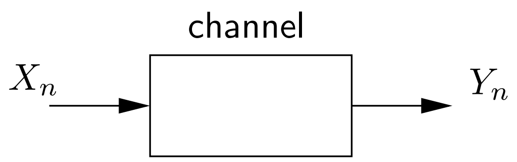
Information Theory

Lecture 9

- **Error Exponents**
  - The part on discrete channels of
    - R. Gallager, "A Simple Derivation of the Coding Theorem and Some Applications," *IEEE Trans. on Inform. Theory*, Jan. 1965
  - In addition some concepts found in
    - R. Gallager, *Information Theory and Reliable Communication*, Wiley 1968

Discrete Channels (recap)

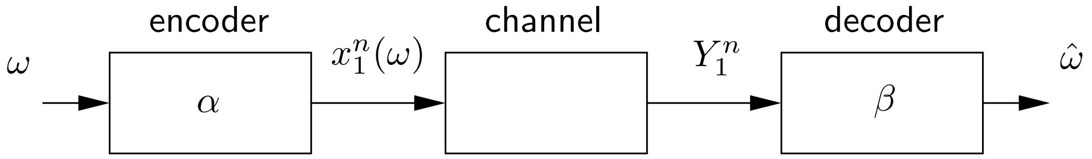


- Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets. A *discrete channel* is a random mapping from $\mathcal{X}^n$ to $\mathcal{Y}^n$ described by the conditional pmfs $p_n(y_1^n|x_1^n)$ for all $n \geq 1$, $x_1^n \in \mathcal{X}^n$ and $y_1^n \in \mathcal{Y}^n$.
- The channel is (stationary and) *memoryless* if

$$p_n(y_1^n|x_1^n) = \prod_{m=1}^{n} p(y_m|x_m), \quad n = 2, 3, \ldots$$

- A *discrete memoryless channel* (DMC) is completely described by the triple $(\mathcal{X}, p(y|x), \mathcal{Y})$

Block Channel Codes (recap)



- Define an (M, n) *block channel code* for a DMC $(\mathcal{X}, p(y|x), \mathcal{Y})$ by
  1. An *index set* $\mathcal{I}_M \triangleq \{1, \ldots, M\}$
  2. An *encoder mapping* $\alpha : \mathcal{I}_M \to \mathcal{X}^n$. The set

$$\mathcal{C} \triangleq \left\{ x_1^n : x_1^n = \alpha(i), \ \forall \, i \in \mathcal{I}_M \right\}$$

     of *codewords* is called the *codebook*.
  3. A *decoder mapping* $\beta : \mathcal{Y}^n \to \mathcal{I}_M$, as characterized by the *decoding subsets*

$$\mathcal{Y}^n(i) = \{ y_1^n \in \mathcal{Y}^n \ : \ \beta(y_1^n) = i \}, \ \ i = 1, \ldots, M$$

- The *rate* of the code is

$$R \triangleq \frac{\log M}{n} \quad \text{[bits per channel use]}$$

- A code is often represented by its *codebook only*; the decoder can often be derived from the codebook using a specific rule (joint typicality, maximum a posteriori, maximum likelihood,. . .)
- Assume, in the following, that $\omega \in \mathcal{I}_M$ is drawn according to $p(m) = \Pr(\omega = m)$

Error Probabilities (recap)

- For a given code
  - Conditional

$$P_{e,m} = \sum_{y_1^n \in (\mathcal{Y}^n(m))^c} p_n(y_1^n | x_1^n(m)) \quad (= \lambda_m \text{ in CT})$$

  - Maximal

$$P_{e,\max} = P_{e,\max}^{(n)} = \max_m P_{e,m} \quad \left(= \lambda^{(n)} \text{ in CT}\right)$$

  - Overall/average/total

$$P_e = P_e^{(n)} = \sum_{m=1}^{M} p(m) P_{e,m}$$

"Random Coding" (recap)

- Assume that the M codewords $x_1^n(m)$, $m = 1, \ldots, M$, of a codebook $\mathcal{C}$ are drawn independently according to $q_n(x_1^n)$, $x_1^n \in \mathcal{X}^n \implies P(\mathcal{C}) = q_n\big(x_1^n(1)\big) \cdots q_n\big(x_1^n(M)\big)$.
- Error probabilities over an ensemble of codes,
  - Conditional

$$\bar{P}_{e,m} = \sum_{\mathcal{C}} P(\mathcal{C}) P_{e,m}(\mathcal{C})$$

  - Overall/average/total

$$\bar{P}_e = \sum_{\mathcal{C}} P(\mathcal{C}) P_e(\mathcal{C})$$

    - *Note*: In addition to $\mathcal{C}$ a *decoder* needs to be specified

The Channel Coding Theorem (recap)

- A rate R is *achievable* if there exists a sequence of (M, n) codes, with $M = \lceil 2^{nR} \rceil$, such that $P_{e,\max}^{(n)} \to 0$ as $n \to \infty$. Capacity C is the supremum of all achievable rates.
- For a discrete memoryless channel,

$$C = \max_{p(x)} I(X;Y)$$

- Previous proof (in CT) based on typical sequences $\implies$ limited insight, e.g., into *how fast* $P_{e,\max}^{(n)} \to 0$ as $n \to \infty$ for $R < C \ldots$
  - In fact, for any $n > 0$,

$$P_{e,\max}^{(n)} < 4 \cdot 2^{-nE_r(R)}$$

  where $E_r(R)$ is the *random coding exponent*

Exponential Bounds

- A code $\mathcal{C}(n, R)$ of length n and rate R
- Assume $p(m) = M^{-1}$, a DMC and consider the average error probability

$$P_e^{(n)} = \frac{1}{M} \sum_{m=1}^{M} P_{e,m}$$

  - Bounds easily extended to $P_{e,\max}^{(n)}$
  - Non-zero lower bound may not exist for arbitrary $p(m)$
- Upper-bounds (there exists a code)

$$P_e^{(n)} \leq 2^{-nE_{\min}(R)}, \quad \text{any } n > 0$$

- Lower-bounds (for all codes)

$$P_e^{(n)} \geq 2^{-nE_{\max}(R)}, \quad \text{as } n \to \infty$$

Reliability Function, Error Exponents

- The *reliability function* of a channel,

$$E(R) = \lim_{n\to\infty} \frac{-\log P_e^*(n,R)}{n},$$

  where $P_e^*(n,R)$ is the minimum over all codes $\mathcal{C}(n,R)$
- Lower bounds to $E(R)$ yield upper bounds to $P_e^{(n)}$ (as $n \to \infty$)
  - "random coding" $E_r(R)$ and "expurgated" $E_{ex}(R)$ exponents
- Upper bounds to $E(R)$ yield lower bounds to $P_e^{(n)}$ (as $n \to \infty$)
  - "sphere-packing" $E_{sp}(R)$ and "straight-line" $E_{sl}(R)$ exponents

- With $E_{\max} = \max(E_r, E_{ex})$ and $E_{\min} = \min(E_{sp}, E_{sl})$

$$E_{\max}(R) \leq E(R) \leq E_{\min}(R)$$

- The *critical rate* R_{cr} is the smallest R in $[0, C]$ such that $E_{\max}(R) = E_{\min}(R) = E(R)$ for $R_{\text{cr}} \leq R \leq C$;
  - For $R \in [R_{\text{cr}}, C)$ the exponent $E(R) > 0$ in

$$P_e^{(n)} \approx 2^{-nE(R)} \text{ as } n \to \infty$$

    for the best possible existing code is known!

Decoding Rules

- Joint typicality ($A_\epsilon^{(n)}$ jointly typical set)

$$\mathcal{Y}^n(m) = \{y_1^n \in \mathcal{Y}^n \,:\, (x_1^n(m'), y_1^n) \in A_\epsilon^{(n)} \iff m' = m\}$$

- Maximum a posteriori (minimum error probability)

$$\mathcal{Y}^n(m) = \{y_1^n \in \mathcal{Y}^n \,:\, m = \operatorname*{argmax}_{m'} \Pr(m'|y_1^n)\}$$

- Maximum likelihood (a priori unknown / unmeaningful / uniform)

$$\mathcal{Y}^n(m) = \{y_1^n \in \mathcal{Y}^n \,:\, m = \operatorname*{argmax}_{m'} p_n(y_1^n|x_1^n(m'))\}$$

- To derive existence results it suffices to consider a specific rule

Two Codewords

- Two codewords, $\mathcal{C} = \{x_1^n(1), x_1^n(2)\}$, and any channel $p_n(y_1^n|x_1^n)$
- Assume maximum likelihood decoding,

$$\mathcal{Y}^n(1) = \{y_1^n \in \mathcal{Y}^n : p_n(y_1^n|x_1^n(1)) > p_n(y_1^n|x_1^n(2))\}$$

Hence, for any $s \in (0,1)$ it holds that

$$\begin{aligned} P_{e,1} &= \sum_{y_1^n \in \mathcal{Y}^n(1)^c} p_n(y_1^n|x_1^n(1)) \\ &\le \sum_{y_1^n \in \mathcal{Y}^n(1)^c} p_n(y_1^n|x_1^n(1))^{1-s} p_n(y_1^n|x_1^n(2))^s \\ &\le \sum_{y_1^n \in \mathcal{Y}^n} p_n(y_1^n|x_1^n(1))^{1-s} p_n(y_1^n|x_1^n(2))^s \end{aligned}$$

- An equivalent bound applies to $P_{e,2}$

- For a *memoryless* channel we get (with $\bar{m} = (m \mod 2) + 1$)

$$P_{e,m} \leq \prod_{i=1}^{n} \sum_{y_i \in \mathcal{Y}} p(y_i|x_i(m))^{1-s} p(y_i|x_i(\bar{m}))^s = \prod_{i=1}^{n} g_n(s), \quad m = 1, 2$$

- For a BSC(ϵ) with two codewords at distance d

$$P_{e,m} \leq \min_{s \in (0,1)} \prod_{i=1}^{n} g_n(s) = \left(2\sqrt{\epsilon(1-\epsilon)}\right)^d \qquad m = 1, 2$$

$\Rightarrow$ For a "best" pair of codewords ($d = n$)

$$P_{e,m} \leq \left(2\sqrt{\epsilon(1-\epsilon)}\right)^n \qquad m = 1, 2$$

$\Rightarrow$ For a "typical" pair of codewords ($d = n/2$)

$$P_{e,m} \leq \left(2\sqrt{\epsilon(1-\epsilon)}\right)^{n/2} \qquad m = 1, 2$$

Ensemble Average – Two Codewords

- Pick a probability assignment q_n on $\mathcal{X}^n$, and choose M codewords in $\mathcal{C} = \{x_1^n(1), \ldots, x_1^n(M)\}$ independently;

$$P(\mathcal{C}) = \prod_{m=1}^{M} q_n(x_1^n(m))$$

- For memoryless channels, we take q_n of the form

$$q_n(x_1^n) = \prod_{i=1}^{n} q_1(x_i)$$

- Thus, for $m = 1, 2$ (with $\bar{m} = (m \mod 2) + 1$)

$$\begin{aligned}\bar{P}_{e,m} &= \sum_{x_1^n(1)\in\mathcal{X}^n} \sum_{x_1^n(2)\in\mathcal{X}^n} q_n(x_1^n(1)) q_n(x_1^n(2)) P_{e,m} \\ &\le \sum_{y_1^n\in\mathcal{Y}^n} \left[\sum_{x_1^n(m)\in\mathcal{X}^n} q_n(x_1^n(m)) p_n(y_1^n|x_1^n(m))^{1-s} \right] \\ &\quad \times \left[\sum_{x_1^n(\bar{m})\in\mathcal{X}^n} q_n(x_1^n(\bar{m})) p_n(y_1^n|x_1^n(\bar{m}))^{s} \right]\end{aligned}$$

Minimum over $s \in (0, 1)$ at $s = 1/2 \implies$

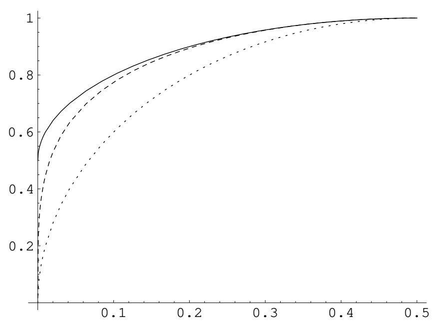
$$\bar{P}_{e,m} \le \sum_{y_1^n\in\mathcal{Y}^n} \left[\sum_{x_1^n\in\mathcal{X}^n} q_n(x_1^n) \sqrt{p_n(y_1^n|x_1^n)} \right]^2$$

- For a memoryless channel

$$\bar{P}_{e,m} \le \left\{ \sum_{y\in\mathcal{Y}^n} \left(\sum_{x\in\mathcal{X}} q_1(x) \sqrt{p_1(y|x)} \right)^2 \right\}^n \qquad m = 1, 2$$

- In particular, for a BSC(ϵ) with $q_1(x) = 1/2$

$$\bar{P}_{e,m} \le \left\{ \frac{1}{2} \left(\sqrt{\epsilon} + \sqrt{1-\epsilon} \right)^2 \right\}^n \qquad m = 1, 2$$



  - Solid: $\frac{1}{2}\left(\sqrt{\epsilon} + \sqrt{1-\epsilon}\right)^2$ (random)
  - Dashed: $\left(2\sqrt{\epsilon(1-\epsilon)}\right)^{1/2}$ (typical)
  - Dotted: $2\sqrt{\epsilon(1-\epsilon)}$ (best)

Alternative Derivation — Still Two Codewords

- Examine the ensemble average directly

$$\bar{P}_{e,1} = \sum_{x_1^n(1)\in\mathcal{X}^n} q_n(x_1^n(1)) \sum_{y_1^n\in\mathcal{Y}^n} p_n(y_1^n|x_1^n(1)) \Pr(y_1^n \in \mathcal{Y}^n(1)^c)$$

- Since the codewords are randomly chosen

$$\begin{aligned}\Pr(y_1^n \in \mathcal{Y}^n(1)^c) &= \sum_{x_1^n(2):\, p_n(y_1^n|x_1^n(1))\le p_n(y_1^n|x_1^n(2))} q_n(x_1^n(2)) \\ &\le \sum_{x_1^n(2)\in\mathcal{X}^n} q_n(x_1^n(2)) \left[\frac{p_n(y_1^n|x_1^n(2))}{p_n(y_1^n|x_1^n(1))}\right]^s\end{aligned}$$

- Substituting this into the first equation yields the result
- This method generalizes more easily!

Bound on $\bar{P}_{e,m}$ – Many Codewords

- As before,

$$\bar{P}_{e,m} = \sum_{x_1^n(m)\in\mathcal{X}^n} q_n(x_1^n(m)) \sum_{y_1^n\in\mathcal{Y}^n} p_n(y_1^n|x_1^n(m)) \Pr(y_1^n \in \mathcal{Y}^n(m)^c)$$

- For $M \ge 2$ codewords, any $\rho \in [0,1]$ and $s > 0$

$$\begin{aligned}\Pr(y_1^n \in \mathcal{Y}^n(m)^c) &\le \Pr\Big(\bigcup_{m'\ne m} \{y_1^n \in \mathcal{Y}^n(m')\}\Big) \\ &\le \left[\sum_{m'\ne m} \Pr(y_1^n \in \mathcal{Y}^n(m'))\right]^\rho \\ &\le \left[(M-1)\sum_{x_1^n\in\mathcal{X}^n} q_n(x_1^n)\frac{p_n(y_1^n|x_1^n)^s}{p_n(y_1^n|x_1^n(m))^s}\right]^\rho\end{aligned}$$

- Substitute back into the first equation

$$\bar{P}_{e,m} \leq (M-1)^\rho \sum_{y_1^n \in \mathcal{Y}^n} \left[\sum_{x_1^n \in \mathcal{X}^n} q_n(x_1^n) p_n(y_1^n | x_1^n)^s \right]^\rho$$
$$\times \left[\sum_{x_1^n(m) \in \mathcal{X}^n} q_n(x_1^n(m)) p_n(y_1^n | x_1^n(m))^{1-s\rho} \right]$$

Minimize over $s > 0$ (see HW prob.) $\Longrightarrow$

$$\bar{P}_{e,m} \leq (M-1)^\rho \sum_{y_1^n \in \mathcal{Y}^n} \left[\sum_{x_1^n \in \mathcal{X}^n} q_n(x_1^n) p_n(y_1^n | x_1^n)^{1/(1+\rho)} \right]^{1+\rho}$$

- For memoryless channels

$$\bar{P}_{e,m} \leq (M-1)^\rho \left(\sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} q_1(x) p_1(y|x)^{1/(1+\rho)} \right]^{1+\rho} \right)^n$$

- Define

$$E_0(\rho, q_1) \triangleq -\log \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} q_1(x) p_1(y|x)^{1/(1+\rho)} \right)^{1+\rho}$$

- Using $M - 1 < 2^{nR}$, we get

$$\bar{P}_{e,m} \leq 2^{-n[E_0(\rho, q_1) - \rho R]}$$

Random Coding Exponent

- To minimize the upper-bound on $\bar{P}_{e,m}$, define the *random coding (Gallager) exponent*

$$E_r(R) = \max_{\rho, q_1} \left(E_0(\rho, q_1) - \rho R \right)$$

- Thus, for the ensemble average error probabilities

$$\bar{P}_{e,m} \le 2^{-nE_r(R)} \implies \bar{P}_e^{(n)} \le 2^{-nE_r(R)}$$

- Since at least one code in the ensemble has error probability $\bar{P}_e^{(n)}$ (or less), there exists a "good" code satisfying

$$P_e^{(n)} \le 2^{-nE_r(R)}$$

- But, this says nothing about $P_{e,m}$!

- To bound $P_{e,m}$ take a code with $2M$ $\left(= 2\lceil 2^{nR} \rceil\right)$ codewords, which satisfies the inequality for equiprobable messages

$$P_e^{(n)} = \frac{1}{2M} \sum_{m=1}^{2M} P_{e,m} \le 2^{-nE_r(\frac{\log 2M}{n})}$$

- Throw away the worst M codewords including all that satisfy

$$P_{e,m} \ge 2 \cdot 2^{-nE_r(\frac{1+\log M}{n})}$$

- Since the decoding subsets didn't get smaller, the remaining M codewords satisfy (since $\rho \in [0,1]$)

$$P_{e,m} \le 2 \cdot 2^{-nE_r(R+\frac{1}{n})} \le 2 \cdot 2^{-n\left[E_r(R) - \frac{1}{n}\right]}$$

$\Rightarrow$ There exists at least one code such that for any $n > 0$

$$\forall m \colon P_{e,m} \le 4 \cdot 2^{-nE_r(R)} \implies P_{e,\max} \le 4 \cdot 2^{-nE_r(R)}$$

The Coding Theorem Based on $E_r(R)$

- *Theorem*: For any DMC $(\mathcal{X}, p(y|x), \mathcal{Y})$ the random coding exponent $E_r(R)$ is a convex, decreasing and positive function of R for $0 \le R < C$ where

$$C = \max_{p(x)} I(X;Y)$$

where

$$I(X;Y) = \sum_{x,y} p(y|x)p(x) \log \frac{p(y|x)}{p(y)}$$

with $p(y) = \sum_x p(y|x)p(x)$, and where the maximum is over all possible pmf's on $\mathcal{X}$.

Examples of $E_r(R)$

- Binary symmetric channel with crossover probability ϵ

$$E_r(R) = \begin{cases} 1 - 2\log\left(\sqrt{\epsilon} + \sqrt{1-\epsilon}\right) - R & R \le R_{\mathsf{cr}} \\ d(h^{-1}(1-R)\|\epsilon) & R_{\mathsf{cr}} \le R \le C \\ 0 & C \le R \end{cases}$$

where

$R_{\mathsf{cr}} = 1 - h\left(\frac{\sqrt{\epsilon}}{\sqrt{\epsilon}+\sqrt{1-\epsilon}}\right)$ *(critical rate)*

$C = 1 - h(\epsilon)$ *(capacity)*

$h(\epsilon) = -\epsilon\log\epsilon - (1-\epsilon)\log(1-\epsilon)$ *(binary entropy)*

$d(\delta\|\epsilon) = \delta\log\frac{\delta}{\epsilon} + (1-\delta)\log\frac{1-\delta}{1-\epsilon}$ *(binary relative entropy)*

- Very noisy channels

$$p_1(y|x) = p(y)(1 + \epsilon_{x,y}), \qquad |\epsilon_{x,y}| \ll 1$$

Using second-order approximation in $\epsilon_{x,y}$

$$E_r(R) \approx \begin{cases} \frac{C}{2} - R & R < \frac{C}{4} \\ \left(\sqrt{C} - \sqrt{R}\right)^2 & \frac{C}{4} \le R \le C \\ 0 & C \le R \end{cases}$$

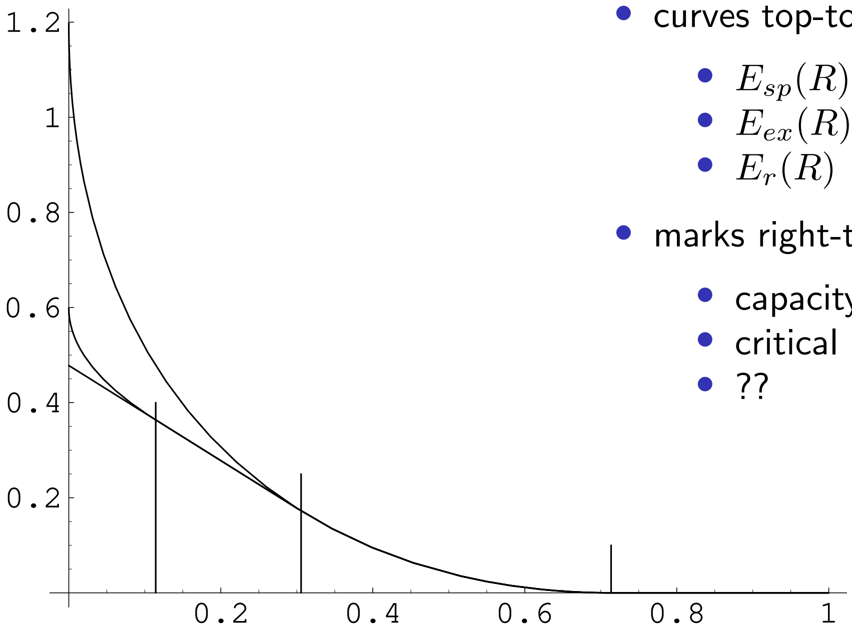
Some Comments on Other Error Exponents

- Expurgated exponent E_{ex}
  - strengthens E_r for small rates
  - generally agrees with E_r on part of its linear portion ($R < R_{\text{cr}}$)
  - can be infinite!
- Sphere-packing exponent E_{sp}
  - agrees with E_r on its non-linear part ($R > R_{\text{cr}}$)
  - can also be infinite!
- Straight-line exponent E_{sl}
  - line through $(0, E_{ex}(0))$, tangent to E_{sp} (when $E_{ex}(0) < \infty$)

Large Deviations Theory...

- Tight connections to large deviations theory, Chernoff bounds,...

A Typical Scenario – BSC(0.05)



- curves top-to-bottom
  - $E_{sp}(R)$
  - $E_{ex}(R)$
  - $E_r(R)$
- marks right-to-left
  - capacity
  - critical rate
  - ??

Rates Above Capacity

Theorem (Wolfowitz (1957)) For an arbitrary DMC of capacity C bits and any length n, rate $R > C$ code

$$P_e^{(n)} \geq 1 - \frac{4A}{n(R-C)^2} - 2^{-\frac{n(R-C)}{2}}$$

where A is a constant depending on the channel but not on n or R.

- Check, e.g., for $R = C + \delta/\sqrt{n}$ with any $\delta > \sqrt{8A} + 2 \implies P_e^{(n)} > 0, \ \forall n > 0$