Überblick zu Kapitel 3 von „Einführung in die Informationstheorie"

Im Mittelpunkt dieses Kapitels steht die **Transinformation** $I(X; Y)$ zwischen zwei Zufallsgrößen X und Y, wofür auch andere Begriffe wie *Mutual Information* oder *gegenseitige Entropie*, üblich sind. Bei statistischer Abhängigkeit ist $I(X; Y)$ kleiner als die Einzelentropien $H(X)$ bzw. $H(Y)$. Beispielsweise wird die Unsicherheit hinsichtlich der Zufallsgröße X $\Rightarrow$ Entropie $H(X)$ durch die Kenntnis von Y vermindert, und zwar um den Betrag $H(X|Y)$ $\Rightarrow$ bedingte Entropie von X, falls Y bekannt ist. Der verbleibende Rest ist die Transinformation $I(X; Y)$. Gleichzeitig gilt aber auch $I(X; Y) = H(Y) – H(Y|X)$. Das Semikolon weist auf die Gleichberechtigung der beiden betrachteten Zufallsgrößen X und Y hin.

Im Einzelnen werden im Kapitel 3 behandelt:

- der Zusammenhang zwischen Wahrscheinlichkeiten und Entropie bei *2D–Zufallsgrößen*,
- die Berechnung der *relativen Entropie*, auch als *Kullback–Leibler–Distanz* bekannt,
- die Definition der *Verbundentropie* $H(XY)$ und der *bedingten Entropien* $H(X|Y)$ bzw. $H(Y|X)$,
- die *Transinformation* zwischen zwei Zufallsgrößen (englisch: *Mutual Information*),
- die *Informationstheorie der Digitalsignalübertragung* und das dazugehörige Modell,
- Definition und Bedeutung der *Kanalkapazität*; Zusammenhang mit der Transinformation,
- die Kapazitätsberechnung für *digitale gedächtnislose Kanäle* wie BSC, BEC und BSEC,
- das *Kanalcodierungstheorem*, eines der Highlights der Shannonschen Informationstheorie.

Geeignete Literatur: [AM90] – [Bla87] – [CT06] – [Fan61] – [For72] – [Fri96] – [Gal68] – [Har28] – [Joh92b] – [Kra13] – [McE77] – [Meck09] – [PS02] – [Sha48] – [WZ73]

Die Theorie zum Thema „Transinformation" wird auf 32 Seiten dargelegt. Außerdem beinhaltet dieses dritte Kapitel 61 Grafiken, 14 Aufgaben und sieben Zusatzaufgaben mit insgesamt 106 Teilaufgaben, sowie vier Lernvideos und drei Interaktionsmodule.

Zusammenstellung der **Lernvideos** (LV) zu den Grundlagen und zu Kapitel 3:

- **Klassische Definition der Wahrscheinlichkeit** (Grundlagen, Dauer 5:19)
- **Statistische Abhängigkeit & Unabhängigkeit** (Grundlagen, 3 Teile – Dauer 11:53)
- **Berechnung der Momente bei diskreten Zufallsgrößen** (zu Kapitel 3.3, Dauer 6:32)
- **Übertragungskanal – Eigenschaften, Beschreibungsgrößen** (zu Kapitel 3.3, Dauer 5:50)

Zusammenstellung der **Interaktionsmodule** (IM) zu den Grundlagen und Kapitel 3:

- **Entropien von Nachrichtenquellen** (Grundlage zu diesem Kapitel)
- **Ereigniswahrscheinlichkeiten der Binomialverteilung** (zu Kapitel 3.1)
- **Transinformation zwischen wertdiskreten Zufallsgrößen** (zu Kapitel 3.3)

Weitere Informationen zu diesem Thema sowie Simulationsprogramme mit Grafikausgaben und Aufgaben mit ausführlichen Musterlösungen finden Sie im letzten Versuch „Wertdiskrete Informationstheorie" des Praktikums *Simulation digitaler Übertragungssysteme* von Prof. Günter Söder an der TU München:

Herunterladen des Windows–Programms „WDIT" (Zip–Version)

Herunterladen der dazugehörigen Texte (PDF–Datei)

Einführungsbeispiel: Statistische Abhängigkeit von Zufallsgrößen

Wir gehen vom Experiment „Würfeln mit zwei Würfeln" aus, wobei beide Würfel unterscheidbar sind. Die untere Tabelle zeigt als Ergebnis die ersten $N = 18$ Wurfpaare dieses exemplarischen Zufallsexperiments:

- In Zeile 2 sind die Augenzahlen des roten Würfels (R) angegeben. Der Mittelwert dieser begrenzten Folge $\langle R_1, \text{ ... }, R_{18}\rangle$ ist mit 3.39 etwas kleiner als der Erwartungswert $\text{E}[R] = 3.5$.
- Die Zeile 3 zeigt die Augenzahlen des blauen Würfels (B). Die Folge $\langle B_1, \text{ ... }, B_{18}\rangle$ hat mit 3.61 einen etwas größeren Mittelwert als die unbegrenzte Folge $\Rightarrow \text{E}[B] = 3.5$.
- Zeile 4 beinhaltet die Summe $S_\nu = R_\nu + B_\nu$. Der Mittelwert der Folge $\langle S_1, \text{ ... }, S_{18}\rangle$ ist 3.39 + 3.61 = 7. Dieser ist hier (zufällig) gleich dem Erwartungswert $\text{E}[S] = \text{E}[R] + \text{E}[B]$.

Hinweis: Entsprechend der auf der **nachfolgenden Seite** erklärten Nomenklatur sind hier R_ν, B_ν und S_ν als Zufallsgrößen zu verstehen. Die Zufallsgröße $R_3 = \{1, 2, 3, 4, 5, 6\}$ gibt beispielsweise die Augenzahl des roten Würfels beim dritten Wurf als Wahrscheinlichkeitsereignis an. Die Angabe „$R_3 = 6$" sagt aus, dass bei der dokumentierten Realisierung der rote Würfel im dritten Wurf eine „6" gezeigt hat.

Lfd. Nummer (ν)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Roter Würfel (R_ν)	5	4	6	3	1	2	5	1	3	3	6	3	2	5	1	2	4	5
Blauer Würfel (B_ν)	6	2	2	4	3	1	5	6	2	3	6	4	3	4	1	5	5	3
Würfelsumme (S_ν)	11	6	8	7	4	3	10	7	5	6	12	7	5	9	2	7	9	8

Nun stellt sich die Frage, zwischen welchen Zufallsgrößen es statistische Abhängigkeiten gibt:

- Setzt man faire Würfel voraus, so bestehen zwischen den Folgen $\langle R\rangle$ und $\langle B\rangle$ – ob begrenzt oder unbegrenzt – keine statistischen Bindungen: Auch wenn man R_ν kennt, sind für B_ν weiterhin alle möglichen Augenzahlen 1, ... , 6 gleichwahrscheinlich.
- Kennt man aber S_ν, so sind sowohl Aussagen über R_ν als auch über B_ν möglich. Aus $S_{11} = 12$ (siehe obige Tabelle) folgt direkt $R_{11} = B_{11} = 6$ und die Summe $S_{15} = 2$ zweier Würfel ist nur mit zwei Einsen möglich. Solche Abhängigkeiten bezeichnet man als *deterministisch*.
- Aus $S_7 = 10$ lassen sich zumindest Bereiche für R_7 und B_7 angeben: $R_7 \geq 4$, $B_7 \geq 4$. Möglich sind dann nur die drei Wertepaare $(R_7 = 4) \cap (B_7 = 6)$, $(R_7 = 5) \cap (B_7 = 5)$ sowie $(R_7 = 6) \cap (B_7 = 4)$. Hier besteht zwar kein deterministischer Zusammenhang zwischen den Zufallsgrößen S_ν und R_ν (bzw. B_ν), aber eine so genannte *statistische Abhängigkeit*.
- Solche statistische Abhängigkeiten gibt es für alle $S_\nu \in \{3, 4, 5, 6, 8, 9, 10, 11\}$. Ist dagegen die Summe $S_\nu = 7$, so kann daraus nicht auf R_ν und B_ν zurückgeschlossen werden. Für beide Würfel sind dann alle möglichen Augenzahlen (1, ... , 6) gleichwahrscheinlich. In diesem Fall bestehen auch keine statistischen Bindungen zwischen S_ν und R_ν bzw. S_ν und B_ν.

Voraussetzungen und Nomenklatur

Im gesamten Kapitel 3 betrachten wir wertdiskrete Zufallsgrößen der Form

$$X = \{x_1, x_2, \ \text{...} \ , x_\mu, \ \text{...} \ , x_M\} ,$$

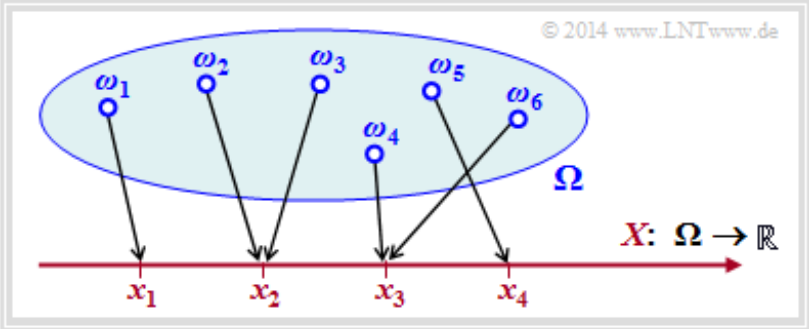
und verwenden folgende Nomenklatur:

- Die Zufallsgröße selbst wird stets mit einem Großbuchstaben bezeichnet, und der Kleinbuchstabe x weist auf eine mögliche Realisierung der Zufallsgröße X hin.
- Alle Realisierungen x_μ (mit $\mu = 1, \ \text{...} \ , M$) sind reellwertig. M gibt den Symbolumfang (englisch: *Symbol Set Size*) von X an. Anstelle von M verwenden wir manchmal auch $|X|$.

Die Zufallsgröße X kann zum Beispiel durch die Transformation $\Omega \ \rightarrow \ X$ entstanden sein, wobei Ω für den Wahrscheinlichkeitsraum eines Zufallsexperiments steht. Die nachfolgende Grafik verdeutlicht eine solche Transformation:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \ \text{...} \ \} \quad \longmapsto \quad X = \{x_1, x_2, x_3, x_4\} \subset \mathcal{R} .$$

Jedes Zufallsereignis $\omega_i \ \in \Omega$ wird eindeutig einem reellen Zahlenwert $x_\mu \ \in X \subset \mathbb{R}$ zugeordnet. Im betrachteten Beispiel gilt für die Laufvariable $1 \leq \mu \leq 4$, das heißt, der Symbolumfang beträgt $M = |X| = 4$. Die Abbildung ist aber nicht eineindeutig: Die Realisierung $x_3 \in X$ könnte sich im Beispiel aus dem Elementarereignis ω_4 ergeben haben, aber auch aus ω_6 (oder aus einem anderen der unendlich vielen, in der Grafik nicht eingezeichneten Elementarereignisse ω_i).



Oft verzichtet man auf die Indizierung sowohl der Elementarereignisse ω_i als auch der Realisierungen x_μ. Damit ergeben sich beispielsweise folgende Kurzschreibweisen:

$$\{X = x\} \quad \equiv \quad \{\omega \in \Omega : \quad X(\omega) = x\} ,$$
$$\{X \leq x\} \quad \equiv \quad \{\omega \in \Omega : \quad X(\omega) \leq x\} .$$

Mit dieser Vereinbarung gilt für die Wahrscheinlichkeiten der diskreten Zufallsgröße:

$$\Pr(X = x_\mu) = \sum_{\omega \ \in \{X = x_\mu\}} \Pr\left(\{\omega\}\right) .$$

Wahrscheinlichkeitsfunktion und Wahrscheinlichkeitsdichtefunktion

Definition: Fasst man die M Wahrscheinlichkeiten einer diskreten Zufallsgröße $X \Rightarrow \Pr(X = x_\mu)$ ähnlich wie bei einem Vektor zusammen, so kommt man zur **Wahrscheinlichkeitsfunktion** (englisch: *Probability Mass Function*, PMF):

$$P_X(X) = [\, P_X(x_1), P_X(x_2), \, \text{...} \, , P_X(x_\mu), \, \text{...} \, , P_X(x_M) \,] \, .$$

Das μ–te Element dieses „Vektors" gibt dabei die folgende Wahrscheinlichkeit an:

$$P_X(x_\mu) = \Pr(X = x_\mu) \, .$$

Im Buch „Stochastische Signaltheorie" haben wir mit der **Wahrscheinlichkeitsdichtefunktion** (WDF, englisch: *Probability Density Function*, PDF) eine ähnliche Beschreibungsgröße definiert und diese mit $f_X(x)$ bezeichnet.

Zu beachten ist aber:

- Die PDF eignet sich eher zur Charakterisierung kontinuierlicher Zufallsgrößen, wie zum Beispiel bei einer Gaußverteilung oder einer Gleichverteilung. Erst durch die Verwendung von Diracfunktionen wird die PDF auch für diskrete Zufallsgrößen anwendbar.
- Die PMF liefert weniger Information über die Zufallsgröße als die PDF und kann zudem nur für diskrete Größen angegeben werden. Für die wertdiskrete Informationstheorie ist sie ausreichend.

Beispiel: Wir betrachten eine Wahrscheinlichkeitsdichtefunktion (abgekürzt WDF bzw. PDF) ohne großen Praxisbezug:

$$f_X(x) = 0.2 \cdot (x + 2) + 0.3 \cdot (x - 1.5) + 0.5 \cdot (x - \pi) \, .$$

Für die diskrete Zufallsgröße gilt somit $x \in X = \{–2, +1.5, +\pi\} \Rightarrow$ Symbolumfang $M = |X| = 3$, und die Wahrscheinlichkeitsfunktion (PMF) lautet:

$$P_X(X) = [\, 0.2 \, , 0.3 \, , 0.5 \,] \, .$$

Man erkennt:

- Die PMF liefert nur Informationen über die Wahrscheinlichkeiten $\Pr(x_1)$, $\Pr(x_2)$, $\Pr(x_3)$. Aus der PDF sind dagegen auch die möglichen Realisierungen x_1, x_2, x_3 der Zufallsgröße X ablesbar.
- Die einzige Voraussetzung für die Zufallsgröße ist, dass sie reellwertig ist. Die möglichen Werte x_μ müssen weder positiv, ganzzahlig, äquidistant noch rational sein.

Wahrscheinlichkeitsfunktion und Entropie (1)

In der wertdiskreten Informationstheorie genügt im Gegensatz zu übertragungstechnischen Problemen schon die Kenntnis der Wahrscheinlichkeitsfunktion $P_X(X)$, zum Beispiel zur Berechnung der **Entropie.**

Die **Entropie** einer diskreten Zufallsgröße X – also deren Unsicherheit für einen Beobachter – kann man mit der Wahrscheinlichkeitsfunktion $P_X(X)$ wie folgt darstellen:

$$H(X) = \mathrm{E}\left[\log \frac{1}{P_X(X)}\right] = \sum_{\mu=1}^{M} P_X(x_\mu) \cdot \log \frac{1}{P_X(x_\mu)}.$$

Verwendet man den Logarithmus zur Basis 2, also $\log_2$ (...) = ld (...) ⇒ *Logarithmus dualis*, so wird der Zahlenwert mit der Pseudo–Einheit „bit" versehen. E[...] gibt den *Erwartungswert* an.

Beispielsweise erhält man

- für $P_X(X) = [0.2, 0.3, 0.5]$:

$$H(X) = 0.2 \cdot \log_2 \frac{1}{0.2} + 0.3 \cdot \log_2 \frac{1}{0.3} + 0.5 \cdot \log_2 \frac{1}{0.5} \approx 1.485\,\text{bit},$$

- für $P_X(X) = [1/3, 1/3, 1/3]$:

$$H(X) = 3 \cdot 1/3 \cdot \log_2 (3) = \log_2 (3) \approx 1.585\,\text{bit}.$$

Das zweite Beispiel liefert das Maximum der Entropiefunktion für den Symbolumfang $M = 3$. Für ein allgemeines M lässt sich dieses Ergebnis beispielsweise wie folgt herleiten – siehe **[Meck09]:**

$$H(X) = \mathrm{E}\left[\log \frac{1}{P_X(X)}\right] \le \log\left[\mathrm{E}\left[\frac{1}{P_X(X)}\right]\right].$$

Diese Abschätzung (Jensens's Ungleichung) ist zulässig, da der Logarithmus eine konkave Funktion ist. Entsprechend **Aufgabe A3.2** gilt:

$$\mathrm{E}\left[\frac{1}{P_X(X)}\right] \le M \quad \Rightarrow \quad H(X) \le \log (M).$$

Das Gleichheitszeichen ergibt sich nach der oberen Rechnung für gleiche Wahrscheinlichkeiten, also für $P_X(x_\mu) = 1/M$ für alle μ. In der **Aufgabe A3.3** soll der gleiche Sachverhalt unter Verwendung der Abschätzung

$$\ln (x) \le x - 1$$

nachgewiesen werden. Das Gleichheitszeichen gilt nur für $x = 1$.

Ist eine der M Wahrscheinlichkeiten $P_X(x_\mu)$ der Wahrscheinlichkeitsfunktion $P_X(X)$ gleich 0 ist, so lässt sich für die Entropie eine engere Schranke angeben:

$$H(X) \le \log (M - 1).$$

Wahrscheinlichkeitsfunktion und Entropie (2)

Für das folgende Beispiel und die nächsten Seiten vereinbaren wir die folgende Nomenklatur:

- Die Entropie $H(X)$ bezieht sich stets auf die tatsächliche Wahrscheinlichkeitsfunktion $P_X(X)$ der diskreten Zufallsgröße. Experimentell erhält man diese Größen erst nach $N \to \infty$ Versuchen.
- Ermittelt man die Wahrscheinlichkeitsfunktion aus einer endlichen Zufallsfolge, so bezeichnen wir diese mit $Q_X(X)$ und die daraus resultierende Entropie versehen wir mit dem Zusatz „N = ...".
- Diese Entropie–Näherung basiert nicht auf Wahrscheinlichkeiten, sondern nur auf den **relativen Häufigkeiten.** Erst für $N \to \infty$ stimmt diese Näherung mit $H(X)$ überein.

Beispiel A: Kommen wir auf unser *Würfel–Experiment* zurück. Die nachfolgende Tabelle zeigt die Wahrscheinlichkeitsfunktionen $P_R(R)$ und $P_B(B)$ für den roten und den blauen Würfel sowie die Näherungen $Q_R(R)$ und $Q_B(B)$, jeweils basierend auf dem Zufallsexperiment mit $N = 18$ Würfen. Die relativen Häufigkeiten $Q_R(R)$ und $Q_B(B)$ ergeben sich aus den **beispielhaften Zufallsfolgen** vom Beginn dieses Kapitels.

Augenzahl r_μ	1	2	3	4	5	6
$P_R(R = r_\mu)$	1/6	1/6	1/6	1/6	1/6	1/6
$Q_R(R = r_\mu)$	3/18	3/18	4/18	2/18	4/18	2/18

Augenzahl b_μ	1	2	3	4	5	6
$P_B(B = b_\mu)$	1/6	1/6	1/6	1/6	1/6	1/6
$Q_B(B = b_\mu)$	2/18	3/18	4/18	3/18	3/18	3/18

Für die Zufallsgröße R gilt mit dem *Logarithmus dualis* (zur Basis 2):

$$H(R) = H(R)\big|_{N\to\infty} = \sum_{\mu=1}^{6} 1/6 \cdot \log_2\,(6) = \log_2\,(6) = 2.585\,\text{bit}\,,$$

$$H(R)\big|_{N=18} = 2 \cdot \frac{2}{18} \cdot \log_2\,\frac{18}{2} \;+\; 2 \cdot \frac{3}{18} \cdot \log_2\,\frac{18}{3} \;+\; 2 \cdot \frac{4}{18} \cdot \log_2\,\frac{18}{4} \;= 2.530\,\text{bit}\,.$$

Der blaue Würfel hat natürlich die gleiche Entropie: $H(B) = H(R) = 2.585$ bit. Hier erhält man für die auf $N = 18$ basierende Näherung einen etwas größeren Wert, da nach obiger Tabelle $Q_B(B)$ von der diskreten (M=6)–Gleichverteilung $P_B(B)$ weniger abweicht als $Q_R(R)$ von $P_R(R)$.

$$H(B)\big|_{N=18} = 1 \cdot \frac{2}{18} \cdot \log_2\,\frac{18}{2} \;+\; 4 \cdot \frac{3}{18} \cdot \log_2\,\frac{18}{3} \;+\; 1 \cdot \frac{4}{18} \cdot \log_2\,\frac{18}{4} \;= 2.558\,\text{bit}\,.$$

Man erkennt aus den angegebenen Zahlenwerten, dass trotz des eigentlich viel zu kleinen Experimentparameters N die Verfälschungen hinsichtlich der Entropie nicht sehr groß sind.

Es soll nochmals erwähnt werden, dass bei endlichem N stets gilt:

$$H(R)\big|_N < H(R) = \log_2\,(6)\,, \quad H(B)\big|_N < H(B) = \log_2\,(6)\,.$$

Relative Entropie – Kullback–Leibler–Distanzen (1)

Wir betrachten die beiden Wahrscheinlichkeitsfunktionen $P_X(\cdot)$ und $P_Y(\cdot)$ über dem gleichen Alphabet $X = \{x_1, x_2, \ldots, x_M\}$, und definieren nun die **relative Entropie** (englisch: *Informational Divergence*) zwischen diesen:

$$D(P_X \,||\, P_Y) = \mathrm{E}\left[\log \frac{P_X(X)}{P_Y(X)}\right] = \sum_{\mu=1}^{M} P_X(x_\mu) \cdot \log \frac{P_X(x_\mu)}{P_Y(x_\mu)}.$$

Bei Verwendung des Logarithmus zur Basis 2 ist wieder die Pseudo–Einheit „bit” hinzuzufügen. Man bezeichnet $D(P_X \,||\, P_Y)$ auch als die **Kullback–Leibler–Distanz** (kurz KL–Distanz). Diese liefert ein Maß für die „Ähnlichkeit” zwischen den beiden Wahrscheinlichkeitsfunktionen $P_X(\cdot)$ und $P_Y(\cdot)$:

In ähnlicher Weise lässt sich auch eine zweiteVariante der Kullback–Leibler–Distanz angeben:

$$D(P_Y \,||\, P_X) = \mathrm{E}\left[\log \frac{P_Y(X)}{P_X(X)}\right] = \sum_{\mu=1}^{M} P_Y(x_\mu) \cdot \log \frac{P_Y(x_\mu)}{P_X(x_\mu)}.$$

Gegenüber der ersten Variante ist jede Funktion $P_X(\cdot)$ durch $P_Y(\cdot)$ ersetzt und umgekehrt. Da sich im allgemeinen $D(P_X \,||\, P_Y)$ und $D(P_Y \,||\, P_X)$ unterscheiden, ist der Begriff „Distanz” eigentlich irreführend. Wir wollen es aber bei dieser Namensgebung belassen.

Wertet man die beiden obigen Gleichungen aus, so erkennt man folgende Eigenschaften:

- Liegt genau die gleiche Verteilung vor $\Rightarrow$ $P_Y(\cdot) \equiv P_X(\cdot)$, so ist $D(P_X \,||\, P_Y) = 0$. In allen anderen Fällen ist $D(P_X \,||\, P_Y) > 0$. Gleiches gilt für die zweite Variante $D(P_Y \,||\, P_X)$.
- Gilt $P_X(x_\mu) \neq 0$ und $P_Y(x_\mu) = 0$ (es genügt ein einziges und ein beliebiges μ), so ergibt sich für die Kullback–Leibler–Distanz $D(P_X \,||\, P_Y)$ ein unendlich großer Wert.
- In diesem Fall ist $D(P_Y \,||\, P_X)$ nicht notwendigerweise ebenfalls unendlich. Diese Aussage macht nochmals deutlich, dass im allgemeinen $D(P_X \,||\, P_Y)$ ungleich $D(P_Y \,||\, P_X)$ sein wird.

Auf der nächsten Seite werden diese beiden Definitionen an unserem Standardbeispiel *Würfel–Experiment* verdeutlicht. Gleichzeitig verweisen wir auf folgende Aufgaben:

A3.4: Kullback–Leibler–Distanz zur Binomialverteilung
Z3.4: Nochmals Kullback–Leibler–Distanz
A3.5: Partitionierungsungleichung

Relative Entropie – Kullback–Leibler–Distanzen (2)

Beispiel B: Für unser Würfel–Experiment haben wir **folgende** Wahrscheinlichkeitsfunktionen $P_R(\cdot)$ und $P_B(\cdot)$ sowie deren Näherungen $Q_R(\cdot)$ und $Q_B(\cdot)$ definiert. Die Zufallsgröße R bezeichnet hierbei die Augenzahl des roten Würfels und B die Augenzahl des blauen Würfels. Die Näherungen $Q_R(\cdot)$ und $Q_B(\cdot)$ ergeben sich aus dem früher beschriebenen **Experiment** mit lediglich $N = 18$ Doppelwürfen.

Augenzahl r_μ	1	2	3	4	5	6
$P_R(R = r_\mu)$	1/6	1/6	1/6	1/6	1/6	1/6
$Q_R(R = r_\mu)$	3/18	3/18	4/18	2/18	4/18	2/18

Augenzahl b_μ	1	2	3	4	5	6
$P_B(B = b_\mu)$	1/6	1/6	1/6	1/6	1/6	1/6
$Q_B(B = b_\mu)$	2/18	3/18	4/18	3/18	3/18	3/18

- Da die Wahrscheinlichkeitsfunktionen $P_R(\cdot)$ und $P_B(\cdot)$ identisch sind, erhält man für die oben definierten Kullback–Leibler–Distanzen $D(P_R \,||\, P_B)$ und $D(P_B \,||\, P_R)$ jeweils 0.
- Der Vergleich von $P_R(\cdot)$, $Q_R(\cdot)$ ergibt für die erste Variante der Kullback–Leibler–Distanz

$$D(P_R||Q_R) = \mathrm{E}\left[\log_2 \frac{P_R(\cdot)}{Q_R(\cdot)}\right] = \sum_{\mu=1}^{6} P_R(r_\mu) \cdot \log \frac{P_R(r_\mu)}{Q_R(r_\mu)} =$$
$$= \frac{1}{6} \cdot \left[2 \cdot \log_2 \frac{1/6}{2/18} + 2 \cdot \log_2 \frac{1/6}{3/18} + 2 \cdot \log_2 \frac{1/6}{4/18}\right] =$$
$$= 1/6 \cdot \left[2 \cdot 0.585 + 2 \cdot 0 - 2 \cdot 0.415\right] \approx 0.0570\,\text{bit}.$$

  Hierbei wurde bei der vorzunehmenden Erwartungswertbildung die Tatsache ausgenutzt, dass wegen $P_R(r_1) = \ldots = P_R(r_6)$ der Faktor 1/6 ausgeklammert werden kann. Da hier der Logarithmus zur Basis 2 verwendet wurde, ist die Pseudo–Einheit „bit” angefügt.
- Für die zweite Variante der Kullback–Leibler–Distanz ergibt sich ein etwas anderer Wert:

$$D(Q_R||P_R) = \mathrm{E}\left[\log_2 \frac{Q_R(\cdot)}{P_R(\cdot)}\right] = \sum_{\mu=1}^{6} Q_R(r_\mu) \cdot \log \frac{Q_R(r_\mu)}{P_R(r_\mu)} =$$
$$= 2 \cdot \frac{2}{18} \cdot \log_2 \frac{2/18}{1/6} + 2 \cdot \frac{3}{18} \cdot \log_2 \frac{3/18}{1/6} + 2 \cdot \frac{4}{18} \cdot \log_2 \frac{4/18}{1/6} =$$
$$= 4/18 \cdot (-0.585) + 3/18 \cdot 0 + 8/18 \cdot 0.415 \approx 0.0544\,\text{bit}.$$

- Für den blauen Würfel erhält man $D(P_B \,||\, Q_B) \approx 0.0283$ bit und $D(Q_B \,||\, P_B) \approx 0.0271$ bit, also etwas kleinere KL–Distanzen, da sich die Approximation $Q_B(\cdot)$ von $P_B(\cdot)$ weniger unterscheidet als $Q_R(\cdot)$ von $P_R(\cdot)$.
- Vergleicht man die Häufigkeiten $Q_R(\cdot)$ und $Q_B(\cdot)$, so erhält man $D(Q_R \,||\, Q_B) \approx 0.0597$ bit und $D(Q_B \,||\, Q_R) \approx 0.0608$ bit. Hier sind die Distanzen am größten, da die Unterschiede zwischen Q_B und Q_R größer sind als zwischen Q_R und P_R oder zwischen Q_B und P_B.

Verbundwahrscheinlichkeit und Verbundentropie (1)

Für den Rest von Kapitel 3 betrachten wir stets zwei diskrete Zufallsgrößen $X = \{x_1, x_2, \ ... \ , x_M\}$ und $Y = \{y_1, y_2, \ ... \ , y_K\}$, deren Wertebereiche nicht notwendigerweise übereinstimmen müssen. Das heißt: $K \neq M$ (in anderer Notation: $|Y| \neq |X|$) ist durchaus erlaubt.

Die Wahrscheinlichkeitsfunktion hat somit eine $K \times M$–Matrixform mit den Elementen

$$P_{XY}(X = x_\mu, Y = y_\kappa) = \Pr\big[(X = x_\mu) \cap (Y = y_\kappa)\big] .$$

Als Kurzschreibweise verwenden wir $P_{XY}(X, Y)$, wobei XY als neue Zufallsgröße zu interpretieren ist, die sowohl die Eigenschaften von X als auch diejenigen von Y beinhaltet.

Definition: Die **Verbundentropie** (englisch: *Joint Entropy*) lässt sich als ein Erwartungswert mit der 2D–Wahrscheinlichkeitsfunktion $P_{XY}(X, Y)$ wie folgt darstellen:

$$H(XY) = \mathrm{E}\left[\log \frac{1}{P_{XY}(X, Y)}\right] = \sum_{\mu=1}^{M} \sum_{\kappa=1}^{K} P_{XY}(x_\mu, y_\kappa) \cdot \log \frac{1}{P_{XY}(x_\mu, y_\kappa)} .$$

Im Folgenden verwenden wir durchgehend den Logarithmus zur Basis 2 $\Rightarrow$ $\log(x) \rightarrow \log_2(x) = \mathrm{ld}(x)$ $\Rightarrow$ *Logarithmus dualis*. Der Zahlenwert ist somit mit der Pseudo–Einheit „bit” zu versehen.

Allgemein kann für die Verbundentropie die folgende **obere Schranke** angegegeben werden:

$$H(XY) \leq H(X) + H(Y) .$$

Diese Ungleichung drückt folgenden Sachverhalt aus:

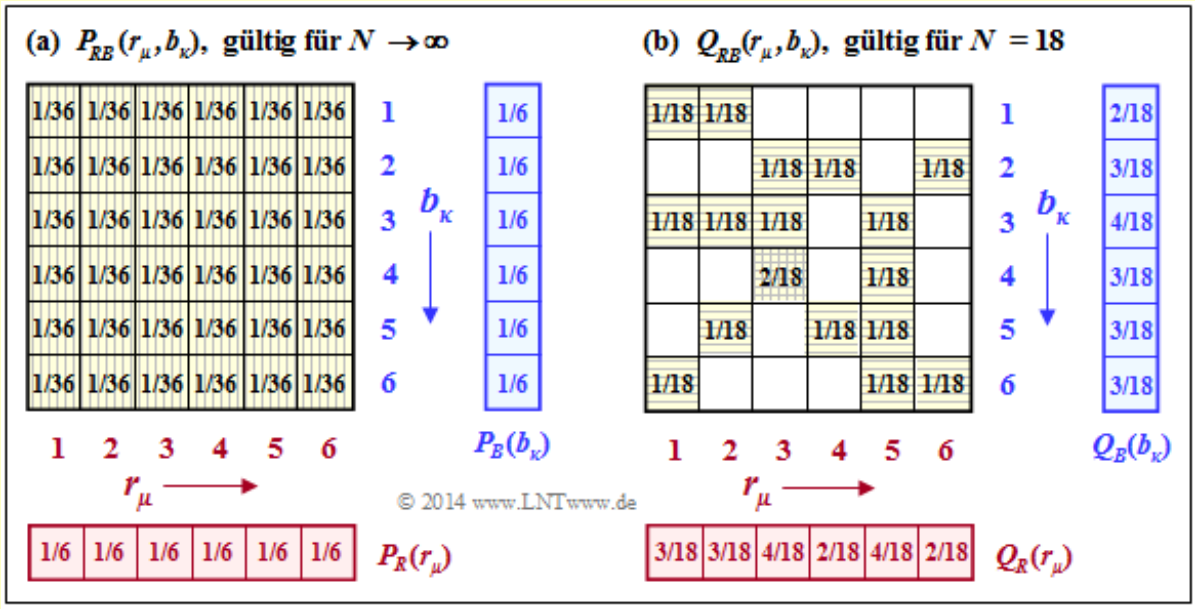
- Das Gleichheitszeichen gilt nur für den Sonderfall statistisch unabhängiger Zufallsgrößen, wie im **Beispiel C** auf der nächsten Seite anhand der Zufallsgrößen R und B demonstriert wird. Hierbei bezeichnen R und B die Augenzahlen eines roten und eines blauen Würfels.

- Gibt es dagegen statistische Abhängigkeiten wie im **Beispiel D** zwischen den Zufallsgrößen R und $S = R + B$, so gilt in obiger Gleichung das „<”–Zeichen: $H(RS) < H(R) + H(S)$.

In den Beispielen wird auch gezeigt, in wie weit sich die Verbundentropien $H(RB)$ und $H(RS)$ ändern, wenn man beim Würfel–Experiment nicht unendlich viele Wurfpaare ermittelt, sondern lediglich $N = 18$.

Verbundwahrscheinlichkeit und Verbundentropie (2)

Beispiel C: Wir kommen wieder auf das **Würfel–Experiment** zurück: Die Zufallsgrößen sind die Augenzahlen des roten und des blauen Würfels: $R = \{1, 2, 3, 4, 5, 6\}$, $B = \{1, 2, 3, 4, 5, 6\}$.



Die linke Grafik zeigt die Wahrscheinlichkeiten $P_{RB}(\cdot)$, die sich für alle $\mu = 1, \ldots, 6$ und für alle $\kappa = 1, \ldots, 6$ gleichermaßen zu 1/36 ergeben. Damit erhält man für die Verbundentropie:

$$H(RB) = H(RB)\big|_{N\to\infty} = \log_2\,(36) = 5.170\,\text{bit}.$$

Man erkennt aus obiger Grafik und der hier angegebenen Gleichung:

- Da R und B statistisch voneinander unabhängig sind, gilt $P_{RB}(R, B) = P_R(R) \cdot P_B(B)$.
- Die Verbundentropie ist die Summe der beiden Einzelentropien: $H(RB) = H(R) + H(B)$.

Die rechte Grafik zeigt die angenäherte 2D–Wahrscheinlichkeitsfunktion $Q_{RB}(\cdot)$, basierend auf den nur $N = 18$ Wurfpaaren unseres Experiments:

- Hier ergibt sich keine quadratische Form der Verbundwahrscheinlichkeit $Q_{RB}(\cdot)$, und die daraus abgeleitete Verbundentropie ist deutlich kleiner als $H(RB)$:

$$H(RB)\big|_{N=18} = 16 \cdot \frac{1}{18} \cdot \log_2 \frac{18}{1} + 1 \cdot \frac{2}{18} \cdot \log_2 \frac{18}{2} = 4.059\,\text{bit}.$$

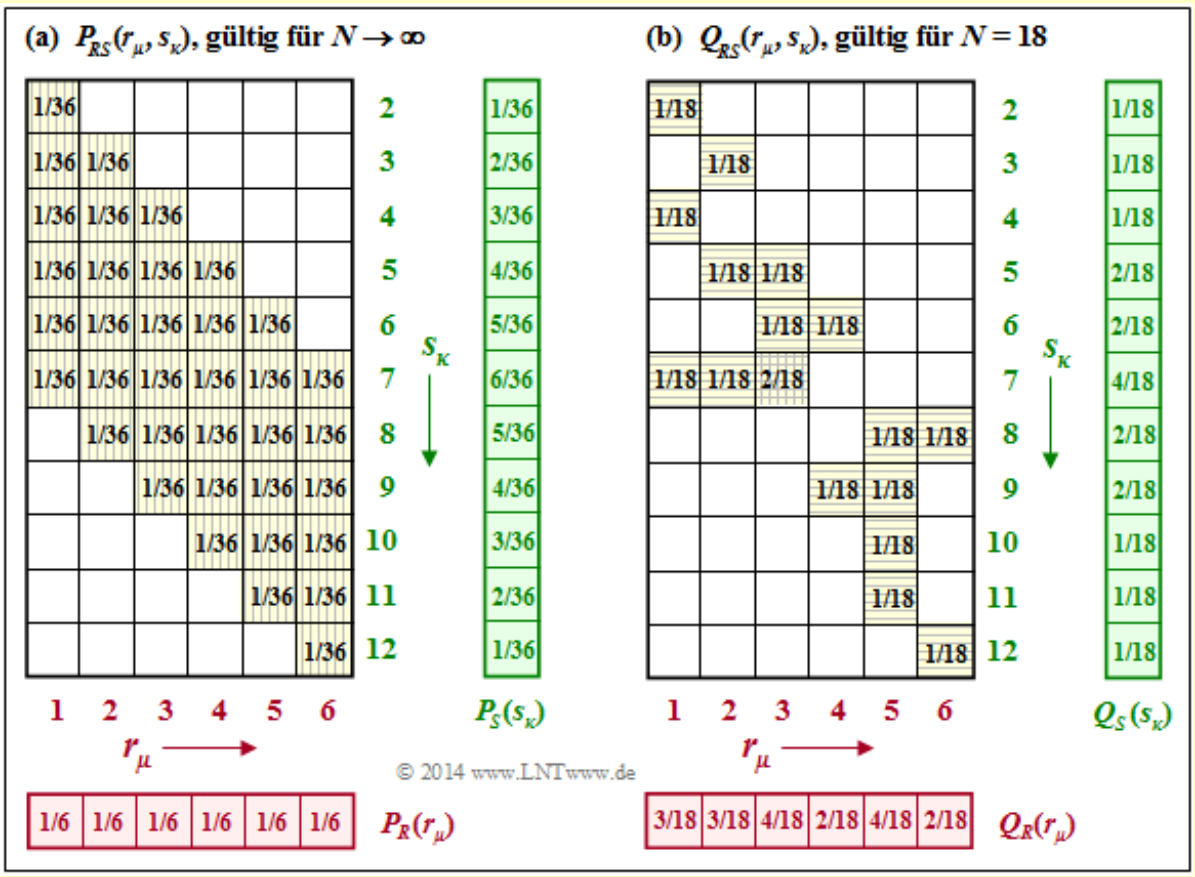
Verbundwahrscheinlichkeit und Verbundentropie (3)

Beispiel D: Bei unserem **Würfel–Experiment** haben wir neben den beiden Zufallsgrößen R (roter Würfel) und B (blauer Würfel) auch die Summe $S = R + B$ betrachtet. Die linke Grafik zeigt, dass $P_{RS}(\cdot)$ nicht als Produkt von $P_R(\cdot)$ und $P_S(\cdot)$ geschrieben werden kann. Mit den Wahrscheinlichkeitsfunktionen

$$P_R(R) = [\,1/6, 1/6, 1/6, 1/6, 1/6, 1/6\,],$$
$$P_S(S) = [\,1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36\,]$$

erhält man für die Entropien:

$$H(R) = \log_2(6) \approx 2.585\,\text{bit},$$
$$H(S) = 2 \cdot \frac{1}{36} \cdot \log_2 \frac{36}{1} + 2 \cdot \frac{2}{36} \cdot \log_2 \frac{36}{2} + 2 \cdot \frac{3}{36} \cdot \log_2 \frac{36}{3} + 2 \cdot \frac{4}{36} \cdot \log_2 \frac{36}{4} +$$
$$+ 2 \cdot \frac{5}{36} \cdot \log_2 \frac{36}{5} + 1 \cdot \frac{6}{36} \cdot \log_2 \frac{36}{6} \approx 3.274\,\text{bit},$$
$$H(RS) = \log_2(36) \approx 5.170\,\text{bit}.$$



Aus diesen Zahlenwerten erkennt man:

- Aufgrund der statistischen Abhängigkeit zwischen dem roten Würfel und der Summe ist die Verbundentropie $H(RS) \approx 5.170$ bit kleiner als $H(R) + H(S) \approx 5.877$ bit.
- Der Vergleich mit **Beispiel C** zeigt, dass $H(RS)$ gleich $H(RB)$ ist. Der Grund ist, dass bei Kenntnis von R die Zufallsgrößen B und S genau die gleichen Informationen liefern.

Rechts dargestellt ist der Fall, dass die 2D–PMF aus nur $N = 18$ Wurfpaaren empirisch ermittelt

wurde $\Rightarrow Q_{RS}(\cdot)$. Obwohl sich aufgrund des sehr kleinen N–Wertes ein völlig anderes Bild ergibt, liefert die Näherung für $H(RS)$ den exakt gleichen Wert wie die Näherung für $H(RB)$:

$$H(RS)\big|_{N=18} = H(RB)\big|_{N=18} = 4.059\ \text{bit}.$$

Definition der Entropie unter Verwendung von supp(P_{XY})

Wir fassen die Ergebnisse des letzten Abschnitts nochmals kurz zusammen, wobei wir von der zweidimensionalen Zufallsgröße XY mit der Wahrscheinlichkeitsfunktion $P_{XY}(X, Y)$ ausgehen. Gleichzeitig verwenden wir die Schreibweise

$$\text{supp}(P_{XY}) = \left\{ (x, y) \in XY\,, \quad \text{wobei } P_{XY}(X, Y) \neq 0 \right\}.$$

Mit dieser Teilmenge supp(P_{XY}) $\subset P_{XY}$ gilt für

- die **Verbundentropie** (englisch: *Joint Entropy*):

$$H(XY) = \text{E}\left[\log_2 \frac{1}{P_{XY}(X, Y)}\right] = \sum_{(x,y)\,\in\,\text{supp}(P_{XY})} P_{XY}(x, y) \cdot \log_2 \frac{1}{P_{XY}(x, y)}.$$

- die **Entropien der 1D–Zufallsgrößen** X und Y:

$$H(X) = \text{E}\left[\log_2 \frac{1}{P_X(X)}\right] = \sum_{x\,\in\,\text{supp}(P_X)} P_X(x) \cdot \log_2 \frac{1}{P_X(x)},$$

$$H(Y) = \text{E}\left[\log_2 \frac{1}{P_Y(Y)}\right] = \sum_{y\,\in\,\text{supp}(P_Y)} P_Y(y) \cdot \log_2 \frac{1}{P_Y(y)}.$$

Beispiel E: Bei der zweidimensionalen (2D) **Wahrscheinlichkeitsfunktion $P_{RS}(R, S)$** unseres Würfel–Experimentes mit

- R: Augenzahl des roten Würfels,
- S: Summe der beiden Würfel R und B

gibt es $6 \cdot 11 = 66$ Felder, von denen viele leer sind ⇒ Wahrscheinlichkeit 0. Die Teilmenge supp(P_{RS}) beinhaltet dagegen nur die 36 schraffierten Felder mit von 0 verschiedenen Wahrscheinlichkeiten.

Die Entropie bleibt gleich, ganz egal, ob man die Mittelung über alle Elemente von P_{RS} oder nur über die Elemente von supp(P_{RS}) erstreckt, da $x \cdot \log_2 (1/x)$ für $x \to 0$ gleich 0 ergibt.

Dagegen sind bei der **2D–Wahrscheinlichkeitsfunktion $P_{RB}(R, B)$** mit

- R: Augenzahl des roten Würels,
- B: Augenzahl des blauen Würfels

die Mengen P_{RB} und supp(P_{RB}) identisch. Hier sind alle $6^2 = 36$ Felder mit Werten ≠ 0 belegt.

Bedingte Wahrscheinlichkeit und bedingte Entropie (1)

Im Buch „Stochastische Signaltheorie" wurden für den Fall zweier Ereignisse X und Y die folgenden **bedingten Wahrscheinlichkeiten** angegeben ⇒ Satz von **Bayes:**

$$\Pr(X \mid Y) = \frac{\Pr(X \cap Y)}{\Pr(Y)}, \quad \Pr(Y \mid X) = \frac{\Pr(X \cap Y)}{\Pr(X)}.$$

Angewendet auf Wahrscheinlichkeitsfunktionen erhält man somit:

$$P_{X|Y}(X \mid Y) = \frac{P_{XY}(X, Y)}{P_Y(Y)}, \quad P_{Y|X}(Y \mid X) = \frac{P_{XY}(X, Y)}{P_X(X)}.$$

Analog zur **Verbundentropie** $H(XY)$ lassen sich hier folgende Entropiefunktionen ableiten:

Definition: Die **bedingte Entropie** (englisch: *Conditional Entropy*) der Zufallsgröße X lautet unter der Bedingung Y:

$$H(X \mid Y) = \mathrm{E}\left[\log_2 \frac{1}{P_{X|Y}(X \mid Y)}\right] = \sum_{(x,y) \,\in\, \mathrm{supp}(P_{XY})} P_{XY}(x, y) \cdot \log_2 \frac{1}{P_{X|Y}(x \mid y)} =$$

$$= \sum_{(x,y) \,\in\, \mathrm{supp}(P_{XY})} P_{XY}(x, y) \cdot \log_2 \frac{P_Y(y)}{P_{XY}(x, y)}.$$

In gleicher Weise erhält man für die zweite bedingte Entropie:

$$H(Y \mid X) = \mathrm{E}\left[\log_2 \frac{1}{P_{Y|X}(Y \mid X)}\right] = \sum_{(x,y) \,\in\, \mathrm{supp}(P_{XY})} P_{XY}(x, y) \cdot \log_2 \frac{1}{P_{Y|X}(y \mid x)} =$$

$$= \sum_{(x,y) \,\in\, \mathrm{supp}(P_{XY})} P_{XY}(x, y) \cdot \log_2 \frac{P_X(x)}{P_{XY}(x, y)}.$$

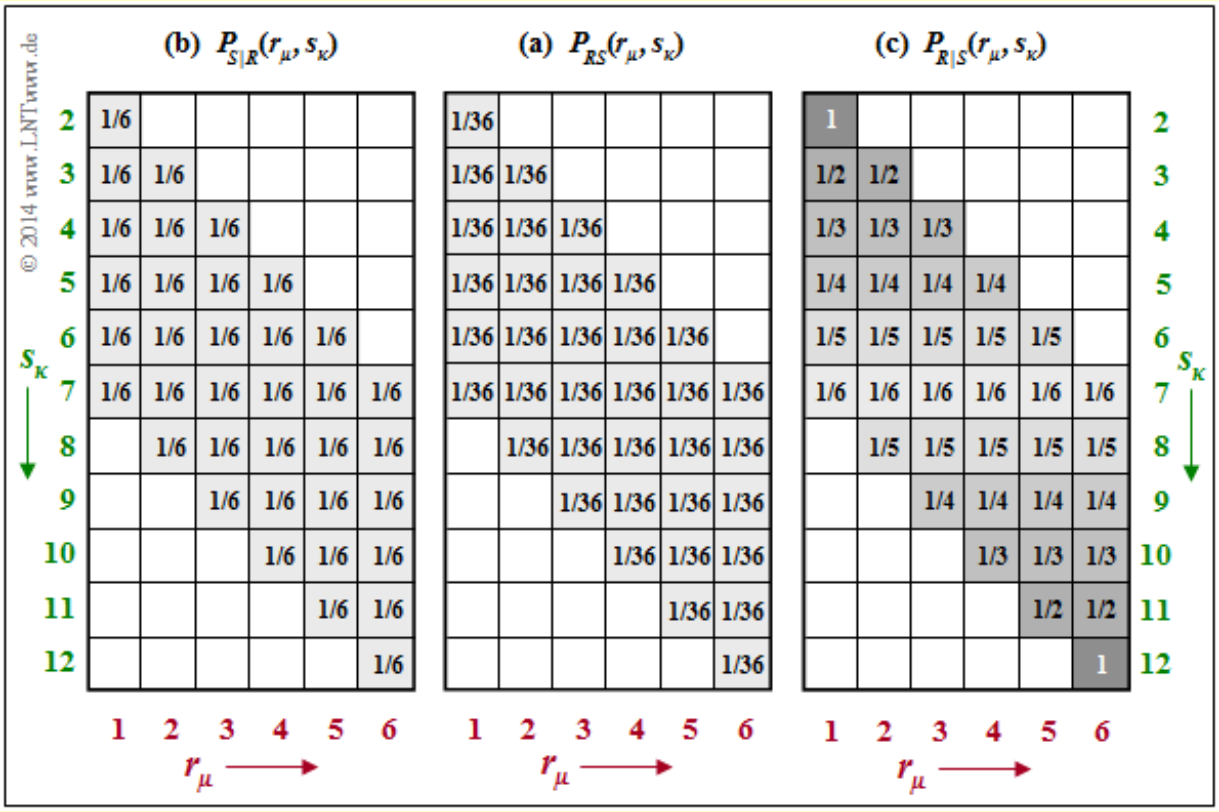
Im Argument der Logarithmusfunktion steht stets eine bedingte Wahrscheinlichkeitsfunktion ⇒ $P_{X|Y}(\cdot)$ bzw. $P_{Y|X}(\cdot)$, während zur Erwartungswertbildung die Verbundwahrscheinlichkeit $P_{XY}(\cdot)$ benötigt wird.

Für die bedingten Entropien gibt es folgende Begrenzungen:

- Sowohl $H(X|Y)$ als auch $H(Y|X)$ sind stets größer oder gleich 0. Aus $H(X|Y) = 0$ folgt direkt auch $H(Y|X) = 0$. Beides ist nur für **disjunkte Mengen** X und Y möglich.
- Es gilt stets $H(X|Y) \le H(X)$ sowie $H(Y|X) \le H(Y)$. Diese Aussage ist einleuchtend, wenn man sich bewusst macht, dass man für *Entropie* synonym auch *Unsicherheit* verwenden kann.
- Denn: Die Unsicherheit bezüglich X kann nicht dadurch größer werden, dass man Y kennt. Außer bei statistischer Unabhängigkeit ⇒ $H(X|Y) = H(X)$ gilt stets $H(X|Y) < H(X)$.
- Wegen $H(X) \le H(XY)$, $H(Y) \le H(XY)$ gilt somit auch $H(X|Y) \le H(XY)$ und $H(Y|X) \le H(XY)$. Eine bedingte Entropie kann also nie größer werden als die Verbundentropie.

Bedingte Wahrscheinlichkeit und bedingte Entropie (2)

Beispiel F: Wir betrachten die Verbundwahrscheinlichkeiten $P_{RS}(\cdot)$ unseres Würfelexperiments, die im **Beispiel D** ermittelt wurden. In der Mitte der folgenden Grafik ist $P_{RS}(\cdot)$ nochmals angegeben.



Außen sind die beiden bedingten Wahrscheinlichkeitsfunktionen gezeichnet:

- Links dargestellt ist die bedingte Wahrscheinlichkeitsfunktion $P_{S|R}(\cdot) = P_{SR}(\cdot)/P_R(\cdot)$. Wegen $P_R(R) = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6]$ steht hier in allen schraffierten Feldern $\Rightarrow$ $\text{supp}(P_{S|R}) = \text{supp}(P_{R|S})$ der gleiche Wahrscheinlichkeitswert 1/6. Daraus folgt für die bedingte Entropie:

$$H(S|R) = \sum_{(r,s)\,\in\,\text{supp}(P_{RS})} P_{RS}(r,s) \cdot \log_2 \frac{1}{P_{S|R}(s\,|\,r)} = 36 \cdot \frac{1}{36} \cdot \log_2\,(6) = 2.585\,\text{bit}\,.$$

- Für die andere bedingte Wahrscheinlichkeitsfunktion $P_{R|S}(\cdot) = P_{RS}(\cdot)/P_S(\cdot)$ mit $P_S(\cdot)$ gemäß **Beispiel D** ergeben sich die gleichen Felder ungleich 0 $\Rightarrow$ $\text{supp}(P_{R|S}) = \text{supp}(P_{S|R})$. Die Wahrscheinlichkeitswerte nehmen nun aber von der Mitte (1/6) zu den Rändern hin bis zur Wahrscheinlichkeit 1 in den Ecken kontinuierlich zu. Daraus folgt:

$$H(R|S) = \frac{1}{36} \cdot \log_2\,(6) + \frac{2}{36} \cdot \sum_{i=1}^{5} [i \cdot \log_2\,(i)] = 1.896\,\text{bit}\,.$$

Dagegen ergibt sich für die Zufallsgröße RB gemäß **Beispiel C** wegen $P_{RB}(\cdot) = P_R(\cdot) \cdot P_B(\cdot)$:

$$H(B|R) = H(B) = \log_2\,(6) = 2.585\,\text{bit}\,,$$
$$H(R|B) = H(R) = \log_2\,(6) = 2.585\,\text{bit}\,.$$

Transinformation zwischen zwei Zufallsgrößen (1)

Wir betrachten die Zufallsgröße XY mit der 2D–Wahrscheinlichkeitsfunktion $P_{XY}(X, Y)$. Bekannt seien auch die 1D–Funktionen $P_X(X)$ und $P_Y(Y)$. Nun stellen sich folgende Fragen:

- Wie vermindert die Kenntnis der Zufallsgröße Y die Unsicherheit bezüglich X?
- Wie vermindert die Kenntnis der Zufallsgröße X die Unsicherheit bezüglich Y?

Zur Beantwortung benötigen wir eine für die Informationstheorie substantielle Definition:

Definition: Die **Transinformation** (englisch: *Mutual Information*) zwischen den Zufallsgrößen X und Y – beide über dem gleichen Alphabet – ist gegeben durch den Ausdruck

$$I(X;Y) = \mathrm{E}\left[\log_2 \frac{P_{XY}(X,Y)}{P_X(X)\cdot P_Y(Y)}\right] = \sum_{(x,y)\,\in\,\mathrm{supp}(P_{XY})} P_{XY}(x,y)\cdot\log_2\frac{P_{XY}(x,y)}{P_X(x)\cdot P_Y(y)}.$$

Ein Vergleich mit **Kapitel 3.1** zeigt, dass die Transinformation auch als **Kullback–Leibler–Distanz** zwischen der 2D–PMF $P_{XY}(\cdot)$ und dem Produkt $P_X(\cdot)\cdot P_Y(\cdot)$ geschrieben werden kann:

$$I(X;Y) = D(P_{XY}\,||\,P_X\cdot P_Y).$$

Es ist offensichtlich, dass stets $I(X; Y) \geq 0$ gilt. Wegen der Symmetrie ist auch $I(Y; X) = I(X; Y)$.

Sucht man in einem Wörterbuch die Übersetzung für „mutual”, so findet man unter Anderem die Begriffe „gemeinsam”, „gegenseitig”, „beidseitig” und „wechselseitig”. Und ebenso sind in Fachbüchern für $I(X; Y)$ auch die Bezeichnungen *gemeinsame Entropie* und *gegenseitige Entropie* üblich. Wir sprechen aber im Folgenden durchgängig von der *Transinformation* $I(X; Y)$ und interpretieren nun diese Größe:

- Durch Aufspalten des $\log_2$–Arguments entsprechend

  $$I(X;Y) = \mathrm{E}\left[\log_2\frac{1}{P_X(X)}\right] - \mathrm{E}\left[\log_2\frac{P_Y(Y)}{P_{XY}(X,Y)}\right]$$

  erhält man unter Verwendung von $P_{X|Y}(\cdot) = P_{XY}(\cdot)/P_Y(Y)$:

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

  Das heißt: Die Unsicherheit hinsichtlich der Zufallsgröße X $\Rightarrow$ Entropie $H(X)$ vermindert sich bei Kenntnis von Y um den Betrag $H(X|Y)$. Der Rest ist die Transinformation $I(X; Y)$.

- Bei anderer Aufspaltung kommt man zum Ergebnis:

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

  Ergo: Die Transinformation $I(X; Y)$ ist symmetrisch: X sagt genau so viel über Y aus wie Y über X $\Rightarrow$ *gegenseitige Information*. Das Semikolon weist auf die Gleichberechtigung hin.

Oft werden die hier genannten Gleichungen durch ein Schaubild verdeutlicht, so auch in den folgenden Beispielen. Daraus erkennt man, dass auch folgende Gleichungen zutreffen:

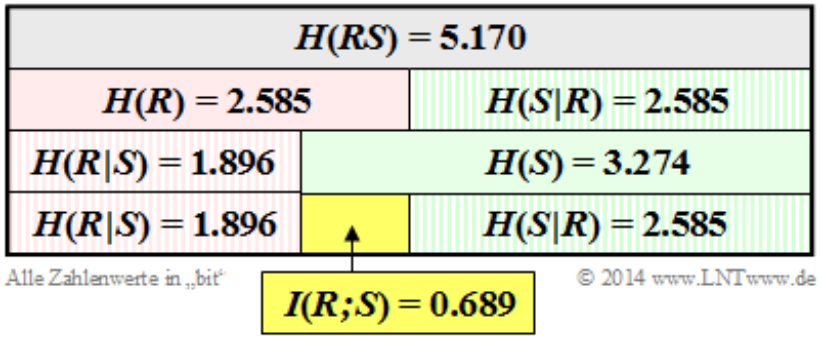
$$I(X;Y) = H(X) + H(Y) - H(XY),$$
$$I(X;Y) = H(XY) - H(X\,|\,Y) - H(Y\,|\,X).$$

Transinformation zwischen zwei Zufallsgrößen (2)

Beispiel F: Wir kommen nochmals auf das **Würfel–Experiment** mit dem roten (R) und dem blauen (B) Würfel zurück. Die Zufallsgröße S gibt die Summe der beiden Würfel an: $S = R + B$.

Wir betrachten hier die 2D–Zufallsgröße RS. In früheren Beispielen haben wir berechnet:

- die Entropien $H(R) = 2.585$ bit und $H(S) = 3.274$ bit ⇒ **Beispiel D,**
- die Verbundentropie $H(RS) = 5.170$ bit ⇒ **Beispiel D,**
- die bedingten Entropien $H(S|R) = 2.585$ bit und $H(R|S) = 1.896$ bit ⇒ **Beispiel F.**



Diese Größen sind in der Grafik zusammengestellt, wobei die Zufallsgröße R durch die Grundfarbe „Rot" und die Summe S durch die Grundfarbe „grün" markiert sind. Bedingte Entropien sind schraffiert.

Man erkennt aus dieser Darstellung:

- Hier ist $H(R) = \log_2(6) = 2.585$ bit genau halb so groß wie die Verbundentropie $H(RS)$. Kennt man R, so liefert S genau die gleiche Information wie die Zufallsgröße B, nämlich $H(S|R) = H(B) = \log_2(6) = 2.585$ bit. *Hinweis:* $H(R) = H(S|R)$ gilt nicht allgemein.
- Die Entropie $H(S) = 3.274$ bit ist im vorliegenden Beispiel erwartungsgemäß größer als $H(R)$. Wegen $H(S) + H(R|S) = H(R) + H(S|R)$ muss deshalb $H(R|S)$ gegenüber $H(S|R)$ um den gleichen Betrag $I(R; S) = 0.689$ bit kleiner sein wie $H(R)$ gegenüber $H(S)$.
- Die Transinformation (englisch: *Mutual Information*) zwischen den Zufallsgrößen R und S ergibt sich aber auch aus der Gleichung

$$I(R;S) = H(R) + H(S) - H(RS) = \\ = 2.585\,\text{bit} + 3.274\,\text{bit} - 5.170\,\text{bit} = 0.689\,\text{bit}.$$

Bedingte Transinformation

Wir betrachten nun drei Zufallsgrößen X, Y und Z, die zueinander in Beziehung stehen (können).

Definition: Die **bedingte Transinformation** (englisch: *Conditional Mutual Information*) zwischen den Zufallsgrößen X und Y bei gegebenem $Z = z$ lautet:

$$I(X;Y\,|\,Z=z) = H(X\,|\,Z=z) - H(X|Y,Z=z).$$

Dagegen bezeichnet man als die **bedingte Transinformation** zwischen den Zufallsgrößen X und Y bei gegebener **Zufallsgröße Z**:

$$I(X;Y\,|\,Z) = H(X\,|\,Z) - H(X|YZ) = \sum_{z\,\in\,\mathrm{supp}(P_Z)} P_Z(z)\cdot I(X;Y\,|\,Z=z).$$

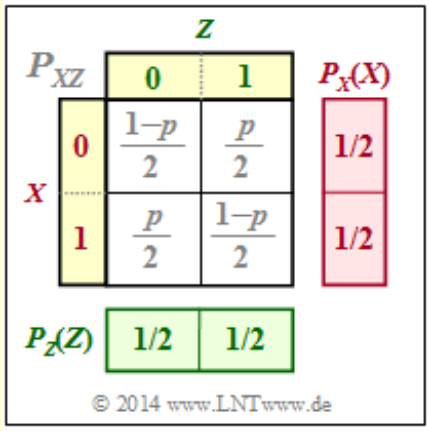
Hierbei ist $P_Z(Z)$ die Wahrscheinlichkeitsfunktion der neben X und Y betrachteten Zufallsgröße Z und $P_Z(z)$ die Wahrscheinlichkeit für $Z = z$.

Bitte beachten Sie: Für die bedingte Entropie gilt bekanntlich die Größenrelation $H(X|Z) \le H(X)$. Für die Transinformation gilt diese Größenrelation nicht unbedingt:

$I(X; Y|Z)$ kann kleiner, gleich, aber auch größer sein als $I(X; Y)$.

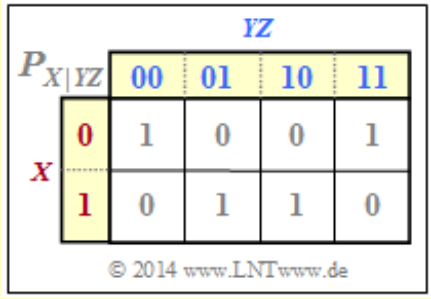
Beispiel: Wir betrachten die binären Zufallsgrößen X, Y und Z mit folgenden Eigenschaften:

- X und Y seien statistisch unabhängig und für ihre Wahrscheinlichkeitsfunktionen gelte: $P_X(X) = [1/2,\ 1/2]$, $P_Y(Y) = [1–p,\ p]$ ⇒ $H(X) = 1$ (bit), $H(Y) = H_{\rm bin}(p)$.
- Z ist die Modulo–2–Summe von X und Y: $Z = X \oplus Y$.



Aus der Verbund–PMF P_{XZ} gemäß der oberen Grafik folgt:

- Durch Summation der Spalten–Wahrscheinlichkeiten ergibt sich $P_Z(Z) = [1/2;\ 1/2] \Rightarrow H(Z) = 1$.
- X und Z sind ebenfalls statistisch unabhängig, da für die 2D–PMF $P_{XZ}(X, Z) = P_X(X) \cdot P_Z(Z)$ gilt.
- Daraus folgt: $H(Z|X) = H(Z)$, $H(X|Z) = H(X)$, $I(X; Z) = 0$.



Aus der bedingten Wahrscheinlichkeitsfunktion $P_{X|YZ}$ gemäß der unteren Grafik lassen sich berechnen:

- $H(X|YZ) = 0$, da alle $P_{X|YZ}$–Einträge entweder 0 oder 1 ⇒ *bedingte Entropie*,
- $I(X; YZ) = H(X) – H(X|YZ) = H(X)$ ⇒ *Transinformation*,
- $I(X; Y|Z) = H(X|Z) = H(X)$ ⇒ *bedingte Transinformation*.

Im vorliegenden Beispiel ist also $I(X; Y|Z) = 1$ (bit) größer als $I(X; Y) = 0$ (bit).

Kettenregel der Transinformation

Bisher haben wir die Transinformation nur zwischen zwei eindimensionalen Zufallsgrößen betrachtet. Nun erweitern wir die Definition auf insgesamt $n + 1$ Zufallsgrößen, die wir aus Darstellungsgründen mit X_1, ..., X_n sowie Z bezeichnen. Dann gilt

Kettenregel der Transinformation:
Die Transinformation zwischen der n–dimensionalen Zufallsgröße $X_1\,X_2\,\ldots\,X_n$ und der Zufallsgröße Z lässt sich wie folgt darstellen und berechnen:

$$I(X_1 X_2 \ldots X_n; Z) = I(X_1; Z) + I(X_2; Z|X_1) + \ldots + I(X_n; Z|X_1 X_2 \ldots X_{n-1}) = \sum_{i=1}^{n} I(X_i; Z|X_1 X_2 \ldots X_{i-1}).$$

Für den Beweis beschränken wir uns hier auf den Fall $n = 2$, also auf insgesamt drei Zufallsgrößen, und ersetzen X_1 und X_2 durch X und Y. Damit erhalten wir:

$$\begin{aligned} I(X\,Y; Z) &= H(XY) - H(XY|Z) = \\ &= \big[H(X) + H(Y|X)\big] - \big[H(X|Z) + H(Y|XZ)\big] = \\ &= \big[H(X) - H(X|Z)\big] - \big[H(Y|X) + H(Y|XZ)\big] = \\ &= I(X; Z) + I(Y; Z|X). \end{aligned}$$

Aus dieser Gleichung erkennt man, dass die die Größenrelation $I(X\,Y; Z) \geq I(X; Z)$ immer gegeben ist. Gleichheit ergibt sich für die bedingte Transinformation $I(Y; Z|X) = 0$, also dann, wenn die Zufallsgrößen Y und Z für ein gegebenes X statistisch unabhängig sind.

Beispiel: Wir betrachten die **Markovkette** $X \to Y \to Z$. Für eine solche Konstellation gilt stets das *Data Processing Theorem* mit der folgenden Konsequenz, die sich aus der Kettenregel der Transinformation ableiten lässt:

$$I(X; Z) \leq I(X; Y),$$
$$I(X; Z) \leq I(Y; Z).$$

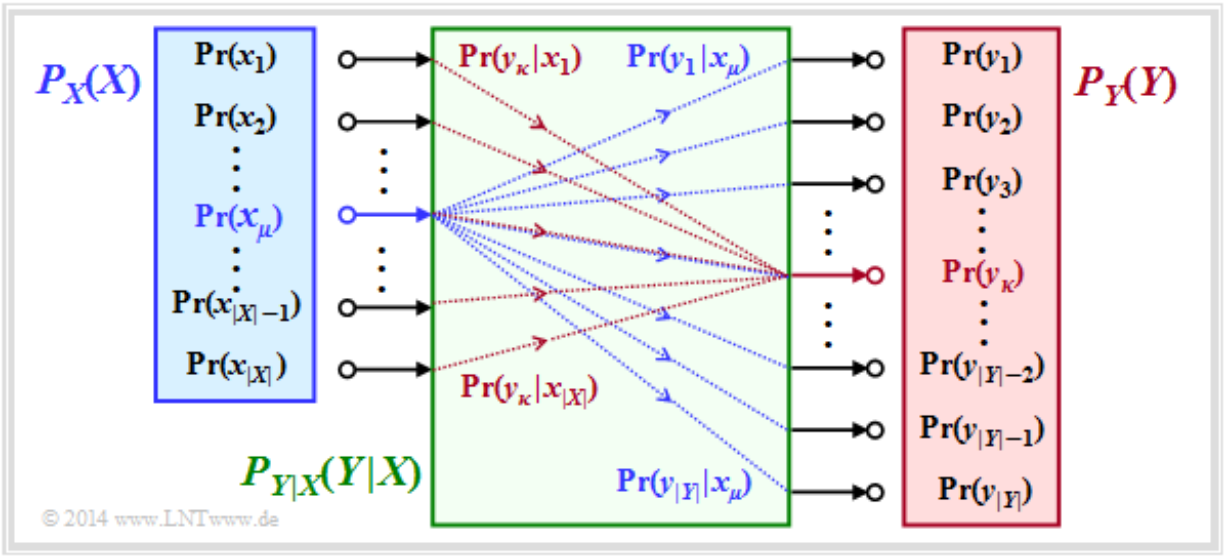
Das Theorem besagt somit:

- Man kann durch Manipulation (*Processing Z*) der Daten Y keine zusätzliche Information über den Eingang X gewinnen.
- Die Datenverarbeitung $Y \to Z$ (durch einen zweiten Prozessor) dient nur dem Zweck, die Information über X besser sichtbar zu machen.

Weitere Informationen zum *Data Processing Theorem* finden Sie in der **Aufgabe A3.14.**

Informationstheoretisches Modell der Digitalsignalübertragung (1)

Die bisher allgemein definierten Entropien werden nun auf die Digitalsignalübertragung angewendet, wobei wir von einem **digitalen Kanalmodell ohne Gedächtnis** (englisch: *Discrete Memoryless Channel*, DMC) entsprechend der nachfolgenden Grafik ausgehen:



- Die Menge der möglichen Quellensymbole wird durch die diskrete Zufallsgröße X charakterisiert, wobei $|X|$ den Quellensymbolumfang angibt:

$$X = \{ x_1, x_2, \ \text{...}\ , x_\mu, \ \text{...}\ , x_{|X|} \}.$$

- Entsprechend kennzeichnet Y die Menge der Sinkensymbole mit dem Symbolvorrat $|Y|$:

$$Y = \{ y_1, y_2, \ \text{...}\ , y_\kappa, \ \text{...}\ , Y_{|Y|} \}.$$

- Meist gilt $|Y| = |X|$. Möglich ist aber auch $|Y| > |X|$, zum Beispiel beim **Binary Erasure Channel** (BEC) mit $X = \{0, 1\}$ und $Y = \{0, 1, \text{E}\}$ $\Rightarrow$ $|X| = 2$, $|Y| = 3$.

- Das Sinkensymbol E kennzeichnet eine Auslöschung (englisch: *Erasure*). Das Ereignis $Y = \text{E}$ gibt an, dass eine Entscheidung für 0 oder für 1 zu unsicher wäre.

- Die Symbolwahrscheinlichkeiten der Quelle und der Sinke sind in der oberen Grafik durch die Wahrscheinlichkeitsfunktionen $P_X(X)$ und $P_Y(Y)$ berücksichtigt, wobei gilt:

$$P_X(x_\mu) = \Pr(X = x_\mu), \quad P_Y(y_\kappa) = \Pr(Y = y_\kappa).$$

- Es gelte: $P_X(X)$ und $P_Y(Y)$ enthalten keine Nullen $\Rightarrow$ $\text{supp}(P_X) = P_X$, $\text{supp}(P_Y) = P_Y$. Diese Voraussetzung erleichtert die Modellbeschreibung, ohne Verlust an Allgemeingültigkeit.

- Alle Übergangswahrscheinlichkeiten des digitalen gedächtnislosen Kanals (DMC) werden durch die *bedingte Wahrscheinlichkeitsfunktion* $P_{Y|X}(Y|X)$ erfasst. Mit $x_\mu \in X$ und $y_\kappa \in Y$ gelte hierfür folgende Definition:

$$P_{Y|X}(y_\kappa|x_\mu) = \Pr(Y = y_\kappa | X = x_\mu).$$

In obiger Grafik ist $P_{Y|X}(\cdot)$ als ein Block mit $|X|$ Eingängen und $|Y|$ Ausgängen dargestellt. Die blauen Verbindungen markieren die Übergangswahrscheinlichkeiten $\Pr(y_i \,|\, x_\mu)$ ausgehend von x_μ mit $1 \le i \le |Y|$,

während alle roten Verbindungen bei y_κ enden: $\Pr(y_\kappa \mid x_i)$ mit $1 \le i \le |X|$.

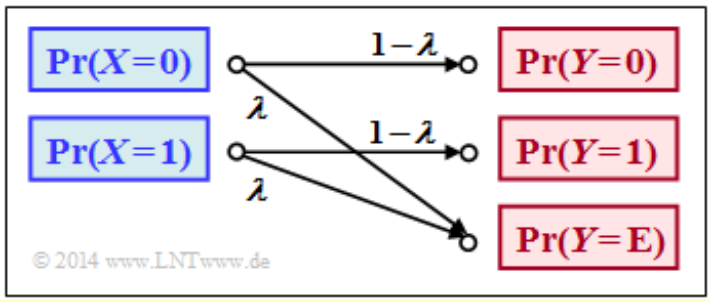
Informationstheoretisches Modell der Digitalsignalübertragung (2)

Bevor wir die Entropien für die einzelnen Wahrscheinlichkeitsfunktionen angeben, nämlich

$$P_X(X) \Rightarrow H(X),\ P_Y(Y) \Rightarrow H(Y),\ P_{XY}(X) \Rightarrow H(XY),\ P_{Y|X}(Y|X) \Rightarrow H(Y|X),\ P_{X|Y}(X|Y) \Rightarrow H(X|Y),$$

sollen die Aussagen der letzten Seite an einem Beispiel verdeutlicht werden.

Beispiel: Im Buch „Einführung in die Kanalcodierung" behandeln wir den **Binary Erasure Channel** (BEC), der rechts in etwas modifizierter Form skizziert ist.



Dabei gelten folgende Voraussetzungen:

- Das Eingangsalphabet ist binär: $X = (0, 1) \Rightarrow |X| = 2$, während am Ausgang drei Werte möglich sind: $Y = (0, 1, \text{E}) \Rightarrow |Y| = 3$.
- „E" kennzeichnet den Fall, dass sich der Empfänger aufgrund von zu großen Kanalstörungen nicht für eines der Binärsymbole 0 oder 1 entscheiden kann. „E" steht hierbei für *Erasure* (Auslöschung).
- Beim BEC gemäß obiger Skizze werden sowohl eine gesendete „0" als auch eine „1" mit der Wahrscheinlichkeit λ ausgelöscht, während die Wahrscheinlichkeit einer richtigen Übertragung jeweils $1 - \lambda$ beträgt.
- Dagegen werden Übertragungsfehler durch das BEC–Modell ausgeschlossen $\Rightarrow$ die bedingten Wahrscheinlichkeiten $\Pr(Y = 1 \,|\, X = 0)$ sowie $\Pr(Y = 0 \,|\, X = 1)$ sind jeweils 0.

Beim Sender seien die Nullen und Einsen nicht unbedingt gleichwahrscheinlich. Vielmehr verwenden wir die beiden Wahrscheinlichkeitsfunktionen

$$P_X(X) = \big(\Pr(X = 0), \Pr(X = 1)\big),$$
$$P_Y(Y) = \big(\Pr(Y = 0), \Pr(Y = 1), \Pr(Y = \text{E})\big).$$

Aus obigem Modell erhalten wir dann:

$$P_Y(0) = \Pr(Y = 0) = P_X(0) \cdot (1 - \lambda),$$
$$P_Y(1) = \Pr(Y = 1) = P_X(1) \cdot (1 - \lambda),$$
$$P_Y(\text{E}) = \Pr(Y = \text{E}) = P_X(0) \cdot \lambda + P_X(1) \cdot \lambda.$$

Fassen wir nun $P_X(X)$ und $P_Y(Y)$ als Vektoren auf, so lässt sich das Ergebnis wie folgt darstellen:

$$P_Y(Y) = P_X(X) \cdot P_{Y|X}(Y|X),$$

wobei die Übergangswahrscheinlichkeiten $\Pr(y_\kappa \,|\, x_\mu)$ durch folgende Matrix berücksichtigt sind:

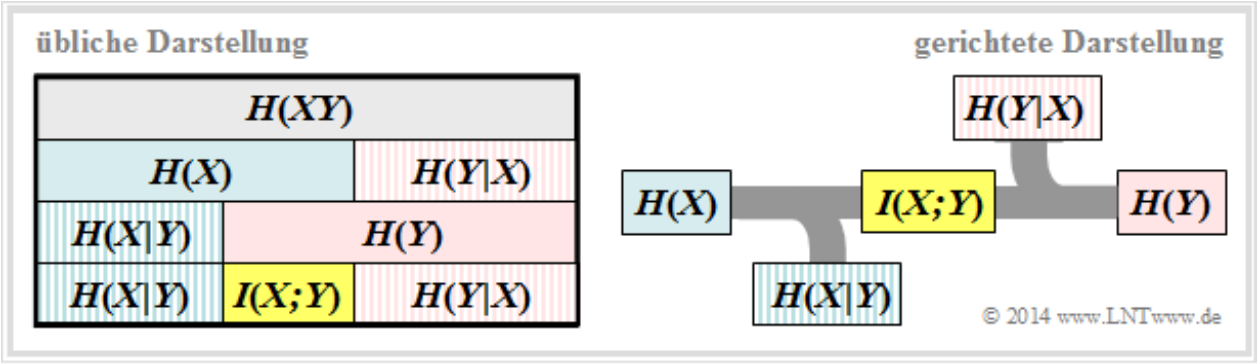
$$P_{Y|X}(Y|X) = \begin{pmatrix} 1-\lambda & 0 & \lambda \\ 0 & 1-\lambda & \lambda \end{pmatrix}.$$

Beachten Sie: Wir haben diese Darstellung nur gewählt, um die Beschreibung zu vereinfachen. $P_X(X)$

und $P_Y(Y)$ sind keine Vektoren im eigentlichen Sinne und $P_{Y|X}(Y|X)$ ist keine Matrix.

Informationstheoretisches Modell der Digitalsignalübertragung (3)

Alle in **Kapitel 3.2** definierten Entropien gelten auch für die Digitalsignalübertragung. Es ist aber zweckmäßig, anstelle des bisher verwendeten Schaubildes (linke Grafik) die rechte Darstellung zu wählen, bei der die Richtung von der Quelle X zur Sinke Y erkennbar ist.



Interpretieren wir nun ausgehend vom allgemeinen **DMC–Kanalmodell** die rechte Grafik:

- D i e **Quellenentropie** (englisch: *Source Entropy*) $H(X)$ bezeichnet den mittleren Informationsgehalt der Quellensymbolfolge. Mit dem Symbolumfang $|X|$ gilt:

$$H(X) = \mathrm{E}\left[\log_2 \frac{1}{P_X(X)}\right] = -\mathrm{E}\left[\log_2 P_X(X)\right] = \sum_{\mu=1}^{|X|} P_X(x_\mu) \cdot \log_2 \frac{1}{P_X(x_\mu)}.$$

- Die **Äquivokation** (auch *Rückschlussentropie* genannt, englisch: *Equivocation*) $H(X|Y)$ gibt den mittleren Informationsgehalt an, den ein Betrachter, der über die Sinke Y genau Bescheid weiß, durch Beobachtung der Quelle X gewinnt:

$$H(X|Y) = \mathrm{E}\left[\log_2 \frac{1}{P_{X|Y}(X|Y)}\right] = \sum_{\mu=1}^{|X|}\sum_{\kappa=1}^{|Y|} P_{XY}(x_\mu, y_\kappa) \cdot \log_2 \frac{1}{P_{X|Y}(x_\mu|y_\kappa)}.$$

- Die Äquivokation ist der Anteil der Quellenentropie $H(X)$, der durch Kanalstörungen (bei digitalem Kanal: Übertragungsfehler) verloren geht. Es verbleibt die **Transinformation** (englisch: *Mutual Information*) $I(X; Y)$, die zur Sinke gelangt:

$$I(X;Y) = \mathrm{E}\left[\log_2 \frac{P_{XY}(X,Y)}{P_X(X)\cdot P_Y(Y)}\right] = H(X) - H(X|Y).$$

- Die **Irrelevanz** (manchmal auch *Streuentropie* genannt, englisch: *Irrelevance*) $H(Y|X)$ gibt den mittleren Informationsgehalt an, den ein Betrachter, der über die Quelle X genau Bescheid weiß, durch Beobachtung der Sinke Y gewinnt:

$$H(Y|X) = \mathrm{E}\left[\log_2 \frac{1}{P_{Y|X}(Y|X)}\right] = \sum_{\mu=1}^{|X|}\sum_{\kappa=1}^{|Y|} P_{XY}(x_\mu, y_\kappa) \cdot \log_2 \frac{1}{P_{Y|X}(y_\kappa|x_\mu)}.$$

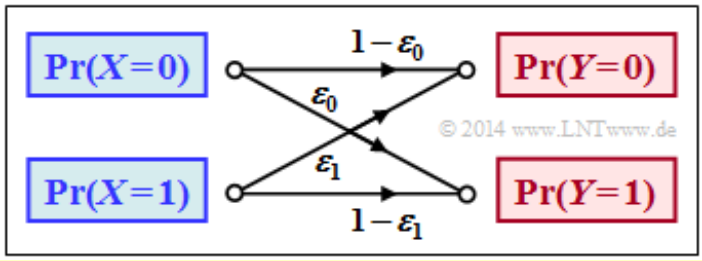
- Die **Sinkenentropie** $H(Y)$, der mittlere Informationsgehalt der Sinke, ist die Summe aus der nützlichen Transinformation $I(X; Y)$ und der Irrelevanz $H(Y|X)$, die ausschließlich von Kanalfehlern herrührt:

$$H(Y) = \mathrm{E}\left[\log_2 \frac{1}{P_Y(Y)}\right] = -\mathrm{E}\left[\log_2 P_Y(Y)\right] = I(X;Y) + H(Y|X)\,.$$

Transinformationsberechnung für den Binärkanal (1)

Die Definitionen der letzten Seite sollen nun an einem Beispiel verdeutlicht werden, wobei wir bewusst vermeiden, die Berechnungen durch die Ausnutzung von Symmetrien zu vereinfachen.

Beispiel: Wir betrachten den allgemeinen Binärkanal (englisch: *Binary Channel*) ohne Gedächtnis gemäß der Skizze mit den Verfälschungswahrscheinlichkeiten.



$$\varepsilon_0 = \Pr(Y=1|X=0) = 0.01\,,$$
$$\varepsilon_1 = \Pr(Y=0|X=1) = 0.2$$
$$\Rightarrow \quad P_{Y|X}(Y|X) = \begin{pmatrix} 1-\varepsilon_0 & \varepsilon_0 \\ \varepsilon_1 & 1-\varepsilon_1 \end{pmatrix} = \begin{pmatrix} 0.99 & 0.01 \\ 0.2 & 0.8 \end{pmatrix}.$$

Außerdem gehen wir von nicht gleichwahrscheinlichen Quellensymbolen aus:

$$P_X(X) = (p_0, p_1) = (0.1, 0.9)\,.$$

Mit der **binären Entropiefunktion** erhält man so für die Quellenentropie:

$$H(X) = H_{\rm bin}(0.1) = 0.4690\ {\rm bit}\,.$$

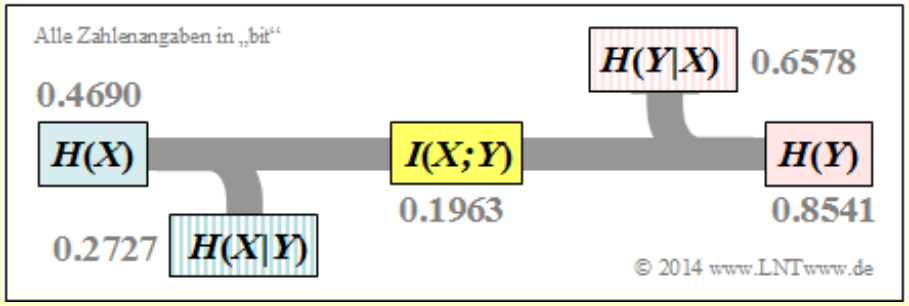
Für die Wahrscheinlichkeitsfunktion der Sinke sowie für die Sinkenentropie ergibt sich somit:

$$P_Y(Y) = \big(\Pr(Y=0), \Pr(Y=1)\big) = (p_0, p_1) \cdot \begin{pmatrix} 1-\varepsilon_0 & \varepsilon_0 \\ \varepsilon_1 & 1-\varepsilon_1 \end{pmatrix}$$
$$\Rightarrow \quad \Pr(Y=0) = p_0 \cdot (1-\varepsilon_0) + p_1 \cdot \varepsilon_1 = 0.1 \cdot 0.99 + 0.9 \cdot 0.2 = 0.279\,,$$
$$\Pr(Y=1) = 1 - \Pr(Y=0) = 0.721$$
$$\Rightarrow \quad H(Y) = H_{\rm bin}(0.279) = 0.8541\ {\rm bit}\,.$$

Auf der nächsten Seite werden noch berechnet:

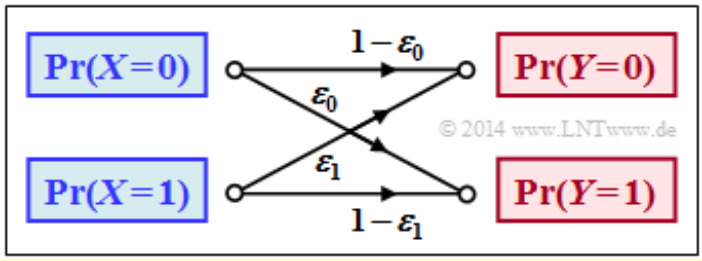
- die Verbundentropie $H(XY)$,
- die Transinformation $I(X;\,Y)$,
- die Rückschlussentropie $H(X|Y)$ ⇒ Äquivokation,
- die Streuentropie $H(Y|X)$ ⇒ Irrelevanz.

Diese Ergebnisse sind in der folgenden zusammenfassenden Grafik bereits mit aufgenommen.



Transinformationsberechnung für den Binärkanal (2)

Fortsetzung des Beispiels:
Wir betrachten weiter den allgemeinen Binärkanal (englisch: *Binary Channel*) ohne Gedächtnis gemäß der Skizze, und es gelte weiterhin:



$$P_X(X) = (p_0, p_1) = (0.1, 0.9)\,,$$
$$\varepsilon_0 = \Pr(Y = 1\,|\,X = 0) = 0.01\,, \quad \varepsilon_1 = \Pr(Y = 0\,|\,X = 1) = 0.2\,.$$

Die Verbundwahrscheinlichkeiten $p_{\mu\kappa} = \Pr[(X = \mu) \cap (Y = \kappa)]$ zwischen Quelle und Sinke sind:

$$p_{00} = p_0 \cdot (1 - \varepsilon_0) = 0.099\,, \quad p_{01} = p_0 \cdot \varepsilon_0 = 0.001\,,$$
$$p_{10} = p_1 \cdot (1 - \varepsilon_1) = 0.180\,, \quad p_{11} = p_1 \cdot \varepsilon_1 = 0.720\,.$$

Daraus erhält man für

- die **Verbundentropie** (englisch *Joint Entropy*):

$$H(XY) = p_{00} \cdot \log_2 \frac{1}{p_{00}} + p_{01} \cdot \log_2 \frac{1}{p_{01}} + p_{10} \cdot \log_2 \frac{1}{p_{10}} + p_{11} \cdot \log_2 \frac{1}{p_{11}} = 1.1268\,\text{bit}\,,$$

- die **Transinformation** (englisch *Mutual Information*):

$$I(X;Y) = H(X) + H(Y) - H(XY) = 0.4690 + 0.8541 - 1.1268 = 0.1963\,\text{bit}\,,$$

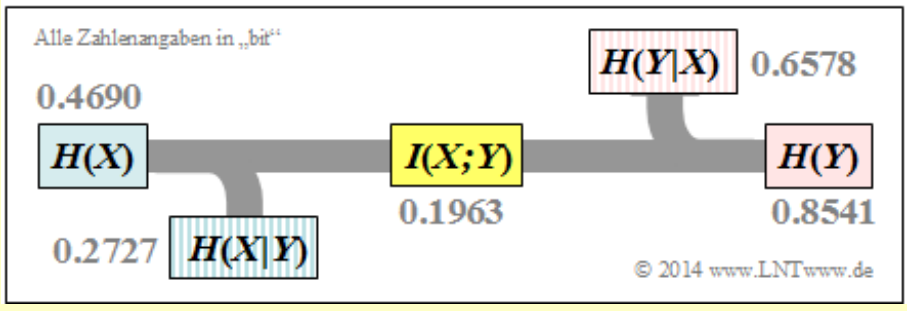
- die **Äquivokation** (oder Rückschlussentropie):

$$H(X|Y) = H(X) - I(X;Y) = 0.4690 - 0.1963 = 0.2727\,\text{bit}\,,$$

- die **Irrelevanz** (oder Streuentropie):

$$H(Y|X) = H(Y) - I(X;Y) = 0.8541 - 0.1963 = 0.6578\,\text{bit}\,.$$

Die Ergebnisse sind in der folgenden Grafik nochmals zusammengefasst.



Anmerkung: Äquivokation und Irrelevanz hätte man auch direkt (aber mit Mehraufwand) aus den entsprechenden Wahrscheinlichkeitsfunktionen berechnen können. Zum Beispiel die Irrelevanz:

$$H(Y|X) = \sum_{(x,y)\,\in\,XY} P_{XY}(x,\,y) \cdot \log_2 \frac{1}{P_{Y|X}(y|x)} =$$
$$= p_{00} \cdot \log_2 \frac{1}{1 - \varepsilon_0} + p_{01} \cdot \log_2 \frac{1}{\varepsilon_0} + p_{10} \cdot \log_2 \frac{1}{1 - \varepsilon_1} + p_{11} \cdot \log_2 \frac{1}{\varepsilon_1} = 0.6578\,\text{bit}\,.$$

Definition und Bedeutung der Kanalkapazität

Wir betrachten weiter einen diskreten gedächtnislosen Kanal (englisch: *Discrete Memoryless Channel*, kurz DMC) mit einer endlichen Anzahl an Quellensymbolen $\Rightarrow$ $|X|$ und ebenfalls nur endlich vielen Sinkensymbolen $\Rightarrow$ $|Y|$, wie auf der **ersten Seite** dieses Kapitels dargestellt. Berechnet man die Transinformation $I(X,\ Y)$ wie zuletzt an einem Beispiel ausgeführt, so hängt diese auch von der Quellenstatistik $\Rightarrow$ $P_X(X)$ ab. Ergo: ***I(X, Y) ist keine reine Kanalkenngröße***.

Definition: Die von **Claude E. Shannon** eingeführte **Kanalkapazität** (englisch: *Channel Capacity*) lautet entsprechend seinem Standardwerk **[Sha48]:**

$$C = \max_{P_X(X)} I(X;Y).$$

Da nach dieser Definition stets die bestmögliche Quellenstatistik zugrunde liegt, hängt C nur von den Kanaleigenschaften $\Rightarrow P_{Y|X}(Y|X)$ ab, nicht jedoch von $P_X(X)$. Oft wird die Zusatzeinheit „bit/Kanalzugriff" hinzugefügt, bei englischen Texten „bit/use".

C. E. Shannon benötigte diese Kanalbeschreibungsgröße C, um das Kanalcodierungstheorem formulieren zu können – eines der Highlights der von ihm begründeten Informationstheorie.

Shannons Kanalcodierungstheorem: Zu jedem Übertragungskanal mit der Kanalkapazität $C > 0$ existiert (mindestens) ein (k, n)–Blockcode, dessen (Block–)Fehlerwahrscheinlichkeit gegen Null geht, so lange die Coderate $R = k/n$ kleiner oder gleich der Kanalkapazität ist: $\boldsymbol{R \le C}$. Voraussetzung hierfür ist allerdings, dass für die Blocklänge dieses Codes gilt: $n \to \infty$.

Den Beweis dieses Theorems finden Sie zum Beispiel in **[CT06], [Kra13]** und **[Meck09].** Der Beweis würde den Rahmen unseres Lerntutorials sprengen.

Im **Kapitel 4.3** wird im Zusammenhang mit dem wertkontinuierlichen **AWGN–Kanalmodell** ausgeführt, welche phänomenal große Bedeutung Shannons informationstheoretisches Theorem für die gesamte Informationstechnik besitzt, nicht nur für ausschließlich theoretisch Interessierte, sondern ebenso auch für Praktiker.

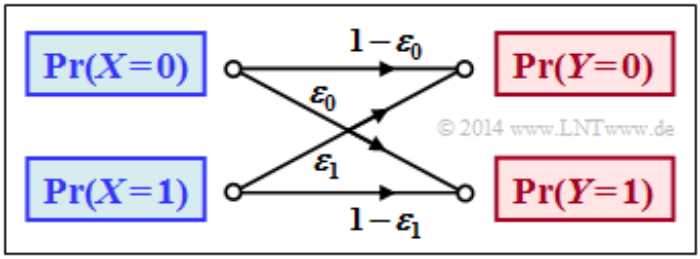
Wie in **Aufgabe A3.12** gezeigt werden soll, gilt auch der Umkehrschluss:

Ist die Rate des verwendeten (n, k)–Blockcodes größer als die Kanalkapazität $\Rightarrow$ $\boldsymbol{R = k/n > C}$, so kann **niemals eine beliebig kleine Blockfehlerwahrscheinlichkeit** erreicht werden.

Auch diesen Beweis finden Sie zum Beispiel wieder in **[CT06], [Kra13]** und **[Meck09].**

Kanalkapazität eines Binärkanals

Die Transinformation des allgemeinen (unsymmetrischen) Binärkanals gemäß der nebenstehenden Grafik wurde auf der **vorletzten Seite** berechnet. Bei diesem Modell werden die Eingangssymbole „0” und „1” unterschiedlich stark verfälscht:



$$P_{Y|X}(Y|X) = \begin{pmatrix} 1-\varepsilon_0 & \varepsilon_0 \\ \varepsilon_1 & 1-\varepsilon_1 \end{pmatrix}.$$

Die Transinformation lässt sich mit $P_X(X) = (p_0, p_1)$ in folgender Form kompakt darstellen:

$$I(X;Y) = \sum_{\mu=1}^{2}\sum_{\kappa=1}^{2} \Pr(y_\kappa|x_\mu)\cdot\Pr(x_\mu)\cdot\log_2\frac{\Pr(y_\kappa|x_\mu)}{\Pr(y_\kappa)} =$$

$$= (1-\varepsilon_0)\cdot p_0\cdot\log_2\frac{1-\varepsilon_0}{(1-\varepsilon_0)\cdot p_0+\varepsilon_1\cdot p_1} + \varepsilon_0\cdot p_0\cdot\log_2\frac{\varepsilon_0}{(1-\varepsilon_0)\cdot p_0+\varepsilon_1\cdot p_1} +$$

$$+\ \varepsilon_1\cdot p_1\cdot\log_2\frac{\varepsilon_1}{\varepsilon_0\cdot p_0+(1-\varepsilon_1)\cdot p_1} + (1-\varepsilon_1)\cdot p_1\cdot\log_2\frac{1-\varepsilon_1}{\varepsilon_0\cdot p_0+(1-\varepsilon_1)\cdot p_1}.$$

Im Folgenden setzen wir $\varepsilon_0 = 0.01$ und $\varepsilon_1 = 0.2$. In der vierten Spalte der nebenstehenden Tabelle (grün hinterlegt) ist die Transinformation $I(X; Y)$ dieses unsymmetrischen Binärkanals abhängig von der Quellensymbolwahrscheinlichkeit $p_0 = \Pr(X = 0)$ angegeben. Man erkennt:

p_0	$H(X)$	$H(X\|Y)$	$I(X;Y)$	$H(X\|Y)$	$H(Y)$
0	0.0000	0.0000	0.0000	0.0000	0.0000
0.10	0.4690	0.2727	0.1963	0.6578	0.8541
0.20	0.7219	0.3746	0.3473	0.5937	0.9410
0.30	0.8813	0.4224	0.4589	0.5296	0.9885
0.40	0.9710	0.4372	0.5338	0.4655	0.9993
0.50	1.0000	0.4276	0.5724	0.4014	0.9738
0.55	0.9928	0.4194	0.5779	0.3693	0.9472
0.60	0.9710	0.3974	0.5735	0.3372	0.9108
0.70	0.8813	0.3479	0.5334	0.2731	0.8065
0.80	0.7219	0.2778	0.4441	0.2090	0.6531
0.90	0.4690	0.1808	0.2882	0.1449	0.4331
1	0.0000	0.0000	0.0000	0.0000	0.0000

Alle Angaben in „bit“ © 2014 www.LNTwww.de

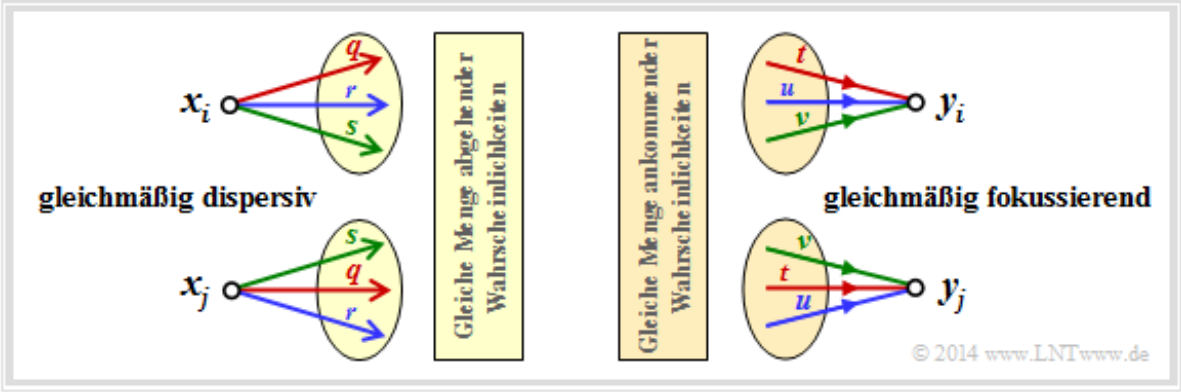
- Die Transinformation $I(X; Y)$ hängt von den Symbolwahrscheinlichkeiten p_0 und $p_1 = 1 – p_0$ ab.
- Der Maximalwert der Transinformation ergibt sich für $p_0 \approx 0.55 \Rightarrow p_1 \approx 0.45$.
- Das Optimierungsergebnis $p_0 > p_1$ folgt aus der Relation $\varepsilon_0 < \varepsilon_1$ (die „0” wird weniger verfälscht).
- Die Kanalkapazität ist somit für $\varepsilon_0 = 0.01$, $\varepsilon_1 = 0.2$ gleich $C = 0.5779$ bit/Kanalzugriff.

In obiger Gleichung ist als Sonderfall auch der **Binary Symmetric Channel** (BSC) mit dem Parameter $\varepsilon = \varepsilon_0 = \varepsilon_1$ mitenthalten. In **Aufgabe A3.9** wird die Transinformation des BSC–Kanals für $\varepsilon = 0.1$, $p_0 = 0.2$ berechnet und in **Aufgabe Z3.9** seine Kanalkapazität wie folgt angegeben:

$$C_{\rm BSC} = 1 - H_{\rm bin}(\varepsilon).$$

Eigenschaften symmetrischer Kanäle (1)

Die Kapazitätsberechnung des (allgemeinen) **diskreten gedächtnislosen Kanals** ist oftmals aufwändig. Sie vereinfacht sich entscheidend, wenn Symmetrien des Kanals ausgenutzt werden. Die Grafik zeigt zwei Beispiele.



- Beim *gleichmäßig dispersiven* Kanal (englisch: *Uniformly Dispersive Channel*) ergibt sich für alle Quellensymbole $x \in X$ die genau gleiche Menge an Übergangswahrscheinlichkeiten $\Rightarrow$ $\{P_{Y\vert X}(y_\kappa\vert x)\}$ mit $1 \le \kappa \le \vert Y\vert$. In der linken Grafik ist dies durch die Werte q, r und s mit $q + r + s = 1$ angedeutet.
- Beim gleichmäßig fokussierenden Kanal (englisch: *Uniformely Focusing Channel*) ergibt sich für alle Sinkensymbole $y \in Y$ die gleiche Menge an Übergangswahrscheinlichkeiten $\Rightarrow$ $\{P_{Y\vert X}(y\vert x_\mu)\}$ mit $1 \le \mu \le \vert X\vert$. Hier muss nicht notwendigerweise $t + u + v = 1$ gelten (siehe rechte Grafik).

Definition: Ist ein diskreter gedächtnisloser Kanal sowohl gleichmäßig dispersiv als auch gleichmäßig fokussierend, so liegt ein **streng symmetrischer Kanal** (englisch: *Strongly Symmetric Channel*) vor. Bei gleichverteiltem Quellenalphabet besitzt dieser die Kapazität

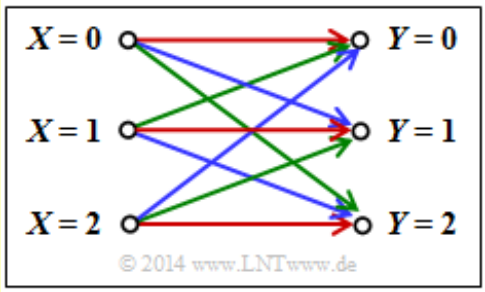
$$C = \log_2 \vert Y\vert + \sum_{y \in Y} P_{Y\vert X}(y\vert x) \cdot \log_2 P_{Y\vert X}(y\vert x).$$

Für diese Gleichung kann jedes beliebige $x \in X$ herangezogen werden.

Diese Definition soll durch ein Beispiel verdeutlicht werden.

Beispiel: Beim betrachteten Kanal bestehen Verbindungen zwischen allen $\vert X\vert = 3$ Eingängen und allen $\vert Y\vert = 3$ Ausgängen:

- Eine rote Verbindung steht für $P_{Y\vert X}(y_\kappa\vert x_\mu) = 0.7$.
- Eine blaue Verbindung steht für $P_{Y\vert X}(y_\kappa\vert x_\mu) = 0.2$.
- Eine grüne Verbindung steht für $P_{Y\vert X}(y_\kappa\vert x_\mu) = 0.1$.



Nach obiger Gleichung gilt dann für die Kanalkapazität:

$$C = \log_2 (3) + 0.7 \cdot \log_2 (0.7) + 0.2 \cdot \log_2 (0.2) + 0.1 \cdot \log_2 (0.1) = 0.4282 \text{ bit}.$$

Hinweis: Der Zusatz „die gleiche Menge an Übergangswahrscheinlichkeiten" bedeutet nicht, dass $P_{Y\vert X}(y_\kappa\vert x_1) = P_{Y\vert X}(y_\kappa\vert x_2) = P_{Y\vert X}(y_\kappa\vert x_3)$ gelten muss. Vielmehr geht in diesem Beispiel von jedem

Eingang ein roter, ein blauer und ein grüner Pfeil ab und an jeden Ausgang kommt ein roter, ein blauer und ein grüner Pfeil an. Die jeweiligen Reihenfolgen permutieren. R – G – B, B – R – G, G – B – R.

Eigenschaften symmetrischer Kanäle (2)

Ein Beispiel für einen streng symmetrischen Kanal ist der **Binary Symmetric Channel** (BSC). Dagegen ist der **Binary Erasure Channel** (BEC) nicht streng symmetrisch, da er

- zwar gleichmäßig dispersiv ist,
- aber nicht gleichmäßig fokussierend.

Nachfolgende Definition ist weniger restriktiv als die vorherige des streng symmetrischen Kanals.

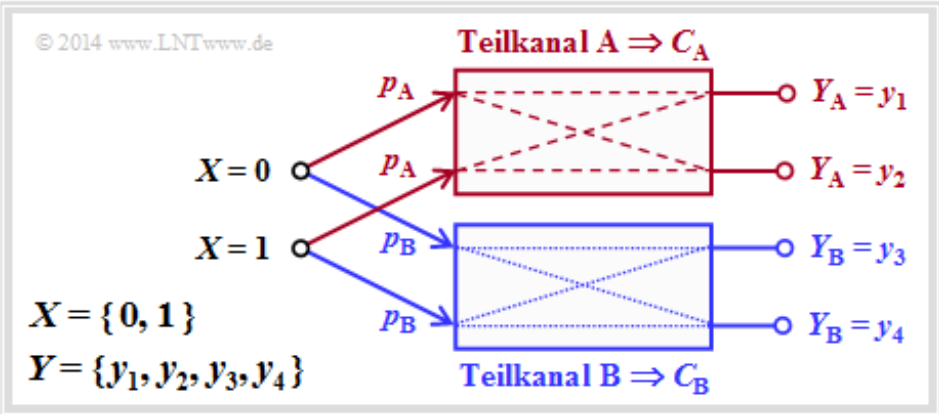
Definition: Ein **symmetrischer Kanal** (englisch: *Symmetric Channel*) liegt vor, wenn er in mehrere (allgemein L) streng symmetrische Teilkanäle aufgeteilt werden kann, indem das Ausgangsalphabet Y in L Teilmengen Y_1, ..., Y_L aufgespalten wird. Ein solcher symmetrischer Kanal besitzt folgende Kapazität:

$$C = \sum_{l=1}^{L} p_l \cdot C_l \,.$$

Hierbei sind folgende Bezeichnungen verwendet:

- p_l gibt die Wahrscheinlichkeit an, dass der l–te Teilkanal ausgewählt wird,
- C_l ist die Kanalkapazität dieses l–ten Teilkanals.

Die Grafik verdeutlicht diese Definition für $L = 2$, wobei die Teilkanäle mit A und B bezeichnet sind. An den unterschiedlich gezeichneten Übergängen (gestrichelt oder gepunktet) erkennt man, dass die zwei Teilkanäle durchaus verschieden sind, so dass $C_\text{A} \neq C_\text{B}$ gelten wird.



Für die Kapazität des Gesamtkanals erhält man somit allgemein:

$$C = p_\text{A} \cdot C_\text{A} + p_\text{B} \cdot C_\text{B} \,.$$

Über die Struktur der beiden Teilkanäle wird hier keine Aussage gemacht. Im Beispiel auf der nächsten Seite wird sich zeigen, dass auch der BEC durch diese Grafik grundsätzlich beschreibbar ist. Allerdings müssen dann die zwei Ausgangssysmbole y_3 und y_4 zu einem einzigen Symbol zusammengefasst werden.

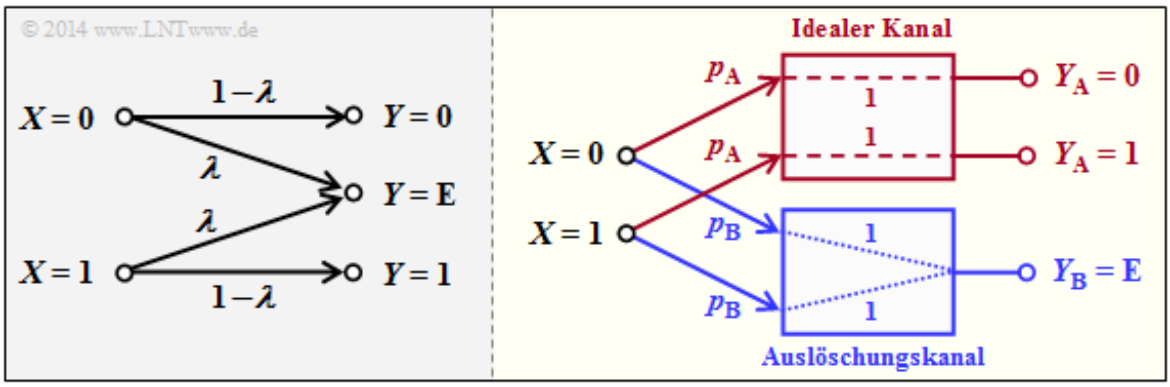
Eigenschaften symmetrischer Kanäle (3)

Beispiel: Die linke Grafik zeigt den **Binary Erasure Channel** (BEC) mit Eingang $X = \{0, 1\}$ und Ausgang $Y = \{0, 1, E\}$, wie er meistens gezeichnet wird. Teilt man diesen entsprechend der rechten Grafik auf in

- einen idealen Kanal ($y = x$) mit $y \in Y_A = \{0, 1\} \Rightarrow C_A = 1$ bit,
- einen Auslöschungskanal mit $y \in Y_B = \{E\} \Rightarrow C_B = 0$,

so ergibt sich mit den Teilkanalgewichtungen $p_A = 1 - \lambda$ und $p_B = \lambda$ für die Kanalkapazität:

$$C_{\rm BEC} = p_{\rm A} \cdot C_{\rm A} = 1 - \lambda.$$



Beide Kanäle sind streng symmetrisch. Für den (idealen) Kanal A gilt gleichermaßen

- für $X = 0$ und $X = 1$: $\Pr(Y = 0|X) = \Pr(Y = 1|X) = 1 - \lambda \Rightarrow$ gleichmäßig dispersiv,
- für $Y = 0$ und $Y = 1$: $\Pr(Y|X = 0) = \Pr(Y|X = 1) = 1 - \lambda \Rightarrow$ gleichmäßig fokussierend.

Entsprechendes gilt für den Auslöschungskanal B.

In **Aufgabe A3.11** wird sich zeigen, dass die Kapazität des Kanalmodells **Binary Symmetric Error & Erasure Channel** (BSEC) in gleicher Weise berechnet werden kann. Mit

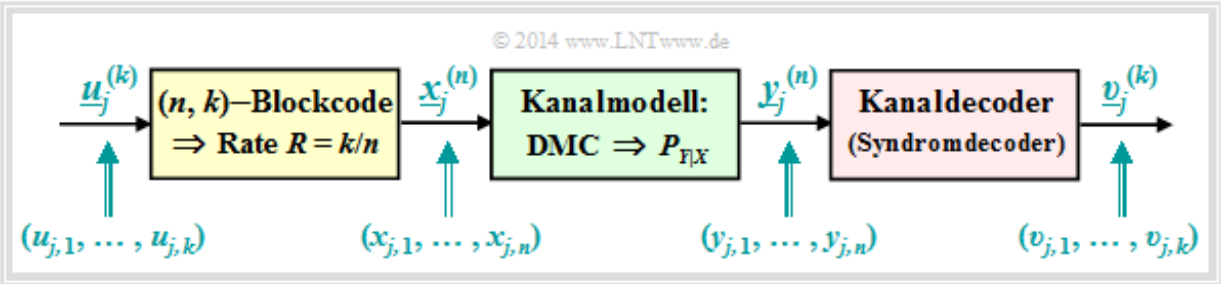
- der Verfälschungswahrscheinlichkeit ε und
- der Auslöschungswahrscheinlichkeit λ

erhält man in diesem Fall:

$$C_{\rm BSEC} = (1 - \lambda) \cdot \left[1 - H_{\rm bin}\left(\frac{\varepsilon}{1 - \lambda}\right)\right].$$

Einige Grundlagen der Kanalcodierung (1)

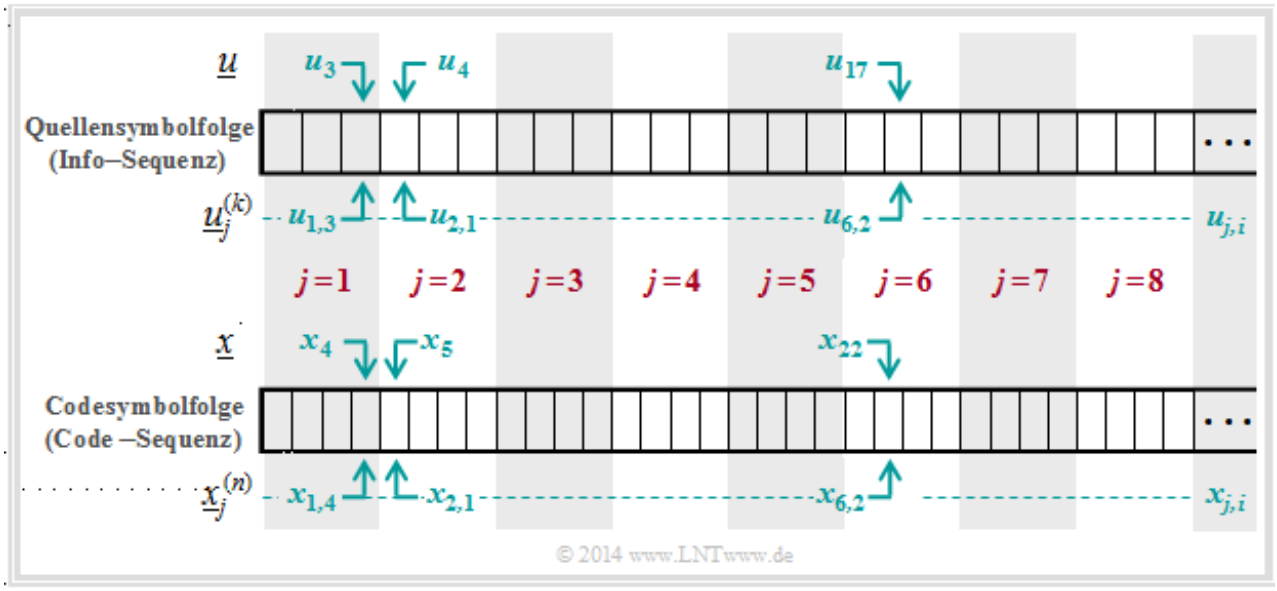
Um das Kanalcodierungstheorem richtig interpretieren zu können, sind einige Grundlagen der Kanalcodierung (englisch: *Channel Coding*) erforderlich. Dieses äußerst wichtige Gebiet der Nachrichtentechnik wird in einem eigenen **LNTwww–Buch** behandelt. Die nachfolgende Beschreibung bezieht sich auf das stark vereinfachte Modell für **binäre Blockcodes:**



Zu diesem Blockschaltbild ist anzumerken:

- Die unendlich lange Quellensymbolfolge $\underline{u}$ (hier nicht dargestellt) wird in Blöcke zu jeweils k bit unterteilt. Wir bezeichnen den Informationsblock mit der laufenden Nummerierung j mit $\underline{u}_j^{(k)}$.
- Jeder Informationsblock j mit $\underline{u}_j^{(k)}$ wird durch den gelb hinterrlegten Kanalcoder in ein Codewort $\underline{x}_j^{(n)}$ umgesetzt, wobei $n > k$ gelten soll. Das Verhältnis $R = k/n$ bezeichnet man als die *Coderate*.
- Der *Discrete Memoryless Channel* (DMC) wird durch die Übergangswahrscheinlichkeit $P_{Y|X}(\cdot)$ berücksichtigt. Dieser grün hinterlegte Block bewirkt Fehler auf Bitebene $\Rightarrow y_{j,\,i} \neq x_{j,\,i}$.
- Damit unterscheiden sich auch die aus n Bit bestehenden Empfangsblöcke $\underline{y}_j^{(n)}$ von den Codeworten $\underline{x}_j^{(n)}$. Ebenso gilt im allgemeinen für die Blöcke nach dem Deoder: $\underline{v}_j^{(k)} \neq \underline{u}_j^{(k)}$.

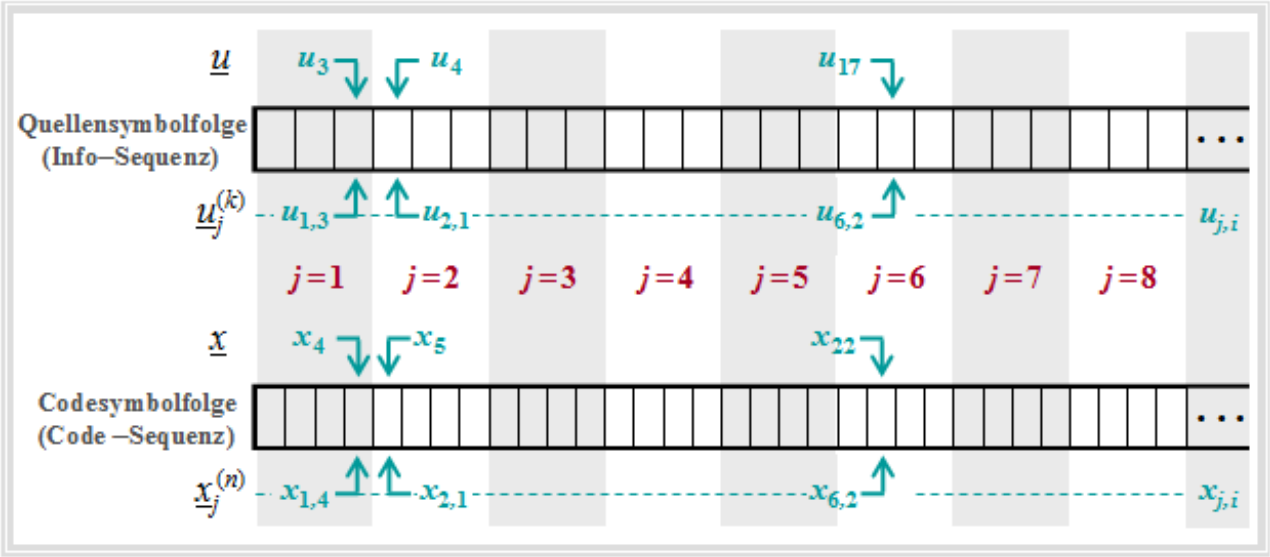
Die Grafik soll die hier verwendete Nomenklatur am Beispiel $k = 3$, $n = 4$ verdeutlichen. Dargestellt sind die jeweils ersten acht Blöcke der Informationssequenz und der Codesequenz.



Die Bildbeschreibung folgt auf der nächsten Seite.

Einige Grundlagen der Kanalcodierung (2)

Die Grafik soll die hier verwendete Nomenklatur am Beispiel $k = 3$ und $n = 4$ verdeutlichen.



Dargestellt sind die jeweils ersten acht Blöcke der Informationssequenz $\underline{u}$ und der Codesequenz $\underline{x}$. Man erkennt folgende Zuordnung zwischen der geblockten und der ungeblockten Beschreibung:

- Bit 3 des 1. Info–Blocks ⇒ $u_{1,\,3}$ entspricht dem Symbol u_3 in ungeblockter Darstellung.
- Bit 1 des 2. Info–Blocks ⇒ $u_{2,\,1}$ entspricht dem Symbol u_4 in ungeblockter Darstellung.
- Bit 2 des 6. Info–Blocks ⇒ $u_{6,\,2}$ entspricht dem Symbol u_{17} in ungeblockter Darstellung.
- Bit 4 des 1. Codewortes ⇒ $x_{1,\,4}$ entspricht dem Symbol x_4 in ungeblockter Darstellung.
- Bit 1 des 2. Codewortes ⇒ $x_{2,\,1}$ entspricht dem Symbol x_5 in ungeblockter Darstellung.
- Bit 2 des 6. Codewortes ⇒ $x_{6,\,2}$ entspricht dem Symbol x_{22} in ungeblockter Darstellung.

Zur Interpretation des Kanalcodierungstheorems benötigen wir noch verschiedene Definitionen für „Fehlerwahrscheinlichkeiten”. Aus dem **Systemmodell** lassen sich folgende Größen ableiten:

- Die **Kanalfehlerwahrscheinlichkeit** ergibt sich beim vorliegenden Kanalmodell zu

  $$\Pr(\text{Kanalfehler}) = \Pr\left(y_{j,\,i} \neq x_{j,\,i}\right) .$$

  Beispielsweise ist beim BSC–Modell Pr(Kanalfehler) = ε für alle $j = 1, 2, \text{...}$ und $1 \le i \le n$.
- Die **Blockfehlerwahrscheinlichkeit** bezieht sich auf die zugeordneten Informationsblöcke am Codereingang ⇒ $\underline{u}_j^{(k)}$ und am Decoderausgang ⇒ $\underline{v}_j^{(k)}$, jeweils in Blöcken zu k Bit:

  $$\Pr(\text{Blockfehler}) = \Pr\left(\underline{v}_j^{(k)} \neq \underline{u}_j^{(k)}\right) .$$

- Die **Bitfehlerwahrscheinlichkeit** bezieht sich ebenfalls auf den Eingang und den Ausgang des betrachteten Codiersystems, allerdings auf Bitebene:

  $$\Pr(\text{Bitfehler}) = \Pr\left(v_{j,\,i} \neq u_{j,\,i}\right) .$$

  Hierbei ist vereinfachend vorausgesetzt, dass alle k Bit $u_{j,\,i}$ des Informationsblockes j mit gleicher Wahrscheinlichkeit verfälscht werden ($1 \le i \le k$). Andernfalls müsste über die k Bit gemittelt

werden.

Einige Grundlagen der Kanalcodierung (3)

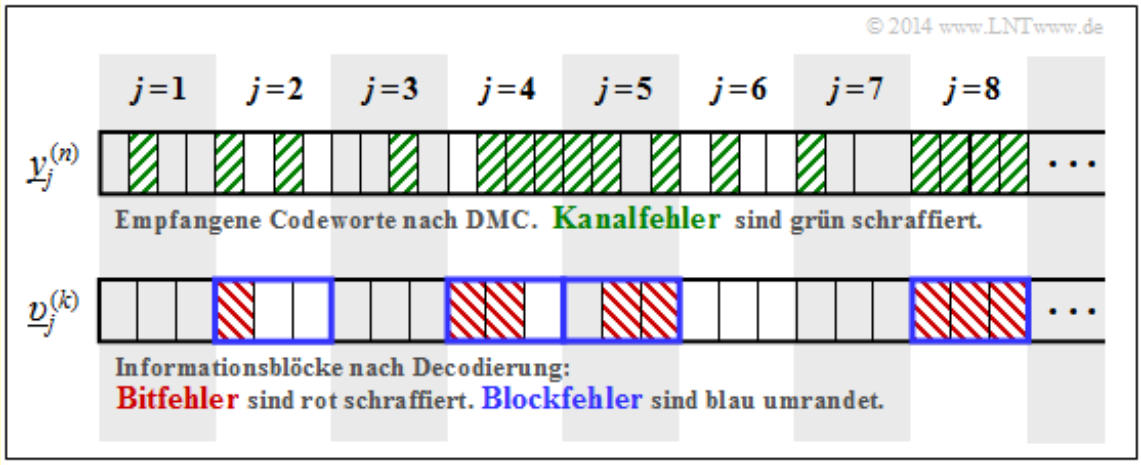
Zwischen Blockfehler– und Bitfehlerwahrscheinlichkeit besteht allgemein der Zusammenhang:

$$1/k \cdot \Pr(\text{Blockfehler}) \le \Pr(\text{Bitfehler}) \le \Pr(\text{Blockfehler}) .$$

- Die untere Schranke ergibt sich, wenn bei allen fehlerhaften Blöcken alle Bit falsch sind.
- Gibt es in jedem fehlerhaften Block genau nur einen einzigen Bitfehler, dann ist die Bitfehlerwahrscheinlichkeit Pr(Bitfehler) identisch mit der Blockfehlerwahrscheinlichkeit Pr(Blockfehler).

Beispiel: Die Grafik zeigt oben die ersten acht Empfangsblöcke $\underline{y}_j^{(n)}$ mit $n = 4$. Kanalfehler sind grün schraffiert. Unten ist die Ausgangsfolge $\underline{v}$ skizziert, unterteilt in Blöcke $\underline{v}_j^{(k)}$ zu je $k = 3$ Bit:

- Bitfehler sind im unteren Diagramm rot schraffiert.
- Blockfehler erkennt man an der blauen Umrahmung.



Hierzu einige (aufgrund der kurzen Folge) vage Angaben zu den Fehlerwahrscheinlichkeiten:

- Die Hälfte der Empfangsbits sind grün schraffiert. Daraus folgt:

  $$\Pr(\text{Kanalfehler}) = 16/32 = 1/2.$$

- Die Bitfehlerwahrscheinlichkeit lautet mit der beispielhaften Codierung & Decodierung:

  $$\Pr(\text{Bitfehler}) = 8/24 = 1/3.$$

- Dagegen würde bei uncodierter Übertragung gelten:

  $$\Pr(\text{Bitfehler}) = \Pr(\text{Kanalfehler}) = 1/2.$$

- Die Hälfte der decodierten Blöcke sind blau umrandet. Daraus folgt:

  $$\Pr(\text{Blockfehler}) = 4/8 = 1/2.$$

Mit Pr(Blockfehler) = 1/2 und $k = 3$ liegt die Bitfehlerwahrscheinlichkeit in folgendem Bereich:

$$1/6 \le \Pr(\text{Bitfehler}) \le 1/2 .$$

- Die obere Schranke ergibt sich, wenn in jedem der vier verfälschten Blöcke alle Bit falsch sind: Pr(Bitfehler) = 12/24 = 1/2.
- Die untere Schranke beschreibt den Fall, dass in jedem der vier verfälschten Blöcke jeweils nur

ein Bit falsch ist: Pr(Bitfehler) = 4/24 = 1/6.

Rate, Kanalkapazität und Bitfehlerwahrscheinlichkeit

Durch Kanalcodierung wird die Zuverlässigkeit (englisch: *Reliability*) der Datenübertragung von der Quelle zur Sinke erhöht. Vermindert man die Coderate $R = k/n$ und erhöht so die hinzugefügte Redundanz $(1 - R)$, so wird im allgemeinen die Datensicherheit verbessert und damit die Bitfehlerwahrscheinlichkeit herabgesetzt, die wir im Weiteren kurz $p_{\rm B}$ nennen:

$$p_{\rm B} = {\rm Pr(Bitfehler)} = {\rm Pr}\left(v_{j,\,i} \neq u_{j,\,i}\right) .$$

Das folgende Theorem basiert auf dem **Data Processing Theorem** und *Fano's Lemma*. Die Herleitung kann in den Standardwerken zur Informationstheorie nachgelesen werden, zum Beispiel in **[CT06]:**

Umkehrung des Shannonschen Kanalcodierungstheorems:
Benutzt man zur Datenübertragung mit Rate R einen Kanal mit unzureichender Kanalkapazität $C < R$, so kann auch bei bestmöglicher Kanalcodierung die Bitfehlerwahrscheinlichkeit $p_{\rm B}$ eine untere Schranke nicht unterschreiten:

$$p_{\rm B} \geq H_{\rm bin}^{-1} \cdot (1 - C/R) > 0 .$$

$H_{\rm bin}(\cdot)$ bezeichnet hierbei die **binäre Entropiefunktion.**

Da die Wahrscheinlichkeit der Blockfehler nie kleiner sein kann als die der Bitfehler, ist für $R > C$ auch die Blockfehlerwahrscheinlichkeit „0" nicht möglich. Aus dem angegebenen Bereich für die Bitfehler,

$$1/k \cdot {\rm Pr(Blockfehler)} \leq {\rm Pr(Bitfehler)} \leq {\rm Pr(Blockfehler)} ,$$

lässt sich auch ein Bereich für die Blockfehlerwahrscheinlichkeit angeben:

$${\rm Pr(Bitfehler)} \leq {\rm Pr(Blockfehler)} \leq k \cdot {\rm Pr(Bitfehler)} .$$

Beispiel: Verwendet man einen Kanal mit der Kapazität $C = 1/3$ (bit) zur Datenübertragung mit der Coderate $R < 1/3$, so ist prinzipiell die Bitfehlerwahrscheinlichkeit $p_{\rm B} = 0$ möglich.

- Allerdings ist aus dem Kanalcodierungstheorem der spezielle (k, n)–Blockcode nicht bekannt, der dieses Wunschergebnis ermöglicht. Shannon macht hierzu keine Aussagen.
- Bekannt ist nur, dass ein solcher bestmöglicher Code mit unendlich langen Blöcken arbeitet. Bei gegebener Coderate $R = k/n$ gilt somit sowohl $k \to \infty$ als auch $n \to \infty$.
- Deshalb ist die Aussage „Die Bitfehlerwahrscheinlichkeit ist 0" nicht identisch mit „Es treten keine Bitfehler auf": Auch bei endlich vielen Bitfehlern und $k \to \infty$ gilt $p_{\rm B} = 0$.

Mit der Coderate $R = 1$ (uncodierte Übertragung) erhält man:

$$p_{\rm B} \geq H_{\rm bin}^{-1} \cdot \left(1 - \frac{1/3}{1.0}\right) = H_{\rm bin}^{-1}(2/3) \approx 0.174 > 0 .$$

Mit der Coderate $R = 1/2 > C$ ist die Bitfehlerwahrscheinlichkeit zwar kleiner, aber nicht 0:

$$p_{\rm B} \geq H_{\rm bin}^{-1} \cdot \left(1 - \frac{1/3}{1/2}\right) = H_{\rm bin}^{-1}(1/3) \approx 0.062 > 0 .$$

Aufgabenhinweis: **A3.12: Coderate und Zuverlässigkeit** – **A3.13: Kanalcodierungstheorem**