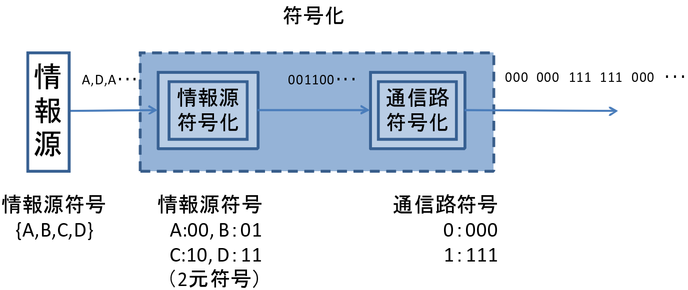
情報理論
第 11 回 通信路符号化の基礎と情報源符号化

堀田 政二
工学部 情報工学科

符号化 (encoding) のモデル (再掲)



- **情報源符号化**
  - **情報源からの信号を 0 と 1 からなる符号記号に変換すること**
  - **符号記号の組み合わせでできたもの: 符号語**
  - **符号語の記号: 情報源符号**
- **通信路符号化**
  - **通信路における誤りに対処できるような別の符号語に変更すること**

(2)

通信路符号化

高信頼化を目的として誤り検出や誤り訂正を可能にする符号化

- 情報源符号化
  - 情報源の記号に対応する符号をエントロピー限界まで短縮
  - この場合，符号には冗長度 (無駄，余裕) がなく，一つでもビット誤りが生じると，受信側で復号できない
- 通信路符号化
  - 送信側で情報を担うビットに加えて，情報を持たない余分なビット (冗長ビット) を符号に付加
  - いかに少ない冗長度で誤りを検出・訂正できるかが通信路符号化の課題

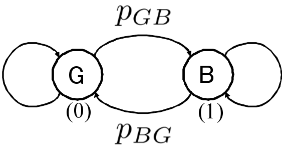
誤りの種類

誤りの原因となる要因を総称して雑音 (noise) **と呼ぶ**

- **デジタル通信路の品質はビット誤り率** (BER, bit error rate) **で評価**
  - **品質の低い伝送路 (例:無線)**： BER$\sim 10^{-2}$
  - **品質の高い伝送路 (光ファイバ)**: BER$\sim 10^{-9}$
- BER **は非常に長いビット系列を測定した結果の平均値**
  - **ランダム誤り** (random error): **伝送ビット列中で不規則 (バラバラ) に発生する誤りで，数は多いが連続はしない**
  - **バースト誤り** (burst error): **長時間に見て誤りの数は同じでも，一部に集中して連続的に発生する誤り**

バースト誤りのモデル

誤り源が単純マルコフモデルで表現できる場合を考える



- G (Good) と B (Bad) の状態があり，B の状態で必ず誤りが発生すると仮定する．二つの状態間は確率的に遷移する
  - 例として $p_{GB} = 0.001$, $p_{BG} = 0.2$ とすると，定常確率は $P_g \sim 0.995$, $p_B = 0.005$
  - この通信路のビット誤り率は 0.005
  - ただし，連続するエラービットの平均数は

$$\sum_{n=1}^{\infty} n(1 - p_{BG})^{n-1} p_{BG} = 1/p_{BG} = 5$$

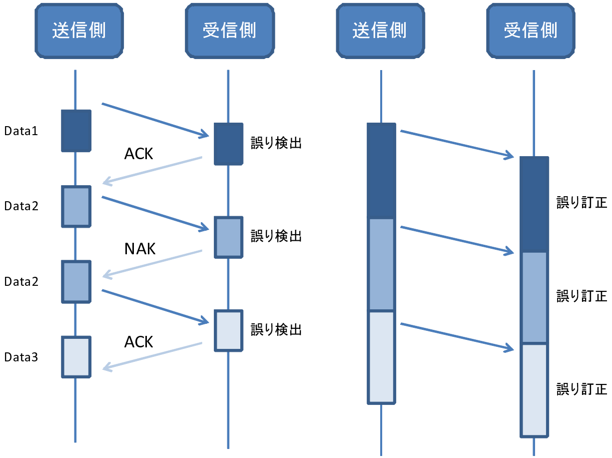
- 誤り率 0.005 の二元対称通信路とすれば，連続するエラービットの平均数は $1/0.995 \sim 1$

誤り制御

- **再送要求方式** (ARQ, automatic repeat request)
  - **送信データに誤りが無ければ** ACK (**肯定応答**, acknowledge) **を送信側に返送**
  - **誤りがある場合は** NAK (**否定応答**, negative acknowledge) **を返送し，そのデータの再送を要求**
  - **誤りがなくなるまで再送するので非常に高い信頼性．ただし，処理に時間がかかるので，伝送効率** (**スループット**) **は低下**
  - **誤りが許されないテキストやデータ通信などに適用**
- **誤り訂正方式** (FEC, forward error correction)
  - **大きな冗長度を付加すれば，受信側で誤りの検出だけでなく，正しい符号に訂正可能**
  - **誤り訂正は受信側のみの処理で実行するので再送は不要．片方向のみの伝送なので誤訂正する可能性がある**
  - **誤りが少々許される音声，映像など，信頼性を若干犠牲にしても，リアルタイムが要求される場合に適用**

(6)

誤り制御方式 (左: ARQ，右: FEC)



誤り検出・訂正の原理

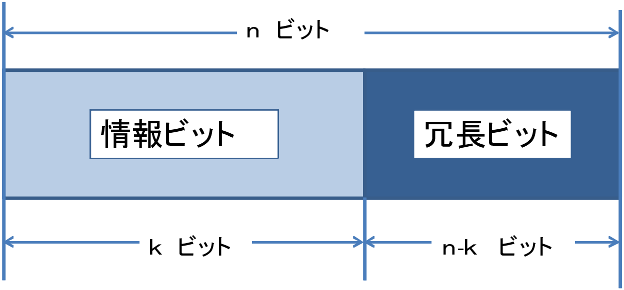
- 符号語と非符号語
  - 符号語: 情報源の記号に対応する符号．送信側から送られる符号はすべて符号語
  - 非符号語: 符号語以外の符号

- 誤りを検出できる条件
  - 誤りがあれば受信符号は符号語に一致せず，必ず非符号語になること
- 誤りを訂正するためには
  - 誤りビットの位置を特定する必要．検出よりもさらに大きな冗長度

誤りを検出・訂正するには，非符号語が必要．非符号語は記号に対応しておらず，情報伝送の観点からは冗長

- 誤り検出のための符号: 誤り検出符号 (error detecting code)
- 訂正のための符号: 誤り訂正符号 (ECC, error correcting code)

通信路符号の構成，(n, k) 符号



情報源記号の種類: 2^k **個，記号は** k **ビット長の符号．送信に用いる符号はこれよりも長い全長が** n **ビット** $(n > k)$ **の符号**

- **記号に割り当てられた符号語は** 2^k **個．それ以外の** $2^n - 2^k$ **個の符号は非符号語**
- **全ビット長に含まれる情報ビットの割合 (符号化率) は** $\eta = k/n$ **と表される．そのため，冗長度は** $\rho = 1 - \eta$

符号化の基本方針

平均符号長を短くする

- 発生する確率の高い記号に短い符号を割り当てる
- 発生する確率の低い記号に長い符号を割り当てる

N 種類の記号があり，各記号の発生確率と符号長をそれぞれ p_i，L_i bit $(i = 1, \ldots, N)$ とする

- 情報源のエントロピー: $H = -\sum_{i=1}^{N} p_i \log_2 p_i$ bit/記号
- 平均符号長: $L = \sum_{i=1}^{N} p_i L_i$ bit/記号

符号化の効率 e

$$e = H/L \ (0 \leq e \leq 1)$$

平均符号長 L がエントロピー H にどれだけ近いかを示す指標．
$L = H$ の時，効率が最大の 1 となる

情報源符号化 (圧縮) 法の分類例

- 非可逆符号化※1
  - 画像符号化
  - 音声符号化
- 可逆符号化※2
  - 静的符号化
    - シャノン・ファノ符号
    - ハフマン符号
    - 算術符号
  - 動的符号化
    - ユニバーサル符号化
      - LZ77符号（スライド辞書）
      - LZ78符号（動的辞書）
      - LZW符号
    - 動的ハフマン符号

※1：歪あり
※2：歪なし、無損失

シャノン・ファノの符号化法

2 元符号において生成される情報源を $\{A_1, A_2, ..., A_m\}$，記号 A_i の生起確率を p_i とする

- Step1: 記号 A_i を生起確率 p_i の大きい記号から順に並べる
- Step2: 生起確率 p_i の和がほぼ等しくなるように記号を 2 グループに分割
- Step3: 各グループを更に二つに分割．このとき分割できなければ終了
- Step4: 分割をツリー表現し，たとえばルートの左の枝に 0，右の枝に 1 を割り当て

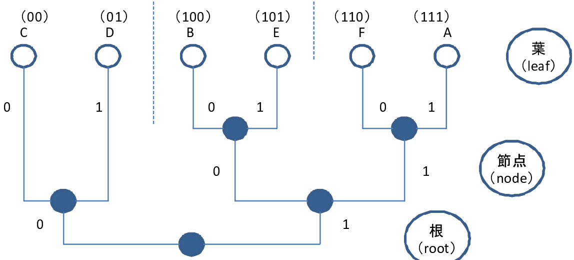
ルートから葉へいたるまでの 0 と 1 の並びが求める符号

シャノン・ファノの符号化法の例

下記の情報源の記号を符号化することを考える

$$\begin{pmatrix} A & B & C & D & E & F \\ 0.09 & 0.14 & 0.40 & 0.15 & 0.12 & 0.10 \end{pmatrix}$$

- **順番に並べる**: $C > D > B > E > F > A$
- **二分割**: CD **と** BEFA
- **更に二分割**: BE **と** FA



ハフマンの符号化法

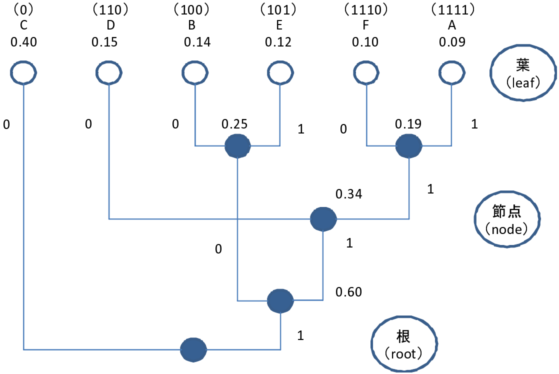
2 元符号において生成される情報源を $\{A_1, A_2, ..., A_m\}$，記号 A_i の生起確率を p_i とする

- Step1: 記号 A_i を生起確率 p_i の大きい記号から順に並べる
- Step2: 生起確率 p_i の最も小さい二つの記号を統合して 1 つのグループとする．この生起確率は二つの生起確率の和とする
- Step3: 全体が 1 つになるまで上の Step1，Step2 を繰り返す
- Step4: 全体をツリー表現し，たとえばルートの左の枝に 0，右の枝に 1 を割り当て

ルートから葉へいたるまでの 0 と 1 の並びが求める符号

ハフマン符号化法の例

$$\begin{pmatrix} A & B & C & D & E & F \\ 0.09 & 0.14 & 0.40 & 0.15 & 0.12 & 0.10 \end{pmatrix}$$



この例での符号化効率の比較

エントロピー H**，平均符号長** L**，符号化効率** e **とする**:

- **シャン・ファノの符号化**
  - $H = 2.35$ bit/**記号**
  - $L = 2.45$ bit/**記号**
  - $e = H/L = 0.96$
- **ハフマンの符号化**
  - $H = 2.35$ bit/**記号**
  - $L = 2.39$ bit/**記号**
  - $e = H/L = 0.98$

一般的にもハフマン符号化が最も効率が優れており，最適符号 (optimum code) **またはコンパクト符号** (compact code) **と呼ばれる**

シャノンの第 1 基本定理 (Shannon's first fundamental theorem)

シャノンの第 1 基本定理

情報源の１記号あたりのエントロピーを H とし，これを r 元符号を用いて符号化する場合，平均符号長 L は

$$\frac{H}{\log_2 r}$$

にいくらでも近づけることが出来るが，これより短くはできない

通信路容量 C を用いて，C/H にいくらでも近い速度で情報を送る符号化法が存在する，という良い方も可能．この定理は以下の別名もある:

- 情報源符号化定理
- 雑音のない場合の符号化定理

問題

- 【11.1】次の情報源 S をハフマン符号化により符号化せよ:

$$S = \begin{pmatrix} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \\ 0.28 & 0.22 & 0.15 & 0.14 & 0.12 & 0.09 \end{pmatrix}$$

- 【11.2】S のエントロピー，ハフマン符号化後の平均符号長 L，符号化効率 e を求めよ．
- 【11.3】上記の例題において，$H(S) \leq L < H(S) + 1$ が満たされていることを確認せよ．