Unit 13: Channel Coding-2

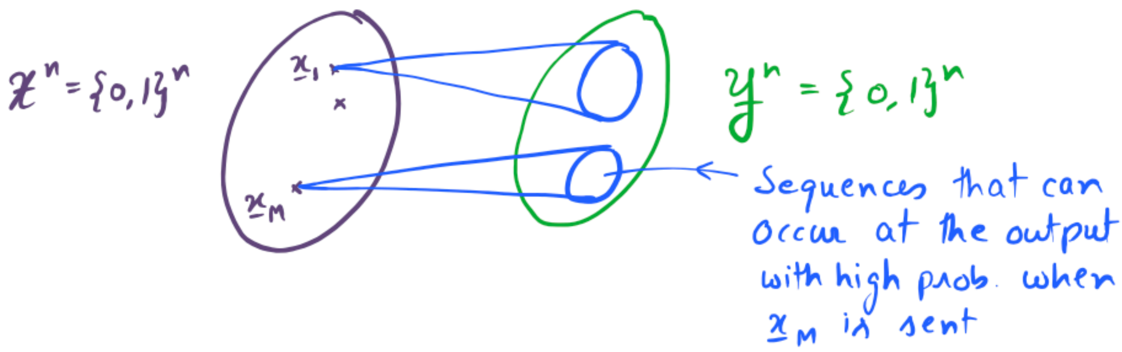
(Proof of Coding Theorem)

A. Sphere packing bound for BSC

We now present a heuristically simple bound for $C(W)$ where $W \equiv BSC(\delta)$. When a codeword $\underline{x}_m$ is sent over a BSC, with high prob., the received vector $\underline{y}$ differs from $\underline{x}_m$ at $\approx n\delta$ locations, i.e., $d_H(\underline{y}, \underline{x}_m) = \sum_{t=1}^{n} \mathbb{1}_{\{y_t \neq x_{m,t}\}} \equiv$ Hamming dist. b/w $\underline{x}_m$ and $\underline{y}$

We can treat these balls of radius $n\delta$ as roughly the decoding sets $D_m = d^{-1}(m)$. They are all disjoint. Further, these balls have cardinality more than $\binom{n}{n\delta} \gtrsim 2^{nh(\delta)}$. Thus,

$$2^n \geq \sum_{m=1}^{M} |B_{n\delta}(\underline{x}_m)| \gtrsim M\, 2^{nh(\delta)} \Rightarrow \frac{1}{n}\log M \leq 1 - h(\delta).$$



We now formalize this proof. The main ingredient is a lower bound for the cardinality of a large prob. set derived in Unit 8 (when we derived a strong converse for source coding theorem).

Denote by D_m, $1 \le m \le M$, the decoding sets $d^{-1}(m)$.

Since the avg. prob. of err. is less than ε, we get

$$\frac{1}{M} \sum_{m=1}^{M} W^n(D_m^c \mid e(m)) \le \varepsilon .$$

It is easy to see that we can find a subset $\mathcal{M}_0 \subseteq [M]$ with $|\mathcal{M}_0| \le M/2$ and such that

$$W^n(D_m^c \mid e(m)) \le 2\varepsilon, \quad \forall\, m \in \mathcal{M}_0 .$$

We will denote $|\mathcal{M}_0| = M'$ and assume that elements of $\mathcal{M}_0$ are $\{1, \dots, M'\}$. Thus, we assume a maximum prob. of err. bound, i.e.,

$$\max_{1 \le m \le M'} W^n(D_m^c \mid e(m)) \le 2\varepsilon .$$

Next, for $Z_1, \dots, Z_n \sim$ iid $\mathrm{Ber}(\delta)$, we have

$$2\varepsilon \le W^n(D_m \mid e(m)) = P_{Z^n}(D_m \oplus e(m)) .$$

Thus, $D_m \oplus e(m)$ is a "large prob." set under an iid distribution P_{Z^n}. Using the lower bound for $\frac{1}{n} L_\varepsilon(X^n)$ derived in Unit 8D, we get

$$\frac{1}{n} \log |D_m \oplus e(m)| \ge h(\delta) - \sqrt{\frac{V}{2n\varepsilon}} - \frac{1}{n} \log \frac{2}{1-2\varepsilon},$$

where

$$V = \mathrm{Var}[-\log P_Z(Z)] .$$

But $|D_m| = |D_m \oplus e(m)|$. Thus,

③

$$2^n \geq \left| \bigcup_{m=1}^{M'} D_m \right| = \sum_{m=1}^{M'} |D_m|$$

$$\geq M' \, 2^{n h(\delta) - \sqrt{nV/2\varepsilon} - \log 2/1-2\varepsilon}$$

Therefore, $\frac{1}{n} \log M' \leq 1 - h(\delta) + \sqrt{\frac{V}{2n\varepsilon}}$

$$+ \frac{1}{n} \log \frac{2}{1-2\varepsilon},$$

which gives

$$\frac{1}{n} \log \frac{M}{2} \leq 1 - h(\delta) + \sqrt{\frac{V}{2n\varepsilon}} + \frac{1}{n} \log \frac{2}{1-2\varepsilon}.$$

It follows that $\boxed{C(W) \leq 1 - h(\delta)}$

B Schemes for Binary Symmetric Channel

Attempt 1: (Gilbert-Varshamov bound)

We simply try to invert the sphere-packing bound above. The first observation is that when $\underline{x} \in \{0,1\}^n$ is sent over BSC(δ) then we observe $\underline{y} \in B_\rho(\underline{x})$ with large prob. where

$$B_r(\underline{x}) = \{\underline{y} : d_H(\underline{x}, \underline{y}) \leq \rho\}$$

$$d_H(\underline{x}, \underline{y}) = \sum_{t=1}^{n} \mathbb{1}_{\{x_t \neq y_t\}}$$

for $\rho \approx n\delta$. Thus, we can simply find balls $B_\rho(\underline{x}_1), \ldots, B_\rho(\underline{x}_M)$ s.t. $B_\rho(\underline{x}_i) \cap B_\rho(\underline{x}_j) = \phi$, $i \neq j$.

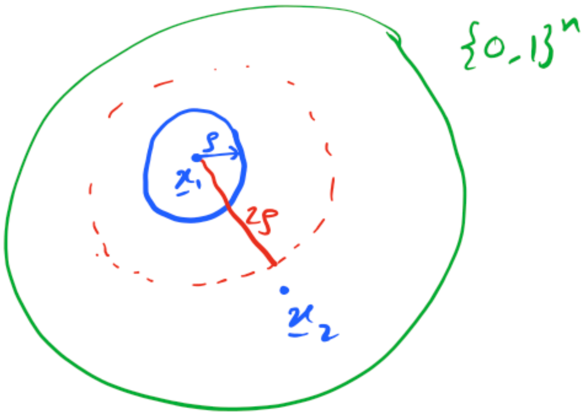
$(e(m) = \underline{x}_m,\ d(\underline{y}) = m$ if $d_H(\underline{x}_m, \underline{y}) \leq \rho)$

Recall that $|B_r(\underline{x})| = |B_r(\underline{0})| = \sum_{t=1}^{r} \binom{n}{t} \leq r\, 2^{n h(\frac{r}{n})}$ if $r \leq \frac{n}{2}$.

Therefore, for $\rho \approx n\delta$, $|B_\rho(\underline{x})| \lesssim n \cdot 2^{n h(\delta)}$. Thus, our goal is to simply disjoint balls of radius ρ. We can do this using the following greedy procedure:

<u>Initialize:</u> $A = \mathcal{X}^n$, $i = 1$

1) Choose $\underline{x}_i \in A$.
2) Update $A \leftarrow A \setminus B_{2\rho}(\underline{x}_i)$
3) Update $i \leftarrow i+1$.
4) Go to 1 if $A \neq \phi$.



Note that we decode $\underline{y}$ to m if $d_H(\underline{y}, \underline{x}_m) \leq \rho$ but we remove all points $\underline{x}' \in B_{2\rho}(\underline{x}_m)$ once an $\underline{x}_m$ is selected. We do this so that $B_\rho(\underline{x}_i) \cap B_\rho(\underline{x}_j) = \phi$ holds.

How many such balls can we find?

Noting that $|B_{2\rho}(\underline{x})| \approx 2^{n h(2\delta)}$, we can find a code of size $\approx 2^{n(1 - h(2\delta))}$. Thus, this code has a rate $\approx \underline{1 - h(2\delta)}$, which falls short of our upper bound of $\underline{1 - h(\delta)}$.

<u>Attempt 2</u> (Random Code Construction)

We now present a very different method for constructing the channel code similar to the one above. We note that our construction fails since we insisted on the balls to be disjoint. Instead, we can choose the centers carefully and try to ensure that the balls do not have a significant intersection. But how do we choose these centers? Randomly!

Specifically, we generate a codebook $\mathcal{C} = \{\underline{x}_1, \ldots, \underline{x}_M\}$ randomly and show that the expected value of its avg. error is small if M is not too large. Thus, there exists one choice of $\mathcal{C} = \{\underline{x}_1, \ldots, \underline{x}_M\}$ s.t. prob. of error for it is small.

Such a random codebook is called a code ensemble.

Consider $\underline{X}_1, \ldots, \underline{X}_M$ generated independently with each $\underline{X}_i = (X_{i1}, \ldots, X_{in}) \sim$ iid $\mathrm{Ber}(1/2)$. Consider the code (e, d) given by $e(m) = \underline{X}_m$ and

$$d(y^n) = \begin{cases} m, & \text{if there is a unique } m \text{ s.t. } d_H(\underline{X}_m, \underline{y}) \le \rho \\ \perp, & \text{o.w.} \end{cases}$$

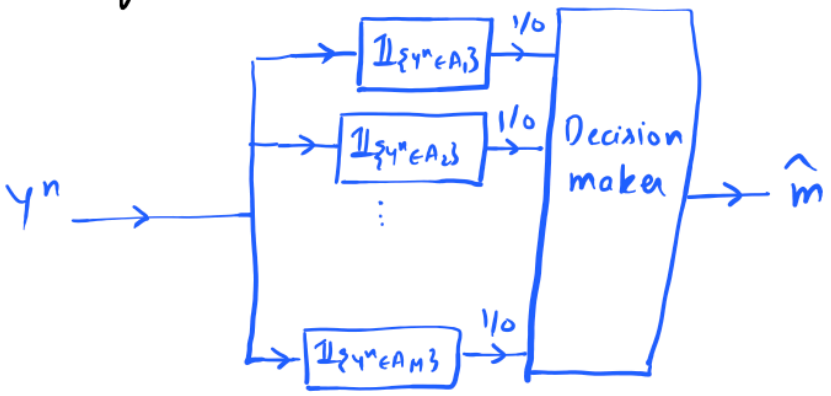
Note that we have allowed the decoder to declare an error $\perp$. But we define the prob. of error exactly as before:
$$P_e(\underline{X}_1, \ldots, \underline{X}_M) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{P}\left(d(Y^n) \neq m \,\middle|\, m \text{ sent}\right).$$

Denoting $A_m = \{\underline{y} : d_H(\underline{X}_m, \underline{y}) \le \rho\}$, we have

$$\mathbb{P}(d(Y^n) \neq m \mid m \text{ sent}) \le \mathbb{P}(Y^n \notin B_\rho(\underline{X}_m) \mid m \text{ sent}) + \mathbb{P}(\exists\, m' \neq m \text{ s.t. } Y^n \in B_\rho(\underline{X}_{m'}) \mid m \text{ sent})$$

$$\le W^n(A_m^c \mid \underline{X}_m) + \sum_{m' \neq m} W^n(A_{m'} \mid \underline{X}_m)$$

The structure of our decoder is depicted in the figure below:



We can use this decoder with any subsets A_m for the intermediate decision making. From the bound above, we get

$$\mathbb{E}\left[P_e(\underline{X}_1, \dots, \underline{X}_M)\right] \le \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} W^n(A_m^c \mid \underline{X}_m)\right] + \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M}\sum_{m' \neq m} W^n(A_{m'} \mid \underline{X}_m)\right]$$

$$= \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}\left[W^n(A_m^c \mid \underline{X}_m)\right] + \frac{1}{M}\sum_{m=1}^{M}\sum_{m' \neq m} \mathbb{E}\left[W^n(A_{m'} \mid \underline{X}_m)\right]$$

Note that $\mathbb{E}\left[W^n(A_m^c \mid \underline{X}_m)\right] = \sum_{\underline{x}} P_{X^n}(\underline{x}) \sum_{\underline{y} : d_H(\underline{y}, \underline{x}) > \rho} W^n(\underline{y} \mid \underline{x})$

$= P_{X^nY^n}(d_H(X^n, Y^n) > \rho)$;

it is important to note that this term does not depend on m. For the second term,

$$\mathbb{E}\left[W^n(A_{m'} \mid \underline{X}_m)\right] = \sum_{\underline{x}_m, \underline{x}_{m'}} P_{X^n}(\underline{x}_m) P_{X^n}(\underline{x}_{m'}) \sum_{\underline{y} : d_H(\underline{y}, \underline{x}_{m'}) \le \rho} W^n(\underline{y} \mid \underline{x}_m)$$

$$= \sum_{\underline{x}_{m'}} P_{X^n}(\underline{x}_{m'}) \sum_{\underline{y} : d_H(\underline{y}, \underline{x}_{m'}) \le \rho} \underbrace{\sum_{\underline{x}_m} P_{X^n}(\underline{x}_m) W^n(\underline{y} \mid \underline{x}_m)}_{P_{Y^n}(\underline{y})}$$

$$= \sum_{(\underline{x}, \underline{y}) : d_H(\underline{x}, \underline{y}) \le \rho} P_{X^n}(\underline{x}) P_{Y^n}(\underline{y}) = P_{X^n} P_{Y^n}\left(d_H(X^n, Y^n) \le \rho\right).$$

Thus, we have shown,

$$\mathbb{E}\left[P_e(\underline{X}_1, \dots, \underline{X}_M)\right] \le P_{X^n Y^n}\left(d_H(X^n, Y^n) > \rho\right) + (M-1) P_{X^n} P_{Y^n}\left(d_H(X^n, Y^n) \le \rho\right).$$

An important remark

The decoder above and its analysis is very generic. Suppose we generate $\underline{X}_m = (X_{m1}, \dots, X_{mn}) \sim$ iid P_X and $\underline{X}_1, \dots, \underline{X}_M \sim$ iid. For $\mathcal{A} \subseteq \mathcal{X}^n \times \mathcal{Y}^n$, let $A_m = \{\underline{y} : (\underline{X}_m, \underline{y}) \in \mathcal{A}\}$. The analysis above extends to this general setting to get

$$\mathbb{E}\left[P_e(\underline{X}_1, \dots, \underline{X}_M)\right] \le P_{X^n Y^n}(\mathcal{A}^c) + (M-1) P_{X^n} P_{Y^n}(\mathcal{A}).$$

We can view this A as an acceptance region for the null hypothesis in the binary hypothesis testing problem of $\mathcal{H}_0 \equiv P_{X^n Y^n}$ vs $\mathcal{H}_1 \equiv P_{X^n} P_{Y^n}$. Thus, we can find A s.t.

$$P_{X^n Y^n}(A^c) \le \frac{\varepsilon}{2} \text{ and } P_{X^n} P_{Y^n}(A) \le \beta_{\frac{\varepsilon}{2}}(P_{X^n Y^n}, P_{X^n} P_{Y^n}).$$

(Recall $\beta_\varepsilon(P, Q) = \min\{Q(A) : P(A^c) \le \varepsilon,\ A \subseteq \mathcal{X}\}$)

We will revisit this general approach later.

→ For now, we come back to our specific choice of A for BSC, namely $A = \{(\underline{x}, \underline{y}) : d_H(\underline{x}, \underline{y}) \le \rho\}$.

We choose $\rho = \rho_n = n\delta + \sqrt{\frac{2n\delta(1-\delta)}{\varepsilon}}$ to get

$$P_{X^n Y^n}\left(d_H(X^n, Y^n) > \rho_n\right) = \mathbb{P}\left(\sum_{t=1}^{n} Z_t > n\delta + \sqrt{\frac{2n\delta(1-\delta)}{\varepsilon}}\right)$$

(where $Z_1, \dots, Z_n \sim$ iid with $Z_t = \mathbb{1}_{\{X_t \ne Y_t\}} \sim \mathrm{Ber}(\delta)$)

$$\le \frac{\varepsilon}{2}.$$

Under $P_{X^n} P_{Y^n}$, $\mathbb{1}_{\{X_t \ne Y_t\}} \sim \mathrm{Ber}(1/2)$, whereby

$$P_{X^n} P_{Y^n}\left(d_H(X^n, Y^n) \le \rho_n\right) = \mathbb{P}\left(\sum_{t=1}^{n} \tilde{Z}_t \le n\delta + \sqrt{\frac{2n\delta(1-\delta)}{\varepsilon}}\right)$$

where $\tilde{Z}_t \sim \mathrm{Ber}(1/2)$.

The right-side equals $\displaystyle\sum_{i=1}^{\rho} \binom{n}{i} \frac{1}{2^n} \le \rho \binom{n}{\rho} 2^{-n} \le \frac{n\, 2^{n h(\frac{\rho}{n})}}{2^n}$,

if we assume $\rho \le \frac{n}{2}$.

If $\delta < \frac{1}{2}$ and n is suff. large, we get (9)

$$P_{X^n} P_{Y^n}\left(d_H(X^n, Y^n) \leq \rho_n \right) \leq n\, 2^{-n\left(1-h\left(\delta+\sqrt{\frac{2\delta}{\varepsilon n}}\right)\right)}.$$

Overall,

$$\mathbb{E}\left[P_e(\underline{X}_1, \ldots, \underline{X}_M)\right] \leq \frac{\varepsilon}{2} + M n\, 2^{-n(1-h(\delta+\sqrt{2\delta/\varepsilon n}))}$$
$$\leq \varepsilon$$

$$\text{if } \frac{1}{n}\log M \leq 1 - h\left(\delta + \sqrt{\frac{2\delta}{\varepsilon n}}\right) - \frac{1}{n}\log n - \frac{1}{n}\log\frac{2}{\varepsilon}.$$

In particular, there exists a specific choice $\mathcal{C} = \{\underline{x}_1, \ldots, \underline{x}_M\}$ for which $P_e(\underline{x}_1, \ldots, \underline{x}_M) \leq \varepsilon$ and $\lim_{n\to\infty} \frac{1}{n}\log M = 1 - h(\delta)$. Thus, taking limit $n \to \infty$, the rate $1 - h(\delta)$ is achievable for $BSC(\delta)$.

$\boxed{C}$ Converse for Shannon's channel coding theorem

For $U \sim \text{unif}(\{1, \ldots, 2^{nR}\})$ and a code (e,d) of rate R, let $X^n = e(U)$. Further, the corresponding output be Y^n and let $\hat{U} = d(Y^n)$. If the prob. of error of (e,d) is less than ε, $\mathbb{P}(U \neq \hat{U}) \leq \varepsilon$. Then,

$$nR = H(U) = I(U \wedge \hat{U}) + \underbrace{H(U|\hat{U})}_{\leq\, \varepsilon n R + 1 \text{ (by Fano's ineq.)}}$$

$$\Rightarrow \quad R \leq \frac{1}{(1-\varepsilon)}\left[\frac{1}{n}\cdot I(U \wedge \hat{U}) + \frac{1}{n}\right]$$

The first term $I(U \wedge \hat{U})$ is bounded as follows:

$$I(U \wedge \hat{U}) \le I(X^n \wedge Y^n)$$ (by the data processing inequality)

$$= \sum_{t=1}^{n} I(X^n \wedge Y_t \mid Y^{t-1})$$

$$\le \sum_{t=1}^{n} I(X^n, Y^{t-1} \wedge Y_t)$$

$$= \sum_{t=1}^{n} I(X_t \wedge Y_t) + \underbrace{I(X^{t-1}, X_{t+1}^{n}, Y^{t-1} \wedge Y_t \mid X_t)}_{=0}$$

$= 0$ (by the memoryless property)

$X^{t-1}, X_{t+1}^{n}, Y^{t-1} - X_t - Y_t$

Thus, $$R \le \frac{1}{1-\varepsilon}\left[\frac{1}{n}\sum_{t=1}^{n} I(X_t \wedge Y_t) + \frac{1}{n}\right]$$

$$= \frac{1}{1-\varepsilon}\left[\frac{1}{n}\sum_{t=1}^{n} I(P_{X_t}; W) + \frac{1}{n}\right]$$

$$\le \frac{1}{1-\varepsilon}\left[I\Big(\underbrace{\frac{1}{n}\sum_{t=1}^{n} P_{X_t}}; W\Big) + \frac{1}{n}\right]$$

by concavity of $I(P;W)$ in P

$$= \frac{1}{1-\varepsilon}\left[I(\bar{P}_X; W) + \frac{1}{n}\right]$$

$$\le \frac{1}{1-\varepsilon}\left[\max_{P_X} I(P_X; W) + \frac{1}{n}\right].$$

By taking limit $n \to \infty$ and $\varepsilon \to 0$,

$$R \le \max_{P_X} I(P_X; W) \Rightarrow \boxed{C(W) \le \max_{P_X} I(P_X; W)}$$

D Achievability for Shannon's channel coding theorem

We already presented a sketch for this proof in Section B, where we showed that expected error is less than

$$\frac{\varepsilon}{2} + M\,\beta_\varepsilon\left(P_{X^nY^n}, P_{X^n}P_{Y^n}\right)$$

$$\leq \frac{\varepsilon}{2} + M\,2^{-n\underbrace{D(P_{XY}\|P_XP_Y)}_{I(X\wedge Y)}} \quad \text{(by Stein's lemma)}$$

$$\leq \varepsilon \quad \text{if} \quad \frac{1}{n}\log M \leq I(X\wedge Y) - \frac{1}{n}\log\frac{2}{\varepsilon}.$$

This completes the proof of achievability. But we will give a more explicit construction. In fact, we note that the overall proof is "single-shot" — it doesn't use the memoryless property till the end.

Consider a channel $W: \mathcal{X} \to \mathcal{Y}$ and a code (e,d) of size M where $e: \{1,\dots,M\} \to \mathcal{X}$ and $d: \mathcal{Y} \to \{1,\dots,M\}$. The prob. of error of this code is given by $\frac{1}{M}\sum_{m=1}^{M} W(D_m^c \mid e(m))$ where $D_m = d^{-1}(m)$.

<u>Theorem</u> Given a channel $W: \mathcal{X} \to \mathcal{Y}$ and a distribution P_X on $\mathcal{X}$, let $P_{Y|X} = W$. Suppose that

$$P_{XY}\left(\left\{(x,y) : \log\frac{P_{XY}(x,y)}{P_X(x)\,P_Y(y)} > \lambda\right\}\right) \geq 1-\varepsilon.$$

Then, there exists a code of size $\lfloor \varepsilon 2^\lambda \rfloor$ with prob. of error less than 2ε.

Proof. Consider $X_1, \ldots, X_M \sim$ iid P_X and the encoder $e: \{1, \ldots, M\} \to \mathcal{X}$ given by $e(m) = X_m$. Let

$$\mathcal{A} = \left\{(x,y): \log \frac{P_{XY}(x,y)}{P_X(x) P_Y(y)} > \lambda\right\}, \text{ and define}$$

$$A_m = \{y : (X_m, y) \in \mathcal{A}\}, \quad 1 \le m \le M.$$

Further, define the decoder as follows

$$d(y) = \begin{cases} \hat{m}, & \text{if } y \in A_m \text{ and } y \notin A_{m'} \ \forall\, m' \neq m, \\ \perp, & \text{o.w.} \end{cases}$$

Our analysis earlier applies and we get

$$\begin{aligned} \mathbb{E}\left[P_e(X_1, \ldots, X_M)\right] &\le P_{XY}(\mathcal{A}^c) + M\, P_X P_Y(\mathcal{A}) \\ &\le \varepsilon + M\, P_X P_Y(\mathcal{A}), \end{aligned}$$

where the second inequality is by the assumption for $\mathcal{A}$.

Also,

$$\begin{aligned} P_X P_Y(\mathcal{A}) &= \sum_{(x,y)\in\mathcal{A}} \frac{P_X P_Y(x,y)}{P_{XY}(x,y)} \cdot P_{XY}(x,y) \\ &\le 2^{-\lambda} \sum_{(x,y)\in\mathcal{A}} P_{XY}(x,y) \\ &= 2^{-\lambda} P_{XY}(\mathcal{A}) \le 2^{-\lambda}. \end{aligned}$$

Thus, for $M \le \varepsilon 2^{\lambda}$, there exists $(x_1, \ldots, x_M)$ such that

$$P_e(x_1, \ldots, x_M) \le 2\varepsilon. \qquad \blacksquare$$

By Chebyshev's inequality, for iid P_{X^n} and memoryless W^n, a choice of λ is $n I(X \wedge Y) - \sqrt{nV/\varepsilon}$, which shows $C(W) \ge I(P_X; W)$.