Objects of study so far:

1. $P_X$ - Single distribution, Compression
2. $P_X$ vs $Q_X$ - Comparing two distributions, Hypothesis testing
3. Now: $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ (called a *random transformation*) - A collection of distributions

## 14.1 Channel Coding

**Definition 14.1.** An $M$-code for $P_{Y|X}$ is an encoder/decoder pair $(f, g)$ of (randomized) functions[1]

- encoder $f : [M] \to \mathcal{X}$
- decoder $g : \mathcal{Y} \to [M] \cup \{\mathsf{e}\}$

**Notation:** $[M] \triangleq \{1, \ldots, M\}$.

In most cases $f$ and $g$ are deterministic functions, in which case we think of them (equivalently) in terms of codewords, codebooks, and decoding regions

- $\forall i \in [M] : c_i = f(i)$ are *codewords*, the collection $\mathcal{C} = \{c_1, \ldots, c_M\}$ is called a *codebook*.
- $\forall i \in [M], D_i = g^{-1}(\{i\})$ is the *decoding region* for $i$.

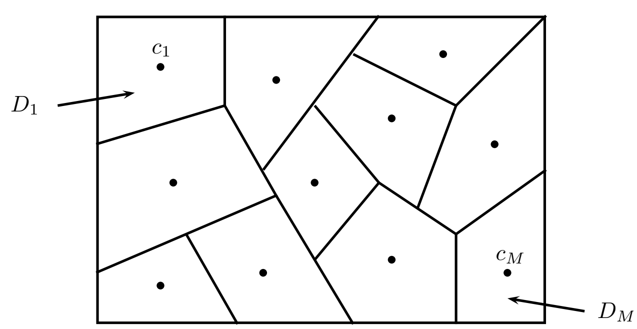



Figure 14.1: When $\mathcal{X} = \mathcal{Y}$, the decoding regions can be pictured as a partition of the space, each containing one codeword.

**Note**: The underlying probability space for channel coding problems will always be

$$W \xrightarrow{f} X \xrightarrow{P_{Y|X}} Y \xrightarrow{g} \hat{W}$$

[1]For randomized encoder/decoders, we identify $f$ and $g$ as probability transition kernels $P_{X|W}$ and $P_{\hat{W}|Y}$.

When the source alphabet is $[M]$, the joint distribution is given by:

$$\text{(general) } P_{WXY\hat{W}}(m,a,b,\hat{m}) = \frac{1}{M}P_{X|W}(a|m)P_{Y|X}(b|a)P_{\hat{W}|Y}(\hat{m}|b)$$

$$\text{(deterministic } f,g\text{) } P_{WXY\hat{W}}(m,c_m,b,\hat{m}) = \frac{1}{M}P_{Y|X}(b|c_m)\mathbf{1}\{b \in D_{\hat{m}}\}$$

Throughout the notes, these quantities will be called:

- $W$ - Original message
- $X$ - (Induced) Channel input
- $Y$ - Channel output
- $\hat{W}$ - Decoded message

### 14.1.1 Performance Metrics

Three ways to judge the quality of a code in terms of error probability:

1. $P_e \triangleq \mathbb{P}[W \neq \hat{W}]$ - Average error probability.
2. $P_{e,\max} \triangleq \max_{m\in[M]} \mathbb{P}[\hat{W} \neq m|W = m]$ - Maximum error probability.
3. In the special case when $M = 2^k$, think of $W = S^k \in \mathbb{F}_2^k$ as a length $k$ bit string. Then the bit error rate is $P_b \triangleq \frac{1}{k}\sum_{j=1}^{k}\mathbb{P}[S_j \neq \hat{S}_j]$, which means the average fraction of errors in a $k$-bit block. It is also convenient to introduce in this case the Hamming distance

   $$d_H(S^k, \hat{S}^k) \triangleq \#\{i : S_i \neq \hat{S}_j\}\,.$$

   Then, the bit-error rate becomes the normalized expected Hamming distance:

   $$P_b = \frac{1}{k}\mathbb{E}[d_H(S^k, \hat{S}^k)]\,.$$

   To distinguish the bit error rate $P_b$ from the previously defined $P_e$ and $P_{e,\max}$, we will also call the latter the average (resp. max) block error rate.

The most typical metric is average probability of error, but the others will be used occasionally in the course as well. By definition, $P_e \le P_{e,\max}$. Therefore maximum error probability is a more stringent criterion which offers uniform protection for all codewords.

### 14.1.2 Fundamental Limit of $P_{Y|X}$

**Definition 14.2.** A code $(f,g)$ is an $(M,\epsilon)$-code for $P_{Y|X}$ if $f : [M] \to \mathcal{X}$, $g : \mathcal{Y} \to [M] \cup \{e\}$, and $P_e \le \epsilon$. Similarly, an $(M,\epsilon)_{\max}$-code must satisfy $P_{e,\max} \le \epsilon$.

Then the fundamental limits of channel codes are defined as

$$M^*(\epsilon) = \max\{M : \exists (M,\epsilon) - \mathit{code}\}$$
$$M^*_{\max}(\epsilon) = \max\{M : \exists (M,\epsilon)_{\max} - \mathit{code}\}$$

**Remark:** $\log_2 M^*$ gives the maximum number of bits that we can pump through a channel $P_{Y|X}$ while still having the error probability (in the appropriate sense) at most $\epsilon$.

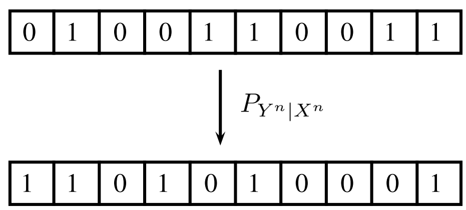
**Example**: The random transformation BSC$(n,\delta)$ (binary symmetric channel) is defined as

$$\mathcal{X} = \{0,1\}^n$$
$$\mathcal{Y} = \{0,1\}^n$$

where the input $X^n$ is contaminated by additive noise $Z^n \perp\!\!\!\perp X^n$ and the channel outputs

$$Y^n = X^n \oplus Z^n$$

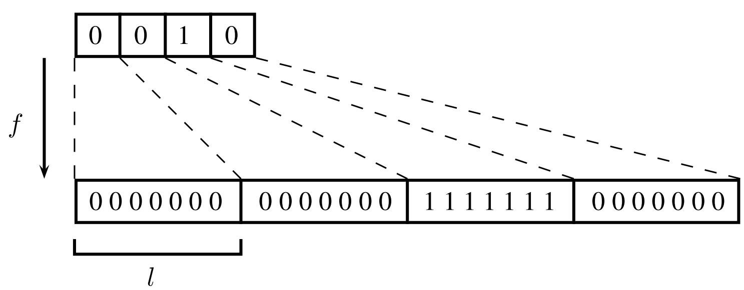
where $Z^n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\delta)$. Pictorially, the BSC$(n,\delta)$ channel takes a binary sequence length $n$ and flips the bits independently with probability $\delta$:




**Question:** When $\delta = .11$, $n = 1000$, what is the max number of bits you can send with $P_e \le 10^{-3}$?
**Ideas:**

0. Can one send 1000 bits with $P_e \le 10^{-3}$? No and apparently the probability that at least one bit is flipped is $P_e = 1 - (1-\delta)^n \approx 1$. This implies that uncoded transmission does not meet our objective and coding is necessary – tradeoff: reduce number of bits to send, increase probability of success.

1. Take each bit and repeat it $l$ times ($l$-repetition code).




With majority decoding, the probability of error of this scheme is $P_e \approx k\mathbb{P}[\text{Binom}(l,\delta) > l/2]$ and $kl \le n = 1000$, which for $P_e \le 10^{-3}$ gives $l = 21$, $k = 47$ bits.

2. Reed-Muller Codes $(1,r)$. Interpret a message $a_0,\ldots,a_{r-1} \in \mathbb{F}_2^r$ as the polynomial (in this case, a degree-1 and $(r-1)$-variate polynomial) $\sum_{i=1}^{r-1} a_i x_i + a_0$, then codewords are formed by evaluating the polynomial at all possible $x^{r-1} \in \mathbb{F}_2^{r-1}$. This code, which maps $r$ bits to $2^{r-1}$ bits, has minimum distance $2^{r-2}$. For $r = 7$, there is a $[64,7,32]$ Reed-Muller code and it can be shown that the MAP decoder of this code passed over the $BSC(n = 64, \delta = 0.11)$ achieves probability of error $\le 6\cdot 10^{-6}$. Thus, we can use 16 such blocks (each carrying 7 data bits and occupying 64 bits on the channel) over the BSC$(1024,\delta)$, and still have (union bound) overall $P_e \lesssim 10^{-4}$. This allows us to send $7\cdot 16 = 112$ bits in 1024 channel uses, more than double that of the repetition code.

3. Shannon's theorem (to be shown soon) tells us that over memoryless channel of blocklength $n$ the fundamental limit satisfies

$$\log M^* = nC + o(n) \tag{14.1}$$

as $n \to \infty$ and for arbitrary $\epsilon \in (0,1)$. Here $C = \max_X I(X_1; Y_1)$ is the capacity of the single-letter channel. In our case we have

$$I(X;Y) = \max_{P_X} I(X; X+Z) = \log 2 - h(\delta) \approx \frac{1}{2} \text{ bit}$$

So Shannon's expansion (14.1) can be used to predict (non-rigorously, of course) that it should be possible to send around 500 bits reliably. As it turns out, for this blocklength this is not quite possible.

4. Even though calculating $\log M^*$ is not computationally feasible (searching over all codebooks is doubly exponential in block length $n$), we can find bounds on $\log M^*$ that are easy to compute. We will show later in the course that in fact, for BSC(1000, .11)

$$414 \le \log M^* \le 416$$

5. The first codes to approach the bounds on $\log M^*$ are called *Turbo codes* (after the turbocharger engine – where exhaust is fed back in to power the engine). This class of codes is known as *sparse graph codes*, of which LDPC codes are particularly well studied. As a rule of thumb, these codes typically approach $80 \ldots 90\%$ of $\log M^*$ when $n \approx 10^3 \ldots 10^4$.

## 14.2 Basic Results

Recall that the object of our study is $M^*(\epsilon) = \max\{M : \exists (M,\epsilon) - code\}$.

### 14.2.1 Determinism

1. Given any encoder $f : [M] \to \mathcal{X}$, the decoder that minimizes $P_e$ is the *Maximum A Posteriori (MAP)* decoder, or equivalently, the *Maximal Likelihood (ML)* decoder, since the codewords are equiprobable:

$$\begin{aligned} g^*(y) &= \operatorname*{argmax}_{m\in[M]} \mathbb{P}[W = m | Y = y] \\ &= \operatorname*{argmax}_{m\in[M]} \mathbb{P}[Y = y | W = m] \end{aligned}$$

Furthermore, for a fixed $f$, the MAP decoder $g$ is deterministic

2. For given $M$, $P_{Y|X}$, the $P_e$-minimizing encoder is deterministic.

*Proof.* Let $f : [M] \to \mathcal{X}$ be a random transformation. We can always represent randomized encoder as deterministic encoder with auxiliary randomness. So instead of $f(a|m)$, consider the deterministic encoder $\tilde{f}(m,u)$, that receives external randomness $u$. Then looking at all possible values of the randomness,

$$P_e = P[W \neq \hat{W}] = \mathbb{E}_U[\mathbb{P}[W \neq \hat{W}|U] = \mathbb{E}_U[P_e(U)]$$

Each $u$ in the expectation gives a deterministic encoder, hence there is a deterministic encoder that is at least as good as the average of the collection, i.e., $\exists u_0$ s.t. $P_e(u_0) \le \mathbb{P}[W \neq \hat{W}]$ □

**Remark:** If instead we use maximal probability of error as our performance criterion, then these results don't hold; randomized encoders and decoders may improve performance. Example: consider $M = 2$ and we are back to the binary hypotheses testing setup. The optimal decoder (test) that minimizes the maximal Type-I and II error probability, i.e., $\max\{1-\alpha,\beta\}$, is not deterministic, if $\max\{1-\alpha,\beta\}$ is not achieved at a vertex of the region $\mathcal{R}(P,Q)$.

### 14.2.2 Bit Error Rate vs Block Error Rate

Now we give a bound on the average probability of error in terms of the bit error probability.

**Theorem 14.1.** *For all* $(f,g)$, $M = 2^k \implies P_b \le P_e \le kP_b$

**Remark:** The most often used direction $P_b \ge \frac{1}{k}P_e$ is rather loose for large $k$.

*Proof.* Recall that $M = 2^k$ gives us the interpretation of $W = S^k$ sequence of bits.

$$\frac{1}{k}\sum_{i=1}^{k}\mathbf{1}\{S_i \ne \hat{S}_i\} \le \mathbf{1}\{S^k \ne \hat{S}^k\} \le \sum_{i=1}^{k}\mathbf{1}\{S_i \ne \hat{S}_i\}$$

Where the first inequality is obvious and the second follow from the union bound. Taking expectation of the above expression gives the theorem. □

**Theorem 14.2** (Assouad)**.** *If* $M = 2^k$ *then*

$$P_b \ge \min\{\mathbb{P}[\hat{W} = c'|W = c] : c, c' \in \mathbb{F}_2^k, d_H(c,c') = 1\}\,.$$

*Proof.* Let $e_i$ be a length $k$ vector that is 1 in the $i$-th position, and zero everywhere else. Then

$$\sum_{i=1}^{k}\mathbf{1}\{S_i \ne \hat{S}_i\} \ge \sum_{i=1}^{k}\mathbf{1}\{S^k = \hat{S}^k + e_i\}$$

Dividing by $k$ and taking expectation gives

$$\begin{aligned} P_b &\ge \frac{1}{k}\sum_{i=1}^{k}\mathbb{P}[S^k = \hat{S}^k + e_i] \\ &\ge \min\{\mathbb{P}[\hat{W} = c'|W = c] : c, c' \in \mathbb{F}_2^k, d_H(c,c') = 1\}\,. \end{aligned}$$

□

Similarly, we can prove the following generalization:

**Theorem 14.3.** *If* $A, B \in \mathbb{F}_2^k$ *(with arbitrary marginals!) then for every* $r \ge 1$ *we have*

$$P_b = \frac{1}{k}\mathbb{E}[d_H(A,B)] \ge \binom{k-1}{r-1}P_{r,\min} \tag{14.2}$$

$$P_{r,\min} \triangleq \min\{\mathbb{P}[B = c'|A = c] : c, c' \in \mathbb{F}_2^k, d_H(c,c') = r\} \tag{14.3}$$

*Proof.* First, observe that

$$\mathbb{P}[d_H(A,B) = r|A = a] = \sum_{b:d_H(a,b)=r} P_{B|A}(b|a) \ge \binom{k}{r}P_{r,\min}\,.$$

Next, notice

$$d_H(x,y) \ge r1\{d_H(x,y) = r\}$$

and take the expectation with $x \sim A$, $y \sim B$. □

**Remark:** In statistics, Assouad's Lemma is a useful tool for obtaining lower bounds on the minimax risk of an estimator. Say the data $X$ is distributed according to $P_\theta$ parameterized by $\theta \in \mathbb{R}^k$ and let $\hat{\theta} = \hat{\theta}(X)$ be an estimator for $\theta$. The goal is to minimize the maximal risk $\sup_{\theta\in\Theta} \mathbb{E}_\theta[\|\theta - \hat{\theta}\|_1]$. A lower bound (Bayesian) to this worst-case risk is the average risk $\mathbb{E}[\|\theta - \hat{\theta}\|_1]$, where $\theta$ is distributed to any prior. Consider $\theta$ uniformly distributed on the hypercube $\{0, \epsilon\}^k$ with side length $\epsilon$ embedded in the space of parameters. Then

$$\inf_{\hat{\theta}} \sup_{\theta\in\{0,\epsilon\}^k} \mathbb{E}[\|\theta - \hat{\theta}\|_1] \geq \frac{k\epsilon}{4} \min_{d_H(\theta,\theta')=1} (1 - \mathrm{TV}(P_\theta, P_{\theta'}))\,. \tag{14.4}$$

This can be proved using similar ideas to Theorem 14.2. WLOG, assume that $\epsilon = 1$.

$$\begin{aligned}
\mathbb{E}[\|\theta - \hat{\theta}\|_1] &\overset{(a)}{\geq} \frac{1}{2}\mathbb{E}[\|\theta - \hat{\theta}_{dis}\|_1] = \frac{1}{2}\mathbb{E}[d_H(\theta, \hat{\theta}_{dis})] \\
&\geq \frac{1}{2}\sum_{i=1}^k \min_{\hat{\theta}_i = \hat{\theta}_i(X)} \mathbb{P}[\theta_i \neq \hat{\theta}_i] \overset{(b)}{=} \frac{1}{4}\sum_{i=1}^k (1 - \mathrm{TV}(P_{X|\theta_i=0}, P_{X|\theta_i=1})) \\
&\overset{(c)}{\geq} \frac{k}{4} \min_{d_H(\theta,\theta')=1} (1 - \mathrm{TV}(P_\theta, P_{\theta'}))\,.
\end{aligned}$$

Here $\hat{\theta}_{dis}$ is the discretized version of $\hat{\theta}$, i.e. the closest point on the hypercube to $\hat{\theta}$ and so (a) follows from $|\theta_i - \hat{\theta}_i| \geq \frac{1}{2}\mathbf{1}_{\{|\theta_i - \hat{\theta}_i| > 1/2\}} = \frac{1}{2}\mathbf{1}_{\{\theta_i \neq \hat{\theta}_{dis,i}\}}$, (b) follows from the optimal binary hypothesis testing for $\theta_i$ given $X$, (c) follows from the convexity of TV: $\mathrm{TV}(P_{X|\theta_i=0}, P_{X|\theta_i=1}) = \mathrm{TV}(\frac{1}{2^{k-1}}\sum_{\theta:\theta_i=0} P_{X|\theta}, \frac{1}{2^{k-1}}\sum_{\theta:\theta_i=1} P_{X|\theta}) \leq \frac{1}{2^{k-1}}\sum_{\theta:\theta_i=0} \mathrm{TV}(P_{X|\theta}, P_{X|\theta\oplus e_i}) \leq \max_{d_H(\theta,\theta')=1} \mathrm{TV}(P_\theta, P_{\theta'})$. Alternatively, (c) also follows from by providing the extra information $\theta^{\setminus i}$ and allowing $\hat{\theta}_i = \hat{\theta}_i(X, \theta^{\setminus i})$ in the second line.

## 14.3 General (Weak) Converse Bounds

**Theorem 14.4** (Weak converse)**.**

1. *Any $M$-code for $P_{Y|X}$ satisfies*

$$\log M \leq \frac{\sup_X I(X;Y) + h(P_e)}{1 - P_e}$$

2. *When $M = 2^k$*

$$\log M \leq \frac{\sup_X I(X;Y)}{\log 2 - h(P_b)}$$

*Proof.* **(1)** Since $W \to X \to Y \to \hat{W}$, we have the following chain of inequalities, cf. Fano's inequality Theorem 5.4:

$$\begin{aligned}
\sup_X I(X;Y) \geq I(X;Y) &\geq I(W;\hat{W}) \\
&\geq d(\mathbb{P}[W = \hat{W}] \| \frac{1}{M}) \\
&\geq -h(\mathbb{P}[W \neq \hat{W}]) + \mathbb{P}[W = \hat{W}]\log M
\end{aligned}$$

Plugging in $P_e = \mathbb{P}[W \neq \hat{W}]$ finishes the first proof.

**(2)** Now $S^k \to X \to Y \to \hat{S}^k$. Recall from Theorem 5.1 that for iid $S^n$, $\sum I(S_i;\hat{S}_i) \le I(S^k;\hat{S}^k)$. This gives us

$$\begin{aligned}
\sup_X I(X;Y) \ge I(X;Y) &\ge \sum_{i=1}^k I(S_i,\hat{S}_i) \\
&\ge k\frac{1}{k}\sum d\left(\mathbb{P}[S_i = \hat{S}_i]\middle\|\frac{1}{2}\right) \\
&\ge kd\left(\frac{1}{k}\sum_{i=1}^k \mathbb{P}[S_i = \hat{S}_i]\middle\|\frac{1}{2}\right) \\
&= kd\left(1 - P_b\middle\|\frac{1}{2}\right) = k(\log 2 - h(P_b))
\end{aligned}$$

where the second line used Fano's inequality (Theorem 5.4) for binary random variable (or divergence data processing), and the third line used the convexity of divergence. □

## 14.4 General achievability bounds: Preview

**Remark:** Regarding differences between information theory and statistics: in statistics, there is a parametrized set of distributions on a space (determined by the model) from which we try to estimate the underlying distribution from samples. In data transmission, the challenge is to *choose* the structure on the parameter space (channel coding) such that, upon observing a sample, we can estimate the correct parameter with high probability. With this in mind, it is natural to view

$$\log\frac{P_{Y|X=x}}{P_Y}$$

as an LLR of a binary hypothesis test, where we compare the hypothesis $X = x$ to the distribution induced by our codebook: $P_Y = P_{Y|X}\circ P_X$ (so compare $c_i$ to "everything else"). To decode, we ask $M$ different questions of this form. This motivates importance of the random variable (called *information density*):

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

, where $P_Y = P_{Y|X}\circ P_X$. (Note: $I(X;Y) = \mathbb{E}[i(X;Y)]$).

Shortly, we will show a result (**Shannon's Random Coding Theorem**), that states: $\forall P_X$, $\forall\tau$, $\exists(M,\epsilon)-$*code* with

$$\epsilon \le \mathbb{P}[i(X;Y) \le \log M + \tau] + e^{-\tau}$$

Details in the next lecture.