abhängigkeit von X und S die Wahrscheinlichkeit des Verbund-Ereignisses (x_2, s_{11}) gleich null und nicht gleich dem Produkt von $P(x_2)$ und $P(s_{11})$ ist.

▲

▼ Wie gross ist die Wahrscheinlichkeit, dass man mit drei unterscheidbaren Würfeln (X rot, Y grün, Z blau) die Kombination $(1, 2, 3)$ würfelt? Die Würfel sind statistisch unabhängig voneinander, also:

$$P(1, 2, 3) = P(X = 1) \cdot P(Y = 2) \cdot P(Z = 3) = \left(\frac{1}{6}\right)^3 = \frac{1}{216}$$

Etwas komplizierter ist es, wenn die drei Würfel nicht unterscheidbar sind (X_0 rot, X_1 rot, X_2 rot). In diesem Fall ist die Reihenfolge der Ereignisse im Tripel $(1, 2, 3)$ belanglos, weil es nicht möglich ist zu sagen, welcher Würfel beispielsweise die 1 geliefert hat. Es geht also nur noch darum, welche Zahlen im Tripel enthalten sind, ohne Beachtung der Reihenfolge. Demnach wären etwa $(1, 2, 3)$ und $(1, 3, 2)$ identisch.

Um die Wahrscheinlichkeit des Ereignisses $(1, 2, 3)$ bei nicht unterscheidbaren Würfeln zu ermitteln gehen wir so vor: Wir machen die Würfel unterscheidbar, indem wir auf jeden eine kleine Nummer

schreiben oder einen Farbtupfer anbringen. Dann überlegen wir uns, welche Kombination zum gesuchten Resultat führen. Wir erzeugen also wie oben geordnete Tripel und finden die günstigen Fälle:

$$(1, 2, 3) \ (1, 3, 2) \ (2, 1, 3) \ (2, 3, 1) \ (3, 1, 2) \ (3, 2, 1)$$

Wir können es uns auch so überlegen: Wenn ich ein gültiges Tripel aufschreibe, so habe ich an der ersten Position drei Möglichkeiten (1, 2, oder 3). Für die zweite Position verbleiben zwei Möglichkeiten, da ja eine bereits auf der ersten Position vergeben wurde. Für die dritte Position verbleibt nur noch eine Möglichkeit, also total $3 \cdot 2 \cdot 1 = 6$ Möglichkeiten. Die Wahrscheinlichkeit $P_u(1, 2, 3)$ für das ungeordnete Tripel (1, 2, 3) ist demnach:

$$P_u(1, 2, 3) = 6 \cdot \frac{1}{216} = \frac{1}{36}$$

Die Wahrscheinlichkeit für das ungeordnete Tripel ist grösser als jene für das geordnete. ▲

5.1.5 Totale Wahrscheinlichkeit

Sind die Verbund-Wahrscheinlichkeiten $P(x_n, x_m)$ gegeben mit $n = 0 \dots N - 1$ und $m = 0 \dots M - 1$, so lassen sich daraus auch die Wahrscheinlichkeiten $P(x_n)$ und $P(x_m)$ berechnen. Wir sprechen dann von der totalen Wahrscheinlichkeit eines Einzelergebnisses. Wir erhalten sie:

$$P(x_n) = \sum_{m=0}^{M-1} P(x_n, x_m)$$

(16)

Formel 16 gilt für statistisch abhängige und unabhängige Zufallsvariablen.

▼ Beim Werfen von zwei unterscheidbaren Würfeln X und Y ist $P(x_n, y_m) = \frac{1}{36}$. Die Wahrscheinlichkeit, dass der Würfel X eine 1 gibt ist:

$$P(x_1) = \sum_{m=1}^6 P(x_1, y_m)$$

Mit ausgeschriebener Summe folgt:

$$P(x_1) = P(1, 1) + P(1, 2) + P(1, 3) + P(1, 4) + P(1, 5) + P(1, 6) = 6 \cdot \frac{1}{36} = \frac{1}{6}$$

Das Resultat stimmt mit früheren Aussagen überein. ▲

▼ Hier ist ein Beispiel mit statistisch abhängigen Zufallsvariablen: In einer Schachtel sind drei Sorten Nägel durcheinander gemischt. Die erste Sorte ist aus Eisen und lang, die zweite Sorte ist aus Eisen

und kurz. Die dritte Sorte schliesslich ist aus Messing und kurz. Von der ersten Sorte hat 100 Nägel, von der zweiten sind es 200 und von der dritten 300 Nägel. Wenn nun jemand einen zufälligen Nagel aus der Schachtel nimmt, wie gross ist die Wahrscheinlichkeit, dass der Nagel kurz ist?

Zuerst geben wir die Wahrscheinlichkeit jeder Sorte an:

$$P(\text{Eisen, lang}) = \frac{100}{600} = \frac{1}{6}$$

$$P(\text{Eisen, kurz}) = \frac{200}{600} = \frac{2}{6}$$

$$P(\text{Messing, kurz}) = \frac{300}{600} = \frac{3}{6}$$

Nun wenden wir die Formel der totalen Wahrscheinlichkeit an:

$$P(\text{kurz}) = P(\text{Eisen, kurz}) + P(\text{Messing, kurz}) = \frac{5}{6}$$

Die Wahrscheinlichkeit, dass man beim Greifen in die Schachtel einen kurzen Nagel erwischt, ist also $\frac{5}{6}$. Der kurze Nagel kann dann aus Eisen oder aus Messing sein. ▲

5.1.6 Bedingte Wahrscheinlichkeit

Die bedingte Wahrscheinlichkeit $P(x_n|x_m)$ ist die Wahrscheinlichkeit des Ereignisses x_n , wenn wir wissen, dass das Ereignis x_m aufgetreten ist. Wir betrachten also im Grunde ein Verbund-Ereignis (x_n, x_m) , geben aber nur die Wahrscheinlichkeit von x_n an. Aus dieser Überlegung folgt:

$$P(x_n, x_m) = P(x_n|x_m) \cdot P(x_m)$$

Oder anders ausgedrückt:

$$P(x_n|x_m) = \frac{P(x_n, x_m)}{P(x_m)} \quad (17)$$

Wir setzen dabei implizit voraus, dass die beiden Ereignisse x_n und x_m in einer gegenseitigen Abhängigkeit stehen. Wir nehmen an, dass sich die Wahrscheinlichkeit $P(x_n)$ unter dem Einfluss des anderen Ereignisses x_m verändert. Wären die Ereignisse x_n und x_m statistisch unabhängig, so würde gelten:

$$P(x_n|x_m) = \frac{P(x_n, x_m)}{P(x_m)} = \frac{P(x_n) \cdot P(x_m)}{P(x_m)} = P(x_n) \quad (18)$$

▼ Es seien wieder zwei Würfel X und Y gegeben, sowie die Zufallsvariable $S = X + Y$. Wie gross ist die Wahrscheinlichkeit dafür, dass Würfel X eine 3 zeigt, wenn S gleich 5 ist?

$$P(x_3|s_5) = \frac{P(x_3, s_5)}{P(s_5)}$$

Das Verbundereignis (x_3, s_5) tritt genau dann auf, wenn y_2 auftritt. Folglich ist:

$$P(x_3, s_5) = P(x_3, y_2) = \frac{1}{36}$$

Das Ereignis s_5 tritt genau in folgenden Fällen auf:

$$(x_1, y_4) \quad (x_2, y_3) \quad (x_3, y_2) \quad (x_4, y_1)$$

Das sind also 4 von 36 möglichen Fällen. Damit folgt:

$$P(s_5) = \frac{4}{36}$$

Schliesslich erhalten wir:

$$P(x_3|s_5) = \frac{P(x_3, s_5)}{P(s_5)} = \frac{1}{36} \cdot \frac{36}{4} = \frac{1}{4}$$

Die Wahrscheinlichkeit, dass Würfel X eine 3 zeigt, wenn wir wissen, dass die Summe beider Würfel 5 ist, ist genau 0.25. ▲

▼ Wir betrachten nochmals die Schachtel mit den Nägeln und fragen uns, wie gross die Wahrscheinlichkeit ist, dass ein gezogener Nagel aus Eisen ist, wenn ich weiss, dass ich beim Greifen in die Schachtel einen kurzen Nagel erwischt habe? Wir fragen also nach $P(\text{Eisen}| \text{kurz})$. Aus der Aufgabe oben kennen wir bereits folgendes:

$$P(\text{Eisen, kurz}) = \frac{2}{6} \quad P(\text{kurz}) = \frac{5}{6}$$

Damit folgt:

$$P(\text{Eisen}| \text{kurz}) = \frac{P(\text{Eisen, kurz})}{P(\text{kurz})} = \frac{2}{5}$$

Entsprechend findet man $P(\text{Messing}| \text{kurz}) = \frac{3}{5}$. ▲

5.2 Logarithmen

Der Zweier-Logarithmus $\log_2(.)$ lässt sich mit Hilfe jedes anderen Logarithmus $\log(.)$ berechnen. Das ist gelegentlich nützlich, weil nicht jeder Taschenrechner über eine Taste für den Zweier-Logarithmus verfügt. Man geht von folgender Aufgabenstellung aus:

$$2^x = K$$

Dabei ist K eine gegebene Konstante und gesucht ist x . Man fragt sich also: "2 hoch wieviel ist K ?" Der Zweier-Logarithmus löst diese Aufgabe:

$$\begin{aligned}\log_2(2^x) &= \log_2(K) \\ x &= \log_2(K)\end{aligned}$$

Nun wenden wir aber den *anderen* Logarithmus an:

$$\begin{aligned}\log(2^x) &= \log(K) \\ x \cdot \log(2) &= \log(K)\end{aligned}$$

Daraus folgt:

$$x = \frac{\log(K)}{\log(2)}$$

Und wenn wir noch $x = \log_2(K)$ von oben einsetzen erhalten wir schliesslich:

$$\boxed{\log_2(K) = \frac{\log(K)}{\log(2)}}$$

▼ Wir geben als Beispiel einige wichtige Zweier-Logarithmen an:

K	1	2	4	8	16	32	64	128	256	1024	65536
$\log_2(K)$	0	1	2	3	4	5	6	7	8	10	16

Ist zum Beispiel $x = \log_2(100)$ gesucht, so können wir aus der Tabelle schon schliessen, dass $6 < x < 7$ ist. Wir schätzen also $x \approx 6.6$, da 100 etwas näher bei 128 liegt als bei 64. Der korrekte Wert auf drei Stellen gerundet ist $x = 6.644$. ▲

5.3 Information von statistisch abhängigen Ereignissen

Bei statistisch abhängigen Ereignissen sind die Verbund-Information und der bedingte Informationsgehalt von Interesse.

5.3.1 Verbund-Informationsgehalt

Da die Verbund-Information im Kapitel 3 schon behandelt wurde, kommen wir hier gleich zur Definition. Wir betrachten das Verbund-Ereignis (x_n, x_m) mit $n = 0 \dots N - 1$ und $m = 0 \dots M - 1$ und geben dessen Informationsgehalt an:

$$I(x_n, x_m) = \log_2 \frac{1}{P(x_n, x_m)} \quad (19)$$

Ein Vergleich mit der Information eines einzelnen Ereignisses verrät, dass wir das Doppel-Ereignis einfach substituieren können, zum Beispiel mit dem Ereignis $v_r = (x_n, x_m)$. Dabei ist $r = 0 \dots N \cdot M - 1$, denn die erste Zufallsvariable kann N Ereignisse annehmen und die zweite M . Total gibt es also $N \cdot M$ Doppel-Ereignisse. Mit der Substitution lässt sich Formel 3 auf die Definition des Informationsgehalts zurück führen:

$$I(v_r) = \log_2 \frac{1}{P(v_r)}$$

5.3.2 Bedingter Informationsgehalt

Auch hier betrachten wir zwei Ereignisse gemeinsam. Die bedingte Information $I(x_n|x_m)$ ist die Information eines Ereignisses x_n , wenn das andere Ereignis x_m schon bekannt ist.

$$I(x_n|x_m) = \log_2 \frac{1}{P(x_n|x_m)} \quad (20)$$

$P(x_n|x_m)$ ist die bedingte Wahrscheinlichkeit des Ereignisses x_n , wenn das andere Ereignis x_m gegeben ist. Wären x_n und x_m statistisch unabhängig, so könnte man vereinfachen:

$$I(x_n|x_m) = \log_2 \frac{1}{P(x_n|x_m)} = \log_2 \frac{P(x_m)}{P(x_n, x_m)} = \log_2 \frac{P(x_m)}{P(x_n) \cdot P(x_m)} = \log_2 \frac{1}{P(x_n)} = I(x_n) \quad (21)$$

Bei statistischer Abhängigkeit ist das jedoch nicht möglich.

▼ Als Beispiel verwenden wir wieder die Schachtel mit den drei Sorten von Nägeln drin. Die erste Sorte ist aus Eisen und lang, die zweite Sorte ist aus Eisen und kurz. Die dritte Sorte schliesslich ist

aus Messing und kurz. Wenn ich in die Schachtel greife und einen Nagel heraus nehme, so gelten die folgenden Wahrscheinlichkeiten:

$$P(\text{Eisen, lang}) = \frac{1}{6}$$

$$P(\text{Eisen, kurz}) = \frac{2}{6}$$

$$P(\text{Messing, kurz}) = \frac{3}{6}$$

Der Informationsgehalt, wenn ich einen kurzen Nagel ziehe, berechnet sich so: Zuerst muss man die Wahrscheinlichkeit kennen, dass man einen kurzen Nagel erwischt. Sie ist $P(\text{Eisen, kurz}) + P(\text{Messing, kurz}) = \frac{5}{6}$. Folglich:

$$I(\text{kurz}) = \log_2 \left(\frac{5}{6} \right)^{-1} \approx 0.26 \quad (\text{Bit})$$

Nun wollen wir wissen, wie gross der Informationsgehalt eines kurzen Nagels ist, wenn ich weiss, dass ich einen eisernen Nagel gezogen habe. Wir beginnen wieder mit den Wahrscheinlichkeiten:

$$P(\text{kurz}|\text{Eisen}) = \frac{P(\text{Eisen, kurz})}{P(\text{Eisen})} = \frac{2}{6} \cdot \left(\frac{1}{6} + \frac{2}{6} \right)^{-1} = \frac{2}{3}$$

Damit folgt die Information:

$$I(\text{kurz}|\text{Eisen}) = \log_2 \left(\frac{2}{3} \right)^{-1} \approx 0.58 \quad (\text{Bit})$$

Der Informationsgehalt von *kurz* ist also kleiner als von *kurz*, wenn ich gleichzeitig weiss, dass ich einen eisernen Nagel gezogen habe. ▲

5.4 Entropie von statistisch abhängigen Quellen

ToDo