

THE DIRECT CODING THEOREM FOR A DMC VIA JOINT TYPICALITY

Remark: The proof that follows parallels that in the text in most respects. C&T use weak typicality. We use strong (joint) typicality, but its stronger properties are not invoked in this application like they will be when we consider rate-distortion theory later in the semester. The proof given below is self-contained in the sense that it does not refer out to the joint AEP; rather, what is needed along these lines is derived using only bounds on (1) the size of set of n -sequences from an alphabet that are δ -typical of a given distribution, and (2) the probability that an i.i.d. source with this distribution imparts to any δ -typical sequence. For a closely related proof, see Yeung's Chapter 8.

Let $Q = \{Q(y|x), (x, y) \in \mathcal{X} \times \mathcal{Y}\}$ be the transition probability matrix of a DMC, and let $p = \{p(x), x \in \mathcal{X}\}$ be a distribution on the channel input alphabet \mathcal{X} .

Let $\mathcal{C} = \{\underline{X}_1, \dots, \underline{X}_M\}$ be a randomly chosen code of blocklength n and rate $R = n^{-1} \log M$. Choose all the letters of all the code words i.i.d. according to p . Let J be a random message index that is independent of \mathcal{C} and has distribution $\{P_j, 1 \leq j \leq M\}$. If $J = j$, then the components $X_{j,1}, \dots, X_{j,n}$ of \underline{X}_j will be put into the channel in this order during n successive channel uses. Let $\underline{Y} = (Y_1, \dots, Y_n)$ be the resulting channel output vector. Let \hat{J} denote the decoder's estimate of J based on \underline{Y} . Let \bar{P}_e denote the probability that $\hat{J} \neq J$; the value of \bar{P}_e depends on the joint distribution of $J, \mathcal{C}, \underline{Y}$ and the decision rule used by the decoder; we shall use a suboptimum decision rule known as *joint typicality decoding* defined in the next paragraph.

Choose any $\delta > 0$. Let $T_\delta = T_\delta(pQ)$ denote the set of all n -vector pairs $(\underline{x}, \underline{y})$ that are jointly δ -typical of pQ . If there is one, and only one, index j^* such that $(\underline{X}_{j^*}, \underline{Y}) \in T_\delta$, set $\hat{J} = j^*$. If there is no such index, or if there is more than one such index, set $\hat{J} = \phi$, which means "declare a decoding error." This decision rule is called joint-typicality decoding.

Let $T_\delta(p)$ denote the set of n -vectors $\underline{x} \in \mathcal{X}^n$ that are δ -typical with respect to p . Similarly, let $T_\delta(q)$ denote the set of n -vectors $\underline{y} \in \mathcal{Y}^n$ that are δ -typical with respect to q , where $q(y) = \sum_{x \in \mathcal{X}} p(x)Q(y|x)$.

We may write

$$\bar{P}_e = \sum_{j=1}^M P_j \bar{P}_e(j),$$

where $\bar{P}_e(j)$ is the probability that joint typicality decoding fails given that $J = j$; i.e., $\bar{P}_e(j) = P[\hat{J} \neq j | J = j]$. Since \mathcal{C} is chosen independently of J , we see from symmetry considerations that $\bar{P}_e(j)$ does not depend on j . It follows that $\bar{P}_e = \bar{P}_e(1)$; i.e., for purposes of computing or bounding \bar{P}_e , no loss of generality results from assuming that $J = 1$. Henceforth, therefore, we assume that $J = 1$; i.e., henceforth $P[\cdot] = P[\cdot | J = 1]$.

Next, observe that we have the upper bound

$$\bar{P}_e = P(E_1 \cup E_2) \leq P(E_1) + P(E_2),$$

where

$$E_1 = [(\underline{X}_1, \underline{Y}) \notin T_\delta | J = 1] \quad \text{and} \quad E_2 = [(\underline{X}_\ell, \underline{Y}) \in T_\delta \text{ for at least one } \ell > 1 | J = 1].$$

Because

$$P_{\underline{X}_1, \underline{Y}}(\underline{x}, \underline{y}) = \prod_{k=1}^n p(x_k) Q(y_k | x_k) ,$$

we know from the weak law of large numbers that $P[(\underline{X}_1, \underline{Y}) \in T_\delta] \rightarrow 1$ as $n \rightarrow \infty$, so $P(E_1) \rightarrow 0$ as $n \rightarrow \infty$. Since

$$E_2 = \bigcup_{\ell=2}^M [(\underline{X}_\ell, \underline{Y}) \in T_\delta] ,$$

we have the union bound

$$P(E_2) \leq \sum_{\ell=2}^M P[(\underline{X}_\ell, \underline{Y}) \in T_\delta].$$

Observe that $\{\underline{X}_\ell, 2 \leq \ell \leq M\}$ is independent of \underline{Y} ; this is because \underline{Y} is the response of the channel to \underline{X}_1 which is independent of $\{\underline{X}_\ell, \ell \geq 2\}$. It follows that all $M-1$ terms in the sum are equal, so

$$P(E_2) \leq (M-1)P[(\underline{X}_2, \underline{Y}) \in T_\delta] < MP[(\underline{X}_2, \underline{Y}) \in T_\delta].$$

Since M grows exponentially in n at rate R , we will be able to conclude that $P(E_2)$ vanishes in the limit as $n \rightarrow \infty$ if we can show that $P[(\underline{X}_2, \underline{Y}) \in T_\delta]$ decays at an exponential rate that's larger than R . Toward this end we note that \underline{X}_2 and \underline{Y} are statistically independent of one another, that \underline{X}_2 's components are i.i.d. according to $\{p(x)\}$, and that \underline{Y} 's components are i.i.d. according to $\{q(y)\}$ because they are the result of putting the components of \underline{X}_1 through the DMC. It follows that

$$P[(\underline{X}_2, \underline{Y}) \in T_\delta] = \sum_{(\underline{x}, \underline{y}) \in T_\delta} p(\underline{x}) q(\underline{y}),$$

where $p(\underline{x}) = \prod_{k=1}^n p(x_k)$ and $q(\underline{y}) = \prod_{k=1}^n q(y_k)$.

But $(\underline{x}, \underline{y}) \in T_\delta$ implies $\underline{x} \in T_\delta(p)$ and $\underline{y} \in T_\delta(q)$ because each of a pair of jointly δ -typical vectors must be δ -typical of its marginal. Moreover, we know that every $\underline{x} \in T_\delta(p)$ is such that

$$p(\underline{x}) \leq 2^{-nH(X)^-} ,$$

where

$$H(X) = H(\{p(x)\}) = - \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

is the entropy of p , and H^- denotes a quantity that is smaller than H but larger than $H - c\delta$ for some positive constant c that does not depend on n . Similarly, we know that every $\underline{y} \in T_\delta(q)$ is such that

$$p(\underline{y}) \leq 2^{-nH(Y)^-} .$$

It follows that

$$P[(\underline{X}_2, \underline{Y}) \in T_\delta] \leq 2^{-nH(X)^-} \cdot 2^{-nH(Y)^-} \cdot |T_\delta| .$$

We also know that T_δ consists of at most $2^{nH(X,Y)^+}$ $(\underline{x}, \underline{y})$ pairs in all, where $H(X,Y) = H(p,Q)$ is the entropy of the joint distribution pQ , and H^+ denotes a quantity that is larger than H but smaller than $H + c\delta$ for some positive constant c that does not depend on n . It follows that

$$P[(\underline{X}_2, \underline{Y}) \in T_\delta] \leq 2^{-n[H(X)^- + H(Y)^- - H(X,Y)^+]} = 2^{-n[H(X) + H(Y) - H(X,Y)]^-} = 2^{-nI(X;Y)^-} .$$

Accordingly,

$$P(E_2) \leq 2^{n[R - I(X;Y)^-]} .$$

Given any $R < I(X; Y) = I(p, Q)$, we can choose δ sufficiently small that $R < I(p, Q)^-$. Therefore, $P(E_2)$ decays to zero as $n \rightarrow \infty$ for any $R < I(p, Q)$. This conclusion remains valid when we maximize over p , so $P(E_2)$ decays to zero exponentially fast for any $R < \max_p I(p, Q) = C$. It follows that $\bar{P}_e \rightarrow 0$ for any $R < C$.

For each n there must exist at least one code \mathcal{C}_n^* in our ensemble of randomly selected codes of rate R whose average error probability $P_e(\mathcal{C}_n^*)$ is at least as small as the ensemble average error probability. Thus, the above argument assures the existence for any $R < C$ of a specific sequence of codes of rate R and increasing blocklength n along which the error probability decays to zero as $n \rightarrow \infty$.

It is, of course, possible for the average error probability of a code to be small while at the same time certain messages have a high probability of not being recovered correctly. Indeed, most of the codes in our random ensemble have some of their code words so close to others that the decoder cannot distinguish reliably among those words. For $R < C$ we know from the above argument that such confusion does not prevail for the overwhelming majority of the $M = 2^{nR}$ code words because the average error probability goes to zero as $n \rightarrow \infty$. However, in many applications we do not want unequal error protection of the possible messages. Even in applications where unequal error protection might be desirable, we may not want to slug through the details of the code structure to determine which messages are protected well and which are not. Accordingly, we would like a result which ensures that $\max_{1 \leq j \leq M} P_e(j) \rightarrow 0$. The following line of reasoning assures us that there are such codes for any rate $R < C$.

For any given $R < C$, let \mathcal{C}_n denote a sequence of rate- R codes of increasing block lengths n along which $\bar{P}_e \rightarrow 0$. The above random code selection argument assures the existence of such a code sequence regardless of the probabilities P_j of the messages. In particular, $\bar{P}_e \rightarrow 0$ along \mathcal{C}_n when the messages are equiprobable, i.e., when $P_j = 1/M, 1 \leq j \leq M$. For equiprobable messages, at most half of the $P_e(j)$ can be more than twice as big as \bar{P}_e . (Think about it!) It follows that, if we expurgate from \mathcal{C}_n the $M/2$ messages for which $P_e(j)$ is largest, the maximum $P_e(j)$ for any of the $M/2$ remaining messages will be at most $2\bar{P}_e$. (Actually, it usually will be less because expurgation permits "growing" the decode regions for each of the non-expurgated messages, which reduces their $P_e(j)$ -values.) Since $\bar{P}_e \rightarrow 0$, we know that $\max_{1 \leq j \leq M/2} P_e(j) \rightarrow 0$ for the expurgated codes. But the rate of a block length n code with $M/2$ messages is

$$n^{-1} \log(M/2) = n^{-1} \log(2^{nR-1}) = R - n^{-1},$$

so the expurgated codes' rates approach the unexpurgated code rate R as $n \rightarrow \infty$. This gives us the desired conclusion, namely: *For any $R < C$ and any $\epsilon > 0$, we can find a code of rate R and sufficiently large block length n for which $\max_j P_e(j) < \epsilon$.*