Review

- $\frac{d}{dx}(x\ln x) = \ln x + 1 = \ln ex$
- $\frac{d}{dx}(x\log x) = \log x + \frac{1}{\ln 2} = \log ex$
- $\frac{d}{dx}(-x\log x) = -\log x - \frac{1}{\ln 2}$.
- $h(p) = -p\log p - (1-p)\log(1-p)$
    - $\frac{dh}{dp}(p) = \log\frac{1-p}{p}$
    - $\frac{d}{dx}h(f(x)) = \left(\frac{df}{dx}(x)\right)\log\frac{1-f(x)}{f(x)}$
    - $2^{h(p)} = p^{-p}(1-p)^{-(1-p)}$
    - $2^{-h(p)} = p^{p}(1-p)^{1-p}$
    - $h\left(\frac{1}{b}\right) = \log b - \frac{b-1}{b}\log(b-1)$
- **Probability transition matrix**: $\left[Q(y|x)\right]_{r,c} = \Pr\left[y=c|x=r\right]$. (The entry in the x^{th} row and the y^{th} column denotes $Q(y|x)$.)
    - Let input distribution be a row vector $\vec{p}^T$. Then the output distribution would be a row vector $\vec{q}^T = \vec{p}^T Q$.

Markov String/Process: X_1^n: $p(x_1^n) = p(x_1)p(x_2|x_1)\cdots p(x_k|x_{k-1})p(x_{k+1}|x_k)\cdots p(x_n|x_{n-1})$

- Ordered substring of Markov string is Markov

  For $0 \le n_1 < n_2 < \cdots < n_k \le n$, $\left(X_{n_1}, X_{n_2}, \ldots, X_{n_k}\right)$ is also a Markov string.

- Given X_k (the present), we have X_1^{k-1} (the past) and X_{k+1}^n (the future) are independent: $p\left(x_1^{k-1}, x_{k+1}^n | x_k\right) = p\left(x_{k+1}^n | x_k\right)p\left(x_1^{k-1}|x_k\right)$.
    - Given X_k, we have X_{k-1} and X_{k+1} are independent: $p\left(x_{k-1}, x_{k+1}|x_k\right) = p\left(x_{k-1}|x_k\right)p\left(x_{k+1}|x_k\right)$.
- For $1 \le k < n$, and $k+1 < m \le n$, $p\left(x_{k+1}^m | x_k\right) = p\left(x_{k+1}|x_k\right)\cdots p\left(x_m|x_{m-1}\right)$.

- $H\left(X_1^{k-1}, X_{k+1}^n | X_k\right) = H\left(X_1^{k-1}|X_k\right) + H\left(X_{k+1}^n|X_k\right)$
- $I\left(X_1^{k-1}; X_{k+1}^n | X_k\right) = 0$
- $H\left(X_{k+1}^n | X_k\right) = \sum_{i=k+1}^{n} H\left(X_i|X_k\right)$

- $p\left(x_{k+1}^m \middle| x_j^k\right) = p\left(x_{k+1}^m \middle| x_k\right)$
- $H\left(X_{k+1}^n \middle| X_1^k\right) = H\left(X_{k+1}^n \middle| X_k\right)$
- $H\left(X_0 \middle| X_k\right)$ is increasing in k: $H\left(X_0 \middle| X_{n+1}\right) \geq H\left(X_0 \middle| X_n\right)$
- $I\left(X_0; X_k\right)$ is decreasing in k: $I\left(X_0; X_n\right) \geq I\left(X_0; X_{n+1}\right)$
- Markov 3-string: $X - \ominus - Y - \ominus - Z$: $p(x,y,z) = p(x)\, p\left(y \middle| x\right) p\left(z \middle| y\right)$
    - Also Markov in the reverse direction: $p(z,y,x) = p\left(x \middle| y,z\right) p(y,z) = p\left(x \middle| y\right) p\left(y \middle| z\right) p(z)$.
    - $I\left(Z;X \middle| Y\right) = I\left(X;Z \middle| Y\right) = 0$.
    - **Data processing theorem**: $I(X;Y) \geq I(X;Z)$, $I(Z;Y) \geq I(Z;X)$; hence, closer $\Rightarrow$ more I.
    - $I(X;Y) \geq I\left(X;Y \middle| Z\right)$
- $\underline{U} - \ominus - \underline{X} - \ominus - \underline{Y} - \ominus - \underline{V} \Rightarrow I\left(\underline{X};\underline{Y}\right) \geq I\left(\underline{U},\underline{V}\right)$.
- Stationary Markov process: $H\left(X_n\right)$ is constant. (by stationaryness). $H\left(X_n \middle| X_1\right)$ increases with n.

Convexity:

- $H(Y)$ is a concave $\cap$ function of $p(x)$ for fixed $Q(y|x)$.
- $H\left(Y \middle| X\right)$ is a linear function of $p(x)$ for fixed $Q(y|x)$.
- $I(X;Y)$ is a continuous concave $\cap$ function of $p(x)$ for fixed $Q(y|x)$.
- $I(X;Y)$ is a convex $\cup$ function of $Q(y|x)$ for fixed $p(x)$.

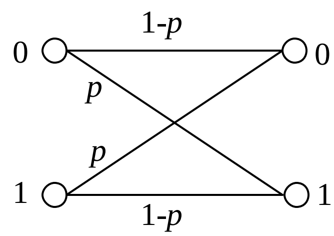
$I(X;Y)$

- $I(X;Y) = 0$ if $|\mathcal{X}| = 1$ or $|\mathcal{Y}| = 1$.
- To find $I(p,Q)$, first find $q(y)$. Then, find $H(Y)$. Next, find $H\left(Y \middle| X\right) = \sum_x p(x) H\left(Y \middle| X = x\right)$. Finally, $I(X;Y) = I(p,Q) = H(Y) - H\left(Y \middle| X\right)$.
- **Parallel d.m.c. channel.** Let Y_1^n be the result of passing X_1^n trough a d.m.c. (n use.)
    - $Q\left(y_1^n \middle| x_1^n\right) = \prod_{i=1}^n Q_i\left(y_i \middle| x_i\right)$.
    - $H\left(Y_1^n \middle| X_1^n\right) = \sum_{i=1}^n H\left(Y_i \middle| X_i\right)$
    - $\forall p\left(x_1^n\right)$ $I\left(X_1^n; Y_1^n\right) \leq \sum_{i=1}^n I\left(X_i; Y_i\right)$ with equality if $p\left(x_1^n\right) = \prod_{i=1}^n p_i\left(x_i\right)$.
- Independent source: U_1^L has independent components $\Rightarrow I\left(U_1^L; V_1^L\right) \geq \sum_{\ell=1}^L I\left(U_\ell; V_\ell\right)$.

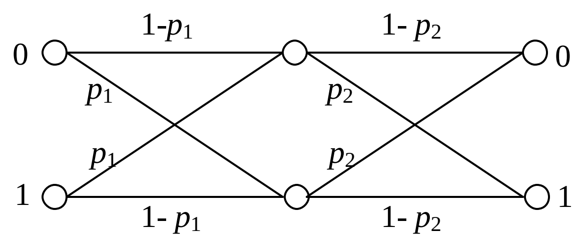
- Let U,V discrete random variables. $M = |\mathcal{U}| = |\mathcal{V}|$. $\mathcal{U} = \mathcal{V} = \{0,1,\ldots,M-1\}$.
    - **Fano Inequality**: $H(U|V) \le h(P_e) + P_e \log(M-1)$.
        - Note: $P_e = 0 \Rightarrow H(U|V) = 0$.
- **Extended Fano inequality**: Let $U_1^L, V_1^L \in \mathcal{U}^L = \mathcal{V}^L$. $\frac{H(U_1^L|V_1^L)}{L} \le h(\overline{P_e}) + \overline{P_e}\log(M-1)$

BSC(p)

- Binary symmetric channel (BSC) with crossover probability p. $\underline{x} \longrightarrow \boxed{BSC(p)} \longrightarrow \underline{y}$.



- $\begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1-2p & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}^{-1}$.
- p also = probability of error.
- $H(Y|X) = h(p)$ regardless of $\{p(x)\}$. (by the "symmetric")
- $C = 1-h(p)$.
- Two BSC's in series is a BSC with transition probability $p^{(2)} = p_1q_2 + p_2q_1$. $q^{(2)} = p_1p_2 + q_1q_2$.



$$Q^{(2)} = Q_1Q_2 = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} (1-2p_1)(1-2p_2) & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}^{-1}.$$

- n BSC's in series is a BSC with transition probability $\boxed{p^{(n)} = \frac{1-\prod_{i=1}^{n}(1-2p_i)}{2}}$.

$$Q^{(n)} = \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \prod_{i=1}^{n} (1-2p_i) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}^{-1}$$

- Two identical BSC's in series is a BSC with transition probability $p^{(2)} = 2pq$.
- n identical BSC's in series is a BSC with transition probability $p^{(n)} = \frac{1-(1-2p)^n}{2}$.

Binary *n*-Sphere

- Binary n-cube: geometric representation of $\{0,1\}^n$. $\underline{x} = (x_1,\ldots,x_n) \in \{0,1\}^n$.
- Combinatoric fact:
  - $\binom{n}{j}$ is increasing for $j \le \left[\frac{n}{2}\right]$ and max when $j = \left[\frac{n}{2}\right]$.
  - **Stirling's approximation**: $\sqrt{2p}\, n^n n^{\frac{1}{2}} e^{-n+\frac{1}{12n+1}} < n! < \sqrt{2p}\, n^n n^{\frac{1}{2}} e^{-n+\frac{1}{12n}} \Rightarrow \ln n! = n \ln n - n + o(n)$ as $n \to \infty$. $\left(\lim_{n\to\infty} \frac{o(n)}{n} = 0\right)$.
  - $\log_D n! = n \log n - \frac{1}{\ln D} n + o(n) = n \log n + an + o(n)$
- The **hamming sphere** of radius r around $\underline{x}$ is $S_r(\underline{x}) = \left\{ \underline{y} \in \{0,1\}^n ; d_H(\underline{x},\underline{y}) \le r \right\}$.
  - $\left|S_r(\underline{x})\right|$ = Volume of radius r sphere = $\sum_{i=0}^{r} \binom{n}{i}$.
  - Let $S_k = S_k(\bar{0}) \subset \{0,1\}^n$.
  - $\binom{n}{k} \le |S_k| \le (k+1)\binom{n}{k}$
  - Let $k = a n$ where $0 < a \le \frac{1}{2}$. Then, $\lim_{n\to\infty} \frac{1}{n} \log \underbrace{|S_{an}|}_{\text{volume}} = \lim_{n\to\infty} \frac{1}{n} \log \underbrace{\sum_{i=0}^{an} \binom{n}{i}}_{\text{volume}} = \lim_{n\to\infty} \frac{1}{n} \log \underbrace{\binom{n}{an}}_{\text{surface shell}} = H(a)$: The rate of growth of exponential of the volume is the same as that of the surface. $|S_{an}| \sim 2^{nH(a)}$.
  - Multinomial extension: for $\sum_{i=1}^{M} a_i n = 1$, $\lim_{n\to\infty} \frac{1}{n} \log \frac{n!}{\prod_{i=1}^{M} (a_i n)!} = H(a_1,\ldots,a_n)$.
- $d_H(\underline{x},\underline{y}) = \|\underline{x} - \underline{y}\|_1 = \left|\{k : 1 \le k \le n,\ x_k \ne y_k\}\right|$
- Given $\underline{x} \in \{0,1\}^n$, $\binom{n}{i}$ = the number of sequences $\underline{y} \in \{0,1\}^n$ with $d_H(\underline{x},\underline{y}) = i$.

Coding

- **Hamming (*n*,*k*) code** consists of 2^k elements in $\{0,1\}^n$
- ***e*-error correcting perfect code**: every $\underline{x} \in \{0,1\}^n$ is in *e*-sphere around some codeword. These *e*-spheres don't overlap.
  - The only binary perfect codes are
    - Hamming
    - Golay code (*e* = 3)
    - Repetition code (*e* = *m*) where *n* = 2*m*+1 (w/ majority vote)
- ***e*-error correcting perfect hamming code**: $\left(\sum_{i=0}^{e}\binom{n}{i}\right)2^{n-k} = 2^n$, or equivalently, $\sum_{i=0}^{e}\binom{n}{i} = 2^k$.
- ***e*-error correcting sphere-packed code**: every $\underline{x} \in \{0,1\}^n$ is in (*e*+1)-sphere around some codeword. These *e*-spheres don't overlap.
- Hamming (7,4) code is
  - not unique unless require $\vec{0}$ in the code.
  - $S_1(\underline{x}) \cap S_1(\underline{y}) = \emptyset$ if $\underline{x}, \underline{y} \in \mathcal{C}$ and $\underline{x} \neq \underline{y}$
  - 1-error correcting perfect code: $\left(\binom{7}{0}+\binom{7}{1}\right)2^{7-4} = 2^7$.
- Hamming (23,11) code
  - 3-error correcting perfect code: $\left(\sum_{i=0}^{3}\binom{23}{i}\right)2^{23-11} = 2^{23}$.
- **Rate of an (*n*,*k*) code** is $R = \frac{k}{n}$. $\Rightarrow$ rate $\frac{k}{n}$ code.
- $\forall R\ \ 0 < R < 1$, $\exists\ n(R)$ such that no sphere-packed codes of blocklength $n > n(R)$ exist.
- Geometric arguments
  - $\forall e > 0$ as $n \to \infty$, received word when $\underline{x}$ is sent falls falls in the spherical shell of width $\pm n\varepsilon$ or normalized width $\pm e$.
  - $\exists R^* > 0$ = max of $R > 0$ such that 2^{nR} words can be put into $\{0,1\}^n$ with sphere of radius *np* around them not overlapping.
  - **Sphere hardening**: For any $e > 0$, as $n \to \infty$, $\underline{y} \in S = S_{np+ne}(\underline{x}) \setminus S_{np-ne}(\underline{x})$ = the sphere of radius *np* with width *2ne* with probability 1.
  - For BSC(*p*), $R^* > C = 1 - H(p) = 1 + p\log p + (1-p)\log(1-p)$??
- **[Block Coding]**

- Block code: $\mathcal{C} \subset \{0,1\}^n$. Blocklength n.
- Size: $|\mathcal{C}|$ = # of codewords.
- **Rate** $R_c = \boxed{R = \frac{\log|\mathcal{C}|}{n}}$ [(info) bits per channel use] if messages are a priori equiprobable.

Capacity for Discrete memoryless channel.

- An **(*n*, *R*, l) code** for a discrete memoryless channel with input alphabet $\mathcal{X}$, and output alphabet $\mathcal{Y}$, is a collection $\mathcal{C}$ of 2^{nR} codewords each belonging to $\mathcal{X}^n$. $\left(2^{nR} \le |\mathcal{X}|^n\right)$. l is the probability of error.
    - $(\text{block length}, \#\text{ codewords}, \min\Pr[\text{error}])$ code.
- Some idea: For BSC(p), we have sphere of radius np with volume (or surface) ~ $2^{nH(p)}$. Whole space has 2^n sequences. So, if can build perfect code for which np-spheres around words don't overlap, then such code would have $\frac{2^n}{2^{nH(p)}} = 2^{n(1-H(p))}$ codewords. Code's rate would be $\frac{1}{n}\log_2 2^{n(1-H(p))} = 1 - H(p)$.
- **Operational definition of capacity**: capacity of a discrete memoryless channel is the sup of all rates R such that there is a sequence of $\left(n, 2^{nR}, l\right)$ codes for which $l \to 0$ as $n \to \infty$.
- Consider discrete memoryless channel with transition matrix $\left[Q(y|x)\right], x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $|\mathcal{X}|, |\mathcal{Y}| < \infty$. Let $I(p,Q) = I(X;Y)$ be the average mutual information between channel input and channel output when input random variable has p.m.f. $\{p(x), x \in \mathcal{X}\}$. Then, $\boxed{C = \max_p I(p,Q)}$ [bits per channel use].
    - C is unique, but $\operatorname{argmax}_p I(p,Q)$ may not unique.

Discrete Memoryless Channel (DMC)

- $\underline{x} = (x_1, \ldots, x_n)$, $\underline{y} = (y_1, \ldots, y_n)$. $\underline{x} \longrightarrow \boxed{\begin{matrix}\text{Channel} \\ Q_{\underline{Y}|\underline{X}}\left(\underline{y}|\underline{x}\right)\end{matrix}} \longrightarrow \underline{y}$.
- $Q_{\underline{Y}|\underline{X}}\left(\underline{y}|\underline{x}\right) = \prod_{k=1}^{n} Q\left(y_k|x_k\right)$. $\{Q(y|x), x \in X, y \in Y\}$ is fixed.
    - **Probability transition matrix**: $\left[Q(y|x)\right]_{r,c} = \Pr\left[y = c|x = r\right]$. (The entry in the x^{th} row and the y^{th} column denotes $Q(y|x)$.)

$$\begin{array}{cc} & Y \\ X & \begin{bmatrix} \ddots & \vdots & \iddots \\ \cdots & Q(Y=y|X=x) & \cdots \\ \iddots & \vdots & \ddots \end{bmatrix} \end{array}$$

- $P\left(x_1^n, y_1^n\right) = p\left(x_1^n\right)\prod_{i=1}^{n} Q\left(y_i \middle| x_i\right)$
- Let $N = \{1,\dots,n\}$, and $A \subset M \subset N$, then $p\left(\vec{x}_M, \vec{y}_A\right) = p\left(x_M\right)\prod_{\substack{\ell \\ \ell \in A}} Q\left(y_\ell \middle| x_\ell\right)$.
- $Q\left(y_i \middle| \underset{i\in I_1}{x_{I_1}}, \underset{i\notin I_2}{y_{I_2}}\right) = Q\left(y_i \middle| x_i\right)$.
- $p\left(x_I, y_I\right) = p\left(x_I\right)\prod_{\substack{\ell \\ \ell \in I}} p\left(y_\ell \middle| x_\ell\right)$
- For $i \in I \subset \{1,\dots,n\}$, $p\left(x_I, y_i\right) = p\left(x_I\right) p\left(y_i \middle| x_i\right)$
- $p\left(y_i \middle| \underset{i\in I}{x_I}\right) = p\left(y_i \middle| x_i\right)$
- $p\left(y_i \middle| x_i, y_k\right) = p\left(y_i \middle| x_i\right) \quad k \neq i$
- $p\left(y_1^k, x_1^n\right) = p\left(x_1^n\right)\prod_{i=1}^{k} p\left(y_i \middle| x_i\right)$
- $p\left(y_i \middle| y_1^{i-1}, x_1^n\right) = \dfrac{p\left(y_1^i, x_1^n\right)}{p\left(y_1^{i-1}, x_1^n\right)} = \dfrac{p\left(x_1^n\right)\prod_{k=1}^{i} p\left(y_k \middle| x_k\right)}{p\left(x_1^n\right)\prod_{k=1}^{i-1} p\left(y_k \middle| x_k\right)} = p\left(y_i \middle| x_i\right)$.
- $I\left(X_1^n; Y_1^n\right) \leq \sum_{i=1}^{n} I\left(X_i; Y_i\right)$

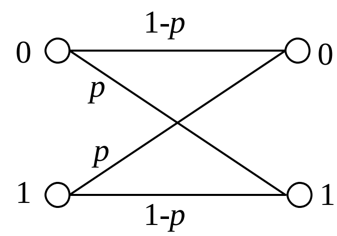
D.M.C. with i.i.d. inputs

- Setup: Channel is d.m.c. $\{Q(y|x)\}$: $Q\left(\vec{y}\middle|\vec{x}\right) = \prod_{k=1}^{n} Q\left(y_k \middle| x_k\right)$. The input X_i's to the channel is i.i.d.

  $\Pr\left[X_1^n = x_1^n\right] = \prod_{k=1}^{n} p\left(x_k\right)$.
- $\left(X_i, Y_i\right)$ is i.i.d. i.e. $P\left(\vec{x}, \vec{y}\right) = \prod_{k=1}^{n} P\left(x_k, y_k\right) = \prod_{k=1}^{n} p\left(x_k\right) Q\left(y_k \middle| x_k\right)$
- Y_i is i.i.d. i.e. $q\left(\vec{y}\right) = \prod_{k=1}^{n} q\left(y_k\right)$

Capacity

- $0 \le C \le \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$.
- Any channel with only one input letter or only one output letter has zero capacity. $\left(I(X;Y) \equiv 0\right)$.
- BSC(p): C = 1- $H(p)$ is achieved when the input is uniform.
  - $H(Y|X) = H(p)$.
  - $I(X;Y) = H(Y) - H(p)$.

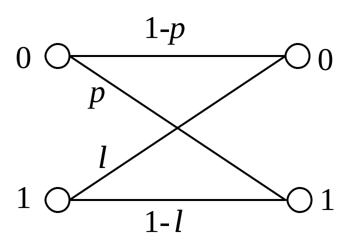


- Cascade of n identical BSC(p): is a BSC with transition probability $p^{(n)} = \dfrac{1 - \prod_{i=1}^{n}(1-2p_i)}{2}$.

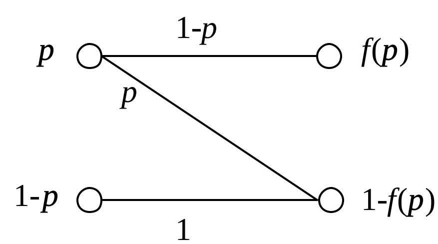
$$X_0 \longrightarrow \boxed{\mathrm{BSC}(p)} \xrightarrow{X_1} \boxed{\mathrm{BSC}(p)} \longrightarrow \boxed{\mathrm{BSC}(p)} \longrightarrow X_n$$

  - For $0 < p < 1$, because $|1-2p| < 1$, $\lim_{n\to\infty} p^{(n)} = \frac{1}{2}$. $\lim_{n\to\infty} C^{(n)} = 1 - H\left(\frac{1}{2}\right) = 0$. $0 \le I(X_0; X_n) \le C^{(n)} \Rightarrow$ $\lim_{n\to\infty} I(X_0; X_n) = 0$ for any initial distribution.

- Binary asymmetric channel:
  - $f(p_1) = (1-p) + (p + l - 1)p_1$.
  - $H(Y) = h(f(p_1))$.
  - $H(Y|X) = (1-p_1)h(p) + p_1 h(l)$.
  - For max $I(X;Y)$, need $\log\dfrac{1-f(p_1)}{f(p_1)} = \dfrac{h(l) - h(p)}{p + l - 1}$.



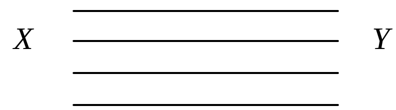
- Z-channel:
  - Let p be the probability of the X that that channel introduces noise.
  - $f(p) = (1-p)p$; $\log\dfrac{1-f(p)}{f(p)} = \dfrac{h(p)}{1-p}$.
  - $H(Y|X) = p\,h(p)$; $H(Y) = h(f(p))$
  - C is achieved when



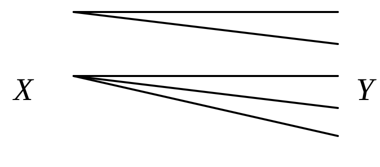
$$p = p^* = \frac{1}{(1-p)\left(1+2^{\frac{h(p)}{1-p}}\right)} = \frac{1}{1-p+p^{-\frac{p}{(1-p)}}}$$

$$C = h\left((1-p)p^*\right) - p^* h(p).$$

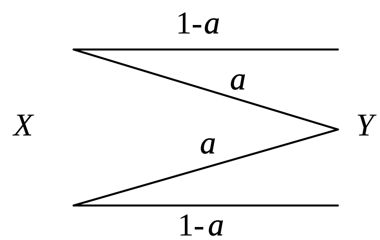
- Noiseless channel: $C = \log|\mathcal{X}|$ is achieved by uniform input distribution.
  - $H(Y|X) = 0$.



- Noisy channel with nonoverlapping outputs: $C = \log|\mathcal{X}|$ is achieved by uniform input distribution..
  - $H(Y|X) = 0$.



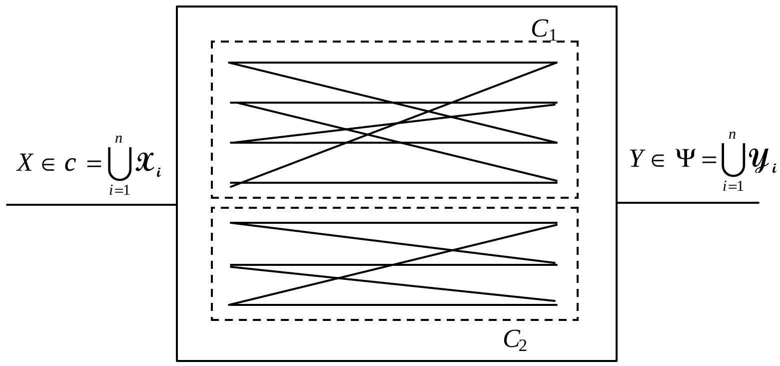
- Binary Erasure Channel: C = 1-a is achieved by uniform input distribution.



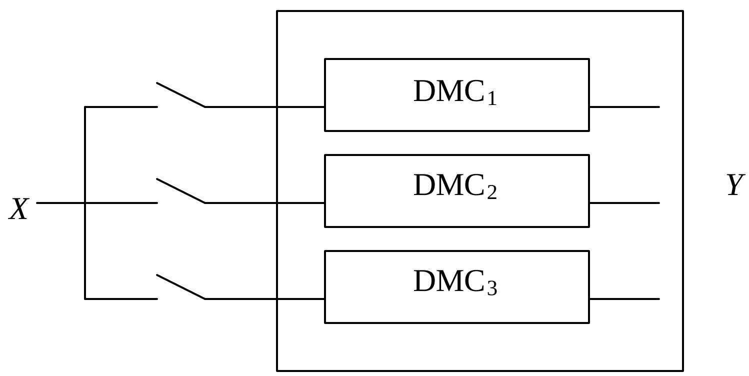
- **Weakly symmetric channel**: 1) every row of the transition matrix are permutations of each other, i.e., $\{Q(y|i)\}$ are permutation of $\{Q(y|j)\}$, and 2) all the column sums $\sum_x Q(y|x)$ are equal.
  - $\sum_x Q(y|x) = \frac{|\mathcal{X}|}{|\mathcal{Y}|}$.
  - Uniform distribution on the input alphabet implies uniform distribution on output.
  - $C = \log|\mathcal{Y}| - H(\text{rows of transition matrix})$ is achieved by a uniform distribution on the input alphabet.
  - $\forall x\ \ H(Y|X=x) = H(\vec{r}^T)$. Hence, $H(Y|X) = H(\vec{r}^T)$ where $\vec{r}^T$ is any row of the transition matrix..
  - $I(X;Y) = H(Y) - H(\vec{r}^T)$
- **Symmetric channels**: All the rows of the probability transition matrix $\left[Q(y|x)\right]_{r,c} = \Pr\left[y=c|x=r\right]$ are permutations of each other and so are the columns, i.e. 1) $\{Q(y|i)\}$ are just permutation of $\{Q(y|j)\}$, and 2) $\{Q(i|x)\}$ are just permutation of $\{Q(j|x)\}$.
  - Ex. $\mathcal{X} = \mathcal{Z} = \{0,1,\ldots,M-1\}$. $Y = (X+Z) \bmod M$.
  - Ex. BSC.

$\Rightarrow$ weakly symmetric. Hence, $C = \log|\mathcal{Y}| - H(\text{rows of transition matrix})$.

- **Sum channel**
    - Consider N DMC's with disjoint input alphabets and disjoint output alphabets and capacities C_1, C_2, …, C_N. Call these DMC's as the **subchannels**. The associated **sum channel** has input and output alphabets that are the unions of those of the sub channel, and for each input x, the transition probabilities $Q(\cdot|x)$ are the same as in the sub channel that has x in its input alphabet. In other words, the sum channel has all N subchannel available but only one subchannel may be used at any given time.



(Or can have common input alphabet but disjoint output alphabet, and selectable channel as shown below)



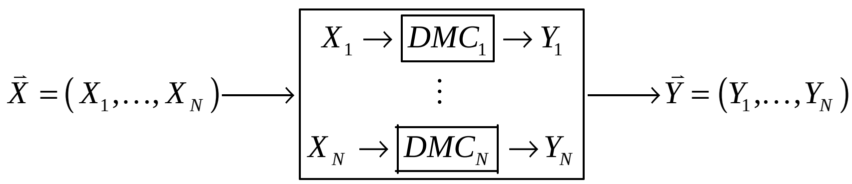
- Let $\mathcal{X}_i$ and $\mathcal{Y}_i$ be the input alphabet and the output alphabet of the i^{th} subchannel. $c = \bigcup_{i=1}^{n} \mathcal{X}_i$, $\Psi = \bigcup_{i=1}^{n} \mathcal{Y}_i$. $\mathcal{X}_i \cap \mathcal{X}_j = \varnothing$, and $\mathcal{Y}_i \cap \mathcal{Y}_j = \varnothing$ for for $i \neq j$. Let $w(n)$ be the probability that $X \in \mathcal{X}_n$, and $p_n(x) = \Pr\left[X = x \middle| X \in \mathcal{X}_n\right]$. Then, for $x \in \mathcal{X}_i$, $p(x) = w(i)\, p_i(x)$.
    - For $x \in \mathcal{X}_i$ and $y \in \mathcal{Y}_j$, $Q(y|x) = Q_i(y|x) d(j,i)$.
- Consider each subchannel,
    - Let $I_j(X;Y) = \sum_{x \in \mathcal{X}_j} p_j(x) \sum_{y \in \mathcal{Y}_j} Q_j(y|x) \log \dfrac{Q_j(y|x)}{\sum_{x \in \mathcal{X}_j} p_j(x) Q_j(y|x)}$. Then, $C_j = \max_{\{p_j(x), x \in \mathcal{X}_j\}} I_j(X;Y)$.

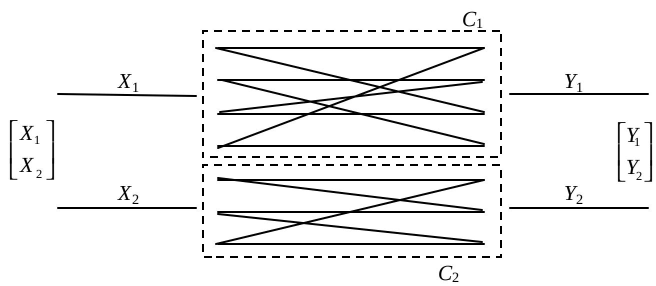
- For $y \in \mathcal{Y}_j$, $q(y) = \sum_{x \in \mathcal{X}_j} w(j) p_j(x) Q_j(y|x)$.

- $I(X;Y) = \sum_{x \in c} p(x) \sum_{y \in \Psi} Q(y|x) \log \dfrac{Q(Y|X)}{q(Y)} = \sum_{j=1}^{N} w(j) I_j(X;Y) - \sum_{j=1}^{N} w(j) \log w(j)$.

- $\boxed{C = \log \sum_{\ell=1}^{N} 2^{C_\ell}}$ is achieved by $\boxed{w(j) = \dfrac{2^{C_j}}{\sum_{\ell=1}^{N} 2^{C_\ell}}}$ and $\{p_j(x), x \in \mathcal{X}_j\} = \arg \max_{\{p_j(x), x \in \mathcal{X}_j\}} I_j(X;Y)$.

- **Parallel channel**:
  - Consider N DMC **subchannels** with capacities $C_1, C_2, \ldots, C_N$. The subchannels are connected in parallel in the sense that once each unit of time an arbitrary symbol is transmitted and received over each subchannel. The output from each subchannel depends only on the input to that channel. $\left(Q(y_1^N | x_1^N) = \prod_{j=1}^{N} Q_j(y_j | x_j) \right)$.





  - $\boxed{C = \sum_{i=1}^{n} C_i}$ is achieved by independent input $p(x_1^n) = \prod_{i=1}^{n} \tilde{p}_i(x_i)$ where $\{\tilde{p}_i(x)\}$ is the distribution that achieves C_i for the i^{th} subchannel.

- Cascade BSC's: A cascade of n identical BSCs each with transition probability p is equivalent to a single BSC with

- Iterative Calculation of C: The **Arimoto-Blahut Algorithm**.
  - Given a DMC with transition probabilities $Q(y|x)$ and any distribution $p_0(x)$.
  - Define a sequence $p_r(x)$, $r = 0, 1, \ldots$ according to the iterative prescription:

$$q_r(y) = \sum_x p_r(x) Q(y|x). \ \log c_r(x) = \sum_y Q(y|x) \log \frac{Q(y|x)}{q_r(y)}. \ p_{r+1}(x) = \frac{p_r(x) c_r(x)}{\sum_x p_r(x) c_r(x)}.$$

- 1) Set $p_0 = \left[p_0(0), p_0(1), \ldots \right]$ (row vector). $Q = \begin{bmatrix} \Pr[0 \to 0] & \Pr[0 \to 1] & \\ \Pr[1 \to 0] & \Pr[1 \to 1] & \\ & & \ddots \end{bmatrix}$. Let $Q(x)$ be the row x of Q.
  2) $q_r = p_r Q$.
  3) For each x, $\log c_r(x) = -H(\text{row } x \text{ of } Q) - \text{sum}\left((\text{row } x \text{ of } Q).*\log(q_r)\right)$.
  4) For each x, $c_r(x) = 2^{\log c_r(x)}$. Form $c_r = c_r = \left[c_r(0), c_r(1), \ldots \right]$.
  5) $temp = \text{sum}(p_r.*c_r)$.
  6) $p_{r+1} = \frac{1}{temp}(p_r.*c_r)$.
  7) Repeat 2) – 6)
- $f\left(p_{r+1}, Q, \hat{P}_r\right) = \max_p f\left(p, Q, \hat{P}_r\right) \geq f\left(p_r, Q, \hat{P}_r\right)$
- $I\left(p_{r+1}(x), Q\right) \geq I\left(p_r(x), Q\right)$; thus, $I(p_r, Q)$ is monotonic increasing with r.
- $\log\left(\sum_x p_r(x) c_r(x)\right) \leq C \leq \log\left(\max_x c_r(x)\right)$
- $\lim_{r \to \infty} I(p_r, Q) = C$ if $\forall x \; p_0(x) > 0$

System

- Feedback: all the received symbols Y_i are sent back immediately and noiselessly to the transmitter, which can then use them to decide which symbol to send next.
  - Feedback can help enormously in simplifying encoding and decoding. However, it can not increase the capacity of d.m.c.
- Source-Channel coding theorem: we can transmit a stationary ergodic source over a channel if and only if its entropy rate is less than the capacity of the channel.

  $$\xrightarrow{U_1^n} \boxed{\text{Encoder}} \xrightarrow{X_1^n} \boxed{Q(y|x)} \xrightarrow{Y_1^n} \boxed{\text{Decoder}} \xrightarrow{V_1^n}$$

  - $P_e^{(n)} = \Pr\left[U_1^n \neq V_1^n\right] = \sum_{y_1^n} \sum_{u_1^n} p(u_1^n) Q\left(y_1^n \middle| \text{Enc}(u_1^n)\right) I_{\left\{\text{Dec}\left(y_1^n\right) \neq u_1^n\right\}}$
  - If $U_1, U_2, \ldots, U_n$ is a finite alphabet stochastic process that satisfies the AEP (ex. stationary ergodic source), then there exists a source channel code with $P_e^{(n)} \to 0$ if source entropy rate $H(\mathcal{U}) < C$.
  - For any stationary stochastic process, if $H(\mathcal{U}) > C$, the probability of error is bounded away from zero, and it is not possible to send the process over the channel with arbitrary low probability of error.

Info transmission theorem with stationary source and DMC

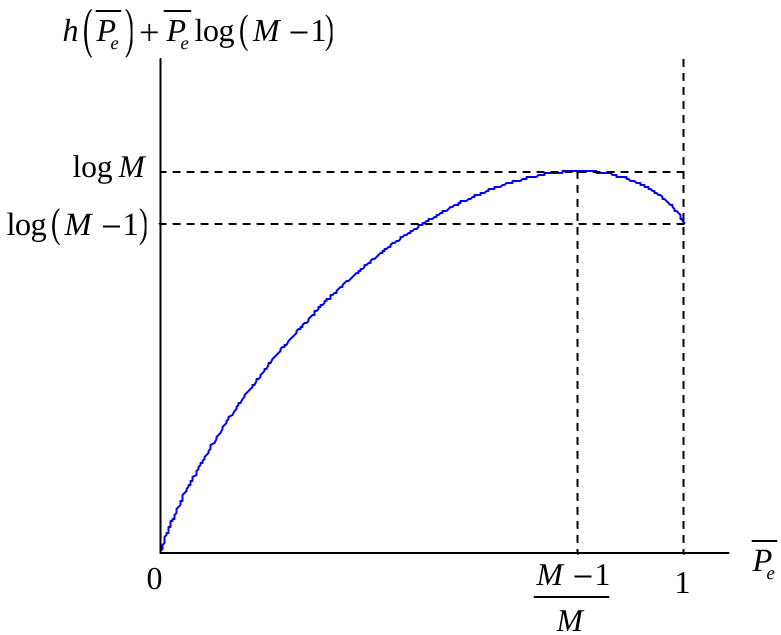
$$\boxed{\text{Source}} \xrightarrow{U_1^L} \boxed{\text{Encoder}} \xrightarrow{X_1^N} \boxed{\text{Channel}} \xrightarrow{Y_1^N} \boxed{\text{Decoder}} \xrightarrow{V_1^L} \boxed{\text{User}}.$$

Fig 1.

- $\underline{U} — \ominus — \underline{X} — \ominus — \underline{Y} — \ominus — \underline{V}$
- Let $\mathcal{U} = \mathcal{V} = \{0,1,\ldots,M-1\}$ M-ary. $\underline{U} = (U_1,\ldots,U_L)$, $\underline{X} = (X_1,\ldots,X_N)$, $\underline{Y} = (Y_1,\ldots,Y_N)$, $\underline{V} = (V_1,\ldots,V_N)$. Source $\{U_\ell\}$ is stationary with entropy rate H. Channel is DMC with $Q = [Q(y|x)]$.
- Define
  - $P_{e,\ell} = \Pr[V_\ell \neq U_\ell]$.
  - Average error probability/frequency: $\overline{P_e} = \frac{1}{L}\sum_{\ell=1}^{L} P_{e,\ell}$ = expected frequency of errors.
  - $C = \max_{\{p(x)\}} I(p,Q)$ = **channel capacity in bits per channel use**.
  - $C' = \frac{N}{L}C$ = **channel capacity in bits per source letter**.
- Information transmission theorem for stationary source and discrete memoryless channel
  1) If $H > C' = \frac{n}{L}C$ [bit per source letter], then $\overline{P_e} > 0$.
  2) For any $R < C$ and any $e > 0$, we can find a code $(M = 2^{nR}, n)$ of rate R and sufficiently large block length n for which $\max_j P_e(j) < e$.

Weak Converse info transmission theorem when channel is DMC

- Let $U_1^L, V_1^L \in \mathcal{U}^L = \mathcal{V}^L$.
- $H - \frac{nC}{L} = H - C' \leq \frac{H\left(U_1^L \middle| V_1^L\right)}{L} \leq h\left(\overline{P_e}\right) + \overline{P_e}\log(M-1) \leq h\left(\overline{P_e}\right) + \overline{P_e} nR_c$
- If $H > C'$, then $\overline{P_e} > 0$.



- Information transmission with an arbitrary small expected frequency of errors is not possible if the source entropy $H = \lim_{L\to\infty}\frac{H\left(U_1^L\right)}{L}$ [bit per source symbol] exceeds the channel capacity C' measured in bit per source symbol. This conclusion holds even if one permits unbounded computation effort and is willing to tolerate enormous coding delay ($L \to \infty$ and $N \to \infty$ with L/N kept fixed).

Typical Sequences for i.i.d. X_i

- Let $\mathcal{X}$ be a discrete alphabet. $\left\{p(x), x\in\mathcal{X}\right\}$ be a p.m.f. $\vec{X} = \left(X_1,\ldots,X_n\right) = X_1^n$. $\vec{x} = \left(x_1,\ldots,x_n\right) = x_1^n \in \mathcal{X}^n$.
- $N\left(x\middle|\vec{x}\right) = \left|\left\{k : 1\le k\le n, x_k = x\right\}\right|$.
  - $\forall \vec{x}\in\mathcal{X}^n, \sum_{x\in\mathcal{X}} N\left(x\middle|\vec{x}\right) = n$.
- The **composition / type** of $\vec{x} = \left\{N\left(x\middle|\vec{x}\right) : \forall x\in\mathcal{X}\right\}$.
- Def. Given d, $\vec{x}$ is **d-typical** of $\left\{p(x), x\in\mathcal{X}\right\}$ if $\forall x\in\mathcal{X}, \left|\frac{N\left(x\middle|\vec{x}\right)}{n} - p(x)\right| < \frac{d}{|\mathcal{X}|}$.
- $T_d = T_d(p) = \left\{\vec{x}\in\mathcal{X}^n : \vec{x} \text{ is } d\text{-typical}\right\}$.
- $p_{\vec{X}}(\vec{x}) = \prod_{k=1}^{n} p(x_k) = \prod_{x\in\mathcal{X}} p(x)^{N\left(x\middle|\vec{x}\right)}$. Thus, word probabilities $p_{\vec{X}}(\vec{x})$ depend only on word type.
- $\vec{x}\in T_d(p) \Rightarrow$
  - $\forall x\in\mathcal{X}, \left|\frac{N\left(x\middle|\vec{x}\right)}{n} - p(x)\right| < \frac{d}{|\mathcal{X}|}$

- $\left| -\frac{\log p_{\vec{X}}(\vec{x})}{n} - H(\{p(x)\}) \right| < d\left|\log p_{\min}\right| = e_d$.
  - $e_d > 0$.
  - e_d can be made arbitrary small by making d small enough. $\left(\lim_{d\to 0} e_d = 0\right)$.
  - Define $H^+ = H(\{p(x)\}) + e_d$, $H^- = H(\{p(x)\}) - e_d$.
  - $2^{-nH^+} < p_{\vec{X}}(\vec{x}) < 2^{-nH^-}$.
- $\left|T_d(p)\right| \le 2^{nH^+}$
- Weak Law of Large Number: $\forall x \in \mathcal{X}\ \forall e > 0\ \lim_{n\to\infty} \Pr\left[\left|\frac{N(x|\vec{X})}{n} - p(x)\right| > e\right] = 0$.
- $\forall d > 0\ \lim_{n\to\infty} \Pr\left[\vec{X} \in T_d\right] = 1$. $\lim_{n\to\infty} \Pr\left[\vec{X} \notin T_d\right] = 0$.

Jointly Typical Sequences for i.i.d. (X_i,Y_i)

- Let $\{P(x,y) = p(x)Q(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$ be the joint pmf over $\mathcal{X}\times\mathcal{Y}$. $|\mathcal{X}\times\mathcal{Y}| = |\mathcal{X}||\mathcal{Y}|$.
- $N(x,y|\vec{x},\vec{y}) = \left|\{k : 1 \le k \le n, (x_k,y_k) = (x,y)\}\right|$
  - $N(x|\vec{x}) = \sum_{y\in\mathcal{Y}} N(x,y|\vec{x},\vec{y})$
- $(\vec{x},\vec{y})$ are **jointly d-typical** of $\{P(x,y) = p(x)Q(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$ iff

  $$\boxed{\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}\ \left|\frac{N(x,y|\vec{x},\vec{y})}{n} - P(x,y)\right| < \frac{d}{|\mathcal{X}||\mathcal{Y}|}}.$$

- $T_d = T_d(pQ) = \{(\vec{x},\vec{y}) : (\vec{x},\vec{y}) \text{ is } d\text{-typical of } \{P(x,y) = p(x)Q(y|x)\}\}$
- $\lim_{n\to\infty} \Pr\left[(\vec{X},\vec{Y}) \notin T_d\right] = 0$
- $(\vec{x},\vec{y}) \in T_d(pQ) \Rightarrow$
  - $2^{-n(H(P)+e_d)} < P(\vec{x},\vec{y}) < 2^{-n(H(P)-e_d)}$ where $e_d = d\left|\log P_{\min}(x,y)\right|$.
  - $\vec{x} \in T_d(p), \vec{y} \in T_d(q)$: Jointly Typicality $\Rightarrow$ Marginal Typicality.

    If $(\vec{x},\vec{y})$ is δ-typical of $\{P(x,y)\}$, then

    $\vec{x}$ is δ-typical of $\left\{p(x) = \sum_y P(x,y)\right\}$, and $\vec{y}$ is δ-typical of $\left\{q(y) = \sum_x P(x,y)\right\}$.

- $\Rightarrow p(\vec{x}) \le 2^{-nH(p)^-}$, $q(\vec{y}) \le 2^{-nH(q)^-}$

- $\left|T_d(P)\right| \le 2^{nH(P)^+}$.

Direct Noisy-Channel Coding Theorem for d.m.c.

- **Block code**: $\mathcal{C} = \left\{\vec{x}^{(1)}, \dots, \vec{x}^{(M)}\right\}$. $\vec{x}^{(i)} \in \mathcal{X}^n$. **Rate of the code** $= R = \frac{\log M}{n}$. $M = 2^{nR}$. Hence, $\left(2^{nR}, n\right)$ code.
- **Random code selection / random coding**: generate $\mathcal{C}$ at random according to the distribution $p = \left\{p(x), x \in \mathcal{X}\right\}$:

  Let $\mathcal{C} = \left\{\vec{X}^{(1)}, \dots, \vec{X}^{(M)}\right\}$ be a randomly chosen block code such that all nM letters (of the M codewords each with n letters) are i.i.d. $\left\{p(x)\right\}$.

  - Note that once code is select, you use it in a deterministic way.
  - $\Pr\left[\vec{X}^{(j)} = x_1^n\right] = \prod_{k=1}^{n} p(x_k)$
  - $\Pr\left[X_1^n = x_1^n\right] = \prod_{k=1}^{n} p(x_k)$
  - Thus, the channel input is i.i.d. with $p = \left\{p(x), x \in \mathcal{X}\right\}$.
  - $\Pr\left[\vec{X}^{(\ell)} = \vec{x} \middle| J = j\right] = \Pr\left[\vec{X}^{(\ell)} = \vec{x}\right] = \prod_{k=1}^{n} p(x_k)$
- Independent of the random code, let J be the **random message index** with pmf $\left\{p_j, 1 \le j \le M\right\}$. If $J = j$, then the components $\vec{X}_1^{(j)}, \vec{X}_1^{(j)}, \dots, \vec{X}_n^{(j)}$ of $\vec{X}^{(j)}$ will be put into the channel in this order during n successive channel uses.
  - So, by knowing J, we know log(M) bits in n channel uses.
- $\vec{Y} = \left(Y_1, \dots, Y_n\right) \equiv$ the resulting channel output vector.
- $\hat{J} \equiv \hat{J}\left(\vec{Y}\right)$ = the decoder's estimate of J based on $\vec{Y}$.
- $\overline{P}_e = \Pr\left[\hat{J} \ne J\right]$ which depends on the joint distribution of J, $\mathcal{C}$, $\vec{Y}$, and D.

- **Joint typicality decoding rule**: Upon receiving $\vec{y}$, if $\exists!\, j^*\ 1 \le j^* \le M\ \left(\vec{x}_{j^*}, \vec{y}\right) \in T_d$, decode $\vec{y}$ by $\hat{J}(\vec{y}) = j^*$. If no such index or if there is more than one such index, declare a decoding error.

- $\overline{P}_e(j) = \Pr\left[\hat{J} \ne J \middle| J = j\right]$ (averaged over all code).
  - By symmetry of code selection scheme, $\forall j\ \overline{P}_e(j) = \overline{P}_e(1)$.
- Overall average error probability: $\overline{P}_e = \Pr\left[\hat{J} \ne J\right] = \sum_{j=1}^{M} p_j \overline{P}_e(j) = \overline{P}_e(1)$.

- Probability facts:
  - $(\vec{x},\vec{y}) \in T_d(pQ) \Rightarrow p_{\vec{X}^{(\ell)}}(\vec{x}) \le 2^{-nH(p)^-}$, and $q_{\vec{Y}}(\vec{y}) \le 2^{-nH(q)^-}$
  - For $\ell = 1$
    - $Q_{\vec{Y}|\vec{X}^{(1)},J}(\vec{y}|\vec{x},1) = \prod_{k=1}^{n} Q(y_k|x_k)$.
    - $P_{\vec{X}^{(1)},\vec{Y}|J}(\vec{x},\vec{y}|1) = \prod_{k=1}^{n} p(x_k)Q(y_k|x_k)$.
  - For $\ell \ne 1$,
    - $\{\vec{X}^{(\ell)}, 2 \le \ell \le M\}$ is independent of $\vec{Y}$.
    - $Q_{\vec{Y}|\vec{X}_\ell,J}(\vec{y}|\vec{x},1) = Q_{\vec{Y}}(\vec{y}) = \prod_{k=1}^{n} q(y_k)$
    - $P_{\vec{X}^{(\ell)},\vec{Y}|J}(\vec{x},\vec{y}|1) = p_{\vec{X}^{(\ell)}}(\vec{x})q_{\vec{Y}}(\vec{y}) = p_{\vec{X}}(\vec{x})q_{\vec{Y}}(\vec{y})$ same for all $\ell \ne 1$.
    - $\Pr\left\{\left(\vec{X}^{(\ell)},\vec{Y}\right) \in T_d(P)\middle|J=1\right\} = \Pr\left\{\left(\vec{X}^{(2)},\vec{Y}\right) \in T_d(P)\middle|J=1\right\} \le 2^{nH(pQ)^+}2^{-nH(p)^-}2^{-nH(q)^-} = 2^{-nI(p,Q)^-}$, where $I(p,Q)^- = I(p,Q) - e_d$ where $e_d = d\left(\left|\log\min_x p(x)\right| + \left|\log\min_y q(y)\right| + \left|\log\min_{x,y} p(x)Q(y|x)\right|\right)$.
- $\overline{P}_e(1) = \Pr[E_1 \cup E_2] \le \Pr[E_1] + \Pr[E_2]$
  - $E_1 = \left\{\left(\vec{X}_1,\vec{Y}\right) \notin T_d(pQ)\middle|J=1\right\}$. $\lim_{n\to\infty}\Pr[E_1] = 0$ because $\left(\vec{X}_1,\vec{Y}\right) \sim pQ$.
  - $E_2 = \bigcup_{\ell>1}\left\{\left(\vec{X}_\ell,\vec{Y}\right) \in T_d(pQ)\middle|J=1\right\}$. $\Pr[E_2] \le 2^{-n\left(I(p,Q)^- - R\right)}$.
    - If $I(p,Q)^- > R$, $\lim_{n\to\infty}\Pr[E_2] = 0$.

- The channel coding theorem:
  - Direct: All rates below capacity C are achievable.
    - $\forall R \;\; R < C$, there exists a sequence of $\left(2^{nR},n\right)$ codes along which the error probability decays to zero as $n \to \infty$. (regardless of the probabilities P_j of the messages.)
    - $\forall R \;\; R < C$, $\forall e > 0 \; \exists$ a $\left(2^{nR},n\right)$ code of rate R and sufficiently large block length n for which $\max_j P_e(j) < e$.
  - Conversely, any sequence of $\left(2^{nR},n\right)$ codes with $\lim_{n\to\infty} l^n = 0$ must have $R \le C$.

Etc

- Separation Theorem for source and channel coding.

Let $\{U_k\}$ be an ergodic stationary information source with entropy rate H. If $H < C$, the capacity of the DMC $Q(y|x)$, then it is possible to convey $\underline{U}$ through the channel with an arbitrary small probability of error. Employ a source code with rate R_s and a channel code with rate R_c such that $H < R_s < R_c < C$ $[\text{bit/sec}]$.

- Asymptotic optimality can be achieved by separating source coding and channel coding.
- The source code and the channel code can be designed separately without losing asymptotic optimality