

entsprechenden Signals sein, und zwar muß der Betrag

$$\left| \frac{\log p^{-1}}{T} - C \right|$$

für alle, außer einer kleinen Untermenge der langen Nachrichten, gering sein.

Wenn eine Quelle nur eine bestimmte Nachricht hervorbringen kann, ist ihre Entropie gleich Null, und es wird kein Kanal benötigt. Zum Beispiel erzeugt eine Rechenmaschine, die aufgestellt wurde, um die Ziffernfolge für den Wert π zu berechnen, eine bestimmte Folge ohne ein Zufallselement. Um dies an einen anderen Ort zu "übertragen", würde kein Kanal benötigt. Man könnte eine zweite Maschine konstruieren, um die gleiche Folge am Ort selbst zu berechnen. Das wäre allerdings etwas unpraktisch. In solch einem Fall können wir entweder einige oder alle statistischen Kenntnisse, die wir über die Quelle haben, ignorieren. Wir könnten die Ziffern von π als Zufallsfolge betrachten, indem wir ein System konstruieren, das jede beliebige Folge von Ziffern übertragen kann. Auf ähnliche Weise könnten wir auch einen gewissen Teil, aber nicht alles, unserer Kenntnisse über die statistische Struktur der englischen Sprache dafür benutzen, einen Code zu konstruieren. In einem solchen Fall nehmen wir an, daß die Quelle mit der maximalen Entropie den statistischen Bedingungen unterworfen ist, die wir annehmen möchten. Die Entropie dieser Quelle legt dann die notwendige und hinreichende Kanalkapazität fest. In unserem π -Beispiel ist die einzige angenommene Information die, daß alle Ziffern aus der Gruppe 0, 1, ..., 9 ausgewählt werden. Im Falle der englischen Sprache könnte man die etwas spärliche statistische Möglichkeit der Buchstaben-Häufigkeiten benutzen wollen, sonst jedoch nichts. Die Quelle mit maximaler Entropie ist dann die erste Näherung an die englische Sprache, und ihre Entropie bestimmt die benötigte Kanalkapazität.

Als einfaches Beispiel einiger dieser Ergebnisse betrachte man eine Quelle, die eine Folge von Buchstaben, ausgewählt aus A, B, C, D , erzeugt, mit den Wahrscheinlichkeiten $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}$, wobei aufeinanderfolgende Zeichen unabhängig voneinander ausgewählt werden. Wir erhalten

$$H = -\left(\frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4} + \frac{2}{8} \log \frac{1}{8}\right) \\ = \frac{7}{4} \text{ bit pro Zeichen.}$$

Daher können wir ein Codiersystem annähern, das Nachrichten von dieser Quelle in Binärziffern codiert, mit einem Durchschnitt von $\frac{7}{4}$

Binärziffern pro Zeichen. In diesem Fall läßt sich tatsächlich der Grenzwert durch den folgenden Code erreichen (der sich durch die Methode des zweiten Beweises von Lehrsatz 9 ergibt):

A	0
B	10
C	110
D	111

Die Durchschnittszahlen der Binärstellen, die dabei benutzt werden, eine Folge von N Zeichen zu codieren, ist dann:

$$N \left(\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{2}{8} \times 3 \right) = \frac{7}{4} N.$$

Es ist leicht ersichtlich, daß die Binärzeichen 0, 1 Wahrscheinlichkeiten von $\frac{1}{2}$ und $\frac{1}{2}$ haben, und so ist H für die codierten Folgen ein bit pro Zeichen. Da wir durchschnittlich $\frac{7}{4}$ Binärzeichen für jeden Originalbuchstaben benötigen, sind die Entropien für eine Zeiteinheit gleich. Die maximal mögliche Entropie für den ursprünglichen Zeichenvorrat ist $\log 4 = 2$, und sie kommt vor, wenn A, B, C, D die Wahrscheinlichkeiten $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}$ und $\frac{1}{8}$ haben. Also ist die relative Entropie gleich $\frac{7}{4}$. Wir können die Binärfolgen übertragen auf den Ursprungssatz der Zeichen mit einer Basis zwei-zu-eins nach der folgenden Tabelle:

00	A'
01	B'
10	C'
11	D'

Dieser doppelte Prozeß codiert nun die Originalnachricht in dieselben Zeichen, jedoch mit einer Durchschnittsverdichtung im Verhältnis $\frac{7}{4}$.

Als zweites Beispiel betrachten wir eine Quelle, die eine Folge von A's und B's erzeugt mit den Wahrscheinlichkeiten p für A und q für B.

Wenn $p < q$ ist, erhalten wir

$$H = -\log p^p (1-p)^{1-p} \\ = -p \log p (1-p)^{(1-p)/p} \\ \cong p \log \frac{e}{p}.$$

In solch einem Falle kann eine einigermaßen gute Codierung auf einem 0-1-Kanal erreicht werden, indem man eine besondere Folge sendet, z.B. 0000 für das sehr seltene Zeichen A und dann eine

Folge, welche die *Anzahl* der folgenden *B*'s angibt. Sie könnte als Binärzahl angegeben werden, für deren Bildung die spezielle 0000-Folge nicht berücksichtigt wird. Alle Zahlen bis 16 werden wie gewöhnlich dargestellt; 16 wird durch die der 16 nächsthöhere Binärzahl dargestellt, die nicht vier Nullen enthält, nämlich 10001 (an sich die Dualzahl für 17) usw.

Es zeigt sich, daß, sobald $p \rightarrow 0$ geht, sich die Codierung dem Ideal nähert, vorausgesetzt, daß die Länge der speziellen Folge richtig angepaßt ist.

II. Der diskrete gestörte Kanal

11. Vorstellung eines diskreten gestörten Kanals

Wir betrachten nun den Fall, in dem das Signal während der Übertragung oder an einer der beiden Stationen gestört wird. Dies bedeutet, daß das empfangene Signal nicht unbedingt dasselbe ist wie jenes, das von dem Sender ausgesendet wurde. Zwei Fälle können unterschieden werden. Wenn ein bestimmtes gesendetes Signal immer dasselbe Empfangssignal erzeugt, d. h. das Empfangssignal ist eine eindeutige Funktion des Sendesignals, dann kann der Effekt "Verzerrung" genannt werden. Wenn es zu dieser Funktion eine inverse Funktion gibt - keine zwei Sendesignale erzeugen dasselbe Empfangssignal - kann die Verzerrung korrigiert werden, wenigstens im Prinzip, indem man das Empfangssignal der inversen funktionalen Operation unterzieht.

Der interessante Fall ist hier derjenige, in dem das Signal bei der Übertragung nicht immer derselben Veränderung unterliegt. In diesem Fall können wir annehmen, daß das Empfangssignal E eine Funktion des Sendesignals S und einer zweiten Variablen, der Störung N , ist.

$$E = f(S, N).$$

Diese Störung wird als eine Zufallsvariable betrachtet, geradeso wie die Nachricht. Im allgemeinen kann sie durch einen passenden stochastischen Prozeß dargestellt werden. Der allgemeinste Typ eines gestörten diskreten Kanals, den wir betrachten, ist eine Verallgemeinerung des störungsfreien Kanals mit einer begrenzten Anzahl von Zuständen, wie er bereits beschrieben wurde. Wir nehmen eine begrenzte Anzahl von Zuständen und einen Satz von Wahrscheinlichkeiten an.

$$p_{\alpha, i}(\beta, j).$$

Dies ist die Wahrscheinlichkeit, wenn sich der Kanal in dem Zustand α befindet und das Zeichen i übertragen wird, daß das Zeichen j empfangen wird und danach der Kanal in den Zustand β übergeht. So stellen α und β die möglichen Zustände dar, i die mögli-

chen Sendesignale und j die möglichen Empfangssignale. Im Falle, daß aufeinanderfolgende Zeichen unabhängig voneinander gestört werden, gibt es nur einen Zustand, und der Kanal ist durch den Satz von Übergangswahrscheinlichkeiten $p_i(j)$ beschrieben, der Wahrscheinlichkeit, mit der das Zeichen i als Zeichen j empfangen wird.

Wenn ein gestörter Kanal von einer Quelle gespeist wird, gibt es zwei statistische Vorgänge: Die Quelle und die Störung. Daher gibt es auch eine Anzahl von Entropien, die errechnet werden können. Zunächst ist da die Entropie $H(x)$ der Quelle oder die der Eingabe in den Kanal (diese werden gleich sein, wenn der Sender nicht-singulär ist). Die Entropie der Ausgabe des Kanals, d. h. des Empfangssignals, wird durch $H(y)$ bezeichnet. Im störungsfreien Fall ist $H(y) = H(x)$. Die gemeinsame Entropie der Ein- und der Ausgabe wird $H(x, y)$ sein. Schließlich gibt es zwei bedingte Entropien $H_x(y)$ und $H_y(x)$, die Entropie der Ausgabe, wenn die Eingabe bekannt ist, und umgekehrt. Unter diesen Größen haben wir die Beziehungen

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x).$$

Alle diese Entropien können entweder "pro Sekunde" oder "pro Zeichen" gemessen werden.

12. Äquivokation und Kanalkapazität

Wenn der Kanal nicht störungsfrei ist, ist es im allgemeinen auch nicht möglich, die ursprüngliche Nachricht oder das Sendesignal durch Umformung des Empfangssignals E mit Sicherheit zu rekonstruieren. Es gibt jedoch Wege, die Information so zu übertragen, daß damit Störungen optimal bekämpft werden. Dieses Problem werden wir nun behandeln.

Angenommen, es gibt zwei mögliche Zeichen 0 und 1 und wir übertragen mit einer Rate von 1000 Zeichen pro Sekunde und mit den Wahrscheinlichkeiten $p_0 = p_1 = \frac{1}{2}$. Auf diese Weise erzeugt unsere Quelle eine Information von 1000 bit pro Sekunde. Während der Übertragung bewirkt die Störung Fehler, so daß im Durchschnitt eines von 100 Zeichen falsch empfangen wird (eine 0 als 1 oder 1 als 0). Wie groß ist die Rate der Informationsübertragung? Sicherlich geringer als 1000 bit pro Sekunde, da etwa 1% der empfangenen Zeichen falsch ist. Im ersten Moment möchte man sagen, die Übertragungsrate sei 990 bit pro Sekunde, indem wir nur die er-

wartete Fehleranzahl subtrahieren. Dies ist jedoch nicht zufriedenstellend, weil dabei versäumt wird, die fehlende Kenntnis des Empfängers darüber, wo die Fehler auftauchen, in Betracht zu ziehen. Wir können das bis zum Extremfall ausdehnen und annehmen, die Störung sei so stark, daß die empfangenen Zeichen ganz unabhängig von den gesendeten Zeichen sind. Die Wahrscheinlichkeit, "1" zu empfangen, ist $\frac{1}{2}$, was immer auch gesendet wurde, und gleichermaßen auch für "0". Dann sind allein durch Zufall etwa die Hälfte aller Zeichen korrekt, und wir würden glauben, daß das System 500 bit pro Sekunde übertrage, während tatsächlich überhaupt keine Information übertragen wird. Eine ebenso "gute" Übertragung erhielte man, wenn man den Kanal vollständig weglasse und am Ort des Empfängers eine Münze werfe.

Offensichtlich wäre die richtige Verbesserung, um auf die Menge der übertragenen Information zu kommen, den Betrag der Information zu benutzen, der im Empfangssignal fehlt, oder entsprechend die Unsicherheit über das empfangene Signal gegenüber dem, das tatsächlich gesendet worden ist. Nach unserer vorhergehenden Betrachtung der Entropie als dem Maß der Unsicherheit scheint es vernünftig, die bedingte Entropie der Nachricht als ein Maß für die fehlende Information zu benutzen, wenn das Empfangssignal bekannt ist. Dies ist in der Tat die richtige Definition, wie wir später sehen werden. Dieser Idee folgend, würde man die tatsächliche Übertragungsrate R dadurch erhalten, daß man von dem Betrag der Erzeugungsrate (d. h. der Entropie der Quelle) die durchschnittliche Rate der bedingten Entropie subtrahiert.

$$R = H(x) - H_y(x).$$

Die bedingte Entropie $H_y(x)$ wird, da dies vorteilhaft ist, als Äquivokation¹⁰⁾ bezeichnet. Sie mißt die durchschnittliche Mehrdeutigkeit des empfangenen Signals. In dem oben erwähnten Beispiel ist, wenn eine 0 empfangen wurde, die Wahrscheinlichkeit *a posteriori*, daß eine 0 auch gesendet wurde, gleich 0,99, und daß eine 1 gesendet wurde, gleich 0,01. Diese Zahlen werden vertauscht, wenn eine 1 empfangen wird. Daher ist

$$\begin{aligned} H_y(x) &= - [0,99 \log 0,99 + 0,01 \log 0,01] \\ &= 0,081 \text{ bit/Zeichen} \end{aligned}$$

oder 81 bit pro Sekunde. Wir können sagen, daß das System mit einer Rate von $1000 - 81 = 919$ bit pro Sekunde überträgt. Im Ex-

¹⁰⁾ Äquivokation: engl. *equivocation* = Zweideutigkeit.

tremfall, in dem es gleich wahrscheinlich ist, ob 0 empfangen wird als 0 oder als 1 und ebenso für 1, sind die Wahrscheinlichkeiten *a posteriori* $\frac{1}{2}$, $\frac{1}{2}$ und

$$H_v(x) = -\left[\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2}\right] \\ = 1 \text{ bit pro Zeichen}$$

oder 1000 bit pro Sekunde. Die Übertragungsrate ist dann 0, wie es sein sollte. Der folgende Lehrsatz gibt eine direkte lehrreiche Interpretation der Äquivokation und dient auch dazu, sie als einzig passendes Maß zu rechtfertigen. Wir betrachten ein Kommunikationssystem und einen Beobachter (oder eine Hilfseinheit), der beides sehen kann, was gesendet und was empfangen wird (mit von Störungen abhängigen Fehlern). Dieser Beobachter bemerkt die Fehler in der empfangenen Nachricht und sendet Daten zu einem Empfangspunkt über einen "Korrekturkanal", um den Empfänger in die Lage zu versetzen, die Fehler zu korrigieren. Die Situation ist schematisch in Abb. 8 dargestellt.

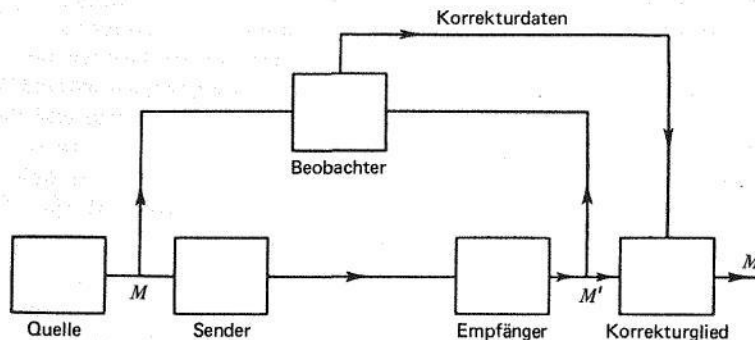


Abb. 8: Schematisches Diagramm eines Korrektursystems

Lehrsatz 10: Wenn der Korrekturkanal eine Kapazität von $H_v(x)$ besitzt, ist es möglich, die Korrekturdaten so zu codieren, daß man sie über den Kanal senden kann, und alle Fehler bis auf einen beliebig kleinen Bruchteil ϵ zu korrigieren. Dies ist nicht möglich, wenn die Kanalkapazität kleiner als $H_v(x)$ ist.

Danach ist $H_v(x)$ ungefähr der Betrag der zusätzlichen Information, der pro Sekunde dem Empfangspunkt zugeteilt werden muß, um die empfangene Nachricht zu korrigieren.

Um den ersten Teil zu beweisen, betrachte man lange Folgen der erhaltenen Nachricht M' und der entsprechenden ursprünglichen Nachricht M . Es wird logarithmisch $TH_v(x)$ von den M -Nachrichten geben, die ziemlich sicher jedes M' erzeugt haben könnten. So haben wir $TH_v(x)$ Binärzeichen alle T Sekunden zu senden. Das kann mit der Fehlerhäufigkeit ϵ bei einer Kanalkapazität $H_v(x)$ durchgeführt werden.

Der zweite Teil kann dadurch bewiesen werden, daß man sich zunächst überlegt, daß für beliebige diskrete Zufallsvariablen x, y, z gilt:

$$H_v(x, z) \geq H_v(x).$$

Die linke Seite kann so ausgeschrieben werden, daß

$$H_v(z) + H_{vz}(x) \geq H_v(x) \\ H_{vz}(x) \geq H_v(x) - H_v(z) \geq H_v(x) - H(z)$$

entsteht.

Wenn wir x als die Ausgabe der Quelle identifizieren, y als das Empfangssignal und z als das über den Korrekturkanal gesendete Signal, dann ist die rechte Seite der Äquivokation, vermindert um die Übertragungsrate über den Korrekturkanal. Wenn die Kapazität dieses Kanals kleiner als die Äquivokation ist, wird die rechte Seite größer als Null sein und $H_{vz}(x) > 0$. Dies ist jedoch die Unsicherheit darüber, was gesendet wurde, wenn man beide, das empfangene Signal und das Korrektursignal kennt. Wenn jener Wert also größer als Null ist, kann die Fehlerfrequenz nicht beliebig klein sein.

Beispiel:

Angenommen, die Fehler tauchen in einer Folge von Binärzeichen zufällig auf, so gibt es die Wahrscheinlichkeit p , daß eine Ziffer falsch ist, und $q = 1 - p$, daß sie richtig ist. Die Fehler können korrigiert werden, wenn ihre Positionen bekannt sind. Daher braucht der Korrekturkanal nur Informationen für diese Positionen zu senden. Das bedeutet, von einer Quelle auszugehen, die Binärzeichen mit der Wahrscheinlichkeit p für 1 (falsch) und q für 0 (richtig) sendet. Dies verlangt einen Kanal mit der Kapazität

$$-[p \log p + q \log q],$$

was die Äquivokation des ursprünglichen Systems darstellt.

Die Übertragungsrate R kann nach den oben beschriebenen Gleichungen

chungen auch in zwei anderen Formen geschrieben werden. Wir haben

$$\begin{aligned} R &= H(x) - H_y(x) \\ &= H(y) - H_x(y) \\ &= H(x) + H(y) - H(x, y). \end{aligned}$$

Der erste Ausdruck wurde bereits als die Informationsmenge interpretiert, die gesendet wurde, vermindert um die Unsicherheit darüber, was gesendet wurde. Der zweite mißt die empfangene Menge minus den Teil davon, der auf Störungen zurückzuführen ist. Der dritte ist die Summe der beiden Mengen, von der die Gesamtentropie abgezogen wird, und daher ist er auch in einem gewissen Sinn die Anzahl der bit pro Sekunde, die beiden gemeinsam ist. So haben alle drei Ausdrücke eine gewisse praktische Bedeutung.

Die Kapazität C eines gestörten Kanals sollte die maximal mögliche Übertragungsrate darstellen, d. h. die Rate, wenn die Quelle richtig an den Kanal angepaßt ist. Wir definieren die Kanalkapazität daher durch

$$C = \text{Max } (H(x) - H_y(x)),$$

wobei der Maximalwert, mit Rücksicht auf alle möglichen Informationsquellen, als Eingabe in den Kanal benützt wird. Wenn der Kanal störungsfrei ist, dann ist $H_y(x) = 0$. Die Definition ist dann gleichbedeutend mit derjenigen, die bereits für einen störungsfreien Kanal gegeben wurde, da nach Lehrsatz 8 die Maximalentropie für einen Kanal gleich seine Kapazität ist.

13. Der fundamentale Lehrsatz für einen diskreten Kanal mit Störung

Es mag erstaunlich erscheinen, daß wir eine bestimmte Kapazität C für einen gestörten Kanal definieren, da doch in solch einem Fall nie sichere Information übertragen werden kann. Es ist jedoch auch klar, daß durch die Sendung der Nachricht in einer redundanten Form die Fehlerwahrscheinlichkeit vermindert werden kann. Wenn man zum Beispiel die Nachricht oft wiederholt, könnte durch eine statistische Studie der verschiedenen empfangenen Versionen der Nachricht die Fehlerwahrscheinlichkeit sehr vermindert werden. Man würde jedoch erwarten, daß, um diese Fehlerwahrscheinlichkeit fast verschwinden zu lassen, die Redundanz beim Codieren unendlich groß werden muß und die Übertragungsrate daher gegen Null

geht. Das ist jedoch keineswegs so, denn wenn es so wäre, gäbe es keine gut definierbare Kapazität, sondern nur eine Kapazität bei einer gegebenen Fehlerhäufigkeit oder bei einer gegebenen Äquivokation, wobei die Kapazität desto geringer wird, je strenger die Maßstäbe an die Fehlerarmut angelegt werden. Tatsächlich hat die oben definierte Kapazität C eine sehr bestimmte Bedeutung. Es ist möglich, durch die richtige Codierung Information mit der Rate C durch den Kanal mit einer beliebig kleinen Fehlerhäufigkeit oder Äquivokation zu übertragen. Diese Aussage stimmt nicht, sobald die Rate größer wird als C . Wenn ein Versuch unternommen wird, eine größere Rate als C zu übertragen, z. B. $C + R_1$, dann wird es notwendigerweise eine Äquivokation geben, die gleich oder größer als die zusätzliche Information R_1 ist. Die Natur trägt dieser Tatsache Rechnung, indem sie gerade so viel Ungewißheit verlangt, daß in der Tat nicht mehr als C korrekt übertragen wird.

Die Situation ist in Abb. 9 dargestellt. Die Rate der Information, die in den Kanal einfließt, ist horizontal aufgetragen, die Äquivokation vertikal. Jeder Punkt über der dicken Linie in dem schraffierten Bereich ist erreichbar und jeder darunter ist unerreichbar. Die Punkte auf der Linie können im allgemeinen nicht erreicht werden, es gibt jedoch normalerweise zwei Punkte darauf, die zu erreichen sind.

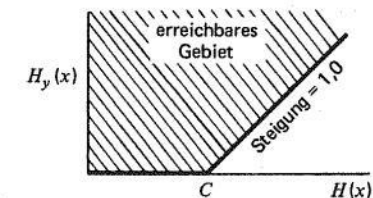


Abb. 9:
Die Äquivokation, die bei einer gegebenen Eingabeentropie für einen Kanal möglich ist

Diese Ergebnisse stellen die Hauptrechtfertigung für die Definition von C dar und werden nun bewiesen.

Lehrsatz 11: Angenommen, ein diskreter Kanal habe die Kapazität C und eine diskrete Quelle die Entropie H pro Sekunde. Wenn $H \leq C$, dann existiert ein Codiersystem, so daß die Ausgabe der Quelle über den Kanal übertragen werden kann mit einer beliebig kleinen Fehlerhäufigkeit (oder einer beliebig geringen Äquivokation). Wenn $H > C$ ist, dann ist es möglich, die Quelle so zu codieren, daß die Äquivokation kleiner als $H - C + \epsilon$ ist, wobei ϵ beliebig klein ist. Es gibt keine Codiermethode, durch die eine geringe Äquivokation als $H - C$ erreicht werden könnte.

Die Methode, den ersten Teil dieses Lehrsatzes zu beweisen, kann nicht darauf hinauslaufen, eine Codiermethode aufzuzeigen, die die gewünschten Eigenschaften hätte, sondern es ist zu zeigen, daß solch ein Code in einer bestimmten Codegruppe vorhanden sein muß.

Tatsächlich werden wir den Durchschnitt der Fehlerhäufigkeit in dieser Gruppe nehmen und zeigen, daß dieser Durchschnitt kleiner als ϵ gemacht werden kann. Wenn aber der Durchschnitt einer Zahlengruppe kleiner als ϵ ist, dann muß mindestens eine Zahl in der Gruppe vorhanden sein, die kleiner als ϵ ist. Dies wird zum gewünschten Ergebnis führen. Die Kapazität C eines gestörten Kanals ist definiert worden als

$$C = \text{Max} (H(x) - H_y(x)),$$

wobei x die Eingabe und y die Ausgabe ist. Die Maximierung betrifft alle Quellen, die für die Eingabe in den Kanal benutzt werden könnten.

Wir nehmen nun an, S_0 sei eine Quelle, die die maximale Kapazität C erreicht. Wenn dieses Maximum durch keine Quelle tatsächlich erreicht, sondern nur als Grenzwert angenähert wird, dann nehmen wir an, daß S_0 eine Quelle ist, die nur beinahe die maximale Rate erreicht. Angenommen, S_0 wird als Eingabe für den Kanal benutzt. Wir nehmen für die möglicherweise gesendeten und empfangenen Folgen eine Zeitspanne T an. Folgendes ist dann richtig:

1. Die übertragenen Folgen lassen sich in zwei Klassen einteilen, eine Gruppe mit hoher Wahrscheinlichkeit, zu der etwa $2^{TH(x)}$ Mitglieder gehören, und die restlichen Folgen mit einer geringen Gesamtwahrscheinlichkeit.
2. Ganz ähnlich hat bei den empfangenen Folgen eine Gruppe von $2^{TH(y)}$ Mitgliedern eine hohe Wahrscheinlichkeit, und die restlichen Folgen haben eine niedrige Wahrscheinlichkeit.
3. Jede Ausgabe hoher Wahrscheinlichkeit könnte durch etwa $2^{TH_y(x)}$ Eingaben erzeugt worden sein. Die Gesamtwahrscheinlichkeit für alle anderen Fälle ist klein.
4. Jede Eingabe hoher Wahrscheinlichkeit könnte ungefähr $2^{TH_x(y)}$ Ausgaben bewirken. Die Gesamtwahrscheinlichkeit aller anderen Ergebnisse ist klein.

Alle die ϵ und δ , die durch die Worte "klein" und "ungefähr" in diesen Feststellungen einbezogen werden, nähern sich Null, wenn

wir T größer werden lassen und S_0 sich der maximalen Quelle annähern läßt.

Die Situation ist in Abb. 10 zusammengefaßt worden, wo die Eingabefolgen Punkte auf der linken Seite und die Ausgabefolgen Punkte auf der rechten Seite sind. Der obere Fächer von Linien stellt den Bereich möglicher Ursachen für eine typische Ausgabe dar. Der untere Fächer weist auf das Gebiet möglicher Ergebnisse hin, ausgehend von einer typischen Eingabe. In beiden Fällen bleiben die Gruppen mit "niedriger Wahrscheinlichkeit" unbeachtet.

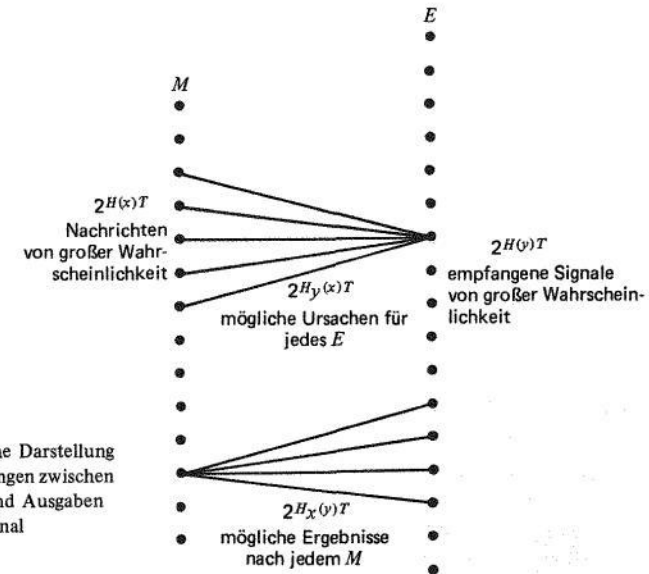


Abb. 10:
Schematische Darstellung
der Beziehungen zwischen
Eingaben und Ausgaben
in einem Kanal

Angenommen, wir haben nun eine andere Quelle S , die Information mit der Rate R produziert, wobei $R < C$ ist. In der Zeit T wird diese Quelle 2^{RT} Nachrichten von hoher Wahrscheinlichkeit erzeugen. Diese wollen wir einer Auswahl von möglichen Kanaleingaben zuordnen, auf eine Weise, die eine kleine Fehlerhäufigkeit erlaubt. Wir werden diese Zuordnung auf jede mögliche Weise herstellen (indem wir jedoch nur die Gruppe der Eingaben mit hoher Wahrscheinlichkeit benutzen, die durch die Quelle S_0 bestimmt wurde) und den Mittelwert der Fehlerhäufigkeit für diese große Klasse von möglichen Codiersystemen bestimmen. Das ist das gleiche, als wenn die Fehlerhäufigkeit für eine zufällige Zuordnung zwischen

Nachrichten und Kanaleingaben mit der Dauer T errechnet wird. Angenommen, eine bestimmte Ausgabe y_1 wird beobachtet. Wie groß ist die Wahrscheinlichkeit dafür, daß mehr als eine Nachricht von S in der Gruppe der möglichen Ursachen für y_1 vorkommt? Es gibt 2^{T^R} Nachrichten, die zufällig auf $2^{T^H(x)}$ Punkte verteilt sind. Die Wahrscheinlichkeit, daß ein bestimmter Punkt eine Nachricht ist, ist dadurch

$$2^{T(R-H(x))},$$

und die Wahrscheinlichkeit, daß keiner dieser Punkte im Fächer eine Nachricht ist (außer der tatsächlichen ursprünglichen Nachricht), wird

$$P = [1 - 2^{T(R-H(x))}]^{2^{T^H(y_1)}}.$$

Nun ist $R < H(x) - H_y(x)$, daher wird $R - H(x) = -H_y(x) - \eta$, wobei η positiv ist. Dementsprechend nähert sich

$$P = [1 - 2^{-T(H_y(x) - T\eta)}]^{2^{T^H(y_1)}}$$

(wenn $T \rightarrow \infty$) dem Wert

$$1 - 2^{-T\eta}.$$

Daraus ergibt sich, daß die Wahrscheinlichkeit für einen Fehler gegen Null geht, und damit ist der erste Teil des Lehrsatzes bewiesen.

Der zweite Teil des Lehrsatzes ist leicht erklärt, wenn man bemerkt, daß lediglich C bit pro Sekunde von der Quelle gesendet werden können und dabei die Reste der erzeugten Information vollständig vernachlässigt werden. Am Empfänger ergibt der vernachlässigte Teil eine Äquivokation von $H(x) - C$, und zum übertragenen Teil braucht nur der Wert ϵ addiert zu werden. Dieser Grenzwert kann auch auf viele andere Arten erreicht werden, wie noch gezeigt wird, wenn wir den Fall einer kontinuierlichen Nachricht behandeln.

Die letzte Feststellung des Lehrsatzes ist eine einfache Folge unserer Definition von C . Angenommen, wir können eine Quelle mit $H(x) = C + a$ in der Weise codieren, daß wir eine Äquivokation $H_y(x) = a - \epsilon$ erhalten, wobei ϵ positiv ist, dann wäre

$$H(x) - H_y(x) = C + \epsilon,$$

wobei ϵ positiv ist. Dies widerspricht der Definition von C als dem Maximum von $H(x) - H_y(x)$.

Tatsächlich ist mehr bewiesen worden, als im Lehrsatz festgestellt wurde. Wenn der Durchschnitt einer Gruppe positiver Zahlen in-

nerhalb eines Toleranzbereichs von ϵ Null ist, dann kann ein Bruchteil von höchstens $\sqrt{\epsilon}$ Werte größer als $\sqrt{\epsilon}$ enthalten. Da ϵ beliebig klein ist, können wir sagen, daß praktisch alle Systeme beliebig nahe an der Idealvorstellung sein müssen.

14. Diskussion

Die Demonstration von Lehrsatz 11, die kein reiner Existenzbeweis ist, hat einige Mängel, wie sie solchen Beweisen eigen sind. Ein Versuch, eine gute Näherung an die ideale Codierung dadurch zu erreichen, daß man dem Weg des Beweises folgt, ist im allgemeinen unpraktisch. Tatsächlich wurde, abgesehen von einigen ziemlich trivialen Fällen und gewissen Grenzsituationen, keine wirkliche Beschreibung einer Serie von Näherungen an das Ideal gefunden. Wahrscheinlich ist dies kein Zufall, sondern hängt mit der Schwierigkeit zusammen, eine brauchbare Konstruktion für eine gute Näherung an eine zufällige Folge zu finden.

Eine Näherung an das Ideal hätte die Eigenschaft, daß, auch wenn das Signal durch Störungen ziemlich verändert würde, das Original noch herausgelesen werden könnte. Mit anderen Worten, die Veränderung wird es im allgemeinen nicht einem anderen vernünftigen Signal mehr annähern als dem ursprünglichen. Dies wird auf Kosten einer gewissen Redundanz in der Codierung erreicht. Die Redundanz muß in der richtigen Weise eingebaut werden, um die bestimmte Struktur der Störung zu bekämpfen. Jedoch wird jegliche Redundanz der Quelle normalerweise helfen, wenn sie am Empfangspunkt ausgenützt wird. Insbesondere wird, wenn die Quelle bereits eine gewisse Redundanz hat und kein Versuch gemacht wird, sie durch Anpassung an den Kanal auszulöschen, diese Redundanz helfen, Störungen zu bekämpfen. Z. B. könnte man in einem ungestörten Telegrafkanal etwa 50 Prozent Zeit sparen durch richtige Codierung der Nachrichten. Dies wird nicht getan, und der größte Teil der Redundanz der englischen Sprache verbleibt in den Zeichen, die als Signal über den Kanal gehen. Dies hat den Vorteil, daß ein beträchtliches Rauschen im Kanal erlaubt ist. Ein ziemlich großer Teil der Buchstaben darf unkorrekt empfangen werden und kann durch den Kontext noch rekonstruiert werden. Tatsächlich ist dies wahrscheinlich in vielen Fällen keine schlechte Annäherung an das Ideal, da die statistische Struktur des Englischen ziemlich kompliziert ist und vernünftige englische Wort-

folgen nicht weit entfernt (im Sinne der Forderung des Lehrsatzes) von einer zufälligen Auswahl sind.

Wie im ungestörten Fall, wird im allgemeinen eine Verzögerung nötig sein, um die ideale Codierung anzunähern. Sie hat nun die zusätzliche Funktion, einen hohen Störpegel zuzulassen, der das Signal beeinflusst, bevor am Empfangspunkt irgendeine Beurteilung in bezug auf die ursprüngliche Nachricht gemacht wird. Dadurch, daß man den Pegel erhöht, werden die möglichen statistischen Aussagen immer mehr verschärft.

Der Inhalt von Lehrsatz 11 und sein Beweis können auf etwas andere Art formuliert werden, wodurch der Zusammenhang mit dem störungsfreien Fall klarer demonstriert wird. Betrachten wir die möglichen Signale der Dauer T und nehmen wir an, daß nur ein Teil davon für den Gebrauch ausgewählt wurde. Nehmen wir weiter an, daß in dieser Untermenge alle mit derselben Wahrscheinlichkeit benutzt werden und daß der Empfänger so konstruiert ist, gerade den wahrscheinlichsten Fall aus der Untermenge von möglichen Signalen als das originale Signal auszusuchen, falls das Signal gestört ankommt. Wir definieren $N(T, q)$ als die größtmögliche Anzahl von Signalen, die wir für die Untermenge auswählen können, so daß die Wahrscheinlichkeit einer falschen Interpretation kleiner als oder gleich q ist.

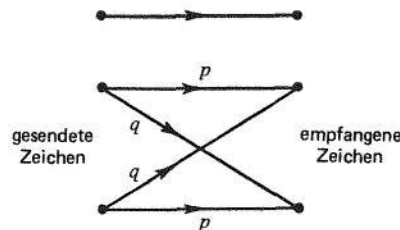
Lehrsatz 12: $\lim_{T \rightarrow \infty} \frac{\log N(T, q)}{T} = C$, wobei C die Kanalkapazität ist, vorausgesetzt, daß q nicht gleich 0 oder 1 ist.

Mit anderen Worten: Gleichgültig, wie wir unsere Zuverlässigkeitsgrenze setzen, können wir in der Zeit T genügend Nachrichten, denen etwa CT bit entsprechen, zuverlässig unterscheiden, wenn T genügend groß ist. Lehrsatz 12 kann mit der Definition der Kapazität eines ungestörten Kanals, die im Kapitel 1 gegeben wurde, verglichen werden.

15. Beispiel eines diskreten Kanals und seiner Kapazität

Ein einfaches Beispiel eines diskreten Kanals ist in Abb. 11 dargestellt. Es gibt drei mögliche Zeichen. Das erste wird niemals durch Störungen beeinflusst. Das zweite und dritte haben beide die Wahr-

Abb. 11:
Beispiel eines diskreten Kanals



scheinlichkeiten p , unverfälscht anzukommen, und q , in das jeweils andere umgewandelt zu werden.

Es sei $\alpha = -[p \log p + q \log q]$, und P, Q und Q seien die Wahrscheinlichkeiten, daß das erste, zweite und dritte Zeichen gesendet wird (wobei die letzten zwei aus Symmetriegründen gleich sind). Wir erhalten:

$$H(x) = -P \log P - 2Q \log Q$$

$$H_y(x) = 2Q\alpha$$

Wir wollen P und Q so wählen, daß $H(x) - H_y(x)$ ein Maximum wird, zusammen mit der Bedingung $P + 2Q = 1$. Dafür betrachten wir

$$U = -P \log P - 2Q \log Q - 2Q\alpha + \lambda(P + 2Q)$$

$$\frac{\partial U}{\partial P} = -1 - \log P + \lambda = 0$$

$$\frac{\partial U}{\partial Q} = -2 - 2 \log Q - 2\alpha + 2\lambda = 0.$$

Es wird λ eliminiert:

$$\log P = \log Q + \alpha$$

$$P = Qe^\alpha = Q\beta$$

$$P = \frac{\beta}{\beta + 2} \quad Q = \frac{1}{\beta + 2}.$$

Daraus folgt die Kanalkapazität

$$C = \log \frac{\beta + 2}{\beta}.$$

Man beachte, wie damit die sofort einleuchtenden Werte in den Fällen $p = 1$ und $p = \frac{1}{2}$ bestätigt werden. Im ersten Fall ist $\beta = 1$ und $C = \log 3$, was richtig ist, da der Kanal dann störungsfrei mit drei möglichen Zeichen arbeitet. Wenn $p = \frac{1}{2}$ ist, dann ist $\beta = 2$ und $C = \log 2$. Hier können das zweite und dritte Zeichen über-

haupt nicht voneinander unterschieden werden und wirken zusammen wie ein einziges Zeichen. Das erste Zeichen wird mit der Wahrscheinlichkeit $P = \frac{1}{2}$ benutzt und das zweite und dritte zusammen mit der Wahrscheinlichkeit $\frac{1}{2}$. Diese kann unter ihnen auf jede gewünschte Art verteilt sein, und es wird immer noch die maximale Kapazität erreicht.

Für dazwischenliegende Werte von p wird die Kanalkapazität zwischen $\log 2$ und $\log 3$ liegen. Die Unterscheidung zwischen dem zweiten und dritten Zeichen bringt einige zusätzliche Information, jedoch nicht so viel wie im ungestörten Fall. Das erste Zeichen wird etwas häufiger benutzt als die beiden anderen, da es nicht gestört wird.

16. Die Kanalkapazität in gewissen Spezialfällen

Wenn die Störung aufeinanderfolgende Kanalzeichen unabhängig voneinander beeinflußt, kann diese durch einen Satz von Übergangswahrscheinlichkeiten p_{ij} beschrieben werden. Dies ist die Wahrscheinlichkeit, daß, wenn das Zeichen i gesendet wird, das Zeichen j empfangen wird. Die Kanalkapazität ist dann gegeben durch das Maximum von

$$-\sum_{i,j} P_i p_{ij} \log \sum_i P_i p_{ij} + \sum_{i,j} P_i p_{ij} \log p_{ij},$$

wobei wir die P_i im Rahmen $\sum P_i = 1$ variieren. Dies führt nach der Methode von Lagrange zu den Gleichungen

$$\sum_j p_{sj} \log \frac{p_{sj}}{\sum_i P_i p_{ij}} = \mu \quad s = 1, 2, \dots$$

Durch Multiplikation mit P_s und Summieren über s zeigt sich, daß $\mu = -C$ ist. Das Inverse von p_{sj} (falls es existiert) sei h_{sj} , so daß $\sum_j h_{sj} p_{sj} = \delta_{ij}$. Dann wird

$$\sum_{i,j} h_{sj} p_{sj} \log p_{sj} - \log \sum_i P_i p_{ij} = -C \sum_i h_{si}$$

und daher

$$\sum_i P_i p_{ii} = \exp [C \sum_i h_{si} + \sum_{i,j} h_{sj} p_{sj} \log p_{sj}]$$

oder

$$P_i = \sum_j h_{ji} \exp [C \sum_j h_{sj} + \sum_{i,j} h_{sj} p_{sj} \log p_{sj}].$$

Dies ist das System von Gleichungen, mit dem die maximierenden Werte von P_i bestimmt werden, wobei C so festgelegt werden muß, daß $\sum P_i = 1$ ist. Wenn dies getan ist, werden C die Kanalkapazität und P_i die geeigneten Wahrscheinlichkeiten für die Kanalzeichen sein, um diese Kapazität zu erreichen.

Wenn jedes Eingabezeichen denselben Wahrscheinlichkeitssatz auf den Linien besitzt, die von ihm ausgehen, und das auch für jedes Ausgabezeichen gilt, kann die Kapazität leicht errechnet werden.

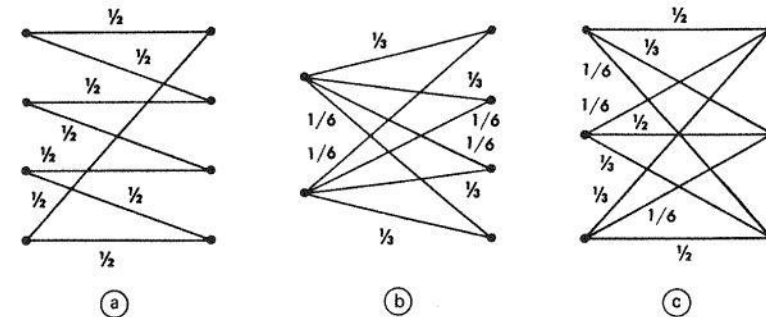


Abb. 12: Beispiel diskreter Kanäle mit den gleichen Streuwahrscheinlichkeiten für jede Eingabe und jede Ausgabe

Drei Beispiele werden in Abb. 12 gezeigt. In solch einem Fall ist $H_s(y)$ unabhängig von der Wahrscheinlichkeitsverteilung bei den Eingabezeichen und ist durch $-\sum p_i \log p_i$ gegeben, wobei die p_i die Werte der Übergangswahrscheinlichkeiten von jedem Eingabezeichen sind. Die Kanalkapazität ist

$$\begin{aligned} \text{Max } [H(y) - H_s(y)] \\ = \text{Max } H(y) + \sum p_i \log p_i. \end{aligned}$$

Das Maximum von $H(y)$ ist eindeutig $\log m$, wobei m die Anzahl von Ausgabezeichen ist, weil es möglich ist, diese alle gleich wahrscheinlich zu machen, indem man auch die Eingabezeichen gleich wahrscheinlich macht. Die Kanalkapazität ist daher

$$C = \log m + \sum p_i \log p_i.$$

In Abb. 12a wäre sie

$$C = \log 4 - \log 2 = \log 2.$$

Dies könnte erreicht werden, indem man nur das erste und das dritte Zeichen benutzt. In Abb. 12b ist

$$\begin{aligned}
 C &= \log 4 - \frac{2}{3} \log 3 - \frac{1}{3} \log 6 \\
 &= \log 4 - \log 3 - \frac{1}{3} \log 2 \\
 &= \log \frac{1}{3} 2^1.
 \end{aligned}$$

In Abb. 12c schließlich ist

$$\begin{aligned}
 C &= \log 3 - \frac{1}{2} \log 2 - \frac{1}{3} \log 3 - \frac{1}{6} \log 6 \\
 &= \log \frac{3}{2^1 3^1 6^1}.
 \end{aligned}$$

Angenommen, die Zeichen sind so in verschiedene Gruppen eingeteilt, daß eine Störung niemals die Ursache dafür ist, daß ein Zeichen einer Gruppe fälschlich für ein Zeichen aus einer anderen Gruppe gehalten wird. Wenn die Kapazität für die n -te Gruppe C_n ist (in bit pro Sekunde), wenn nur die Zeichen dieser Gruppe benutzt werden, dann ist leicht ersichtlich, daß, um alle Gruppen zusammen so gut wie möglich zu benutzen, die Gesamtwahrscheinlichkeit P_n aller Zeichen der n -ten Gruppe

$$P_n = \frac{2^{C_n}}{\sum 2^{C_n}}$$

sein muß.

Innerhalb einer Gruppe ist die Wahrscheinlichkeit geradeso verteilt, als ob diese die einzigen benutzten Zeichen wären. Die Kanalkapazität ist dafür

$$C = \log \sum 2^{C_n}.$$

17. Ein Beispiel effizienter Codierung

Das folgende Beispiel ist, obwohl es etwas künstlich ist, ein Fall, bei dem die genaue Anpassung an einen gestörten Kanal möglich ist. Es gibt zwei Kanalzeichen, 0 und 1, und die Störung beeinflusst sie in Blöcken von sieben Zeichen. Ein solcher Siebener-Block wird entweder fehlerfrei übertragen oder aber es ist genau ein Zeichen der sieben falsch. Diese acht Möglichkeiten sind gleich wahrscheinlich. Wir haben

$$\begin{aligned}
 C &= \text{Max} [H(y) - H_z(y)] \\
 &= \frac{1}{7} \left[7 + \frac{8}{8} \log \frac{1}{8} \right] \\
 &= \frac{4}{7} \text{ bit/Zeichen.}
 \end{aligned}$$

Ein effizienter Code, der die vollständige Berichtigung der Fehler

und die Übertragung mit der Rate C erlaubt, ist der folgende (der nach einer Methode von R. Hamming gefunden wurde):

Ein Block mit sieben Zeichen sei $X_1, X_2, \dots, X_7$. Von diesen sind X_3, X_5, X_6 und X_7 Nachrichtenzeichen und willkürlich durch die Quelle ausgewählt. Die anderen drei sind redundant und werden wie folgt errechnet:

X_4 wird benutzt, um $\alpha = X_1 + X_2 + X_3 + X_4$ geradzahlig zu machen, X_2 wird benutzt, um $\beta = X_2 + X_3 + X_6 + X_7$ geradzahlig zu machen, X_1 wird benutzt, um $\gamma = X_1 + X_3 + X_5 + X_7$ geradzahlig zu machen.

Wenn ein solcher Block empfangen wird, werden α, β und γ berechnet, dabei werden gerade Zahlen "Null" und ungerade Zahlen "Eins" genannt. Die Binärzahl $\alpha \beta \gamma$ ergibt dann den Index des Zeichens X_i , das falsch ist (wenn sich 0 ergibt, dann war kein Fehler vorhanden).¹¹⁾

¹¹⁾ Für einige weiteren Beispiele der selbstkorrigierenden Codes siehe M. J. E. Golay, "Notes on Digital Coding", *Proceedings of the Institute of Radio Engineers*, Bd. 37, No. 6, Juni 1949, Seite 637.

III. Kontinuierliche Information

Wir betrachten nun den Fall, in dem die Signale oder die Nachrichten oder beide kontinuierlich variabel sind, im Gegensatz zu der diskreten Art, die wir bisher angenommen haben. Bis zu einem gewissen Grad kann der kontinuierliche Fall durch einen Grenzprozeß aus dem diskreten Fall erhalten werden, indem man das Kontinuum der Nachrichten und Signale in eine große, jedoch endliche Anzahl kleiner Bereiche aufteilt und die verschiedenen Parameter auf eine diskrete Basis bezieht. Wenn die Größe der Bereiche verringert wird, nähern sich diese Parameter im allgemeinen als Grenzwerte den tatsächlichen Werten für den kontinuierlichen Fall. Es treten jedoch einige neue Effekte auf und auch eine allgemeine Änderung der Beurteilung in Richtung auf eine Spezialisierung der allgemeinen Ergebnisse für besondere Fälle.

Wir werden im kontinuierlichen Fall nicht versuchen, unsere Ergebnisse mit der größten Allgemeinheit zu erzielen oder mit der extremen Strenge der reinen Mathematik, da das einen hohen Grad an abstrakter Theorie der Messungen voraussetzen und den roten Faden der Analyse zerreißen würde. Eine frühere Studie zeigt jedoch, daß die Theorie auf eine vollständig axiomatische, strenge Weise formuliert werden kann, die beide Fälle, den kontinuierlichen und den diskreten, und sogar viele andere Fälle einschließt. Die gelegentlichen Freiheiten, die bei den Grenzwert-Bildungen in der vorliegenden Analyse in Anspruch genommen werden, sind in allen Fällen von praktischem Interesse gerechtfertigt.

18. Mengen und Ensembles von Funktionen

Im kontinuierlichen Fall haben wir mit Mengen und Ensembles von Funktionen zu tun. Eine Menge von Funktionen ist, wie der Name schon sagt, nur eine Klasse oder Sammlung von Funktionen, normalerweise von Funktionen einer Variablen, der Zeit. Sie kann beschrieben werden, indem man eine explizite Darstellung der verschiedenen Funktionen innerhalb der Menge gibt, oder implizit, indem eine Eigenschaft angegeben wird, die die Funktionen der Men-

ge auszeichnet und die andere nicht besitzen. Einige Beispiele sind:

1. Die Menge der Funktionen

$$f_{\theta}(t) = \sin(t + \theta).$$

Jeder einzelne Wert von θ bestimmt eine besondere Funktion innerhalb der Menge.

2. Die Menge aller Funktionen der Zeit, die keine Frequenzen von mehr als W Perioden pro Sekunde enthalten.

3. Die Menge aller Funktionen, die in der Bandbreite W und in der Amplitude durch A begrenzt sind.

4. Die Menge aller englischen Sprachsignale als Funktionen der Zeit.

Ein Ensemble von Funktionen ist eine Menge von Funktionen zusammen mit einem Wahrscheinlichkeitsmaß, wobei wir die Wahrscheinlichkeit einer Funktion der Menge mit gewissen Eigenschaften bestimmen können.¹⁾ Zum Beispiel mit der Menge

$$f_{\theta}(t) = \sin(t + \theta)$$

können wir eine Wahrscheinlichkeitsverteilung für θ , etwa $P(\theta)$, angeben. Die Menge wird dann zu einem Ensemble.

Einige weitere Beispiele für Ensembles von Funktionen sind:

1. Eine endliche Menge von Funktionen $f_k(t)$ ($k = 1, 2, \dots, n$), wobei f_k die Wahrscheinlichkeit p_k hat.

2. Eine endlichdimensionale Familie von Funktionen

$$f(\alpha_1, \alpha_2, \dots, \alpha_n; t)$$

mit einer Wahrscheinlichkeitsverteilung für die Parameter α_i :

$$p(\alpha_1, \dots, \alpha_n).$$

Oder wir könnten das Ensemble betrachten, das durch

$$f(\alpha_1, \dots, \alpha_n, \theta_1, \dots, \theta_n; t) = \sum_{n=1}^n \alpha_n \sin(n\omega t + \theta_n)$$

definiert wird, wobei die Amplituden α_i normal und unabhängig voneinander und die Phasen θ_i einheitlich (von 0 bis 2π) und auch unabhängig voneinander verteilt sind.

¹⁾ In der mathematischen Terminologie gehören die Funktionen zu einem Maßraum, dessen Gesamtmaß die Einheit ist.

3. Das Ensemble

$$f(a_i, t) = \sum_{n=-\infty}^{+\infty} a_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)},$$

wobei die a_i -Werte normal verteilt und unabhängig sind, alle mit derselben Standardabweichung $\sqrt{N}$. Dies ist eine Darstellung des weißen Rauschens, mit der begrenzten Bandbreite zwischen 0 und W Perioden pro Sekunde und einer mittleren Leistung von N .²⁾

4. Nehmen wir an, es seien Punkte auf der t -Achse verteilt, entsprechend einer Poisson-Verteilung. Für jeden ausgewählten Punkt wird die Funktion $f(t)$ verschoben, und die verschiedenen Funktionen werden addiert, die dann die Gruppe

$$\sum_{k=-\infty}^{\infty} f(t + t_k)$$

ergeben, wobei t_k die Punkte der Poisson-Verteilung sind. Diese Gruppe kann als eine Art Impuls- oder Punktstörung betrachtet werden, bei der alle Impulse gleich groß sind.

5. Die Menge der englischen Sprachfunktionen mit dem Wahrscheinlichkeitsmaß, das durch die Häufigkeit des Vorkommens im Normalfall gegeben ist.

Ein Ensemble von Funktionen $f_a(t)$ ist stationär, wenn sich dasselbe Ensemble ergibt, falls alle Funktionen um eine beliebige Zeitspanne verschoben werden. Das Ensemble

$$f_\theta(t) = \sin(t + \theta)$$

ist stationär, wenn θ gleichmäßig zwischen 0 und 2π verteilt ist. Wenn wir jede Funktion um t_1 verschieben, erhalten wir

$$\begin{aligned} f_\theta(t + t_1) &= \sin(t + t_1 + \theta) \\ &= \sin(t + \varphi) \end{aligned}$$

mit einer Phase φ , die einheitlich von 0 bis 2π verteilt ist. Jede Funktion hat sich geändert, jedoch das Ensemble als Ganzes ist in-

²⁾ Diese Darstellung kann als eine Definition von weißem Rauschen mit begrenzter Bandbreite benutzt werden. Sie hat gewisse Vorteile darin, daß hier weniger Grenzwertoperationen einbezogen werden, als es bei Definitionen notwendig ist, die früher benutzt worden sind. Der Name "weißes Rauschen", der bereits in der Literatur seinen festen Platz hat, ist vielleicht etwas unglücklich. In der Optik bedeutet weißes Licht entweder jedes kontinuierliche Spektrum als Gegensatz zu einem Linien-Spektrum, oder ein Spektrum, das mit zunehmender Wellenlänge abfällt (was nicht dasselbe ist wie ein Spektrum, das mit zunehmender Frequenz abfällt).

variant gegenüber der Verschiebung. Die anderen Beispiele, die oben gegeben wurden, sind ebenfalls stationär.

Ein Ensemble ist *ergodisch*, wenn es stationär ist und es keine stationäre Untermenge der Funktionen in der Menge gibt, mit einer Wahrscheinlichkeit, die verschieden ist von 0 und 1. Das Ensemble

$$\sin(t + \theta)$$

ist ergodisch. Keine Untermenge dieser Funktion mit der Wahrscheinlichkeit 0 oder 1 läßt sich bei beliebigen Zeitverschiebungen in sich selbst überführen. Andererseits ist das Ensemble

$$a \sin(t + \theta),$$

wobei a normal und θ gleichmäßig verteilt ist, zwar stationär, aber nicht ergodisch. Die Untermenge dieser Funktionen mit a zwischen 0 und 1 ist beispielsweise stationär und hat eine Wahrscheinlichkeit, die nicht gleich 0 oder 1 ist.

Von den gegebenen Beispielen sind 3 und 4 ergodisch und 5 kann vielleicht als ergodisch betrachtet werden. Wenn ein Ensemble ergodisch ist, können wir grob sagen, daß jede Funktion aus der Menge typisch für das Ensemble ist. Genauer gesagt: Bei einem ergodischen Ensemble ist der Durchschnitt einer beliebigen Statistik über das Ensemble gleich (mit der Wahrscheinlichkeit 1) dem Durchschnitt über alle Zeitverschiebungen einer bestimmten Funktion aus dem Ensemble.³⁾ Vereinfacht ausgedrückt: Von jeder Funktion kann erwartet werden, daß sie, über eine bestimmte Zeit, aus einer Faltung mit einer beliebigen Funktion aus diesem Ensemble mit ihrer eigenen Frequenz hervorgeht.

Genauso, wie wir verschiedene Operationen an Zahlen oder Funktionen vornehmen können, um neue Zahlen oder Funktionen zu erhalten, können wir auch Operationen an Ensembles durchführen, um neue Ensembles zu erhalten. Angenommen, wir haben ein En-

³⁾ Dies ist der berühmte Ergoden-Lerhsatz, oder eher ein Aspekt dieses Lehrsatzes, der etwas anders formuliert von Birkhoff bewiesen wurde sowie durch Neumann und Koopman und anschließend verallgemeinert von Wiener, Hopf, Hurewicz und anderen. Die Literatur über die Ergodentheorie ist ziemlich umfangreich, und der Leser wird auf die Schriften dieser Autoren hingewiesen, wenn er genaue und allgemeine Formulierungen erwartet; z. B. E. Hopf "Ergodentheorie", *Ergebnisse der Mathematik und ihrer Grenzgebiete*, Bd. 5; "On Causality Statistics and Probability", *Journal of Mathematics and Physics*, Bd. XIII, No. 1, 1934; N. Wiener "The Ergodic Theorem", *Duke Mathematical Journal*, Bd. 5, 1939.

semble von Funktionen $f_a(t)$ und einen Operator T , der aus jeder Funktion $f_a(t)$ eine Funktion $g_a(t)$ erzeugt:

$$g_a(t) = Tf_a(t).$$

Das Wahrscheinlichkeitsmaß ist für die Menge $g_a(t)$ durch jenes für das Ensemble $f_a(t)$ definiert. Die Wahrscheinlichkeit für eine bestimmte Untergruppe der $g_a(t)$ -Funktionen ist gleich derjenigen für die Untergruppe der $f_a(t)$ -Funktionen, die die betreffenden g -Funktionen unter der Operation T erzeugen. Physikalisch bedeutet das, daß das Ensemble von Funktionen irgendeine technische Einrichtung passiert, z. B. einen Filter, einen Gleichrichter oder einen Modulator. Die Ausgabefunktionen der Einrichtung erzeugen das Ensemble $g_a(t)$.

Eine solche Einrichtung oder ein Operator T wird invariant genannt, wenn, falls die Eingabephasen verschoben werden, auch die Ausgabe nur eine Verschiebung erfährt, d. h. wenn die Gleichung

$$g_a(t) = Tf_a(t)$$

die folgende Gleichung mit einbezieht:

$$g_a(t + t_1) = Tf_a(t + t_1),$$

und zwar für jedes $f_a(t)$ und jedes t_1 . Es ist leicht ersichtlich (siehe Anhang 5), daß, wenn T invariant und das Eingabeensemble stationär ist, auch das Ausgabeensemble stationär ist. Ebenso wird, wenn die Eingabe ergodisch ist, auch die Ausgabe ergodisch sein.

Ein Filter oder ein Gleichrichter ist gegenüber allen Zeitverschiebungen invariant. Die Operation der Modulatoren ist es nicht, da die Trägerphase eine gewisse Zeitstruktur vorgibt. Allerdings ist die Modulation für Zeitverschiebungen, die ein Vielfaches der Periode der Trägerfunktion sind, invariant.

Wiener hat die enge Beziehung zwischen der Invarianz physikalischer Einrichtungen unter Zeitverschiebung und der Fourier-Theorie hervorgehoben.⁴⁾ Er hat tatsächlich gezeigt, daß, wenn eine Einrich-

⁴⁾ Die Kommunikationstheorie verdankt Wiener einen Großteil ihrer fundamentalen Philosophie und Theorie. Sein klassischer NDRC-Report, *The Interpolation, Extrapolation, and Smoothing of Stationary Time Series* (Wiley, 1949), enthält die erste klare Formulierung der Kommunikationstheorie als ein statistisches Problem, eine Studie über Operationen an Zeitfolgen. Dieses Werk, obwohl es sich hauptsächlich mit der linearen Vorhersage und mit Filterproblemen beschäftigt, ist eine wichtige zusätzliche Beziehung in Verbindung mit diesem Buch. Wir können auch auf Wieners *Cybernetics* (Wiley, 1948) Bezug nehmen, das die allgemeinen Probleme von Kommunikation und Regelung behandelt.



tung sowohl linear als auch invariant ist, die Fourier-Analyse dann das richtige mathematische Werkzeug ist, um das Problem zu behandeln.

Ein Ensemble von Funktionen ist die geeignete mathematische Darstellung der Nachrichten, die von einer kontinuierlichen Quelle (z. B. bei der Sprache) erzeugt werden, der von einem Sender gelieferten Signale und der Störungen. Die Kommunikationstheorie behandelt eigentlich, wie von Wiener betont worden ist, nicht die Operationen an speziellen Funktionen, sondern die Operationen an Ensembles von Funktionen. Ein Kommunikationssystem wird nicht für eine einzelne Sprachfunktion und noch weniger für eine Sinusfunktion entworfen, sondern für das Ensemble der Sprachfunktionen.

19. Ensembles von Funktionen begrenzter Bandbreite

Wenn eine Funktion der Zeit $f(t)$ eine Bandbreite von 0 bis W Perioden pro Sekunde besitzt, so ist sie durch die im folgenden Lehrsatz beschriebene Angabe ihrer Ordinaten als eine Reihe aus diskreten Punkten im Abstand von $\frac{1}{2W}$ Sekunden vollständig bestimmt.⁵⁾

Lehrsatz 13: Wenn $f(t)$ keine Frequenzen größer als W enthält, dann ist

$$f(t) = \sum_{-\infty}^{\infty} X_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

mit

$$X_n = f\left(\frac{n}{2W}\right).$$

In dieser Entwicklung ist $f(t)$ als eine Summe von orthogonalen Funktionen beschrieben. Die Koeffizienten X_n der verschiedenen Glieder können als Koordinaten in einem unendlich dimensional "Funktionenraum" betrachtet werden. In diesem Raum entspricht jede Funktion genau einem Punkt und jeder Punkt einer Funktion.

Eine Funktion kann nachweislich als auf eine Zeit T begrenzt betrachtet werden, wenn alle Ordinaten X_n außerhalb dieser Zeitspanne Null sind. In diesem Fall werden alle bis auf $2TW$ der Koordinaten Null sein. Daher entsprechen Funktionen, die durch eine

⁵⁾ Zum Beweis dieses Lehrsatzes und für weitere Diskussionen siehe die Abhandlung des Autors "Communication in the Presence of Noise", die publiziert wurde in den *Proceedings of the Institute of Radio Engineers*, Bd. 37, No. 1, Jan. 1949, Seite 10-21.

Bandbreite W und eine Zeitdauer T begrenzt sind, den Punkten in einem Raum von $2TW$ Dimensionen.

Eine Untermenge der Funktionen der Bandbreite W und Dauer T entspricht einem Gebiet in diesem Raum. Z.B. entsprechen die Funktionen, deren Gesamtenergie kleiner oder gleich E ist, den Punkten in einer $2TW$ -dimensionalen Kugel mit dem Radius $r = \sqrt{2WE}$.

Ein Ensemble von Funktionen mit begrenzter Dauer und Bandbreite wird durch eine Wahrscheinlichkeitsverteilung $p(x_1, \dots, x_n)$ in dem entsprechenden n -dimensionalen Raum dargestellt. Wenn das Ensemble nicht zeitbegrenzt ist, können wir sagen, daß die $2TW$ -Koordinaten in einem gegebenen Zeitintervall T im wesentlichen den Teil der Funktion innerhalb T darstellen und die Wahrscheinlichkeitsverteilung $p(x_1, \dots, x_n)$ die statistische Struktur des Ensembles für Intervalle jener Dauer angibt.

20. Entropie einer kontinuierlichen Verteilung

Die Entropie für einen Satz diskreter Wahrscheinlichkeiten $p_1, \dots, p_n$ ist definiert worden als

$$H = - \sum p_i \log p_i.$$

In einer entsprechenden Weise definieren wir die Entropie einer kontinuierlichen Verteilung mit der Dichte-Verteilungsfunktion $p(x)$ als

$$H = - \int_{-\infty}^{\infty} p(x) \log p(x) dx.$$

Mit einer n -dimensionalen Verteilung $p(x_1, \dots, x_n)$ haben wir

$$H = - \int \dots \int p(x_1 \dots x_n) \log p(x_1, \dots, x_n) dx_1 \dots dx_n.$$

Wenn wir zwei Argumente x und y (die selbst vieldimensional sein können) haben, sind die Gesamtentropien und die bedingten Entropien von $p(x, y)$ gegeben durch

$$H(x, y) = - \iint p(x, y) \log p(x, y) dx dy$$

und

$$H_x(y) = - \iint p(x, y) \log \frac{p(x, y)}{p(x)} dx dy$$

$$H_y(x) = - \iint p(x, y) \log \frac{p(x, y)}{p(y)} dx dy$$

mit

$$p(x) = \int p(x, y) dy$$

$$p(y) = \int p(x, y) dx.$$

Die Entropien kontinuierlicher Verteilungen haben die meisten (jedoch nicht alle) Eigenschaften des diskreten Falles. Im besonderen gilt folgendes:

1. Wenn x auf ein bestimmtes Volumen v in seinem Raum begrenzt ist, so ist $H(x)$ dann ein Maximum und gleich dem $\log v$, wenn $p(x)$ konstant den Wert $\left(\frac{1}{v}\right)$ in dem Volumen hat.

2. Mit zwei beliebigen Variablen x und y wird

$$H(x, y) \leq H(x) + H(y)$$

bei Gleichheit, wenn (und nur dann, wenn) x und y unabhängig voneinander sind, d.h. $p(x, y) = p(x)p(y)$ (abgesehen von einer möglichen Menge von Punkten mit der Wahrscheinlichkeit Null).

3. Man betrachte eine verallgemeinerte Mittelwertoperation der folgenden Art

$$p'(y) = \int a(x, y) p(x) dx$$

mit

$$\int a(x, y) dx = \int a(x, y) dy = 1, \quad a(x, y) \geq 0.$$

Dann ist die Entropie der Durchschnittsverteilung $p'(y)$ gleich oder größer als jene der ursprünglichen Verteilung $p(x)$.

4. Es gilt

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$

und

$$H_x(y) \leq H(y).$$

5. Nehmen wir an, $p(x)$ sei eine eindimensionale Verteilung. Für eine maximale Entropie ergibt sich eine Gaußsche Form für $p(x)$, wenn die Standardabweichung von x auf den Wert σ beschränkt ist.

Um dies zu zeigen, müssen wir

$$H(x) = - \int p(x) \log p(x) dx$$

maximieren mit

$$\sigma^2 = \int p(x) x^2 dx \quad \text{und} \quad 1 = \int p(x) dx$$

als Einschränkungen. Dies erfordert nach der Variationsrechnung die Maximierung von

$$\int [-p(x) \log p(x) + \lambda p(x)x^2 + \mu p(x)] dx.$$

Als Bedingung dafür gilt

$$-1 - \log p(x) + \lambda x^2 + \mu = 0$$

und daraus folgt (indem man die Konstanten anpaßt, um den Beschränkungen zu genügen)

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x^2/2\sigma^2)}.$$

Auf gleiche Weise lautet das für n Dimensionen, wenn die Momente zweiter Ordnung von $p(x_1, \dots, x_n)$ durch A_{ij} festgesetzt sind:

$$A_{ij} = \int \dots \int x_i x_j p(x_1, \dots, x_n) dx_1 \dots dx_n.$$

Dann ergibt sich die maximale Entropie (durch eine gleichartige Berechnung), wenn $p(x_1, \dots, x_n)$ die n -dimensionale Gaußsche Verteilung mit den Momenten zweiter Ordnung A_{ij} ist.

6. Die Entropie einer eindimensionalen Gaußschen Verteilung mit der Standardabweichung σ ist gegeben durch

$$H(x) = \log \sqrt{2\pi e} \sigma.$$

Dies wurde wie folgt berechnet:

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x^2/2\sigma^2)} \\ -\log p(x) &= \log \sqrt{2\pi}\sigma + \frac{x^2}{2\sigma^2} \\ H(x) &= -\int p(x) \log p(x) dx \\ &= \int p(x) \log \sqrt{2\pi}\sigma dx + \int p(x) \frac{x^2}{2\sigma^2} dx \\ &= \log \sqrt{2\pi}\sigma + \frac{\sigma^2}{2\sigma^2} \\ &= \log \sqrt{2\pi}\sigma + \log \sqrt{e} \\ &= \log \sqrt{2\pi e} \sigma. \end{aligned}$$

In ähnlicher Art ist die n -dimensionale Gaußsche Verteilung mit der zugehörigen quadratischen Form a_{ij} gegeben durch

$$p(x_1, \dots, x_n) = \frac{|a_{ij}|^{1/2}}{(2\pi)^{n/2}} \exp \left(-\frac{1}{2} \sum a_{ij} X_i X_j \right)$$

und die Entropie kann berechnet werden:

$$H = \log (2\pi e)^{n/2} |a_{ij}|^{-1/2},$$

wobei $|a_{ij}|$ die Determinante mit den Elementen a_{ij} ist.

7. Wenn x auf eine Halbachse ($p(x) = 0$ für $x \leq 0$) beschränkt ist und das erste Moment von x auf a festgelegt ist

$$a = \int_0^\infty p(x) x dx,$$

dann erscheint die maximale Entropie, wenn

$$p(x) = \frac{1}{a} e^{-(x/a)}$$

und ist gleich dem $\log ea$.

8. Es besteht nun ein wichtiger Unterschied zwischen den kontinuierlichen und den diskreten Entropien. Im diskreten Fall mißt die Entropie auf eine *absolute* Art die Zufälligkeit der Zufallsvariablen. Im kontinuierlichen Fall ist das Maß *relativ zum Koordinatensystem*. Wenn wir die Koordinaten verändern, wird sich die Entropie im allgemeinen auch verändern. Tatsächlich ist die neue Entropie, wenn wir zu Koordinaten $y_1 \dots y_n$ übergehen, gegeben durch

$$\begin{aligned} H(y) &= \int \dots \int p(x_1 \dots x_n) J\left(\frac{x}{y}\right) \\ &\quad \log p(x_1 \dots x_n) J\left(\frac{x}{y}\right) dy_1 \dots dy_n, \end{aligned}$$

wobei $J\left(\frac{x}{y}\right)$ die Jakobi-Determinante der Koordinatentransformation ist. Wenn wir den Logarithmus entwickeln und die Variablen umbenennen in $x_1 \dots x_n$, dann erhalten wir

$$H(y) = H(x) - \int \dots \int p(x_1, \dots, x_n) \log J\left(\frac{x}{y}\right) dx_1 \dots dx_n.$$

Dadurch ist die neue Entropie gleich der alten Entropie minus dem erwarteten Logarithmus der Jakobi-Determinante. Im kontinuierlichen Fall kann die Entropie als ein Maß der Zufälligkeit *relativ zu einem angenommenen Standard* angesehen werden, nämlich zu dem Koordinatensystem, das gewählt wurde, wobei jedes kleine Volumenelement $dx_1 \dots dx_n$ gleiches Gewicht hat. Wenn wir das Koordinatensystem ändern, mißt die Entropie in dem neuen System die Zufälligkeit, wenn gleichen Volumenelementen $dy_1 \dots dy_n$ in dem neuen System gleiches Gewicht gegeben ist.

Trotz dieser Abhängigkeit vom Koordinatensystem ist der Begriff der Entropie im kontinuierlichen Fall genauso wichtig wie im diskreten Fall. Dies ist darauf zurückzuführen, daß die hergeleiteten Begriffe der Informationsrate und der Kanalkapazität aus der *Differenz* zweier Entropien gebildet werden und diese Differenz *nicht* vom Koordinatensystem abhängt, da sich jedes der beiden Glieder um denselben Betrag ändert.

Die Entropie einer kontinuierlichen Verteilung kann negativ sein. Der Maßstab setzt einen willkürlichen Nullpunkt fest, dem eine einheitliche Verteilung über ein Einheitsvolumen entspricht. Eine Verteilung, die mehr als diese beschränkt ist, hat eine geringere Entropie und muß negativ sein. Die Raten und Kapazitäten werden jedoch immer nicht-negativ sein.

9. Ein spezieller Fall einer Veränderung der Koordinaten ist die lineare Transformation

$$y_i = \sum_j a_{ij} x_j.$$

In diesem Fall ist die Jakobi-Determinante einfach $|a_{ij}|^{-1}$ und

$$H(y) = H(x) + \log |a_{ij}|.$$

Im Fall einer Drehung der Koordinaten (oder jeder das Maß erhaltenden Transformation) ist $J = 1$ und $H(y) = H(x)$.

21. Entropie eines Ensembles von Funktionen

Nehmen wir ein ergodisches Ensemble von Funktionen an, die auf eine bestimmte Bandbreite von W Perioden pro Sekunde begrenzt sind. Es sei

$$p(x_1, \dots, x_n)$$

die Funktion der Dichteverteilung für die Amplituden $x_1 \dots x_n$ bei n aufeinanderfolgenden Probepunkten. Wir definieren die Entropie des Ensembles pro Freiheitsgrad mit

$$H' = - \lim_{n \rightarrow \infty} \frac{1}{n} \int \dots \int p(x_1, \dots, x_n) \log p(x_1, \dots, x_n) dx_1 \dots dx_n.$$

Wir können auch eine Entropie H pro Sekunde dadurch definieren, daß wir nicht durch n , sondern durch die Zeit T in Sekunden für n Abtastwerte dividieren. Da $n = 2TW$ ist, wird $H = 2WH'$.

Für weißes thermisches Rauschen entspricht p einer Gaußschen

Verteilungskurve und wir haben

$$H' = \log \sqrt{2\pi e N},$$

$$H = W \log 2\pi e N.$$

Für eine gegebene Durchschnittsleistung N hat das weiße Rauschen die maximal mögliche Entropie. Dies ergibt sich aus den maximierenden Eigenschaften der oben erwähnten Gaußschen Verteilung.

Die Entropie für einen kontinuierlichen stochastischen Prozeß hat viele Eigenschaften, die analog sind zu jenen für diskrete Prozesse. Im diskreten Fall war die Entropie verbunden mit dem Logarithmus der *Wahrscheinlichkeit* langer Folgen und mit der *Anzahl* von einigermaßen wahrscheinlichen Folgen großer Länge. Im kontinuierlichen Fall steht sie auf ähnliche Art in Beziehung zu dem Logarithmus der *Wahrscheinlichkeitsdichte* einer langen Folge von Probenwerten und dem *Volumen* einer einigermaßen hohen Wahrscheinlichkeit im Funktionenraum. Genauer gesagt, wenn wir $p(x_1, \dots, x_n)$ kontinuierlich in den x_i für jedes n annehmen, dann ist für genügend große n

$$\left| \frac{\log p}{n} - H' \right| < \epsilon$$

für alle möglichen $(x_1, \dots, x_n)$, abgesehen von einer Menge, deren Gesamtwahrscheinlichkeit kleiner als δ ist, wobei δ und ϵ beliebig klein sind. Dies ergibt sich aus dem ergodischen Verhalten, wenn wir den Raum in eine große Anzahl kleiner Elemente aufteilen.

Die Beziehung zwischen H und dem Volumen kann wie folgt beschrieben werden: Unter den gleichen Voraussetzungen betrachte man den n -dimensionalen Raum, der der Verteilung $p(x_1, \dots, x_n)$ entspricht. $V_n(q)$ sei das kleinste Volumen in diesem Raum, das in seinem Innern eine Gesamtwahrscheinlichkeit q enthält.

Dann ist

$$\lim_{n \rightarrow \infty} \frac{\log V_n(q)}{n} = H',$$

vorausgesetzt, q ist nicht gleich 0 oder 1.

Diese Ergebnisse zeigen, daß für große n ein ziemlich gut definiertes Volumen (zumindest im logarithmischen Sinn) von hoher Wahrscheinlichkeit existiert und daß innerhalb dieses Volumens die Wahrscheinlichkeitsdichte verhältnismäßig einheitlich ist (wiederum im logarithmischen Sinn).

Im Fall des weißen Rauschens ist die Verteilungsfunktion gegeben durch

$$p(x_1, \dots, x_n) = \frac{1}{(2\pi N)^{n/2}} \exp - \frac{1}{2N} \sum x_i^2.$$

Da dies nur von $\sum x_i^2$ abhängig ist, sind die Flächen gleicher Wahrscheinlichkeitsdichte Kugelflächen und die gesamte Verteilung hat eine kugelförmige Symmetrie. Der Bereich hoher Wahrscheinlichkeit ist eine Kugel mit dem Radius $\sqrt{nN}$. Für $n \rightarrow \infty$ nähert sich die Wahrscheinlichkeit dafür, daß man sich außerhalb einer Kugel mit dem Radius $\sqrt{n(N + \epsilon)}$ befindet, dem Wert Null, unabhängig davon, wie klein ϵ ist, und $\frac{1}{n}$ mal dem Logarithmus vom Volumen der Kugel nähert sich $\log \sqrt{2\pi e N}$.

Im kontinuierlichen Fall ist es vorteilhaft, nicht mit der Entropie H eines Ensembles zu arbeiten, sondern mit einer abgeleiteten Größe, die wir die *Entropieleistung* nennen. Sie wird definiert als die Leistung eines weißen Rauschens, das mit derselben Bandbreite begrenzt ist wie das ursprüngliche Ensemble und das dieselbe Entropie besitzt. In anderen Worten, wenn H' die Entropie eines Ensembles ist, ist seine Entropieleistung

$$N_1 = \frac{1}{2\pi e} \exp 2H'.$$

In der geometrischen Anschauung bedeutet dies, daß das Volumen hoher Wahrscheinlichkeit durch das Quadrat des Radius einer Kugel mit dem gleichen Volumen gemessen wird. Da das weiße Rauschen die maximale Entropie für eine gegebene Leistung besitzt, ist die Entropieleistung einer beliebigen Störung kleiner oder gleich ihrer tatsächlichen Leistung.

22. Entropieverluste in linearen Filtern

Lehrsatz 14: Wenn ein Ensemble, das pro Freiheitsgrad innerhalb der Bandbreite W eine Entropie H_1 besitzt, durch einen Filter mit der Charakteristik $Y(f)$ geschickt wird, hat das Ausgabeensemble eine Entropie

$$H_2 = H_1 + \frac{1}{W} \int_W \log |Y(f)|^2 df.$$

Die Operation des Filters ist im wesentlichen eine lineare Koordinatentransformation. Wenn wir uns an die verschiedenen Frequenz-

komponenten als das ursprüngliche Koordinatensystem erinnern, sind die neuen Frequenzkomponenten nur die alten, die mit bestimmten Faktoren multipliziert wurden. Die Matrix der Koordinatentransformation hat daher im wesentlichen eine Diagonalform für diese Koordinaten. Die Jakobi-Determinante dieser Transformation ist (für n Sinus- und n Cosinuskomponenten)

$$J = \prod_{i=1}^n |Y(f_i)|^2 = \exp \sum \log |Y(f_i)|^2,$$

wobei die f_i gleichmäßig über die Bandbreite W verteilt sind. Dies wird im Grenzfall

$$\exp \frac{1}{W} \int_W \log |Y(f)|^2 df.$$

Da J konstant ist, hat sein Durchschnittswert dieselbe Größe, und wendet man den Lehrsatz über die Änderung der Entropie bei einer Änderung der Koordinaten an, dann folgt dieses Ergebnis.

Wir können es auch im Sinne der Entropieleistung ausdrücken. Dann ergibt sich die Entropieleistung des zweiten Ensembles, wenn die des ersten Ensembles N_1 ist, zu

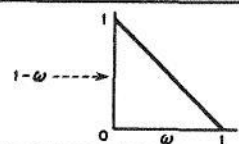
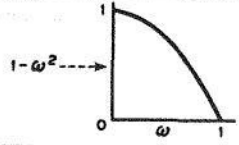
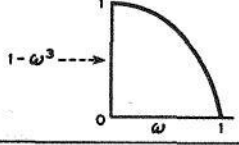
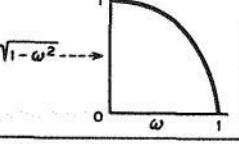
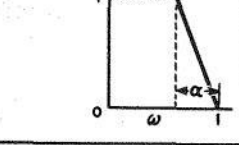
$$N_1 \exp \frac{1}{W} \int_W \log |Y(f)|^2 df.$$

Die endgültige Entropieleistung ist die anfängliche Entropieleistung, multipliziert mit der geometrisch gemittelten Verstärkung durch das Filter. Wenn die Verstärkung in db gemessen wird, dann wird die Ausgabe-Entropieleistung um die arithmetisch gemittelte db -Verstärkung dividiert durch W vergrößert.

In Tabelle I ist der Verlust an Entropieleistung für eine Anzahl von idealisierten Verstärkungs-Charakteristiken errechnet worden (und auch in db ausgedrückt). Auch die Impuls-Durchlaßkurven dieser Filter sind für $W = 2\pi$ gegeben, mit einer Phase, die als Null angenommen wurde.

Der Verlust an Entropie für viele andere Fälle kann aus diesen Ergebnissen abgeleitet werden. Zum Beispiel ist der Entropie-Leistungsfaktor $\frac{1}{e^2}$ des ersten Falles auch auf jede beliebige Verstärkungscharakteristik anzuwenden, die von $1 - \omega$ durch eine das Maß erhaltende Transformation der ω -Achse erzielt wird. Insbesondere hat eine linear anwachsende Verstärkung $G(\omega) = \omega$ oder eine Sägezahn-Charakteristik zwischen 0 und 1 denselben Entropieverlust.

Tabelle I

Verstärkungsfaktor	Faktor der Entropieleistung	Verstärkung der Entropieleistung in Dezibel	Impuls-Durchlaßkurve
	$\frac{1}{e^2}$	-8,68	$\frac{\sin^2 \pi t}{(\pi t)^2}$
	$\left(\frac{2}{e}\right)^4$	-5,32	$2 \left[\frac{\sin t}{t^3} - \frac{\cos t}{t^2} \right]$
	0,384	-4,15	$6 \left[\frac{\cos t - 1}{t^4} - \frac{\cos t}{2t^2} + \frac{\sin t}{t^3} \right]$
	$\left(\frac{2}{e}\right)^2$	-2,66	$\frac{\pi}{2} \frac{J_1(t)}{t}$
	$\frac{1}{e^{2\alpha}}$	-8,68 α	$\frac{1}{\alpha t^2} \left[\cos(1-\alpha)t - \cos t \right]$

Die reziproke Verstärkung hat den reziproken Faktor. Dadurch ergibt sich bei $\frac{1}{\omega}$ der Faktor e^2 . Eine Erhöhung der Verstärkung für irgendeine Leistung erhöht auch den Faktor dieser Leistung.

23. Die Entropie der Summe zweier Ensembles

Wenn wir zwei Ensembles von Funktionen $f_a(t)$ und $g_s(t)$ haben, können wir ein neues Ensemble durch "Addition" bilden. Angenommen, das erste Ensemble habe die Wahrscheinlichkeitsdichte-Funk-

tion $p(x_1, \dots, x_n)$ und das zweite $q(x_1, \dots, x_n)$. Dann ist die Dichtefunktion für die Summe gegeben durch die Faltung der beiden Funktionen:

$$r(x_1, \dots, x_n) = \int \dots \int p(y_1, \dots, y_n) \cdot q(x_1 - y_1, \dots, x_n - y_n) dy_1 dy_2 \dots dy_n.$$

Physikalisch bedeutet das, daß die Störungen oder die Signale, die von den ursprünglichen Ensembles von Funktionen dargestellt werden, addiert werden.

Das folgende Ergebnis ist in Anhang 6 abgeleitet.

Lehrsatz 15: N_1 und N_2 seien die durchschnittlichen Leistungen von zwei Ensembles und ihre Entropieleistungen seien $\bar{N}_1$ und $\bar{N}_2$. Dann ist die Entropieleistung der Summe, $\bar{N}_3$, beschränkt durch:

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2.$$

Das weiße (Gaußsche) Rauschen hat die besondere Eigenschaft, daß es jedes andere Rauschen oder jedes Ensemble von Signalen, die zu ihm addiert werden, absorbiert, mit einer resultierenden Entropieleistung, die näherungsweise gleich der Summe der Leistungen des weißen Rauschens und des Signals ist (ausgehend vom Durchschnittswert des Signals, der normalerweise Null ist), vorausgesetzt, daß die Leistung des Signals in gewisser Hinsicht gering ist, verglichen mit dem Rauschen.

Betrachten wir den zugehörigen Funktionenraum dieses Ensembles mit n Dimensionen. Das weiße Rauschen entspricht der kugelförmigen Gaußschen Verteilung der Wahrscheinlichkeitsdichte in diesem Raum. Das Signalensemble entspricht einer anderen Wahrscheinlichkeitsverteilung, nicht unbedingt der Gaußschen oder kugelförmigen. Die zweiten Momente dieser Verteilung über ihrem Schwerpunkt seien a_{ij} . Das bedeutet, wenn $p(x_1, \dots, x_n)$ die Funktion der Dichteverteilung ist, daß

$$a_{ij} = \int \dots \int p(x_i - \alpha_i) (x_j - \alpha_j) dx_1 \dots dx_n,$$

wobei die α_i die Koordinaten des Schwerpunkts sind. Nun hat a_{ij} eine positiv definite quadratische Form, und wir können unser Koordinatensystem so drehen, daß es mit den Hauptachsen dieser Form übereinstimmt. a_{ij} ist dann auf die diagonale Form b_{ii} reduziert.

Wir verlangen, daß jedes b_{ii} klein ist gegenüber dem Wert N , dem Quadrat des Radius der sphärischen Verteilung.

In diesem Fall ergibt das Faltungsintegral über Störung und Signal eine annähernd Gaußsche Verteilung, deren entsprechende quadratische Form

$$N + b_{ii} \text{ ist.}$$

Die Entropieleistung dieser Verteilung ist

$$[\Pi(N + b_{ii})]^{1/n}$$

oder näherungsweise

$$\begin{aligned} &= [(N)^n + \sum b_{ii}(N)^{n-1}]^{1/n} \\ &= N + \frac{1}{n} \sum b_{ii}. \end{aligned}$$

Der letzte Ausdruck ist die Signalleistung, während der erste die Leistung der Störung ist.

IV. Der kontinuierliche Kanal

24. Die Kapazität eines kontinuierlichen Kanals

In einem kontinuierlichen Kanal werden die Eingabe- oder die übertragenen Signale stetige Funktionen der Zeit $f(t)$ sein, die zu einer bestimmten Menge gehören, und die Ausgabe- oder die empfangenen Signale werden gestörte Formen dieser Funktionen sein. Wir werden nur den einen Fall behandeln, bei dem beide Signale, die gesendeten und die empfangenen, mit einer bestimmten Bandbreite W begrenzt sind. Sie selbst können dann für eine Zeit T durch $2TW$ Zahlenwerte beschrieben werden und ihre statistische Struktur durch endlich dimensionale Verteilungsfunktionen. Daher wird die Statistik der gesendeten Signale bestimmt werden durch

$$P(x_1, \dots, x_n) = P(x)$$

und jene der Störung durch die bedingte Wahrscheinlichkeitsverteilung

$$P_{x_1, \dots, x_n}(y_1, \dots, y_n) = P_x(y).$$

Die Übertragungsrate für die Information über einen kontinuierlichen Kanal ist auf eine Art definiert, die analog zu jener für einen diskreten Kanal ist, nämlich

$$R = H(x) - H_y(x),$$

wobei $H(x)$ die Entropie der Eingabe und $H_y(x)$ die Äquivokation darstellt. Die Kanalkapazität C ist als Maximalwert von R definiert, wenn wir die Eingabe über alle möglichen Gruppen variieren. Das bedeutet, daß wir in einer endlichdimensionalen Näherung $P(x) = P(x_1, \dots, x_n)$ variieren müssen und von

$$- \int P(x) \log P(x) dx + \iint P(x, y) \log \frac{P(x, y)}{P(y)} dx dy$$

das Maximum ermitteln.

Dies kann geschrieben werden als

$$\iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy,$$

wobei die Tatsache genutzt wird, daß

$$\iint P(x, y) \log P(x) dx dy = \int P(x) \log P(x) dx.$$

Die Kanalkapazität wird nun wie folgt ausgedrückt:

$$C = \lim_{T \rightarrow \infty} \max_{P(x)} \frac{1}{T} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy.$$

Es ist bei dieser Form offenkundig, daß R und C vom Koordinatensystem unabhängig sind, da der Zähler und der Nenner in $\log \frac{P(x, y)}{P(x)P(y)}$ mit den gleichen Faktoren multipliziert wird, wenn x und y auf eine beliebige Weise eins-zu-eins transformiert werden. Dieser Integralausdruck für C ist allgemeiner als $H(x) - H_y(x)$. Richtig interpretiert (siehe Anhang 7), wird er immer existieren, während $H(x) - H_y(x)$ in einigen Fällen eine unbestimmte Form $\infty - \infty$ annehmen kann. Dies kommt zum Beispiel vor, wenn x auf eine Fläche beschränkt ist, die weniger Dimensionen als n aufweist, bei einer n -dimensionalen Näherung.

Wenn die logarithmische Basis, die bei der Errechnung von $H(x)$ und $H_y(x)$ benutzt wird, zwei ist, dann ist C die maximale Anzahl von Binärziffern, die pro Sekunde über den Kanal gesendet werden können, mit einer beliebig kleinen Äquivokation, genau wie im diskreten Fall. Physikalisch gesehen bedeutet das, daß man den Signalraum in eine große Anzahl von genügend kleinen Zellen aufteilt, so daß die Wahrscheinlichkeitsdichte $P_x(y)$, daß das Signal x in den Punkt y gestreut wird, innerhalb einer Zelle weitgehend konstant ist (entweder für x oder für y). Wenn die Zellen als einzelne Punkte angesehen werden, ist diese Situation im wesentlichen dieselbe wie bei einem diskreten Kanal, und die dort benutzten Beweise dürfen verwendet werden. Es ist jedoch physikalisch einleuchtend, daß diese "Quantelung" des Volumens in einzelne Punkte in irgendeiner praktischen Situation das endgültige Ergebnis nicht wesentlich verändern wird, vorausgesetzt, die Bereiche sind genügend klein. Auf diese Weise wird die eigentliche Kapazität als Grenzwert der Kapazitäten für die diskreten Unterteilungen angesehen, und das ist genau die oben definierte kontinuierliche Kapazität.

Von der Mathematik her kann nun zuerst gezeigt werden (siehe Anhang 7), daß, wenn u die Nachricht ist, x das Signal, y das empfangene Signal (gestört durch Rauschen) und v die daraus gewonnene Nachricht, dann ist

$$H(x) - H_y(x) \geq H(u) - H_v(u),$$

gleichgültig, welche Operationen an u durchgeführt werden, um x zu erhalten, oder an y , um v zu erhalten. Deshalb wird die diskrete Übertragungsrate für die Binärziffern nicht die von uns definierte Kanalkapazität übersteigen, wobei es gleichgültig ist, wie wir die Binärziffern codieren, um das Signal zu erzeugen, oder wie wir das empfangene Signal decodieren, um die Nachricht wiederzugewinnen. Andererseits ist es unter sehr allgemeinen Bedingungen möglich, ein Codiersystem zu finden, um Binärziffern mit der Rate C und einer beliebig geringen Äquivokation oder Fehlerhäufigkeit zu übertragen. Das ist zum Beispiel dann der Fall, wenn zusammen mit der Näherung eines endlich-dimensionalen Raumes zur Beschreibung der Signalfunktionen $P(x, y)$ in beiden Größen x und y stetig ist, außer in einer Menge von Punkten, deren Wahrscheinlichkeit Null ist.

Ein wichtiger Spezialfall ergibt sich, wenn die Störung zu dem Signal addiert wird und unabhängig von ihm ist (im Sinne der Wahrscheinlichkeit). Dann ist $P_x(y)$ nur eine Funktion der (Vektor-) Differenz $n = (y - x)$:

$$P_x(y) = Q(y - x),$$

und wir können der Störung eine bestimmte Entropie zuweisen (unabhängig von der Statistik des Signals), nämlich die Entropie der Verteilung $Q(n)$. Diese Entropie wird durch $H(n)$ gekennzeichnet.

Lehrsatz 16: Wenn Signal und Störung unabhängig voneinander sind und das empfangene Signal die Summe des gesendeten Signals und der Störung ist, dann ist die Übertragungsrate

$$R = H(y) - H(n),$$

d. h. die Entropie des empfangenen Signals minus der Entropie der Störung. Die Kanalkapazität ist

$$C = \max_{P(x)} H(y) - H(n).$$

Da $y = x + n$ ist, wird

$$H(x, y) = H(x, n).$$

Schreibt man die linke Seite um und benutzt man die Tatsache, daß x und n unabhängig voneinander sind, dann ist

$$H(y) + H_y(x) = H(x) + H(n).$$

Daraus folgt

$$R = H(x) - H_y(x) = H(y) - H(n).$$

Da $H(n)$ von $P(x)$ unabhängig ist, erfordert die Maximierung von R die Maximierung von $H(y)$, der Entropie des empfangenen Signals. Wenn gewisse Beschränkungen für das Ensemble der übertragenen Signale gegeben sind, muß die Entropie des empfangenen Signals in Übereinstimmung mit diesen Beschränkungen maximiert werden.

25. Die Kanalkapazität bei einer Begrenzung der Durchschnittsleistung

Eine einfache Anwendung von Lehrsatz 16 erhält man, wenn die Störung ein weißes thermisches Rauschen ist und die übertragenen Signale auf eine gewisse Durchschnittsleistung P beschränkt sind. Dann haben die empfangenen Signale eine Durchschnittsleistung $P + N$, wobei N die Durchschnittsleistung des Rauschens darstellt. Die maximale Entropie für die empfangenen Signale wird dann erreicht, wenn auch diese ein Ensemble von weißem Rauschen bilden, da das die größtmögliche Entropie für eine Leistung $P + N$ ergibt. Sie ist durch eine geeignete Wahl des Ensembles der übertragenen Signale zu erreichen, eben dann, wenn diese ein Ensemble von weißem Rauschen mit der Leistung P bilden. Die Entropie (pro Sekunde) des empfangenen Ensembles ist dann

$$H(y) = W \log 2\pi e(P + N).$$

Die Entropie des Rauschens ist

$$H(n) = W \log 2\pi eN$$

und die Kanalkapazität ist

$$C = H(y) - H(n) = W \log \frac{P + N}{N}.$$

Zusammenfassend ergibt sich folgendes:

Lehrsatz 17: Die Kapazität eines Kanals der Bandbreite W , gestört durch weißes thermisches Rauschen der Leistung N , ist, wenn die durchschnittliche Übertragungsleistung auf P begrenzt ist, gegeben durch

$$C = W \log \frac{P + N}{N}.$$

Das bedeutet, daß wir bei einem genügend komplizierten Codiersystem die Binärzeichen mit der Rate $W \log_2 \frac{P + N}{N}$ bit pro Sekunde übertragen können, mit einer beliebig kleinen Fehlerhäufigkeit.

Es ist nicht möglich, mit irgendeinem Codiersystem eine größere Rate zu übertragen, ohne eine begrenzte, positive Fehlerhäufigkeit.

Um diesen Grenzwert der Übertragungsrate anzunähern, müssen sich die übertragenen Signale in bezug auf die statistischen Eigenschaften dem weißen Rauschen ⁶⁾ nähern.

Ein System, das diese ideale Übertragungsrate annähert, kann wie folgt beschrieben werden: Man erzeuge $M = 2^s$ Muster von weißem Rauschen, jedes von der Dauer T . Diesen werden Binärzahlen von 0 bis $(M - 1)$ zugeordnet. Am Sender werden die Nachrichtenfolgen aufgeteilt in Gruppen von je s Binärzeichen, und für jede Gruppe wird das entsprechende Muster des Rauschens als Signal übertragen. Am Empfänger sind die M Muster bekannt, und das tatsächlich empfangene Signal (gestört durch Rauschen) wird mit jedem davon verglichen. Das Muster mit dem kleinsten R. M. S.-Abstand ^{*}) gegenüber dem empfangenen Signal wird als das übertragene Signal ausgewählt, und die entsprechende Binärzahl wird rekonstruiert. Dieser Vorgang entspricht der Wahl des am meisten wahrscheinlichen Signals (*a posteriori*). Die Anzahl M von Rauschmustern, die benutzt wurden, wird von der geduldeten Fehlerfrequenz ϵ abhängen, es gilt jedoch für fast alle Beispiele solcher Rauschmuster

$$\lim_{\epsilon \rightarrow 0} \lim_{T \rightarrow \infty} \frac{\log M(\epsilon, T)}{T} = W \log \frac{P + N}{N},$$

so daß, gleichgültig wie klein ϵ gewählt wurde, wenn nur T genügend groß ist, der Wert $TW \log \frac{P + N}{N}$ bit in der Zeit T beliebig genau erreicht wird.

Gleichungen, die ähnlich lauten wie $C = W \log \frac{P + N}{N}$ für den Fall des weißen Rauschens, sind unabhängig voneinander von verschiedenen anderen Autoren entwickelt worden, jedoch mit etwas unterschiedlichen Interpretationen. Hier sollen die Arbeiten von N. Wiener ⁷⁾, W. G. Tuller ⁸⁾ und H. Sullivan erwähnt werden.

⁶⁾ Diese und andere Eigenschaften des weißen Rauschens werden vom geometrischen Standpunkt aus in dem Buch "Communication in the presence of Noise" besprochen, loc. cit.

^{*}) R. M. S. = Root Mean Square, hier in der Bedeutung des mittleren Abstandsquadrates (Anm. d. Übers.)

⁷⁾ Cybernetics, loc. cit.

⁸⁾ "Theoretical Limitations on the Rate of Transmission of Information", *Proceedings of the Institute of Radio Engineers*, Bd. 37, Nr. 5, Mai 1949, Seite 468-478.

Im Fall einer beliebigen Störung (nicht unbedingt des weißen thermischen Rauschens) scheint es nicht möglich zu sein, das Maximierungsproblem, das bei der Bestimmung der Kanalkapazität C einbezogen werden muß, einwandfrei und allgemeingültig zu lösen. Es können jedoch obere und untere Grenzen für C an Hand der durchschnittlichen Störleistung N und der Störentropieleistung N_1 bestimmt werden. In den meisten praktischen Fällen liegen diese beiden Grenzwerte genügend nahe beieinander, um das Problem befriedigend zu lösen.

Lehrsatz 18: Die Kanalkapazität für einen beliebig gestörten Kanal mit der Bandbreite W ist durch die folgende Ungleichung begrenzt:

$$W \log \frac{P + N_1}{N_1} \leq C \leq W \log \frac{P + N}{N_1}$$

mit

P = mittlere Sendeleistung,

N = mittlere Störleistung,

N_1 = Entropieleistung der Störung.

Auch hier wird die mittlere Leistung des gestörten Signals $P + N$ sein. Die maximale Entropie für diese Leistung träte auf, wenn das empfangene Signal ein weißes Rauschen wäre; ihr Wert wäre dann $W \log 2\pi e(P + N)$. Es mag unmöglich sein, diesen zu erreichen, d.h. es kann sein, daß es kein Ensemble von gesendeten Signalen gibt, das, wenn es der Störung hinzugefügt wird, ein weißes Rauschen am Empfänger erzeugt, aber zumindest wird dadurch eine obere Grenze für $H(y)$ festgelegt. Daher ist

$$\begin{aligned} C &= \text{Max } H(y) - H(n) \\ &\leq W \log 2\pi e(P + N) - W \log 2\pi e N_1. \end{aligned}$$

Dies ist die im Lehrsatz angegebene obere Grenze. Die untere Grenze entspricht der Übertragungsrate, wenn das gesendete Signal zu einem weißen Rauschen der Leistung P gemacht wird. In diesem Fall muß die Entropieleistung des empfangenen Signals mindestens so groß sein wie jene des weißen Rauschens der Leistung $P + N_1$, da in Lehrsatz 15 gezeigt wurde, daß die Entropieleistung der Summe zweier Ensembles größer oder gleich der Summe der individuellen Entropieleistungen ist. Daraus folgt

$$\text{Max } H(y) \geq W \log 2\pi e(P + N_1)$$

und

$$\begin{aligned} C &\geq W \log 2\pi e(P + N_1) - W \log 2\pi e N_1 \\ &= W \log \frac{P + N_1}{N_1}. \end{aligned}$$

Wird P größer, so nähern sich nach Lehrsatz 18 die obere und untere Grenze, und so erhalten wir als asymptotischen Wert

$$W \log \frac{P + N}{N_1}.$$

Wenn die Störung selbst ein weißes Rauschen ist, gilt $N = N_1$, und das Ergebnis reduziert sich auf die bereits bewiesene Formel:

$$C = W \log \left(1 + \frac{P}{N} \right).$$

Handelt es sich um ein Gaußsches Rauschen, jedoch mit einem Spektrum, das nicht unbedingt gleichmäßig sein muß, dann ist N_1 das geometrische Mittel der Rauschleistung über die verschiedenen Frequenzen im Band W .

$$N_1 = \exp \frac{1}{W} \int_W \log N(f) df,$$

wobei $N(f)$ die Rauschleistung bei der Frequenz f ist.

Lehrsatz 19: Wenn wir die Kapazität für eine gegebene Sendeleistung P festsetzen mit

$$C = W \log \frac{P + N - \eta}{N_1},$$

dann verringert sich η gleichmäßig, wenn P wächst, und nähert sich dem Grenzwert Null.

Nehmen wir an, daß für eine gegebene Leistung P_1 die Kanalkapazität gegeben ist durch

$$W \log \frac{P_1 + N - \eta_1}{N_1}.$$

Das bedeutet, daß die beste Verteilung der Signale, etwa $p(x)$, wenn sie zu der Verteilung des Rauschens addiert wird, eine Verteilung $r(y)$ ergibt, deren Entropieleistung $(P_1 + N - \eta_1)$ beträgt. Vergrößern wir nun die Leistung bis auf $P_1 + \Delta P$, indem wir ein weißes Rauschen der Leistung ΔP zu dem Signal addieren, dann beträgt die Entropie des empfangenen Signals mindestens

$$H(y) = W \log 2\pi e(P_1 + N - \eta_1 + \Delta P),$$

wenn man den Lehrsatz der minimalen Entropieleistung einer Sum-

me anwendet. Daraus folgt, da wir das angegebene H erreichen können, daß die Entropie der maximierenden Verteilung mindestens genauso groß sein muß und η sich gleichmäßig verringert. Um zu zeigen, daß $\eta \rightarrow 0$, wenn $P \rightarrow \infty$, betrachten wir ein Signal, das ein weißes Rauschen mit einem großen Wert P ist. Wie die Störung auch immer aussehen mag, das empfangene Signal wird näherungsweise ein weißes Rauschen sein, falls P genügend groß ist, in dem Sinne, daß man eine Entropieleistung hat, die sich $P + N$ nähert.

26. Die Kanalkapazität bei einer Begrenzung der Spitzenleistung

Bei einigen Anwendungen ist nicht die mittlere Leistungsabgabe des Senders begrenzt, sondern die kurzzeitige Leistungsspitze. Das Problem bei der Berechnung der Kanalkapazität ist dann die Maximierung (durch die Variation des Ensembles von übertragenen Zeichen) von

$$H(y) - H(n)$$

unter der Nebenbedingung, daß alle Funktionen $f(t)$ des Ensembles kleiner als oder gleich $\sqrt{S}$ für alle Werte t sind. Eine Einschränkung dieser Art ist mathematisch nicht so leicht zu verwirklichen wie die Begrenzung der Durchschnittsleistung. Das Wesentliche, das wir für diesen Fall erreicht haben, ist eine untere Grenze, die für alle Werte $\frac{S}{N}$ gültig ist, eine "asymptotische" obere Grenze (gültig für große $\frac{S}{N}$) und einen asymptotischen Wert von C für kleine $\frac{S}{N}$.

Lehrsatz 20: Für die Kapazität C eines Kanals mit der Bandbreite W , der durch weißes thermisches Rauschen der Leistung N gestört ist, gilt

$$C \geq W \log \frac{2}{\pi e^3} \frac{S}{N},$$

wobei S die Spitze der möglichen Übertragungsleistung darstellt. Für ein genügend großes $\frac{S}{N}$ ist

$$C \leq W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon),$$

wobei ϵ beliebig klein sein darf. Wenn $\frac{S}{N} \rightarrow 0$ (und vorausgesetzt, das Band W beginnt bei Null), dann gilt

$$C/W \log \left(1 + \frac{S}{N} \right) \rightarrow 1.$$

Wir wollen die Entropie des empfangenen Signals maximieren. Wenn $\frac{S}{N}$ groß ist, wird dies sicherlich gut erreicht, wenn die Entropie des übertragenen Ensembles maximiert wird.

Die asymptotische obere Grenze erhalten wir durch die Abschwächung der Bedingungen für das Ensemble. Nehmen wir an, daß die Leistung nicht in jedem Augenblick auf S begrenzt ist, sondern nur zu bestimmten Probezeitpunkten. Die maximale Entropie des übertragenen Ensembles unter diesen abgeschwächten Bedingungen ist sicherlich größer oder gleich gegenüber jener Entropie, die unter den ursprünglichen Bedingungen erzielt wurde. Dies veränderte Problem kann leicht gelöst werden. Die maximale Entropie wird erreicht, wenn die verschiedenen Probepunkte unabhängig voneinander sind und eine Verteilungsfunktion haben, die konstant ist zwischen $-\sqrt{S}$ und $+\sqrt{S}$. Die Entropie kann errechnet werden zu

$$W \log 4S.$$

Das empfangene Signal wird dann eine Entropie haben, die kleiner ist als

$$W \log (4S + 2\pi eN) (1 + \epsilon)$$

mit $\epsilon \rightarrow 0$, wenn $\frac{S}{N} \rightarrow \infty$, und die Kanalkapazität erhält man, indem man die Entropie des weißen Rauschens $W \log 2\pi eN$ subtrahiert:

$$\begin{aligned} W \log (4S + 2\pi eN) (1 + \epsilon) - W \log (2\pi eN) \\ = W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon). \end{aligned}$$

Dies ist die erwartete obere Grenze der Kanalkapazität.

Um eine untere Grenze zu erhalten, betrachte man das gleiche Ensemble von Funktionen. Man lasse diese Funktionen einen idealen Filter mit einer dreiecksförmigen Übertragungscharakteristik passieren. Die Verstärkung soll bei der Frequenz 0 gleich Eins sein und nimmt dann linear bis zur Verstärkung Null bei der Frequenz

W ab. Wir zeigen zunächst, daß die Ausgangsfunktionen des Filters eine Grenze S für die Leistungsspitzen zu allen Zeiten haben (nicht nur zu den Probezeitpunkten). Zuerst stellen wir fest, daß

ein Impuls $\frac{\sin 2\pi Wt}{2\pi Wt}$, der in den Filter hineingeht, in der Ausgabe die Größe

$$\frac{1}{2} \frac{\sin^2 \pi Wt}{(\pi Wt)^2}$$

erzeugt. Diese Funktion ist niemals negativ. Die Eingabefunktion kann man sich (im Normalfall) als die Summe einer Folge von verschobenen Funktionen der Form

$$a \frac{\sin 2\pi Wt}{2\pi Wt}$$

vorstellen, wobei a , die Amplitude des Musters, nicht größer als $\sqrt{S}$ ist. Daraus folgt, daß die Ausgabe die Summe der verschobenen Funktionen der nichtnegativen eben erwähnten Form mit den gleichen Koeffizienten ist. Da diese Funktionen nicht negativ sind, erhält man den größten positiven Wert für ein beliebiges t , wenn alle Koeffizienten a ihren maximalen positiven Wert haben, also $\sqrt{S}$. In diesem Fall war die Eingabefunktion eine Konstante der Amplitude $\sqrt{S}$, und da der Filter für Gleichstrom die Verstärkung Eins hat, ist die Ausgabe gleich der Eingabe. Daraus folgt, daß das Ausgabeensemble eine Spitzenleistung S hat.

Die Entropie des Ausgabeensembles kann aus jener des Eingabeensembles errechnet werden, indem man den Lehrsatz verwendet, der sich auf eine solche Situation bezieht. Die Ausgabeentropie ist gleich der Eingabeentropie plus der geometrisch gemittelten Verstärkung des Filters:

$$\int_0^W \log G^2 df = \int_0^W \log \left(\frac{W-f}{W} \right)^2 df = -2W.$$

Daraus folgt die Ausgabeentropie

$$W \log 4S - 2W = W \log \frac{4S}{e^2}$$

und die Kanalkapazität ist größer als

$$W \log \frac{2}{\pi e^3} \frac{S}{N}.$$

Wir wollen zeigen, daß für kleines $\frac{S}{N}$ (Signalspitzenleistung pro Leistung des weißen Rauschens) die Kanalkapazität approximiert

wird durch

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

Genauer gesagt: $C/W \log \left(1 + \frac{S}{N} \right) \rightarrow 1$, wenn $\frac{S}{N} \rightarrow 0$.

Da die mittlere Signalleistung P gleich oder kleiner als die Spitze S ist, folgt für alle Werte $\frac{S}{N}$

$$C \leq W \log \left(1 + \frac{P}{N} \right) \leq W \log \left(1 + \frac{S}{N} \right).$$

Daher wird das Ergebnis bewiesen sein, wenn wir ein Ensemble von Funktionen finden können, das einer Übertragungsrate entspricht, die nahe bei $W \log \left(1 + \frac{S}{N} \right)$ liegt und auf die Bandbreite W und Spitzenleistung S begrenzt ist. Betrachten wir deshalb das Ensemble von Funktionen der folgenden Art. Eine Folge von t Proben hat denselben Wert, entweder $+\sqrt{S}$ oder $-\sqrt{S}$, dann haben die t nächsten Proben den gleichen Wert usw. Der Wert für eine Folge wird zufällig ausgewählt mit der Wahrscheinlichkeit $\frac{1}{2}$ für $+\sqrt{S}$ und $\frac{1}{2}$ für $-\sqrt{S}$. Wird dieses Ensemble durch ein Filter mit dreiecksförmiger Verstärkungscharakteristik gesendet (Verstärkung Eins für Gleichstrom), dann ist der Spitzenwert der Ausgabe $\pm S$. Außerdem ist die Durchschnittsleistung nahezu S , und dies kann man beliebig annähern, indem t genügend groß gewählt wird. Die Entropie der Summe aus der Filterausgabe und dem thermischen Rauschen kann nun gefunden werden, indem man den Lehrsatz auf die Summe aus einem Rauschen und einem kleinen Signal anwendet. Dieser Lehrsatz wird angesetzt, wenn

$$\sqrt{t} \frac{S}{N}$$

genügend klein ist. Dies kann sicher erreicht werden, indem man einen genügend kleinen Wert $\frac{S}{N}$ (nachdem t gewählt wurde) nimmt.

Die Entropieleistung wird beliebig nahe bei $S+N$ liegen, und daraus folgt, daß die Rate der Übertragung beliebig nahe bei

$$W \log \left(\frac{S+N}{N} \right)$$

liegt.

V. Die Rate einer kontinuierlichen Quelle

27. Bestimmungsfunktionen für die Übertragungstreue

Im Falle einer diskreten Informationsquelle konnten wir eine genaue Rate der erzeugten Information bestimmen, nämlich die Entropie des zugrundeliegenden stochastischen Prozesses. Mit einer kontinuierlichen Quelle wird die Lage beträchtlich komplizierter. Zuerst einmal kann eine kontinuierliche variable Größe eine unendliche Anzahl von Werten annehmen und verlangt daher eine unendliche Anzahl von Binärwerten, um sie genau zu beschreiben. Das würde bedeuten, daß zum Übertragen der Ausgabe einer kontinuierlichen Quelle mit *exakter Rekonstruktion* am Empfangspunkt im allgemeinen ein Kanal mit einer unendlichen Kapazität (in bit pro Sekunde) benötigt würde. Da normalerweise Kanäle ein gewisses Maß an Störungen aufweisen und damit eine endliche Kapazität, wäre die exakte Übertragung unmöglich.

Das geht jedoch am wirklichen Problem vorbei. In der Praxis sind wir nicht an einer exakten Übertragung interessiert, wenn wir eine kontinuierliche Quelle haben, sondern nur an der Übertragung innerhalb einer bestimmten Toleranz. Die Frage ist, können wir eine begrenzte Rate einer stetigen Quelle zuordnen, wenn wir nur eine gewisse Genauigkeit oder Rekonstruktion am Empfänger benötigen, die auf geeignete Weise gemessen wird? Natürlich wird sich die Rate vergrößern müssen, sobald die Forderungen an die Übertragungstreue wachsen. Es wird sich zeigen, daß wir in sehr allgemeinen Fällen solch eine Übertragungsrate definieren können, die die Eigenschaft hat, daß sie bei richtiger Codierung der Information über einen Kanal gesendet werden kann, dessen Kapazität gleich der fraglichen Rate ist, und daß sie die Forderungen an die Übertragungstreue erfüllt. Ein Kanal mit einer geringeren Kapazität wäre ungenügend.

Zunächst ist es notwendig, eine allgemeine mathematische Formulierung des Begriffs der Übertragungstreue zu finden. Betrachten wir die Menge aus Nachrichten von langer Dauer, sagen wir T Sekunden. Die Quelle wird beschrieben, indem man die Wahr-

scheinlichkeitsdichte $P(x)$ im zugehörigen Raum angibt, daß die Quelle die fragliche Nachricht auswählen wird. Ein gegebenes Kommunikationssystem wird dadurch beschrieben (vom Beobachterstandpunkt aus), daß die bedingte Wahrscheinlichkeit $P_2(y)$ angegeben wird, daß, falls die Nachricht x von der Quelle erzeugt wird, die erhaltene Nachricht am Empfangspunkt y ist. Das System als Ganzes (einschließlich Quelle und Übertragungssystem) wird durch die Wahrscheinlichkeitsfunktion $P(x, y)$ beschrieben, daß eine Nachricht x zu einer Ausgabe y führt. Wenn diese Funktion bekannt ist, sind sämtliche Charakteristiken des Systems vom Gesichtspunkt der Übertragungstreue her bekannt. Jede Bestimmung der Übertragungstreue muß mathematisch einer gegenüber $P(x, y)$ angewandten Operation entsprechen. Diese Operation muß zumindest die Eigenschaft einer einfachen Zuordnung von Systemen haben, d. h. es muß möglich sein, von zwei durch $P_1(x, y)$ und $P_2(x, y)$ dargestellten Systemen zu sagen, daß, unserem Treuekriterium entsprechend, entweder 1.) das erste eine höhere Übertragungstreue hat oder 2.) das zweite eine höhere Übertragungstreue aufweist oder 3.) sie dieselbe Übertragungstreue haben. Das bedeutet, daß ein Kriterium der Übertragungstreue durch eine *Bestimmungsfunktion* mit numerischem Wert dargestellt werden kann,

$$v(P(x, y)),$$

deren Argument sich über mögliche Wahrscheinlichkeitsfunktionen $P(x, y)$ erstreckt. Die Funktion $v(P(x, y))$ ordnet Kommunikationssysteme nach der Übertragungstreue, und zur Erleichterung setzen wir fest, daß niedrigere Werte von v "größerer Übertragungstreue" entsprechen.

Wir werden nun zeigen, daß unter sehr allgemeinen und vernünftigen Voraussetzungen die Funktion $v(P(x, y))$ in einer anscheinend viel spezielleren Form geschrieben werden kann, nämlich als der Durchschnitt einer Funktion $\rho(x, y)$ über die Menge von möglichen Werten für x und y :

$$v(P(x, y)) = \iint P(x, y) \rho(x, y) dx dy.$$

Um dies zu erreichen, müssen wir einfach voraussetzen, daß 1.) die Quelle und das System ergodisch sind, so daß eine sehr lange Nachricht mit der Wahrscheinlichkeit von nahezu 1 typisch für das Ensemble sein wird, und 2.) daß die Berechnung "vernünftig" ist in dem Sinne, daß es möglich ist, durch die Beobachtung einer typischen Eingabe und Ausgabe x_1 und y_1 eine vorläufige Bestim-

mung auf der Basis dieser Stichproben zu bilden; und wenn die Dauer dieser Proben größer wird, wird sich die vorläufige Bestimmung mit der Wahrscheinlichkeit 1 der genauen Bestimmung nähern, die auf einer vollständigen Kenntnis von $P(x, y)$ basiert. Die vorläufige Bestimmung sei $\rho(x, y)$. Dann nähert sich die Funktion $\rho(x, y)$ einer Konstanten (wenn $T \rightarrow \infty$) für fast alle Werte (x, y) , die sich entsprechend dem System in dem Gebiet der hohen Wahrscheinlichkeit befinden:

$$\rho(x, y) \rightarrow v(P(x, y)).$$

Und wir können auch schreiben

$$\rho(x, y) \rightarrow \iint P(x, y) \rho(x, y) dx dy,$$

weil

$$\iint P(x, y) dx dy = 1.$$

Dies stellt das gewünschte Ergebnis dar.

Die Funktion $\rho(x, y)$ hat die allgemeinen Eigenschaften eines "Abstands" zwischen x und y ⁹⁾. Sie mißt, wie unerwünscht es ist (gemäß unserem Treuekriterium), den Wert y zu erhalten, wenn x gesendet worden ist. Das allgemeine oben gegebene Ergebnis kann wie folgt noch etwas anders formuliert werden: Jede vernünftige Bestimmung kann als Mittelwert eines Abstands angesehen werden über die Menge von Nachrichten und empfangenen Nachrichten x und y , gewichtet nach der Wahrscheinlichkeit $P(x, y)$, das fragliche Paar zu erhalten, vorausgesetzt, daß die Dauer T der Nachrichten genügend groß angenommen wurde. Es folgen einfache Beispiele von Bestimmungsfunktionen:

1. R.M.S.-Kriterium

$$v = \overline{(x(t) - y(t))^2}.$$

In diesem häufig benutzten Maß der Übertragungstreue ist die Abstands-Funktion $\rho(x, y)$ (abgesehen von einem konstanten Faktor) das Quadrat des gewöhnlichen Euklidischen Abstands zwischen den Punkten x und y im zugehörigen Funktionenraum.

$$\rho(x, y) = \frac{1}{T} \int_0^T [x(t) - y(t)]^2 dt.$$

2. R.M.S.-Kriterium nach Frequenzen gewichtet

Allgemeiner kann man verschiedene Gewichte auf die verschie-

⁹⁾ Sie ist nicht "metrisch" im strengen Sinne, weil sie im allgemeinen weder $\rho(x, y) = \rho(y, x)$ noch $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$ genügt.

denen Frequenzkomponenten anwenden, bevor eine R.M.S.-Übertragungstreuebestimmung durchgeführt wird. Dies ist gleichbedeutend damit, die verschiedenen $x(t) - y(t)$ durch ein formendes Filter zu schicken und dann die Durchschnittsleistung der Ausgabe zu bestimmen.

Nehmen wir daher an

$$e(t) = x(t) - y(t)$$

und

$$f(t) = \int_{-\infty}^{\infty} e(\tau) k(t - \tau) d\tau;$$

dann ist

$$\rho(x, y) = \frac{1}{T} \int_0^T f(t)^2 dt.$$

3. Absolutes Fehlerkriterium.

$$\rho(x, y) = \frac{1}{T} \int_0^T |x(t) - y(t)| dt.$$

4. Die Struktur von Ohr und Gehirn legt implizit eine Anzahl von Bestimmungen nahe, die im Falle der Sprach- oder Musikübertragung angemessen sind. Es gibt z. B. ein "Verständlichkeitskriterium", bei dem $\rho(x, y)$ gleich der relativen Häufigkeit falsch interpretierter Worte ist, wenn die Nachricht $x(t)$ als $y(t)$ empfangen wird. Obwohl wir keine eindeutige Darstellung von $\rho(x, y)$ geben können, könnte sie in diesen Fällen im Prinzip durch genügend Experimente bestimmt werden. Einige der Eigenschaften folgen aus gut bekannten experimentellen Ergebnissen für das Hören; z. B. ist das Ohr relativ unempfindlich für die Phase, und die Empfindlichkeit für Amplituden und Frequenz ist ungefähr logarithmisch.

5. Der diskrete Fall kann als ein Spezialfall betrachtet werden, bei dem wir stillschweigend eine Bestimmung vorausgesetzt haben, die auf der Fehlerhäufigkeit basiert. Die Funktion $\rho(x, y)$ ist dann als die Anzahl der Zeichen in der Folge y definiert, die von den entsprechenden Zeichen der Folge x abweichen, dividiert durch die Gesamtanzahl der Zeichen in x .

28. Die Senderate einer Quelle in Abhängigkeit von der Bestimmung der Übertragungstreue

Wir sind nun in der Lage, die Rate der Informationserzeugung für eine kontinuierliche Quelle zu definieren. $P(x)$ sei gegeben für die Quelle und eine Bestimmung v durch eine Abstandsfunktion $\rho(x, y)$ festgelegt, die als stetig in x und y vorausgesetzt wird. Mit einem besonderen System $P(x, y)$ wird die Qualität gemessen durch

$$v = \iint \rho(x, y) P(x, y) dx dy.$$

Außerdem ist die Übertragungsrate von Binärzeichen entsprechend $P(x, y)$

$$R = \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy.$$

Wir definieren die Rate R_1 , mit der Information mit einer gegebenen Qualität v_1 wieder reproduziert wird, als das Minimum von R , wenn wir bei festem $v = v_1$ $P_*(y)$ variieren. Das bedeutet:

$$R_1 = \min_{P_*(y)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy,$$

abhängig von der Einschränkung

$$v_1 = \iint P(x, y) \rho(x, y) dx dy.$$

Dies bedeutet, daß wir tatsächlich alle Kommunikationssysteme berücksichtigen, welche benutzt werden dürfen und die mit der verlangten Übertragungstreue übertragen. Die Übertragungsrate in bit pro Sekunde ist für jedes einzelne errechnet worden, und wir wählen jenes mit der kleinsten Rate aus. Diese letztere ist die Rate, die wir der Quelle mit der fraglichen Übertragungstreue zuordnen.

Die Berechtigung dieser Definition beruht auf dem folgenden Ergebnis:

Lehrsatz 21: Wenn eine Quelle mit einer Rate R_1 und der Übertragungstreue v_1 sendet, ist es möglich, die Ausgabe der Quelle zu codieren und über einen Kanal der Kapazität C zu übertragen mit einer Übertragungstreue, die beliebig nahe bei v_1 liegt, vorausgesetzt $R_1 \leq C$. Dies ist nicht möglich, wenn $R_1 > C$.

Die letzte Behauptung des Lehrsatzes folgt unmittelbar aus der Definition von R_1 und den vorhergehenden Ergebnissen. Wenn das nicht

wahr wäre, könnten mehr als C bit pro Sekunde über einen Kanal der Kapazität C übertragen werden. Der erste Teil des Lehrsatzes wird mit einer Methode bewiesen, die analog zu jener ist, die für Lehrsatz 11 benutzt wurde. Als erstes können wir den (x, y) -Raum in eine große Anzahl kleiner Zellen aufteilen und die Situation als einen diskreten Fall darstellen. Dies wird aufgrund der für $\rho(x, y)$ vorausgesetzten Stetigkeit die Bestimmungsfunktion nur um einen beliebig kleinen Wert verändern (wenn die Elemente sehr klein sind). Nehmen wir an, daß $P_1(x, y)$ das bestimmte System ist, das die Rate minimiert und R_1 ergibt. Wir wählen aus den sehr wahrscheinlichen y -Werten eine zufällige Menge aus, die

$$2^{(R_1 + \epsilon)T}$$

Elemente enthält, wobei $\epsilon \rightarrow 0$, wenn $T \rightarrow \infty$. Für große T wird jeder ausgewählte Punkt (wie in Abb. 10) durch Linien großer Wahrscheinlichkeit mit einer Menge von x -Werten verbunden sein. Eine Berechnung, die der im Beweis für Lehrsatz 11 ähnlich ist, zeigt, daß für große T fast alle x durch die Fächer der ausgewählten y -Punkte für fast jede Wahl der y belegt sind. Das zu benutzende Kommunikationssystem arbeitet folgendermaßen: Den ausgewählten Punkten werden Binärzahlen zugeordnet. Wenn eine Nachricht x erstellt wurde, wird sie (mit einer Wahrscheinlichkeit, die sich für $T \rightarrow \infty$ dem Wert 1 nähert) innerhalb von mindestens einem der Fächer liegen. Die entsprechende Binärzahl wird über den Kanal übertragen (oder eine willkürlich ausgewählte, wenn es mehrere gibt) mit einer geeigneten Codierung, um eine geringe Fehlerhäufigkeit zu erreichen. Da $R_1 \leq C$ gilt, ist dies möglich. Am Empfangspunkt wird das entsprechende y rekonstruiert und als die wiedererlangte Nachricht benutzt.

Die Übertragungstreue v'_1 für dieses System kann an v_1 beliebig angenähert werden, indem man T genügend groß wählt. Dies ist der Tatsache zu verdanken, daß für jedes lange Muster der Nachricht $x(t)$ und der wiedererlangten Nachricht $y(t)$ die Übertragungstreue sich v_1 (mit der Wahrscheinlichkeit 1) nähert.

Es ist interessant, anzumerken, daß in diesem System die Störung in der wiedererlangten Nachricht tatsächlich durch eine Art allgemeiner Quantisierung beim Sender und nicht durch die Störung im Kanal erzeugt wird. Es ist mehr oder weniger analog zu der quantisierten Störung in der Puls-Code-Modulation.

29. Die Berechnung der Raten

Die Definition der Rate ist in vielen Punkten ähnlich der Definition der Kanalkapazität. Erstere ist

$$R = \text{Min}_{P_s(y)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy,$$

wobei $P(x)$ und $v_1 = \iint P(x, y) \rho(x, y) dx dy$ festgelegt sind, letztere

$$C = \text{Max}_{P_s(y)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy,$$

wobei $P_s(y)$ festgelegt ist und möglicherweise eine oder mehrere andere Einschränkungen (z. B. eine Begrenzung der mittleren Leistung) der Form $K = \iint P(x, y) \lambda(x, y) dx dy$ bestehen.

Eine Teillösung des allgemeinen Maximierungsproblems für die Bestimmung der Rate einer Quelle kann angegeben werden. Nach Lagranges Methode betrachten wir

$$\iint \left[P(x, y) \log \frac{P(x, y)}{P(x)P(y)} + \mu P(x, y) \rho(x, y) + \nu(x) P(x, y) \right] dx dy.$$

Die Variationsgleichung (wenn wir die erste Variation über $P(x, y)$ nehmen) führt zu

$$P_y(x) = B(x) e^{-\lambda \rho(x, y)},$$

wobei λ so bestimmt wird, daß die geforderte Übertragungstreue erreicht wird, und $B(x)$ so gewählt ist, daß es der folgenden Gleichung genügt:

$$\int B(x) e^{-\lambda \rho(x, y)} dx = 1.$$

Dies zeigt, daß bei bester Codierung die bedingte Wahrscheinlichkeit $P_y(x)$ für einen bestimmten Fall bei verschiedenen empfangenen y exponentiell abnimmt mit der Funktion des Abstands $\rho(x, y)$ zwischen den fraglichen Werten x und y .

Im speziellen Fall, wenn die Funktion des Abstands $\rho(x, y)$ nur von der (Vektor-) Differenz zwischen x und y abhängt,

$$\rho(x, y) = \rho(x - y),$$

gilt

$$\int B(x) e^{-\lambda \rho(x-y)} dx = 1.$$

Daraus folgt, daß $B(x)$ eine Konstante, z. B. α , ist und

$$P_y(x) = \alpha e^{-\lambda \rho(x-y)}.$$

Leider sind diese formalen Lösungen in einzelnen Fällen schwierig auszuwerten und scheinen von geringem Wert zu sein. Tatsächlich wurde die reale Berechnung der Raten nur in einigen sehr einfachen Fällen durchgeführt.

Wenn die Funktion des Abstands $\rho(x, y)$ die mittlere quadratische Differenz zwischen x und y darstellt und das Nachrichtenensemble ein weißes Rauschen ist, dann kann die Rate bestimmt werden. In einem solchen Fall ist

$$R = \text{Min} [H(x) - H_y(x)] = H(x) - \text{Max} H_y(x)$$

mit $N = (x - y)^2$. Aber der Maximalwert von $H_y(x)$ wird erreicht, wenn $y - x$ ein weißes Rauschen und gleich $W_1 \log 2\pi eN$ ist, wobei W_1 die Bandbreite des Nachrichtenensembles ist. Deswegen wird

$$\begin{aligned} R &= W_1 \log 2\pi eQ - W_1 \log 2\pi eN \\ &= W_1 \log \frac{Q}{N} \end{aligned}$$

mit Q als der mittleren Nachrichtenleistung. Das beweist den folgenden:

Lehrsatz 22: Die Rate für eine Quelle weißen Rauschens der Leistung Q mit der Bandbreite W_1 ist, relativ zu einem R. M. S.-Maß der Übertragungstreue

$$R = W_1 \log \frac{Q}{N}.$$

Dabei ist N die erlaubte mittlere quadratische Abweichung zwischen der ursprünglichen und der wiedergewonnenen Nachricht.

Allgemeiner gesagt: Für jede Nachrichtenquelle können wir Ungleichungen erhalten, die die Rate relativ zu einem Kriterium für mittlere quadratische Fehler begrenzen.

Lehrsatz 23: Die Rate jeder Quelle der Bandbreite W_1 ist begrenzt durch

$$W_1 \log \frac{Q_1}{N} \leq R \leq W_1 \log \frac{Q}{N},$$

wobei Q die Durchschnittsleistung der Quelle, Q_1 ihre Entropieleistung und N der zulässige mittlere quadratische Fehler sind.

Die untere Grenze folgt aus der Tatsache, daß das Maximum von $H_v(x)$ für ein gegebenes $\overline{(x-y)^2} = N$ für den Fall des weißen Rauschens gilt. Die obere Grenze ergibt sich, wenn wir die Punkte (die im Beweis des Lehrsatzes 21 benutzt werden) nicht auf die beste Art, sondern zufällig auf einer Kugel mit dem Radius $\sqrt{Q-N}$ anordnen.

Anerkennung

Der Verfasser ist seinen Kollegen von den Bell Telephone Laboratories, besonders den Herren Dr. H.W. Bode, Dr. J.R. Pierce, Dr. B. McMillan und Dr. B.M. Oliver sehr verbunden für viele hilfreiche Vorschläge und die Kritik während des Verlaufs dieser Arbeit. Große Anerkennung schuldet er auch Herrn Professor N. Wiener, dessen elegante Lösung der Probleme der Filterung und Vorhersage von stationären Ensembles die Denkweise des Verfassers auf diesem Gebiet beträchtlich beeinflusst hat.

Anhang 1 Das Wachstum einer Zahl von Zeichen-Blöcken mit einer endlichen Zustandsbedingung

Nehmen wir $N_i(L)$ als die Anzahl der Blöcke von Zeichen an, die die Länge L haben und im Zustand i enden. Dann haben wir

$$N_i(L) = \sum_j N_j(L - b_{ij}^{(q)}),$$

wobei $b_{ij}^1, b_{ij}^2, \dots, b_{ij}^m$ die Längen der Zeichen sind, die, im Zustand i ausgewählt, zum Zustand j führen. Dies sind lineare Differenzgleichungen, und sie müssen sich, wenn $L \rightarrow \infty$, nach der Art

$$N_i = A_i W^L$$

verhalten.

Durch eine Substitution in der Differenzgleichung wird

$$A_i W^L = \sum_{j,s} A_j W^{L-b_{ij}^{(q)s}}$$

oder

$$A_i = \sum_{j,s} A_j W^{-b_{ij}^{(q)s}}$$

$$\sum_i \left(\sum_s W^{-b_{ij}^{(q)s}} - \delta_{ij} \right) A_i = 0.$$

Damit das möglich ist, muß die Determinante

$$D(W) = |a_{ij}| = \left| \sum_s W^{-b_{ij}^{(q)s}} - \delta_{ij} \right|$$

verschwinden, und damit liegt W fest, das natürlich die größte reale Wurzel von $D = 0$ ist.

Die Größe C ist dann gegeben durch

$$C = \lim_{L \rightarrow \infty} \frac{\log \sum A_i W^L}{L} = \log W.$$

Damit stellen wir auch fest, daß sich dieselben Eigenschaften der Vergrößerung ergeben, wenn wir fordern, daß alle Blöcke in demselben (willkürlich ausgewählten) Zustand beginnen.

Anhang 2 Herleitung von $H = -\sum p_i \log p_i$

Es sei $H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) = A(n)$. Nach der Bedingung (3) können wir eine Wahl aus s^m gleichwahrscheinlichen Möglichkeiten in eine Reihe von m Auswahlvorgängen zerlegen, jede von ihnen aus s gleichwahrscheinlichen Möglichkeiten. Dann erhalten wir

$$A(s^m) = m A(s).$$

Ganz ähnlich ist

$$A(t^n) = n A(t).$$

Wir können n beliebig groß wählen und finden immer ein m , das der folgenden Ungleichung genügt:

$$s^m \leq t^n < s^{(m+1)}.$$

Daher wird, nimmt man die Logarithmen und dividiert man durch $n \log s$,

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \epsilon,$$

wobei ϵ beliebig klein ist. Nun ergibt sich aus der Tatsache, daß $A(n)$ monoton ist

$$A(s^m) \leq A(t^n) \leq A(s^{m+1}) \\ m A(s) \leq n A(t) \leq (m+1) A(s).$$

Daraus folgt, dividiert man durch $n A(s)$,

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| < \epsilon \\ \left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq 2\epsilon \quad A(t) = -K \log t,$$

wobei K positiv sein muß, um (2) zu erfüllen.

Nun nehmen wir an, daß wir eine Auswahlmenge von n Möglichkeiten haben mit den zusammen meßbaren Wahrscheinlichkeiten

$p_i = \frac{n_i}{\sum n_i}$, wobei n_i ganze Zahlen sind. Wir können eine Auswahl

aus $\sum n_i$ Möglichkeiten in eine Wahl aus n Möglichkeiten mit den

Wahrscheinlichkeiten $p_1, \dots, p_n$ und, wenn die i -te Möglichkeit ausgewählt wurde, eine Wahl aus n_i mit gleichen Wahrscheinlichkeiten aufteilen. Wenn wir wieder die Bedingung 3 verwenden, so können wir die nach zwei Methoden berechneten Ergebnisse für die Wahl aus $\sum n_i$ gleichsetzen:

$$K \log \sum n_i = H(p_1, \dots, p_n) + K \sum p_i \log n_i.$$

Daraus folgt

$$H = K[\sum p_i \log \sum n_i - \sum p_i \log n_i] \\ = -K \sum p_i \log \frac{n_i}{\sum n_i} = -K \sum p_i \log p_i.$$

Wenn die p_i nicht zugleich meßbar sind, können sie durch rationale Zahlen angenähert werden, und es muß derselbe Ausdruck nach unserer Voraussetzung der Stetigkeit gelten. Damit gilt der Ausdruck auch allgemein. Die Wahl des Koeffizienten K ist eine Sache der Bequemlichkeit und entspricht der Wahl einer Maßeinheit.

Anhang 3 Lehrsätze über ergodische Quellen

Wir nehmen an, eine Quelle sei ergodisch, so daß das strenge Gesetz der großen Zahlen angewandt werden kann. Deswegen ist die Anzahl, wie oft ein gegebener Weg p_{ij} im Netzwerk während einer langen Folge der Länge N durchlaufen wird, etwa proportional zu der Wahrscheinlichkeit P_i , an der Stelle i zu sein und nun diesen Weg auszuwählen, $P_i p_{ij} N$. Wenn N genügend groß ist, ist die darin enthaltene Wahrscheinlichkeit eines prozentualen Fehlers $\pm \delta$ kleiner als ϵ , so daß für alle außer einer Menge mit kleiner Wahrscheinlichkeit die tatsächlichen Zahlen innerhalb der folgenden Grenzen liegen:

$$(P_i p_{ij} \pm \delta) N.$$

Dadurch haben fast alle Folgen eine Wahrscheinlichkeit p , die gegeben ist durch

$$p = \prod p_{ij}^{(P_i p_{ij} \pm \delta) N}$$

und $\frac{\log p}{N}$ ist begrenzt durch

$$\frac{\log p}{N} = \sum (P_i p_{ij} \pm \delta) \log p_{ij}$$

oder

$$\left| \frac{\log p}{N} - \sum P_i p_{ij} \log p_{ij} \right| < \eta.$$

Dies beweist Lehrsatz 3.

Lehrsatz 4 folgt unmittelbar daraus durch die Berechnung der oberen und unteren Grenzen für $n(q)$, die auf dem möglichen Wertebereich für p in Lehrsatz 3 basiert.

Im gemischten (nicht ergodischen) Fall, wenn

$$L = \sum p_i L_i$$

und die Entropien der Komponenten $H_1 \geq H_2 \geq \dots \geq H_n$ sind, gilt der

Lehrsatz: $\lim_{N \rightarrow \infty} \frac{\log n(q)}{N} = \varphi(q)$ ist eine sinkende Stufen-Funktion

$$\varphi(q) = H_s \text{ im Intervall } \sum_1^{i-1} \alpha_i < q < \sum_1^i \alpha_i.$$

Um die Lehrsätze 5 und 6 zu beweisen, stellt man zuerst fest, daß F_N monoton fallend ist, weil durch ein wachsendes N ein Index zu einer bedingten Entropie addiert wird. Eine einfache Substitution für $p_{Bi}(S_j)$ in der Definition von F_N zeigt, daß

$$F_N = N G_N - (N - 1) G_{N-1}.$$

Summiert man über alle N , ergibt sich $G_N = \frac{1}{N} \sum F_N$. Daraus

folgt $G_N \geq F_N$, wobei G_N monoton fällt. Auch müssen sie sich demselben Grenzwert nähern. Indem wir Lehrsatz 3 anwenden, stellen wir fest, daß $\lim_{N \rightarrow \infty} G_N = H$ ist.

Anhang 4 Maximierung der Rate für ein System mit Beschränkungen

Angenommen, wir haben eine Menge von Beschränkungen gegenüber einer Folge von Zeichen, die sich durch eine begrenzte Anzahl von Zuständen beschreiben läßt und daher durch einen linearen Graphen dargestellt werden kann, wie in Abb. 2 gezeigt wird. Die $l_{ij}^{(s)}$ seien die Längen der verschiedenen Zeichen, die beim Übergang vom Zustand i in den Zustand j vorkommen können. Welche Verteilung der Wahrscheinlichkeiten P_i für die verschiedenen Zustände und $p_{ij}^{(s)}$ für die Auswahl des Zeichens s im Zustand i und den Übergang in den Zustand j maximiert die Rate der erzeugten Information unter diesen Beschränkungen? Die Beschränkungen definieren einen diskreten Kanal, und die maximale Rate muß kleiner oder gleich der Kapazität C dieses Kanals sein, denn wären alle Blöcke großer Länge gleich wahrscheinlich, ergäbe sich diese Rate, und diese wäre, wenn möglich, die beste. Wir werden zeigen, daß diese Rate durch die richtige Auswahl von P_i und $p_{ij}^{(s)}$ erreicht werden kann. Die fragliche Rate beträgt

$$\frac{-\sum_{i,j,s} P_i p_{ij}^{(s)} \log p_{ij}^{(s)}}{\sum_{i,j,s} P_i p_{ij}^{(s)} l_{ij}^{(s)}}.$$

Es sei

$$p_{ij}^{(s)} = \frac{B_j}{B_i} W^{-l_{ij}^{(s)}},$$

wobei die B_i den Gleichungen

$$B_i = \sum_{j,s} B_j W^{-l_{ij}^{(s)}}$$

genügen.

Dieses homogene Gleichungssystem hat eine nicht verschwindende Lösung, da W so gewählt wurde, daß die Determinante der Koeffizienten gleich 0 ist:

$$\left| \sum_s W^{-l_{ij}^{(s)}} - \delta_{ij} \right| = 0.$$

Die so definierten $p_{ij}^{(s)}$ sind befriedigende Übergangswahrscheinlichkeiten, denn zunächst ist

$$\begin{aligned} \sum_{j,s} p_{ij}^{(s)} &= \sum_{j,s} \frac{B_j}{B_i} W^{-l_{ij}^{(s)}} \\ &= \frac{B_i}{B_i} = 1, \end{aligned}$$

so daß die Summe der Wahrscheinlichkeiten von irgendeinem einzelnen Knotenpunkt Eins ist. Weiterhin sind sie nicht-negativ, was aus einer Betrachtung der Größen A_i ersichtlich wird, die in Anhang 1 angegeben wurden. Die A_i sind notwendigerweise nicht-negativ, und die B_i erfüllen ein ähnliches System von Gleichungen, wobei jedoch i und j vertauscht sind. Dies läuft darauf hinaus, die Laufrichtung der einzelnen Linien des Graphen umzukehren.

Wenn wir diese angenommenen Werte für $p_{ij}^{(s)}$ in der allgemeinen Gleichung für die Rate einsetzen, erhalten wir

$$\begin{aligned} &= \frac{\sum P_i p_{ij}^{(s)} \log \frac{B_j}{B_i} W^{-l_{ij}^{(s)}}}{\sum P_i p_{ij}^{(s)} l_{ij}^{(s)}} \\ &= \frac{\log W \sum P_i p_{ij}^{(s)} l_{ij}^{(s)} - \sum P_i p_{ij}^{(s)} \log B_j + \sum P_i p_{ij}^{(s)} \log B_i}{\sum P_i p_{ij}^{(s)} l_{ij}^{(s)}} \\ &= \log W = C. \end{aligned}$$

Daraus folgt, daß die Rate mit dieser Menge von Übergangswahrscheinlichkeiten C ist, und da diese Größe niemals überschritten werden kann, ist sie das Maximum.

Anhang 5

S_1 sei irgendeine meßbare Untermenge des g -Ensembles und S_2 die Untermenge des f -Ensembles, die S_1 nach der Operation T erzeugt. Dann ist

$$S_1 = TS_2.$$

Die Größe H^λ sei ein Operator, der alle Funktionen einer Menge um den Wert λ in der Zeit verschiebt. Dann wird

$$H^\lambda S_1 = H^\lambda TS_2 = TH^\lambda S_2$$

da T invariant ist und deshalb mit H^λ kommutiert. Daraus folgt, wenn $m[S]$ das Wahrscheinlichkeitsmaß der Menge S ist,

$$\begin{aligned} m[H^\lambda S_1] &= m[TH^\lambda S_2] = m[H^\lambda S_2] \\ &= m[S_2] = m[S_1], \end{aligned}$$

wobei die zweite Gleichung durch Definition des Maßes im g -Raum entsteht, die dritte, weil das f -Ensemble stationär ist, und die letzte wieder durch Definition des Maßstabs für g . Dies beweist, daß das g -Ensemble stationär ist.

Um zu beweisen, daß die ergodische Eigenschaft unter invarianten Operationen erhalten bleibt, seien S_1 eine Untermenge des g -Ensembles, die invariant gegenüber H^λ ist, und S_2 die Menge aller Funktionen f , die sich in S_1 transformieren. Dann ist

$$H^\lambda S_1 = H^\lambda TS_2 = TH^\lambda S_2 = S_1,$$

so daß $H^\lambda S_2$ in S_2 für alle λ inbegriffen wäre.

Weil nun

$$m[H^\lambda S_2] = m[S_2] = m[S_1],$$

bedeutet dies

$$H^\lambda S_2 = S_2$$

für alle Werte λ mit $m[S_2] \neq 0, 1$. Dieser Widerspruch zeigt, daß S_1 nicht existiert.

Anhang 6

Die obere Grenze, $\bar{N}_3 \leq N_1 + N_2$ ist auf die Tatsache zurückzuführen, daß die maximale Entropie für eine Leistung $N_1 + N_2$ vor- kommt, wenn wir ein weißes Rauschen dieser Leistung haben. In diesem Fall ist die Entropieleistung $N_1 + N_2$.

Um die untere Grenze zu erreichen, nehmen wir an, wir haben zwei Verteilungen in n Dimensionen $p(x_i)$ und $q(x_i)$ mit Entropie- leistungen $\bar{N}_1$ und $\bar{N}_2$. Welche Form sollten p und q haben, um die Entropieleistung $\bar{N}_3$ ihrer Faltung $r(x_i)$ zu minimieren?

$$r(x_i) = \int p(y_i)q(x_i - y_i) dy_i.$$

Die Entropie H_3 von r ist gegeben durch

$$H_3 = - \int r(x_i) \log r(x_i) dx_i.$$

Wir wollen dies unter folgenden Beschränkungen minimieren

$$H_1 = - \int p(x_i) \log p(x_i) dx_i$$

$$H_2 = - \int q(x_i) \log q(x_i) dx_i.$$

Wir betrachten deshalb

$$U = - \int [r(x) \log r(x) + \lambda p(x) \log p(x) + \mu q(x) \log q(x)] dx$$

$$\delta U = - \int \{ [1 + \log r(x)] \delta r(x) + \lambda [1 + \log p(x)] \delta p(x) + \mu [1 + \log q(x)] \delta q(x) \} dx.$$

Wenn $p(x)$ um ein bestimmtes Argument $x_i = s_i$ variiert wird, dann ist die Variation in $r(x)$

$$\delta r(x) = q(x_i - s_i)$$

und damit

$$\delta U = - \int q(x_i - s_i) \log r(x_i) dx_i - \lambda \log p(s_i) = 0.$$

Ähnliches gilt, wenn q variiert wird. Daraus folgen die Bedingungen für das Minimum

$$\int q(x_i - s_i) \log r(x_i) = -\lambda \log p(s_i)$$

$$\int p(x_i - s_i) \log r(x_i) = -\mu \log q(s_i).$$

Wenn wir die erste Gleichung mit $p(s_i)$ und die zweite mit $q(s_i)$ multiplizieren und über s integrieren, so erhalten wir

$$H_3 = -\lambda H_1$$

$$H_3 = -\mu H_2$$

oder nach λ und μ aufgelöst und in den Gleichungen eingesetzt

$$H_1 \int q(x_i - s_i) \log r(x_i) dx_i = -H_3 \log p(s_i)$$

$$H_2 \int p(x_i - s_i) \log r(x_i) dx_i = -H_3 \log p(s_i).$$

Nun nehmen wir $p(x_i)$ und $q(x_i)$ als normal an

$$p(x_i) = \frac{|A_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\frac{1}{2} \Sigma A_{ij} x_i x_j$$

$$q(x_i) = \frac{|B_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\frac{1}{2} \Sigma B_{ij} x_i x_j.$$

Dann wird auch $r(x_i)$ normal sein mit quadratischer Form C_{ij} .

Wenn die Inversen dieser Formen a_{ij} , b_{ij} , c_{ij} sind, dann ist

$$c_{ij} = a_{ij} + b_{ij}.$$

Wir möchten zeigen, daß diese Funktionen die minimierenden Bedingungen erfüllen, dann und nur dann, wenn $a_{ij} = Kb_{ij}$ ist und sich so das Minimum von H_3 unter diesen Beschränkungen ergibt. Zuerst haben wir

$$\log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \Sigma C_{ij} x_i x_j$$

$$\int q(x_i - s_i) \log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \frac{1}{2} \Sigma C_{ij} s_i s_j - \frac{1}{2} \Sigma C_{ij} b_{ij}.$$

Dies soll gleichbedeutend sein mit

$$\frac{H_3}{H_1} \left[\frac{n}{2} \log \frac{1}{2\pi} |A_{ij}| - \frac{1}{2} \Sigma A_{ij} s_i s_j \right],$$

was voraussetzt, daß $A_{ij} = \frac{H_1}{H_3} C_{ij}$.

In diesem Fall ist $A_{ij} = \frac{H_1}{H_2} B_{ij}$, und beide Gleichungen werden identisch.

Anhang 7

Mit dem folgenden wird ein allgemeinerer und klarer bezeichneter Schritt auf dem Weg zu den zentralen Definitionen der Kommunikationstheorie gemacht.

Man stelle sich einen Raum für das Wahrscheinlichkeitsmaß vor, dessen Elemente geordnete Paare (x, y) sind. Die Variablen x, y sind als die möglichen gesendeten und empfangenen Signale irgendeiner längeren Dauer T zu identifizieren. Nennen wir die Menge aller Punkte, deren x zu einer Untermenge S_1 von x Punkten gehört, den Streifen über S_1 und gleichermaßen die Menge, deren y zu S_2 gehören, den Streifen über S_2 . Wir teilen x und y in eine Ansammlung von sich nicht überschneidenden meßbaren Untermengen X_i und Y_i auf, die der Übertragungsrate R angenähert sind durch

$$R_1 = \frac{1}{T} \sum_i P(X_i, Y_i) \log \frac{P(X_i, Y_i)}{P(X_i)P(Y_i)},$$

wobei

$P(X_i)$ das Wahrscheinlichkeitsmaß des Streifens über X_i und

$P(Y_i)$ das Wahrscheinlichkeitsmaß des Streifens über Y_i ist.

$P(X_i, Y_i)$ ist das Wahrscheinlichkeitsmaß des Schnittpunkts der Streifen.

Eine weitere Unterteilung kann R_1 niemals verringern. Denn angenommen X_1 wird unterteilt in $X_1 = X'_1 + X''_1$ und

$$P(Y_1) = a \quad P(X_1) = b + c$$

$$P(X'_1) = b \quad P(X'_1, Y_1) = d$$

$$P(X''_1) = c \quad P(X''_1, Y_1) = e$$

$$P(X_1, Y_1) = d + e.$$

Dann haben wir (für den Schnitt von X_1 und Y_1) in der Summe

$$(d + e) \log \frac{d + e}{a(b + c)} \text{ durch } d \log \frac{d}{ab} + e \log \frac{e}{ac}$$

ersetzt.

Es ist leicht ersichtlich, daß mit der Begrenzung, die wir gegenüber b, c, d und e angenommen haben,

$$\left[\frac{d+e}{b+c} \right]^{d+e} \leq \frac{d^d e^e}{b^d c^e}$$

und folglich die Summe vergrößert ist. Auf diese Weise bilden die verschiedenen möglichen Unterteilungen eine geordnete Menge, wobei R monoton wächst mit einer Verfeinerung der Unterteilung. Wir können R eindeutig als die niedrigste obere Grenze für R_i definieren und sie beschreiben mit

$$R = \frac{1}{T} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)} dx dy.$$

Dieses Integral, im obigen Sinne verstanden, enthält beide Fälle, den kontinuierlichen und den diskreten Fall und natürlich noch viele andere, die nicht in einer der beiden Formen dargestellt werden können. Es ist bei dieser Formulierung trivial, daß, wenn x und u sich eins-zu-eins entsprechen, die Rate von u nach y gleich jener von x nach y ist. Wenn v irgendeine Funktion von y ist (nicht unbedingt mit einer Umkehrung), dann ist die Rate von x nach y gleich oder kleiner als jene von x nach v , da bei der Berechnung der Näherungen die Unterteilungen für y notwendigerweise verfeinerte Unterteilungen jener für v sind. Allgemeiner gesagt, wenn y und v zusammenhängen, nicht funktionell, sondern statistisch, d.h. wir haben einen Raum für das Wahrscheinlichkeitsmaß (y, v) , dann wird $R(x, v) \leq R(x, y)$. Dies bedeutet, daß jede beliebige Operation, die gegenüber dem empfangenen Signal angewandt wird, sogar wenn sie statistische Elemente einbezieht, R nicht vergrößert.

Ein anderer Begriff, der exakt in einer abstrakten Formulierung der Theorie definiert werden sollte, ist jener der "Dimensionsrate"; das ist die mittlere Anzahl von Dimensionen, die pro Sekunde notwendig sind, um ein Element eines Ensembles zu spezifizieren. Im Fall der Bandbegrenzung genügen $2W$ Dimensionen pro Sekunde. Eine allgemeine Definition kann wie folgt ausgedrückt werden: Es sei $f_\alpha(t)$ ein Ensemble von Funktionen und $\rho_T[f_\alpha(t), f_\beta(t)]$ ein metrisches Maß für den "Abstand" zwischen f_α und f_β während der Zeit T (beispielsweise der R.M.S.-Abstand in diesem Intervall). $N(\epsilon, \delta, T)$ sei die Mindestanzahl von Elementen f , die so gewählt werden kann, daß sich alle Elemente des Ensembles abgesehen von einer Menge der Größe δ , innerhalb des Abstands ϵ von zumindest einem der ausgewählten befinden. Auf diese Weise über-

decken wir den Raum innerhalb ϵ , abgesehen von einer Menge der geringen Größe δ . Wir definieren die Dimensionsrate λ für das Ensemble durch den dreifachen Grenzwert

$$\lambda = \lim_{\delta \rightarrow 0} \lim_{\epsilon \rightarrow 0} \lim_{T \rightarrow \infty} \frac{\log N(\epsilon, \delta, T)}{T \log \epsilon}.$$

Dies ist eine Verallgemeinerung der Definitionen für den Maß-Typ der Dimension in der Topologie und stimmt mit der intuitiven Dimensionsrate für einfache Ensembles überein, bei denen das gewünschte Ergebnis offensichtlich ist.