Theorem 1 (Shannon)**.** *Let $\mathcal{C}_{AB}$ be a discrete memoryless channel from A to B with capacity $\mathbf{c}$. Let $\mathcal{U}$ be a stationary process on the alphabet C satisfying the a.e.p. asymptotic equipartition property and $I(\mathcal{U}) = r < \mathbf{c}$. Then for any $\delta > 0$ there is an $n \in \mathbb{N}$ and a mapping $m\colon C^n \to A^n$ such that the values of $\mathcal{U}$ can be determined from the outputs of the combined channel $\mathcal{C}_{AB} \circ m$ with an error probability less than δ.*

Proof. For the proof we first consider an input process $\mathcal{X}$ on A, which is i.i.d. and has $\mathcal{T}(\mathcal{X}, \mathcal{Y}) = \mathbf{c}$ (and $\mathcal{C}_{AB}\colon\ \mathcal{X} \rightsquigarrow \mathcal{Y}$).

From the a.e.p. we can infer that for any $\epsilon > 0$ there is an $n \in \mathbb{N}$ such that

1. $p\Big[\Big|\frac{1}{n} \cdot \mathcal{N}\big(\widetilde{X}_1 \cap \ldots \cap \widetilde{X}_n\big) - I(\mathcal{X})\Big| > \epsilon\Big] < \epsilon$
2. $p\Big[\Big|\frac{1}{n} \cdot \mathcal{N}\big(\widetilde{Y}_1 \cap \ldots \cap \widetilde{Y}_n\big) - I(\mathcal{Y})\Big| > \epsilon\Big] < \epsilon$
3. $p\Big[\Big|\frac{1}{n} \cdot \mathcal{N}\big(\widetilde{X}_1 \cap \ldots \cap \widetilde{X}_n \cap \widetilde{Y}_1 \cap \ldots \cap \widetilde{Y}_n\big) - I(\mathcal{X}, \mathcal{Y})\Big| > \epsilon\Big] < \epsilon$
4. $p\Big[\Big|\frac{1}{n} \cdot \mathcal{N}\big(\widetilde{U}_1 \cap \ldots \cap \widetilde{U}_n\big) - I(\mathcal{U})\Big| > \epsilon\Big] < \epsilon$

From Prop. **??** we can also estimate the number of the corresponding high–probability sequences, i.e. $N = \#(A_{n,\epsilon})$, $\#(B_{n,\epsilon})$, $M = \#(C_{n,\epsilon})$, and also the number P of high–probability pairs.

Now the idea is to consider only the high–probability elements in C^n, A^n, B^n and $A^n \times B^n$, and to map each high–probability element $c = (c_1, \ldots, c_n) \in C_{n,\epsilon}$ onto a different randomly chosen $a = (a_1, \ldots, a_n)$ that is the first element in a high–probability pair (a, b). This procedure will work, if there are more such a‘s than there are high–probability c‘s, and if the probability of finding the first element a from the second element b in a high–probability pair (a, b) is sufficiently high. In this case we can guess first a and then c from the channel output b. In order to carry out the proof we now have to estimate the number of these a‘s appearing in high–probability pairs (a, b).

1. Given a high–probability a, we estimate the number N_a of high-probability pairs (a,b) containing a as follows:
   We use the abreviations $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$ and $U = (U_1, \ldots, U_n)$, and consider only high–probability elements $\omega \in \Omega$. Then
   $$p(\widetilde{Y}|\widetilde{X}) = \frac{p(\widetilde{X, Y})}{p(\widetilde{X})} \geq 2^{-n(I(\mathcal{X},\mathcal{Y})-I(\mathcal{X})+2\epsilon)}\,.$$
   Thus
   $$1 \geq \sum_{b \in B^n} p(b|a) \geq N_a 2^{-n(I(\mathcal{X},\mathcal{Y})-I(\mathcal{X})+2\epsilon)} \quad \text{and} \quad N_a \leq 2^{n(I(\mathcal{X},\mathcal{Y})-I(\mathcal{X})+2\epsilon)}\,.$$
   Now we have to make sure that
   $$M \leq \frac{P}{2^{n(I(\mathcal{X},\mathcal{Y})-I(\mathcal{X})+2\epsilon)}} \leq \frac{P}{N_a}\,.$$

Because of the estimates for M and P from Prop. **??** this is true if

$$2^{n(I(\mathcal{U})+\epsilon)} \le (1-\epsilon)2^{n(I(\mathcal{X})-3\epsilon)}\,.$$

Since $\mathcal{I}(\mathcal{X}) \ge \mathcal{T}(\mathcal{X},\mathcal{Y}) = c > r = \mathcal{I}(\mathcal{U})$ this is certainly true for sufficiently small ϵ.

2. Given a high-probability $b \in B^n$, we estimate the number N_b of high–probability pairs (a,b) in $A^n \times B^n$ containing b similarly to i):

$$p(\widetilde{X}|\widetilde{Y}) = \frac{p(\widetilde{X,Y})}{p(\widetilde{Y})} \ge 2^{-n(I(\mathcal{X},\mathcal{Y})-I(\mathcal{Y})+2\epsilon)}\,.$$

Thus

$$1 \ge \sum_{a\in A^n} p(a|b) \ge N_b 2^{-n(\mathcal{I}(\mathcal{X},\mathcal{Y})-\mathcal{I}(\mathcal{X})+2\epsilon)} \quad \text{and} \quad N_b \le 2^{n(\mathcal{I}(\mathcal{X},\mathcal{Y})-\mathcal{I}(\mathcal{X})+2\epsilon)}\,.$$

This number we use to estimate the probability that there is at most one $m(c)$ occurring as first component among the N_b pairs, for each of the high–probability b's at the channel output. More exactly, for a fixed high probability c we take $a = m(c)$ as channel input and obtain b as channel output. Now we ask for the probability p_f that there is another c' such that $(m(c'),b)$ is also a high–probability pair. For fixed b let n_b be the number of codewords $m(c')$ such that $(m(c'),b)$ is a high-probability pair. Now we can estimate

$$\begin{aligned} p_f \le p[n_b \ge 1] &< E(n_b) \\ &= M \cdot \frac{N_b}{N} \\ &\le 2^{n(I(\mathcal{U})+I(\mathcal{X},\mathcal{Y})-I(\mathcal{Y})-I(\mathcal{X})+4\epsilon)} \\ &= 2^{n(4\epsilon+r-c)}\,. \end{aligned}$$

Since $\mathcal{I}(\mathcal{U}) + \mathcal{I}(\mathcal{X},\mathcal{Y}) - \mathcal{I}(\mathcal{Y}) - \mathcal{I}(\mathcal{X}) = r - c < 0$, this probability is sufficiently small for sufficiently large n and sufficiently small ϵ.

This means that a high-probability c will be coded into a channel input a in such a way that with high probability a can be determined from the channel output b, and from a one can determine c. What is the error probability in this procedure?
An error may occur when c is not in the high-probability group, or (a,b) is not in the high-probability group, or b is not in the high-probability group, or if there is more than one $m(c)$ in N_b. Otherwise, we know which a we have chosen to correspond to the output b and we know which c has been mapped by m onto a.
Taking our various estimates together, the probability of error is at most $3\epsilon + 2^{n(4\epsilon+r-c)}$ and it remains to choose ϵ sufficiently small and n sufficiently large to finish the proof of the theorem. □