Notation: in the following proofs, we shall make use of the *independent pairs* $(X,Y) \perp\!\!\!\perp (\overline{X},\overline{Y})$

$$\begin{aligned} X &\to Y \quad (X: \text{ sent codeword}) \\ \overline{X} &\to \overline{Y} \quad (\overline{X}: \text{ unsent codeword}) \end{aligned}$$

The joint distribution is given by:

$$P_{XY\overline{XY}}(a,b,\overline{a},\overline{b}) = P_X(a)P_{Y|X}(b|a)P_X(\overline{a})P_{Y|X}(\overline{b}|\overline{a}).$$

15.1 Information density

Definition 15.1 (Information density)**.** Given joint distribution $P_{X,Y}$ we define

$$i_{P_{XY}}(x;y) = \log\frac{P_{Y|X}(y|x)}{P_Y(y)} = \log\frac{dP_{Y|X=x}(y)}{dP_Y(y)} \tag{15.1}$$

and we define $i_{P_{XY}}(x;y) = +\infty$ for all y in the singular set where $P_{Y|X=x}$ is not absolutely continuous w.r.t. P_Y. We also define $i_{P_{XY}}(x;y) = -\infty$ for all y such that $dP_{Y|X=x}/dP_Y$ equals zero. We will almost always abuse notation and write $i(x;y)$ dropping the subscript $P_{X,Y}$, assuming that the joint distribution defining $i(\cdot;\cdot)$ is clear from the context.
Notice that $i(x;y)$ depends on the underlying P_X and $P_{Y|X}$, which should be understood from the context.

Remark 15.1 (Intuition)**.** Information density is a useful concept in understanding decoding. In discriminating between two codewords, one concerns with (as we learned in binary hypothesis testing) the LLR, $\log\frac{P_{Y|X=c_1}}{P_{Y|X=c_2}}$. In M-ary hypothesis testing, a similar role is played by information density $i(c_1;y)$, which, loosely speaking, evaluates the likelihood of c_1 against the average likelihood, or "everything else", which we model by P_Y.

Remark 15.2 (Joint measurability)**.** There is a measure-theoretic subtlety in (15.1): The so-defined function $i(\cdot;\cdot)$ may not be a measurable function on the product space $\mathcal{X}\times\mathcal{Y}$. For resolution, see Section 2.6* and Remark 2.4 in particular.

Remark 15.3 (Alternative definition)**.** Observe that for discrete $\mathcal{X},\mathcal{Y}$, (15.1) is equivalently written as

$$i(x;y) = \log\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)} = \log\frac{P_{X|Y}(x|y)}{P_X(x)}$$

For the continuous case, people often use the alternative definition, which is symmetric in X and Y and is measurable w.r.t. $\mathcal{X}\times\mathcal{Y}$:

$$i(x;y) = \log\frac{dP_{X,Y}}{dP_X\times P_Y}(x,y) \tag{15.2}$$

Notice a subtle difference between (15.1) and (15.2) for the continuous case: In (15.2) the Radon-Nikodym derivative is only defined up to sets of measure zero, therefore whenever $P_X(x) = 0$ the value of $P_Y(i(x,Y) > t)$ is undefined. This problem does not occur with definition (15.1), and that is why we prefer it. In any case, for discrete $\mathcal{X}$, $\mathcal{Y}$, or under other regularity conditions, all the definitions are equivalent.

Proposition 15.1 (Properties of information density)**.**

1. *$\mathbb{E}[i(X;Y)] = I(X;Y)$. This justifies the name "(mutual) information density".*

2. *If there is a bijective transformation $(X,Y) \to (A,B)$, then almost surely $i_{P_{XY}}(X;Y) = i_{P_{AB}}(A;B)$ and in particular, distributions of $i(X;Y)$ and $i(A;B)$ coincide.*

3. *(Conditioning and unconditioning trick) Suppose that $f(y) = 0$ and $g(x) = 0$ whenever $i(x;y) = -\infty$, then*[1]

$$\mathbb{E}[f(Y)] = \mathbb{E}[\exp\{-i(x;Y)\}f(Y)|X = x], \forall x \tag{15.3}$$

$$\mathbb{E}[g(X)] = \mathbb{E}[\exp\{-i(X;y)\}g(X)|Y = y], \forall y \tag{15.4}$$

4. *Suppose that $f(x,y) = 0$ whenever $i(x;y) = -\infty$, then*

$$\mathbb{E}[f(\overline{X},Y)] = \mathbb{E}[\exp\{-i(X;Y)\}f(X,Y)] \tag{15.5}$$

$$\mathbb{E}[f(X,\overline{Y})] = \mathbb{E}[\exp\{-i(X;Y)\}f(X,Y)] \tag{15.6}$$

Proof. The proof is simply change of measure. For example, to see (15.3), note

$$\mathbb{E}f(Y) = \sum_{y\in\mathcal{Y}} P_Y(y)f(y) = \sum_{y\in\mathcal{Y}} P_{Y|X}(y|x)\frac{P_Y(y)}{P_{Y|X}(y|x)}f(y)$$

notice that by the assumption on $f(\cdot)$, the summation is valid even if for some y we have that $P_{Y|X}(y|x) = 0$. Similarly, $\mathbb{E}[f(x,Y)] = \mathbb{E}[\exp\{-i(x;Y)\}f(x,Y)|X = x]$. Integrating over $x \sim P_X$ gives (15.5). The rest are by interchanging X and Y. □

Corollary 15.1.

$$\mathbb{P}[i(x;Y) > t] \le \exp(-t) \tag{15.7}$$

$$\mathbb{P}[i(\overline{X};Y) > t] \le \exp(-t) \tag{15.8}$$

Proof. Pick $f(Y) = \mathbf{1}\{i(x;Y) > t\}$ in (15.3). □

Remark 15.4. We have used this trick before: For any probability measure P and any measure Q,

$$Q\left[\log\frac{\mathrm{d}P}{\mathrm{d}Q} \ge t\right] \le \exp(-t). \tag{15.9}$$

for example, in hypothesis testing (Corollary 10.1). In data compression, we frequently used the fact that $|\{x : \log P_X(x) \ge t\}| \le \exp(-t)$, which is also of the form (15.9) with Q = counting measure.

[1] Note that (15.3) holds when $i(x;y)$ is defined as $i = \log\frac{dP_{Y|X}}{P_Y}$, and (15.4) holds when $i(x;y)$ is defined as $i = \log\frac{dP_{X|Y}}{P_X}$. (15.5) and (15.6) hold under either of the definitions. Since in the following we shall only make use of (15.3) and (15.5), this is another reason we adopted definition (15.1).

15.2 Shannon's achievability bound

Theorem 15.1 (Shannon's achievability bound)**.** *For a given* $P_{Y|X}$*,* $\forall P_X$*,* $\forall \tau > 0$*,* $\exists(M,\epsilon)$*-code with*

$$\epsilon \le \mathbb{P}[i(X;Y) \le \log M + \tau] + \exp(-\tau). \tag{15.10}$$

Proof. Recall that for a given codebook $\{c_1, \ldots, c_M\}$, the optimal decoder is MAP, or equivalently, ML, since the codewords are equiprobable:

$$\begin{aligned} g^*(y) &= \operatorname*{argmax}_{m\in[M]} P_{X|Y}(c_m|y) \\ &= \operatorname*{argmax}_{m\in[M]} P_{Y|X}(y|c_m) \\ &= \operatorname*{argmax}_{m\in[M]} i(c_m;y). \end{aligned}$$

The step of selecting the maximum likelihood can make analyzing the error probability difficult. Similar to what we did in almost loss compression (e.g., Theorem 7.4), the magic in showing the following two achievability bounds is to consider a suboptimal decoder. In Shannon's bound, we consider a threshold-based suboptimal decoder $g(y)$ as follows:

$$g(y) = \begin{cases} m, & \exists! c_m \text{ s.t. } i(c_m;y) \ge \log M + \tau \\ \mathsf{e}, & \text{o.w.} \end{cases}$$

Interpretation: $i(c_m;y) \ge \log M + \tau \Leftrightarrow P_{X|Y}(c_m|y) \ge M\exp(\tau)P_X(c_m)$, i.e., the likelihood of c_m being the transmitted codeword conditioned on receiving y exceeds some threshold.

For a given codebook $(c_1, \ldots, c_M)$, the error probability is:

$$P_e(c_1, \ldots, c_M) = \mathbb{P}[\{i(c_W;Y) \le \log M + \tau\} \cup \{\exists \overline{m} \ne W, i(c_{\overline{m}};Y) > \log M + \tau\}]$$

where W is uniform on $[M]$.
We generate the codebook $(c_1, \ldots, c_M)$ randomly with $c_m \sim P_X$ i.i.d. for $m \in [M]$. By symmetry, the error probability averaging over all possible codebooks is given by:

$$\begin{aligned} & \mathbb{E}[P_e(c_1, \ldots, c_M)] \\ &= \mathbb{E}[P_e(c_1, \ldots, c_M)|W=1] \\ &= \mathbb{P}[\{i(c_1;Y) \le \log M + \tau\} \cup \{\exists \overline{m} \ne 1, i(c_{\overline{m}},Y) > \log M + \tau\}|W=1] \\ &\le \mathbb{P}[i(c_1;Y) \le \log M + \tau|W=1] + \sum_{m=2}^{M} \mathbb{P}[i(c_m;Y) > \log M + \tau|W=1] && \text{(union bound)} \\ &= \mathbb{P}\left[i(X;Y) \le \log M + \tau\right] + (M-1)\mathbb{P}\left[i(\overline{X};Y) > \log M + \tau\right] && \text{(random codebook)} \\ &\le \mathbb{P}\left[i(X;Y) \le \log M + \tau\right] + (M-1)\exp(-(\log M + \tau)) && \text{(by Corollary 15.1)} \\ &\le \mathbb{P}\left[i(X;Y) \le \log M + \tau\right] + \tau) + \exp(-\tau) \end{aligned}$$

Finally, since the error probability averaged over the random codebook satisfies the upper bound, there must exist some code allocation whose error probability is no larger than the bound. □

Remark 15.5 (Typicality)**.**

- The property of a pair (x, y) satisfying the condition $\{i(x; y) \geq \gamma\}$ can be interpreted as "**joint typicality**". Such version of joint typicality is useful when random coding is done in product spaces with $c_j \sim P_X^n$ (i.e. coordinates of the codeword are iid).

- A popular alternative to the definition of typicality is to require that the empirical joint distribution is close to the true joint distribution, i.e., $\hat{P}_{x^n,y^n} \approx P_{XY}$, where

  $$\hat{P}_{x^n,y^n}(a, b) = \frac{1}{n} \cdot \#\{j : x_j = a, y_j = b\}.$$

  This definition is natural for cases when random coding is done with $c_j \sim$ uniform on the set $\{x^n : \hat{P}_{x^n} \approx P_X\}$ (type class).

15.3 Dependence-testing bound

Theorem 15.2 (DT bound). *$\forall P_X$, $\exists(M, \epsilon)$-code with*

$$\epsilon \leq \mathbb{E}\left[\exp\left\{-\left(i(X;Y) - \log\frac{M-1}{2}\right)^+\right\}\right] \tag{15.11}$$

where $x^+ \triangleq \max(x, 0)$.

Proof. For a fixed γ, consider the following suboptimal decoder:

$$g(y) = \begin{cases} m, & \text{for the smallest } m \text{ s.t. } i(c_m; y) \geq \gamma \\ \mathtt{e}, & \text{o/w} \end{cases}$$

Note that given a codebook $\{c_1, \ldots, c_M\}$, we have by union bound

$$\begin{aligned}\mathbb{P}[\hat{W} \neq j | W = j] &= \mathbb{P}[i(c_j; Y) \leq \gamma | W = j] + \mathbb{P}[i(c_j; Y) > \gamma, \exists k \in [j-1], \text{ s.t. } i(c_k; Y) > \gamma] \\ &\leq \mathbb{P}[i(c_j; Y) \leq \gamma | W = j] + \sum_{k=1}^{j-1} \mathbb{P}[i(c_k; Y) > \gamma | W = j].\end{aligned}$$

Averaging over the randomly generated codebook, the expected error probability is upper bounded by:

$$\begin{aligned}\mathbb{E}[P_e(c_1, \ldots, c_M)] &= \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}[\hat{W} \neq j | W = j] \\ &\leq \frac{1}{M}\sum_{j=1}^{M}\left(\mathbb{P}[i(X;Y) \leq \gamma] + \sum_{k=1}^{j-1}\mathbb{P}[i(\overline{X};Y) > \gamma]\right) \\ &= \mathbb{P}[i(X;Y) \leq \gamma] + \frac{M-1}{2}\mathbb{P}[i(\overline{X};Y) > \gamma] \\ &= \mathbb{P}[i(X;Y) \leq \gamma] + \frac{M-1}{2}\mathbb{E}[\exp(-i(X;Y))\mathbf{1}\{i(X;Y) > \gamma\}] \quad \text{(by (15.3))} \\ &= \mathbb{E}\left[\mathbf{1}\{i(X;Y) \leq \gamma\} + \frac{M-1}{2}\exp(-i(X;Y))\mathbf{1}\{i(X,Y) > \gamma\}\right] \\ &= \mathbb{E}\left[\min\left(1, \frac{M-1}{2}\exp(-i(X;Y))\right)\right] \quad \left(\gamma = \log\frac{M-1}{2} \text{ minimizes the upper bound}\right) \\ &= \mathbb{E}\left[\exp\left\{-\left(i(X;Y) - \log\frac{M-1}{2}\right)^+\right\}\right].\end{aligned}$$

To optimize over γ, note the simple observation that $U\mathbf{1}_E + V\mathbf{1}_{\{E^c\}} \geq \min\{U, V\}$, with equality iff $U \geq V$ on E. Therefore for any x, y, $\mathbf{1}[i(x;y) \leq \gamma] + \frac{M-1}{2} e^{-i(x;y)} \mathbf{1}[i(x;y) > \gamma] \geq \min(1, \frac{M-1}{2} e^{-i(x;y)})$, achieved by $\gamma = \log \frac{M-1}{2}$ regardless of x, y. □

Note: Dependence-testing: The RHS of (15.11) is equivalent to the minimum error probability of the following Bayesian hypothesis testing problem:

$$\begin{aligned} &H_0 : X, Y \sim P_{X,Y} \text{ versus } \quad H_1 : X, Y \sim P_X P_Y \\ &\text{prior prob.: } \pi_0 = \frac{2}{M+1}, \pi_1 = \frac{M-1}{M+1}. \end{aligned}$$

Note that $X, Y \sim P_{X,Y}$ and $\overline{X}, Y \sim P_X P_Y$, where X is the sent codeword and $\overline{X}$ is the unsent codeword. As we know from binary hypothesis testing, the best threshold for the LRT to minimize the weighted probability of error is $\log \frac{\pi_1}{\pi_0}$.

Note: Here we avoid minimizing over τ in Shannon's bound (15.10) to get the minimum upper bound in Theorem 15.1. Moreover, DT bound is stronger than the best Shannon's bound (with optimized τ).

Note: Similar to the random coding achievability bound of almost lossless compression (Theorem 7.4), in Theorem 15.1 and Theorem 15.2 we only need the random codewords to be *pairwise* independent.

15.4 Feinstein's Lemma

The previous achievability results are obtained using *probabilistic* methods (random coding). In contrast, the following achievability due to Feinstein uses a **greedy** construction. Moreover, Feinstein's construction holds for **maximal** probability of error.

Theorem 15.3 (Feinstein's lemma)**.** *$\forall P_X$, $\forall \gamma > 0$, $\forall \epsilon \in (0,1)$, $\exists (M, \epsilon)_{\max}$-code such that*

$$M \geq \gamma(\epsilon - \mathbb{P}[i(X;Y) < \log \gamma]) \tag{15.12}$$

Remark 15.6 (Comparison with Shannon's bound)**.** We can also interpret (15.12) as for fixed M, there exists an $(M, \epsilon)_{\max}$-code that achieves the maximal error probability bounded as follows:

$$\epsilon \leq \mathbb{P}[i(X;Y) < \log \gamma] + \frac{M}{\gamma}$$

Take $\log \gamma = \log M + \tau$, this gives the bound of exactly the same form in (15.10). However, the two are proved in seemingly quite different ways: Shannon's bound is by random coding, while Feinstein's bound is by greedily selecting the codewords. Nevertheless, Feinstein's bound is stronger in the sense that it concerns about the max error probability instead of the average.

Proof. The idea is to construct the codebook of size M in a greedy way.

For every $x \in \mathcal{X}$, associate it with a preliminary decode region defined as follows:

$$E_x \triangleq \{y : i(x;y) \geq \log \gamma\}$$

Notice that the preliminary decoding regions $\{E_x\}$ may be overlapping, and we denote the final decoding region partition regions by $\{D_x\}$.

We can assume that $\mathbb{P}[i(X;Y) < \log \gamma] \leq \epsilon$, for otherwise the R.H.S. of (15.12) is negative and there is nothing to prove. We first claim that there exists some c such that $P_Y[E_c | X = c] \geq 1 - \epsilon$.

Show by contradiction. Assume that $\forall c \in \mathcal{X}$, $\mathbb{P}[i(c;Y) \geq \log\gamma | X = c] < 1 - \epsilon$, then pick $c \sim P_X$, we have $\mathbb{P}[i(X;Y) \geq \log\gamma] < 1 - \epsilon$, which is a contradiction.

Then we construct the codebook in the following greedy way:

1. Pick c_1 to be any codeword such that $P_Y[E_{c_1} | X = c_1] \geq 1 - \epsilon$, and set $D_1 = E_{c_1}$;

2. Pick c_2 to be any codeword such that $P_Y[E_{c_2} \backslash D_1 | X = c_2] \geq 1 - \epsilon$, and set $D_2 = E_{c_2} \backslash D_1$; $\ldots$

3. Pick c_M to be any codeword such that $P_Y[E_{c_M} \backslash \cup_{j=1}^{M-1} D_j | X = c_M] \geq 1 - \epsilon$, and set $D_M = E_{c_M} \backslash \cup_{j=1}^{M-1} D_j$. We stop if no more codeword can be found, i.e., M is determined by the stopping condition:
$$\forall x_0 \in \mathcal{X}, P_Y[E_{x_0} \backslash \cup_{j=1}^{M} D_j | X = x_0] < 1 - \epsilon$$

Averaging over $x_0 \sim P_X$, the stopping condition gives that

$$\mathbb{P}(\{i(X;Y) \geq \log\gamma\} \backslash \{Y \in \cup_{j=1}^{M} D_j\}) < 1 - \epsilon$$

by union bound $P(A \backslash B) \geq P(A) - P(B)$, we have

$$\mathbb{P}(i(X;Y) \geq \log\gamma) - \sum_{j=1}^{M} P_Y(D_j) < 1 - \epsilon$$
$$\Rightarrow \mathbb{P}(i(X;Y) \geq \log\gamma) - \frac{M}{\gamma} < 1 - \epsilon$$

where the last step makes use of the following key observation:

$$P_Y(D_j) \leq P_Y(E_{c_j}) = P_Y(i(c_j;Y) \geq \log\gamma) < \frac{1}{\gamma}, \quad \text{(by Corollary 15.1)}.$$

□

6.441 Information Theory
Spring 2016