Rate-dist. Theory

Motivating Example

$X_1, X_2, \ldots \sim \text{Bernoulli}(1/2)$ iid. (e.g. output of very good lossless compression alg.) Want to compress further, but $H(X)=1$. $\Rightarrow$ Further Lossless comp. is not possible.

Willing to accept up to $\delta = 1/4$ of bits incorrectly reproduced on average (upon decompression).

How good can we do?

Let's encode in blocks: n symbols at a time, into codewords of k symbols. (So, compression rate $= k/n$.)

Case $n=1$, $k=0$:

x	$p(x)$	$C(x)$	$\hat{x}$	$d(x,\hat{x})$
0	1/2	—	0	0
1	1/2	—	0	1

$\leftarrow$ Fraction wrong, in this case

$\Rightarrow D = Ed(X,\hat{X}) = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 = \frac{1}{2} > \frac{1}{4} = \delta$

$R = \frac{k}{n} = 0.$

No good

Case $n=1$, $k=1$:

x	$p(x)$	$C(x)$	$\hat{x}$	$d(x,\hat{x})$
0	1/2	0	0	0
1	1/2	1	1	0

$\Rightarrow D = \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 = 0 \leq 1/4$. OK.

$R = \frac{k}{n} = 1$. (So for $n=1$, $R=1$ is best possible s.t. $D \leq \delta$.)

Let's try $n=2$ at a time...

$\underline{n=2,\ k=0}$

$x_{1:2}$	$p(x_{1:2})$	$C(x_{1:2})$	$\hat{x}_{1:2}$	$d(x_{1:2}, \hat{x}_{1:2})$
00	1/4	–	00	0
01	1/4	–	00	1/2
10	1/4	–	00	1/2
11	1/4	–	00	1

$\Rightarrow D = \frac{1}{4}(0 + \frac{1}{2} + \frac{1}{2} + 1) = \frac{1}{2} > \delta$

No good.

$\underline{n=2,\ k=1}$

$x_{1:2}$	.	.	.	.
00	1/4	0	00	0
01	1/4	0	00	1/2
10	1/4	1	11	1/2
11	1/4	1	11	0

$\Rightarrow D = \frac{1}{4}\cdot\frac{1}{2} + \frac{1}{4}\cdot\frac{1}{2} = \frac{1}{4} \le \delta$

$R = \frac{k}{n} = \frac{1}{2}$.

OK!

So, we are down to $R = 1/2$ using $n=2$ (instead of $n=1$.)

We could keep going, $n = 3, 4, \ldots$ reducing R further, while keeping $D \le \delta$.
(Gets tricky to choose good codes.)

- Turns out, ~~if R_n = best possible R for n is~~ letting R_n denote the rate of best possible code for n (subject to $D \le \delta$) then $R_n \to 0.1887 = 1 - H(\delta)$

(Theoretical result due to Shannon's Rate-Dist Thm.)
(Difficult to choose codes in practice!)

i.e. Can store n bits into $k < n/5$ bits and still get 75% of bits correct upon decompressing!

Note: If we took the naive approach of just discarding all but the first ~~0.1887 n bits~~ $1-\delta$ bits we would have $R = \frac{k}{n} = \frac{(1-\delta)n}{n} = 1-\delta. = 3/4$. Nowhere close to 0.19!

$n=3,\ k=1$

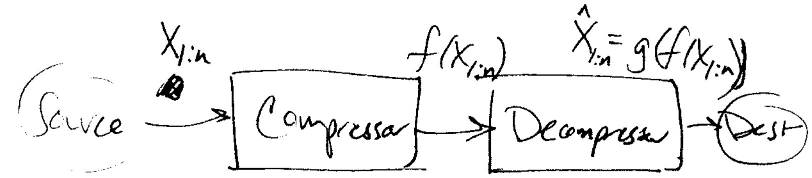
$x_{1:3}$	$p(x_{1:3})$	$C(x_{1:3})$	$\hat{x}_{1:3}$	$d(x_{1:3}, \hat{x}_{1:3})$
000	1/8	0	000	0
001	1/8	0	000	1/3
010	1/8	0	000	1/3
011	.	1	111	1/3
100	.	0	000	1/3
101	.	1	111	1/3
110	.	1	111	1/3
111	.	1	111	0

$\Rightarrow D = \frac{1}{3} \cdot \frac{3}{4} + 0 = \frac{1}{4} \le \delta.$

OK.

$R = \frac{1}{3}.$ Getting closer!

Rate-Distortion Theory (i.e. Lossy Compression)



- Want to compress an image, or video, or song/audio.

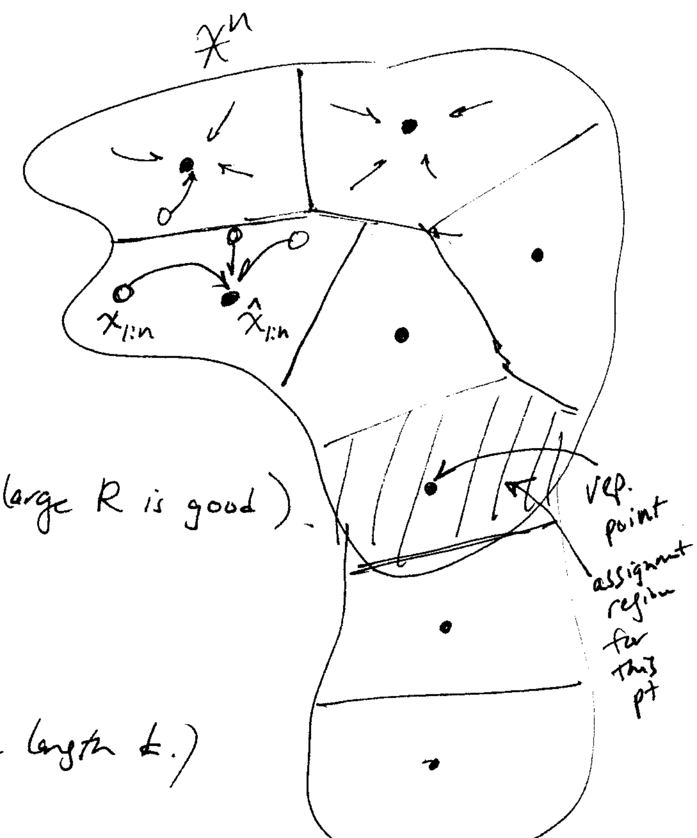
  ~~Have a measure of "distortion" b/w repr~~
- Willing to accept some loss of fidelity
- For a given acceptable level of distortion, what is the best compression that can be achieved?

~~Discrete~~ Memoryless Source: Let $X_1, \ldots, X_n \sim p$ iid (simplification), where p is a pmf on a finite set $\mathcal{X}$ (can generalize, but we'll consider only finite case).

Defn: A (n,k) lossy compression code C (aka rate distortion code) for $\mathcal{X}$ consists of

(a) encoder $f: \mathcal{X}^n \to \mathbb{B}^k$

and (b) decoder $g: \mathbb{B}^k \to \hat{\mathcal{X}}^n$.



8 chosen "reproduction points" $\hat{x}_{1:n}$. $\Rightarrow 8 = 2^k$, $k = 3$.

Rks: (1) Typically, $\hat{\mathcal{X}} = \mathcal{X}$.

(2) Compression rate $= R = k/n$.

(3) Small R is good (as opposed to Error-correction, where large R is good).

(Small R means smaller compressed size)

(4) Connection with lossless compression:

expected codeword length $= L = k$ (since all codewords have length k.)

$$\text{expected encoded bits/symbols per source symbol} = L_n = \frac{L}{n} = \frac{k}{n} = R.$$

~~(5) codebook = {~~

(loss fn)

Defn: A ~~distortion fn~~ distortion fn is a map $d: \mathcal{X} \times \hat{\mathcal{X}} \to [0,\infty]$.

Rks: ① Interpretation: $d(x,\hat{x})$ is larger when $\hat{x}$ is worse (as a reproduction of x.) $d(x,\hat{x})$ is the penalty resulting from representing x as $\hat{x}$.

② $d(x,\hat{x}) = \infty$ is permitted.

③ d is bounded if $\max_{x,\hat{x}} d(x,\hat{x}) < \infty$.

Example: "0-1 distortion" ~~$\mathcal{X} = \hat{\mathcal{X}} = \{0,1\}$, $d(x,\hat{x}) = I$~~

$\mathcal{X} = \hat{\mathcal{X}}$ and $d(x,\hat{x}) = I(x \neq \hat{x}) = \begin{cases} 1 & \text{if "incorrect"} \\ 0 & \text{if "correct"} \end{cases}$

Defn: The distortion between sequences $x_{1:n} \in \mathcal{X}^n$ and $\hat{x}_{1:n} \in \hat{\mathcal{X}}^n$ induced by $d(x,\hat{x})$

is $d(x_{1:n}, \hat{x}_{1:n}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, \hat{x}_i)$.

Defn: The expected distortion of an (n,k) lossy code $C=(f,g)$ wrt the source $X_1,\ldots,X_n \sim p$ (iid)

is $D_C = E\, d(X_{1:n}, \hat{X}_{1:n})$, where $\hat{X}_{1:n} = g(f(X_{1:n}))$.

(i.e. $D_C = \sum_{x_{1:n}} d(x_{1:n}, g(f(x_{1:n})))\, p(x_{1:n})$.)

Defn: The expected distortion of a conditional dist $q(\hat{x}|x)$ on $\hat{X}$ given X (where $X \sim p$)

is $D_q = E\, d(X, \hat{X}) = \sum_{x \in \mathcal{X}, \hat{x} \in \hat{\mathcal{X}}} d(x,\hat{x})\, p(x)\, q(\hat{x}|x)$. (Note: $x \in \mathcal{X}$, $\hat{x} \in \hat{\mathcal{X}}$ are single symbols.)

Given a source dist $p(x)$ on $\mathcal{X}$ and a dist. fn $d: \mathcal{X} \times \hat{\mathcal{X}} \to [0,\infty]$, and

Defn: Given $r \geq 0$, and $\delta \geq 0$, we say (r,δ) is achievable (as a "rate-distortion pair") if $\forall \varepsilon > 0$ there is a lossy code with $R \leq r$ and $D \leq \delta + \varepsilon$.

The rate-distortion problem (i.e. lossy comp. problem) is:

(a) (Theoretically) Given an acceptable level of distortion $\delta \geq 0$, what is the best (smallest) ~~r achievable~~ r that is achievable? (Answered by: Rate-Dist. Thm)

(b) (Practically) How to do this in practice? (Solved for very few problems - very difficult)

Defn: The rate distortion function is $\rho(\delta) = \inf \{ r \geq 0 : (r,\delta) \text{ is achievable.} \}$

Thm (Shannon's Rate-Dist. Thm): For an iid source and a bounded dist. fn,

$$\boxed{\rho(\delta) = \min_{q : D_q \leq \delta} I(X : \hat{X})},$$

(where the min is over cond. dists $q(\hat{x}|x)$, and ~~$(\hat{X}|X) \sim q(\hat{x}|x)$~~
$(X,\hat{X}) \sim p(x)\, q(\hat{x}|x)$.)

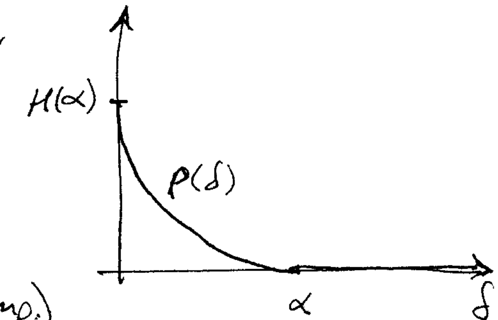
Note: $X \in \mathcal{X}$, $\hat{X} \in \hat{\mathcal{X}}$ are single symbols (not in $\mathcal{X}^n$.)

Rate-dist fn for a Bernoulli source

(Before proving the result, let's use it to compute $\rho(\delta)$ in an important special case.)

Thm: If $X_1, X_2, \ldots \overset{iid}{\sim} \text{Bernoulli}(\alpha)$ with $\alpha \in (0, 1/2]$ and $d(x,\hat{x}) = \mathbb{1}(x \neq \hat{x})$,

then
$$\boxed{\rho(\delta) = \begin{cases} H(\alpha) - H(\delta) & \text{if } 0 \leq \delta \leq \alpha \\ 0 & \text{if } \delta > \alpha. \end{cases}}$$



Rk: Intuition: If $\delta = 0$, this is the Source Coding Thm (zero dist = lossless comp.)
If $\delta \geq \alpha$, can get rate 0 since typical x seqs have $\approx \alpha n \leq \delta n$ ones. (So, take $\hat{x} \equiv 0$.)
If $0 < \delta < \alpha$, expect to interpolate b/w these.

Pf: (I) First, suppose $\delta > \alpha$. Choose $q(\hat{x}|x) = \mathbb{1}(\hat{x} = 0)$. Then $I(X:\hat{X}) = 0$ and

$$D_q = E d(X,\hat{X}) = E\mathbb{1}(X \neq \hat{X}) = \mathbb{P}(X \neq \hat{X}) = \mathbb{P}(X \neq 0) = \mathbb{P}(X=1) = \alpha < \delta. \Rightarrow D_q \leq \delta,\ I(X:\hat{X}) = 0.$$

$$\Rightarrow \underline{\rho(\delta) = 0} \text{ if } \delta > \alpha.$$

(II) Suppose $0 \leq \delta \leq \alpha$.

(a) Claim: $\rho(\delta) \geq H(\alpha) - H(\delta)$.

Pf of claim: Let $q(\hat{x}|x)$ be any cond dist s.t. $D_q \leq \delta$.

Let $Z = \mathbb{1}(X \neq \hat{X})$. Then $\mathbb{P}(Z=1) = EZ = D_q$ (since $Z = d(X,\hat{X})$). Hence, $H(Z) = H(D_q) \leq H(\delta)$. (since $D_q \leq \delta \leq \alpha \leq 1/2$)

$$I(X:\hat{X}) = \underbrace{H(X)}_{= H(\alpha)} - \underbrace{H(X|\hat{X})}_{= H(Z|\hat{X}) \leq H(Z) \leq H(\delta).} \geq H(\alpha) - H(\delta).$$

(since given $\hat{X} = \hat{x}$, Z is a 1-1 fn of X)

So, $\forall q$ s.t. $D_q \leq \delta$, $I(X:\hat{X}) \geq H(\alpha) - H(\delta)$.

$$\Rightarrow \rho(\delta) = \min_{q:\, D_q \leq \delta} I(X:\hat{X}) \geq H(\alpha) - H(\delta)$$

$\square$ claim (a)

(b) Claim: $\exists q$ s.t. $D_q \le \delta$ and $I(X:\hat{X}) = H(\alpha) - H(\delta)$.

Want: $H(X) - H(X|\hat{X}) \overset{?}{=} H(\alpha) - H(\delta) \Leftarrow H(\delta) \overset{?}{=} H(X|\hat{X}) = \sum_{\hat{x}} H(X|\hat{X}=\hat{x})\, \mathbb{P}(\hat{X}=\hat{x})$

$\Uparrow$

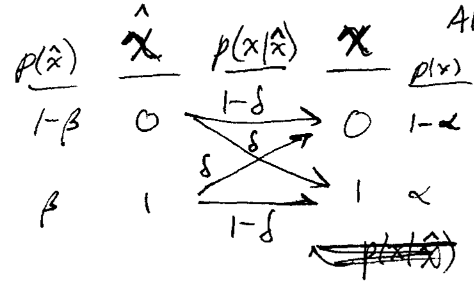
$\alpha = (1-\beta)\delta + \beta(1-\delta)$ (***) $\Leftarrow$ $\mathbb{P}(X=x|\hat{X}=x) = 1-\delta$ (*) $\Longrightarrow$ $H(X|\hat{X}=\hat{x}) \overset{?}{=} H(\delta) \ \forall \hat{x}$

where $\beta = \mathbb{P}(\hat{X}=1)$.

~~(since~~ ~~(since pt~~

since

$$\left(\mathbb{P}(X=1) = \mathbb{P}(X=1|\hat{X}=0)\mathbb{P}(\hat{X}=0) + \mathbb{P}(X=1|\hat{X}=1)\mathbb{P}(\hat{X}=1) \right)$$



Also, if q satisfies this, then

$$D_q = \mathbb{E}d(X,\hat{X}) = \mathbb{P}(X \ne \hat{X}) = \sum_x \underbrace{\mathbb{P}(X \ne x | \hat{X}=x)}_{\delta} \mathbb{P}(\hat{X}=x) = \delta \quad \Rightarrow \quad D_q \le \delta$$

$\Updownarrow$

$$\delta + \beta(1-2\delta)$$

$$\alpha = \delta - \beta\delta + \beta(1-\delta) \iff \beta = \frac{\alpha-\delta}{1-2\delta} \ (**) .$$

Note: $\beta \ge 0$ since $0 \le \delta \le \alpha \le 1/2$, and $\beta \le 1$ since $\alpha - \cancel{\delta} \le 1 - 2\delta = 1 - \delta - \cancel{\delta}$.

This motivates the choice: $q(1|1) = \mathbb{P}(\hat{X}=1|X=1) = \dfrac{\mathbb{P}(X=1|\hat{X}=1)\,\mathbb{P}(\hat{X}=1)}{\mathbb{P}(X=1)} = \dfrac{(1-\delta)\beta}{\alpha} = \dfrac{(1-\delta)(\alpha-\delta)}{\alpha(1-2\delta)}$.

~~and $q(0|1) = 1 - q(1|1)$. (Note: $q(1|1) \ge 0$ since $\delta < \alpha$ and $0 < \delta \le \alpha \le 1/2$.)~~

Note: $q(1|1) \ge 0$ since $\beta \ge 0$.

Similarly, $q(0|1) = \mathbb{P}(\hat{X}=0|X=1) = \dfrac{\delta(1-\beta)}{\alpha}$.

Note: $q(0|1) \ge 0$ since $\beta \le 1$.

Also, $q(0|1) + q(1|1) = 1$ since $\dfrac{\delta(1-\beta)}{\alpha} + \dfrac{(1-\delta)\beta}{\alpha} = \dfrac{1}{\alpha}\underbrace{(\delta - \delta\beta + \beta - \delta\beta)}_{\alpha \text{ by } (**)} = 1$.

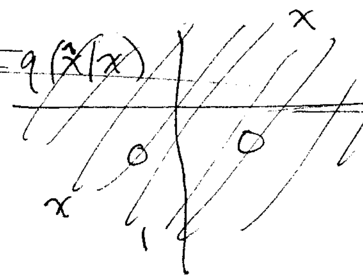
So, this choice of $q(\hat{x}|1)$ is a valid prob. dist.

$$q(0|0) = \mathbb{P}(\hat{X}=0 \mid X=0) = \frac{\mathbb{P}(X=0 \mid \hat{X}=0)\,\mathbb{P}(\hat{X}=0)}{\mathbb{P}(X=0)} = \frac{(1-\delta)(1-\beta)}{1-\alpha} \geq 0.$$

$$q(1|0) = \cdots 1 \cdots = \frac{\delta\beta}{1-\alpha} \geq 0.$$

$$q(0|0) + q(1|0) = \frac{1}{1-\alpha}\Big((1-\delta)(1-\beta) + \delta\beta\Big) = \frac{1}{1-\alpha}\Big(\underbrace{1-\delta-\beta+2\delta\beta}_{= 1-\alpha \text{ by } (***)}\Big) = 1.$$

~~So, choose $q(\hat{x}|x)$~~



~~So, choose~~ $\cancel{q(1|1) = \frac{(1-\delta)(\alpha-\delta)}{\alpha(1-2\delta)}}$, $\cancel{q(0|1) = 1-q(1|1)}$,

$\cancel{q(1|0) =}$

Now, we know what to do:

Define $q(\hat{x}|x)$ by: $q(1|1) = \dfrac{(1-\delta)\beta}{\alpha}$ and $q(1|0) = \dfrac{\delta\beta}{1-\alpha}$ where $\beta = \dfrac{\alpha-\delta}{1-2\delta}$

(and $q(0|1) = 1-q(1|1)$, $q(0|0) = 1-q(1|0)$.)

Then $\mathbb{P}(X=x \mid \hat{X}=x) = 1-\delta$ holds, and we obtain $D_q \leq \delta$ and $\cancel{H(\delta) = H(X|\hat{X})}$ $I(X:\hat{X}) = H(\alpha) - H(\delta)$.

$\square$ claim (b).

Claims (a) + (b) imply that $\rho(\delta) = H(\alpha) - H(\delta)$ when $0 \leq \delta \leq \alpha \leq 1/2$. $\square$ Thm

Rate-Dist. Thm - Sketch of proof (of achievability)

$\forall \varepsilon > 0\ \exists$ code C with rate $\le r$ and exp. dist $D_C \le \delta + \varepsilon$

(Recall:) **Defn**: The rate dist. fn is $\rho(\delta) = \inf\{r \ge 0 : (r,\delta) \text{ is achievable}\}$.

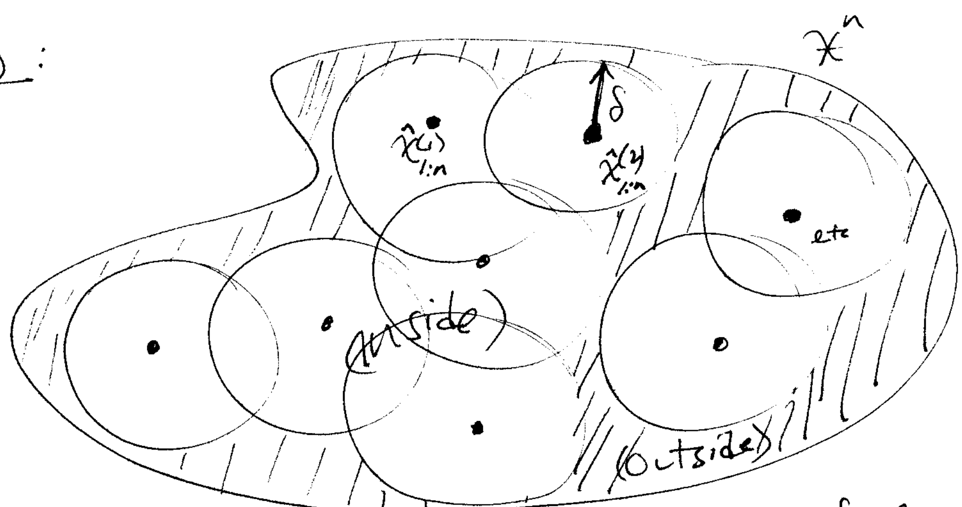
Thm: For an iid source $X_1, \ldots, X_n \sim p$ and bounded dist fn $d: \mathcal{X} \times \hat{\mathcal{X}} \to [0, d_{max}]$, $\rho(\delta) = \min_{q : D_q \le \delta} I(X : \hat{X})$

(where the min is over $q(\hat{x}|x)$ and $(X, \hat{X}) \sim p(x) q(\hat{x}|x)$.)

(Proof has two parts: $\ge$ and $\le$. Showing $\ge$ is relatively straightforward seq. of ineqs - see text.

Here's the basic idea of $\le$.)

Pf sketch ($\le$):



(Note: Purely for visualization purposes... $d(x, \hat{x})$ need not be a metric!)

$$D_C = \sum_{x_{1:n}} d(x_{1:n}, \hat{x}_{1:n}) p(x_{1:n}) = \underbrace{\sum_{\text{inside}} \overbrace{d(\ldots)}^{\le \delta + \varepsilon} p(\ldots)}_{\le \delta + \varepsilon} + \underbrace{\sum_{\text{outside}} \overbrace{d(\ldots)}^{\le d_{max}} p(\ldots)}_{\le d_{max} P(X_{1:n} \text{ "outside" } C)} \le (\delta + \varepsilon) + d_{max} P$$

Random Codebook

$$\Rightarrow \quad \bar{D} := \sum_C D_C\, p(C) \le (\delta + \varepsilon) + d_{max} \mathbb{P}(X_{1:n} \text{ "outside" } \mathbb{C}) \le \delta + 2\varepsilon \text{ for } n \text{ suff large}$$

$\mathbb{P}(X_{1:n} \text{ "outside" } \mathbb{C})$ small if R is big enough since # codeword is 2^{nR}. (Turns out $R > \rho(\delta)$ is good enough.) (The codewords "fill up" the space whp.)

$\Rightarrow \exists\, C$ with rate R s.t. $D_C \le \delta + 2\varepsilon$.

A little more formally:

Setup: Let $X_1, \dots, X_n \sim p$ iid. Let $d: \mathcal{X} \times \hat{\mathcal{X}} \to [0, d_{max}]$. Let $\delta \geq 0$.

(Goal: $\inf\{r \geq 0 : (r, \delta) \text{ is achievable}\} \leq \min_{q : D_q \leq \delta} I(X : \hat{X})$.)

Choose ~~Let~~ $q(\hat{x}|x)$ to attain the min; ~~so that $D_q \leq \delta$ and $\exists$ and let $(X_1$~~

and let $(X, \hat{X}) \sim p(x) q(\hat{x}|x)$. ~~(Goal: $\exists$ lossy code C~~

Let $R > I(X : \hat{X})$. Let $\varepsilon > 0$. (Goal: $\exists$ lossy code C with rate R and exp. dist. $D_C \leq \delta + 2\varepsilon$)

write $p(\hat{x})$ for $p(\hat{x}) = \sum_x p(x) q(\hat{x}|x)$.

Random Codebook: Let $\mathbb{C} = \left(\hat{X}^{(1)}_{1:n}, \dots, \hat{X}^{(2^{nR})}_{1:n}\right)$ where $\hat{X}^{(i)}_j \sim p(\hat{x})$ $\forall i, j$. iid.

(So, $\mathbb{P}(\mathbb{C} = c) = \prod_{i j} p(\hat{x}^{(i)}_j)$.)

Encode by Dist. Typicality: Introduce the dist. typical set:

$$T^{(n)}_\varepsilon = \Big\{ (x_{1:n}, \hat{x}_{1:n}) : x_{1:n} \in A^{(n)}_\varepsilon(X),\ \hat{x}_{1:n} \in A^{(n)}_\varepsilon(\hat{X}),\ (x_{1:n}, \hat{x}_{1:n}) \in A^{(n)}_\varepsilon(X, \hat{X}),$$
$$\text{and } \big| d(x_{1:n}, \hat{x}_{1:n}) - \underbrace{\mathbb{E} d(X, \hat{X})}_{D_q} \big| \leq \varepsilon \Big\}$$

To encode $x_{1:n}$ with codebook $C = (x^{(1)}_{1:n}, \ldots, x^{(2^{nR})}_{1:n})$:

- If $\exists i$ s.t. $(x_{1:n}, \hat{x}^{(i)}_{1:n}) \in T^{(n)}_\varepsilon$ then define $f_c(x_{1:n}) = i$. (If more than one such i, choose the first one, say.)
- Otherwise, define $f(x_{1:n}) = 1$. (or whatever.)

and to Decode: $g_c(i) = \hat{x}^{(i)}_{1:n}$.   <u>Abbreviate</u>: $C(x_{1:n}) = g_c(f_c(x_{1:n}))$.

(This defines a random code with rate R. Next task: show that exp. dist. of random code is small: $\bar{D} \leq \delta + 2\varepsilon$.)

<u>Exp. Dist.</u>: For fixed C, $D_C = E d(X_{1:n}, C(X_{1:n})) = \sum_{x_{1:n}} d(x_{1:n}, C(x_{1:n}))\, p(x_{1:n})$

$$= \underbrace{\sum_{x_{1:n} \text{ s.t. } (x_{1:n}, C(x_{1:n})) \in T^{(n)}_\varepsilon} \overbrace{d(\cdots)}^{\leq D_q + \varepsilon \leq \delta + \varepsilon} p(\cdots)}_{\leq\, \delta + \varepsilon} + \underbrace{\sum_{x_{1:n} \text{ s.t. } (x_{1:n}, C(x_{1:n})) \notin T^{(n)}_\varepsilon} \overbrace{d(\cdots)}^{\leq d_{max}} p(\cdots)}_{\leq\, d_{max} \sum_{\notin T} p(x_{1:n})}$$

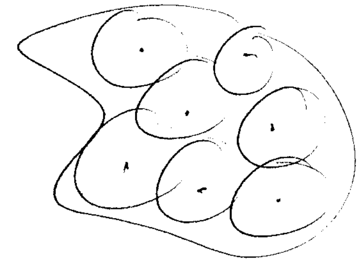
(by defn of $T^{(n)}_\varepsilon$) (by our choice of q)

$$\Rightarrow \bar{D} := \sum_c D_c\, p(c) \leq (\delta + \varepsilon) + d_{max} \underbrace{\sum_c \sum_{x_{1:n} \text{ s.t. } (x_{1:n}, C(x_{1:n})) \notin T} p(x_{1:n})\, p(c)}_{= \sum_c \sum_{x_{1:n}} p(x_{1:n}) p(c)\, I((x_{1:n}, C(x_{1:n})) \in T) = \mathbb{P}((X_{1:n}, \mathbb{C}(X_{1:n})) \in T^{(n)}_\varepsilon)}$$

since $X_{1:n} \perp\!\!\!\perp \mathbb{C}$

(Now, show that this prob. can be made arbitrarily small.)

(Here is a heuristic argument — see text for rigorous pf.)



$$\mathbb{P}((X_{1:n}, C(X_{1:n})) \notin T) = \mathbb{P}((X_{1:n}, \hat{X}^{(i)}_{1:n}) \notin T \quad \forall i=1,\ldots,2^{nR})$$

$$\text{"}\approx\text{"} \prod_{i=1}^{2^{nR}} \mathbb{P}((X_{1:n}, \hat{X}^{(i)}_{1:n}) \notin T) \approx \left(1 - 2^{-nI(X:\hat{X})}\right)^{2^{nR}}$$

Not really correct... but this is the basic idea. (See text for details)

$$\mathbb{P}((X_{1:n}, \hat{X}^{(i)}_{1:n}) \notin T) = 1 - \mathbb{P}((X_{1:n}, \hat{X}_{1:n}) \in T) \approx 1 - \frac{1}{2^{nI(X:\hat{X})}}$$

Joint AEP (same pf works)

$$= \left(1 - \frac{2^{n(R-I)}}{2^{nR}}\right)^{2^{nR}} \approx e^{-2^{n(R-I)}} \xrightarrow[\text{as } n\to\infty]{} 0 \quad \text{since } R > I(X:\hat{X}).$$

$\lim \left(1-\frac{a}{n}\right)^n = e^{-a}$

$\Rightarrow \quad \bar{D} \lesssim (\delta+\varepsilon) + d_{max}\, e^{-2^{n(R-I)}} \le \delta + 2\varepsilon$ for n suff. large.

Since $\bar{D} = \sum_C D_C\, p(C)$, then there is at least one code C s.t. $D_C \le \bar{D}$.

$\Rightarrow$ $\exists\, C$ with rate R and $D_C \le \delta + 2\varepsilon$. $\Rightarrow$ (R,δ) is achievable. $\square$