2010年11月6～7日於：中京大学

経営情報学会2010年秋季全国研究発表大会

「誤り訂正符号を用いた直積ファイルのディスク配置」

平澤　茂一　（サイバー大学 IT総合学部，早稲田大学 理工学術院 総合研究所）
斎藤　友彦　（青山学院大学　理工学部）
稲積　宏誠　（青山学院大学　社会情報学部）

サイバー大学
Cyber University

目次

サイバー大学 Cyber University

1. はじめに

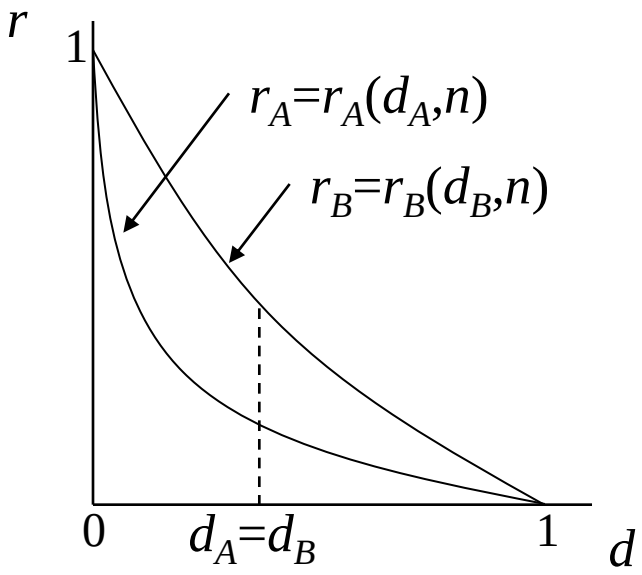
サイバー大学 Cyber University

1. 質問回答システム(QAシステム) J. Pearlと
   A. Crolotte(1979[10])
   – 情報縮約論
   –「フレキシブル(Flexible)」,
   –「エラスティック(Elastic)」
2. システム(トレードオフ)評価モデル[3][6][8]
3. 誤り訂正符号[11]

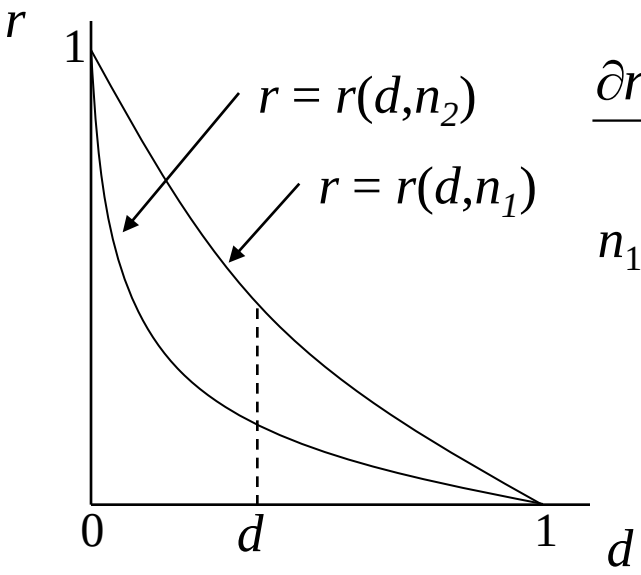
2. トレードオフ評価モデル

- 情報縮約論：レート歪関数　$R = R(D)$
- システム評価モデル：　$r = r(d)$ **(1)**
- $R\,(r)$　：レート
- $D\,(d)$：歪
- n：システムの規模

「フレキシブル」



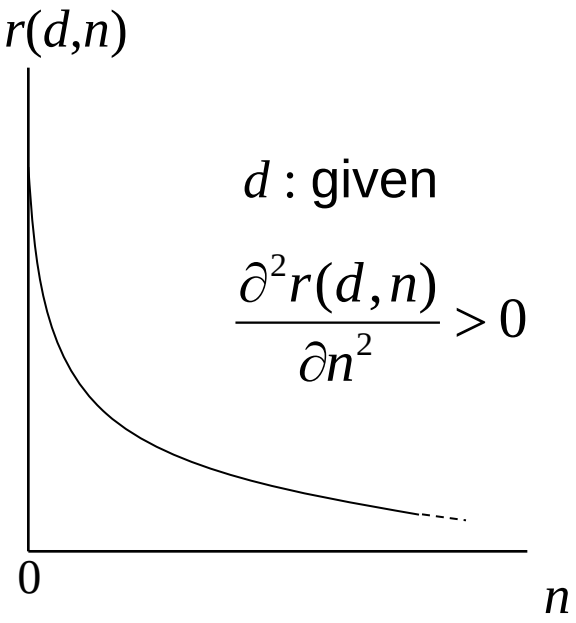
「エラスティック」



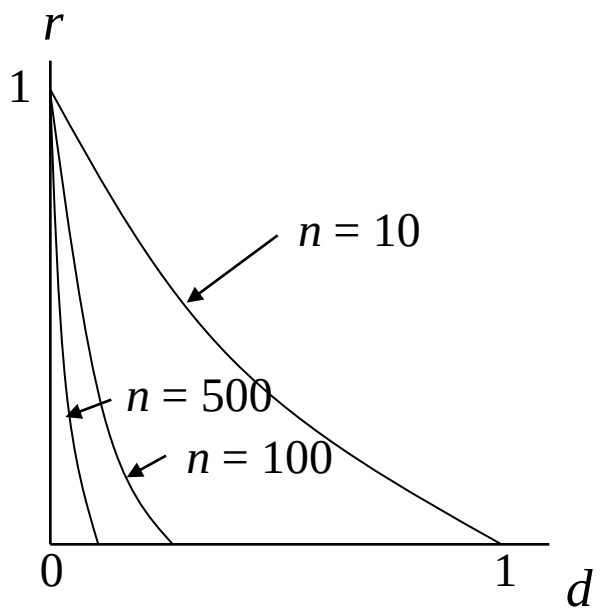
$$\frac{\partial r(d,n)}{\partial d} < 0$$

$n_1 < n_2$

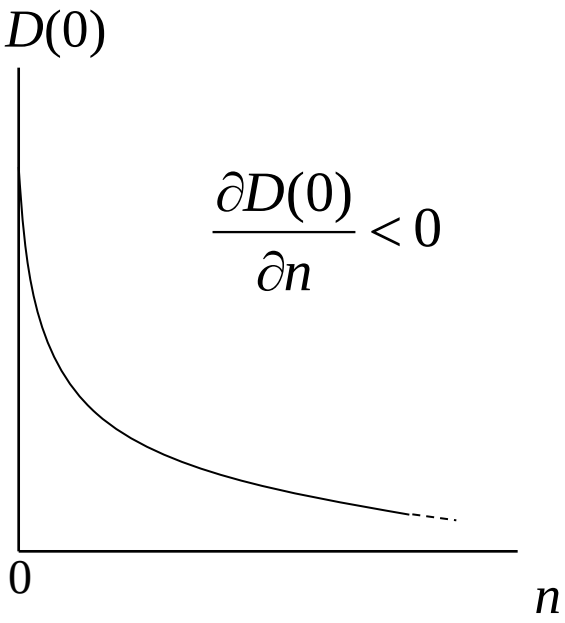
「効果的エラスティック」



「トリビアルエラスティック」



「マージナルエラスティック」

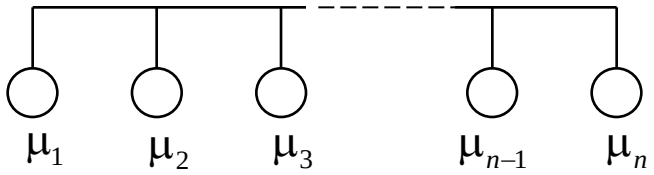


[例] コンピュータネットワークにおけるデータベースの分散配置問題[7][8]

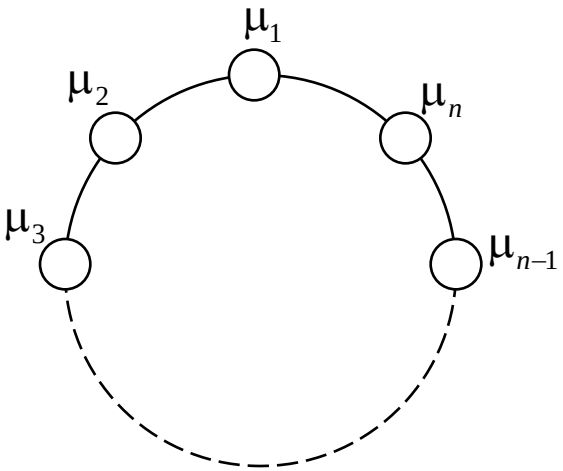
- ノード数：n
- データベースの配置コスト：R　・・・（ファイルの重複配置の冗長度）
- ファイルアクセスコスト：D　・・・（アクセス要求を出したノードからデータベースを持つノードへの最短リンク数に比例するコスト）

ネットワークトポロジー:

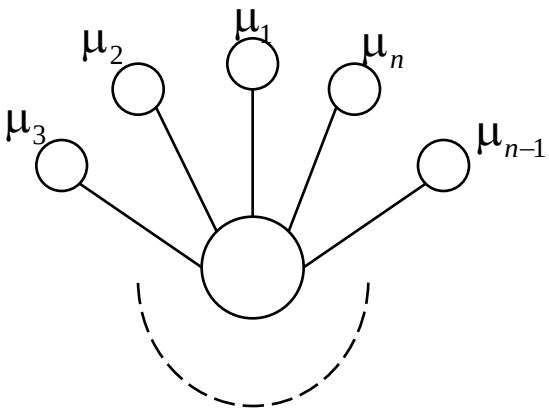
(a) バス型

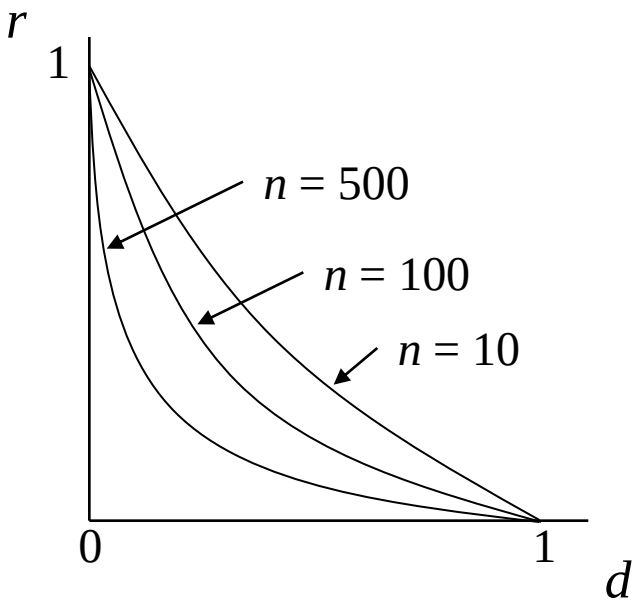


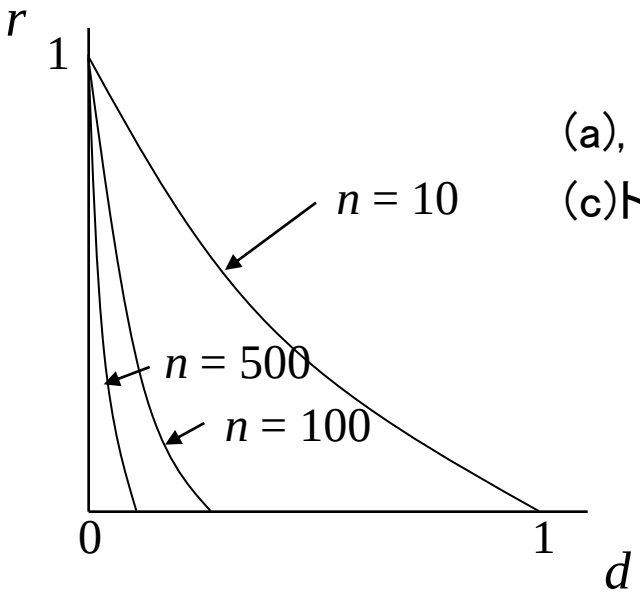
(b) リング型



(c) スター型





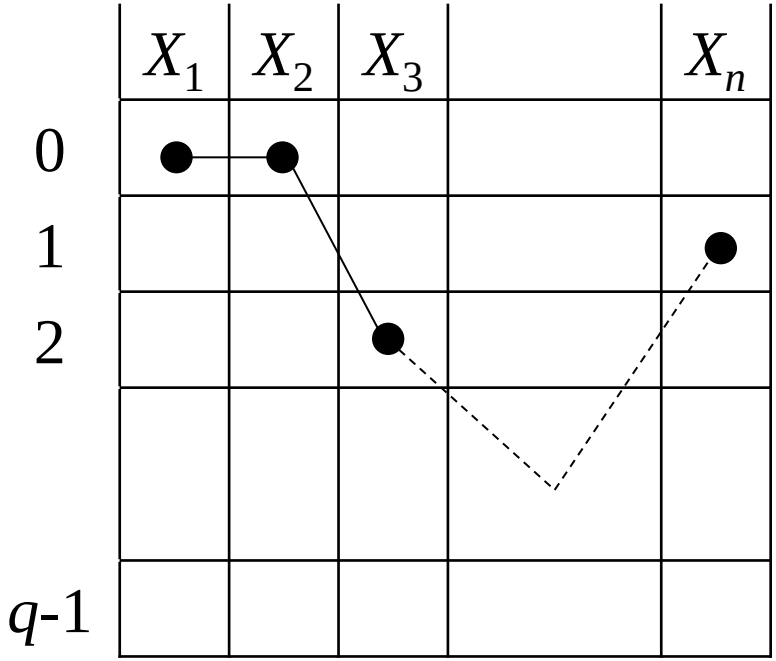


(a), (b)エラスティック,
(c)トリビアルエラスティック

3. 直積ファイルのディスク配置問題

3.1 （q元）直積ファイル

- 属性：$X_1, X_2, \ldots, X_n$
- 領域：$Z_1, Z_2, \ldots, Z_n$,　$Z_i = \{0, 1, 2, \ldots, q\text{-}1\}$



path：bucket

q^n bucket

$(X_1=0, X_2=0, X_3=2, \ldots, X_n=1)$

3.2 部分照合アクセス要求

- $\boldsymbol{Q} = (X_1 = z_1, X_2 = z_2, \ldots, X_n = z_n)$ **(2)**
- $z_i \in \{0, 1, \ldots, q\text{-}1, *\}$ **(3)**

  ＊：don't care（不定）：　＊$=\{0, 1, \ldots, q\text{-}1\}$

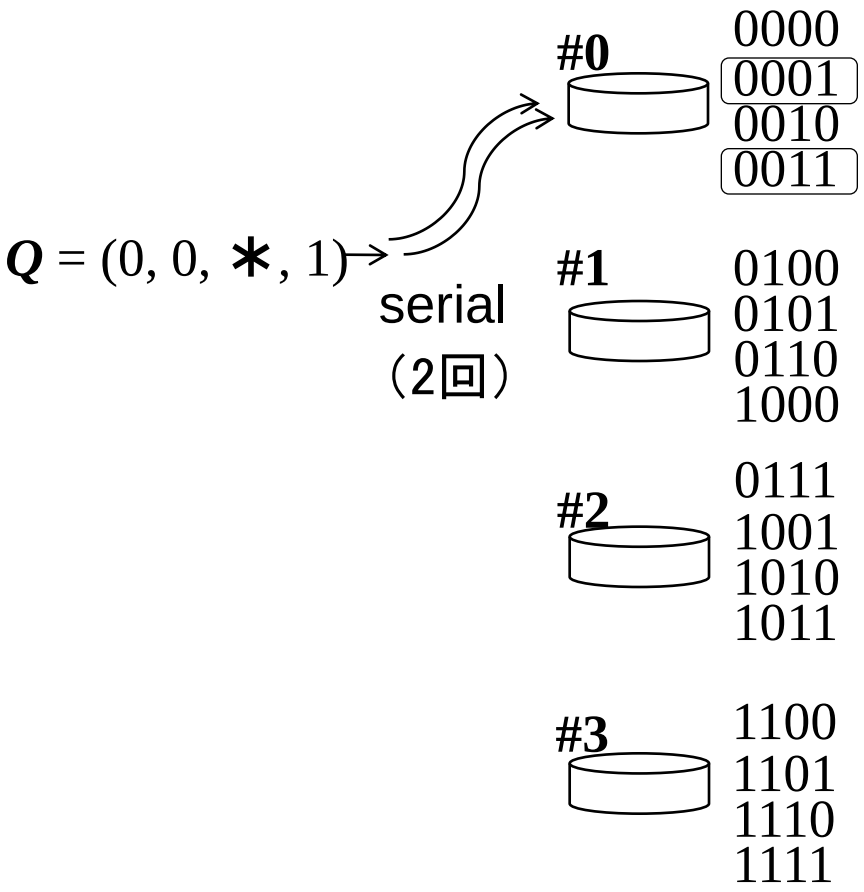
[例1] 部分照合アクセス要求(q=2, n=4, G=4)

X_1(Sex)	X_2(Income ($/year))	X_3(Married)	X_4(Age)
0 (Male)	0 (100K ≤)	0 (No)	0 (< 20)
1 (Female)	1 (< 100K)	1 (Yes)	1 (20 ≤)

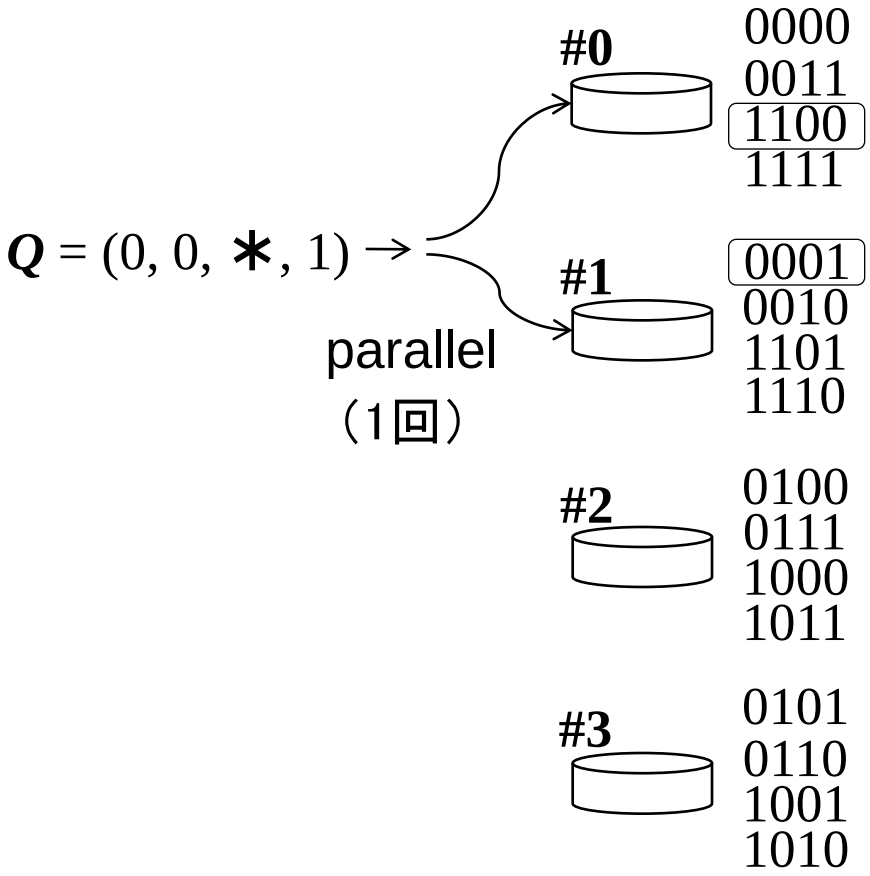
$\boldsymbol{Q} = (0, 0, *, 1)$ → 既婚・未婚によらない

→ $\boldsymbol{Q} = (0, 0, 0, 1)$ and $\boldsymbol{Q} = (0, 0, 1, 1)$

(a) バイナリ配置



(b) 分散配置



3.3 ファイルの分散配置

[例2] 直積ファイルのディスクへの分散配置 (q=2, n=6, G=8, q^n=64)
(6, 3, 3)符号の場合

disk #	bucket #							
0	000000	100110	010101	110011	001111	101001	011010	111100
1	100000	000110	110101	010011	101111	001001	111010	011100
2	010000	110110	000101	100011	011111	111001	001010	101100
3	001000	101110	011101	111011	000111	100001	010010	110100
4	000100	100010	010001	110111	001011	101101	011110	111000
5	000010	100100	010111	110001	001101	101011	011000	111110
6	000001	100111	010100	110010	001110	101000	011011	111101
7	000011	100101	010110	110000	001100	101010	011001	111111

$\boldsymbol{Q}$ = (0,∗,1,∗,0,0)の場合

→ $\boldsymbol{Q}$ = (001000), (001100), (011000), (011100)

- 直積ファイル（バケット）数：q^n
- ディスク数：G
  → (n, k, d)符号C：符号長n, 情報記号数k, 最小距離dの符号
  → 標準配列
- #of $* = w$
- ディスクアクセス回数：J

[補題1] [1]

$0 \leq w < d \Rightarrow q^w$個のバケットを $J = 1$でアクセス可能

4. ファイル配置問題の評価

4.1 ファイル配置問題の定式化

- ディスクアクセス回数: J
- 評価損失: ρ
- $\boldsymbol{Q}$で指定されたバケットの集合: $S(\boldsymbol{Q})$
- 符号Cを用いてJ=1でアクセス可能なバケットの集合: $S(C)$

$$\rho = \begin{cases} 0, & J = 1\,(S(\boldsymbol{Q}) \subseteq S(C)), \\ 1, & J \geq 2\,(S(\boldsymbol{Q}) \supset S(C)), \end{cases} \tag{4}$$

- アクセス性能: ν

$$\nu = 0 \times \Pr(J = 1) + 1 \times \Pr(J \geq 2) \tag{5}$$

- コスト: ディスク数 G

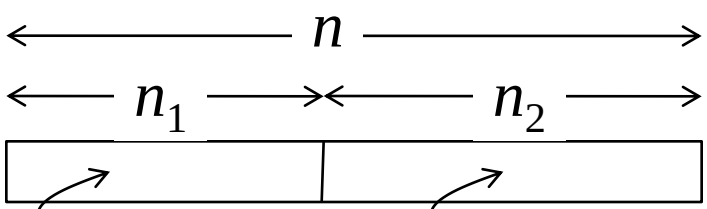
$$G = q^{n-k} \tag{6}$$

$$g = G / G_{\max} = q^{-k} \tag{7}$$

4.2 均一の場合（符号C）

- ＊の生起確率：$\Pr(*) = p$
- $\Pr(J \geq 2) = \Pr(w \geq d)$ ， [補題1]より **(8)**

4.3 不均一の場合（UEP符号C_u）



$(n=n_1+n_2)$

	C_{u1}	C_{u2}	
最小距離	d_1	d_2	$(d_1 > d_2)$
＊の数	w_1	w_2	
Pr(＊)	p_1	p_2	

[補題2] [9]

(1) $w_1=0,\ w_2<d_2 \ \Rightarrow\ J=1$

(2) $w_1 \geq 1,\ w_1+w_2 < d_1 \Rightarrow J=1$

[定理1]

$$\Pr(J \geq 2) = \Pr(w_1 = 0)\Pr(w_2 \geq d_2) + \sum_{s=1}^{n_1} \Pr(w_1 = s)\Pr(w_2 \geq d_1 - s)$$

4.4 評価関数の計算法

- アクセス性能: $\nu = \nu(n, d)$

  ↕ ―――――― $\delta = d/n$ 対 $R = k/n$

- コスト: $g = g(n, k)$

（1）LP上界式[11]: $M \leq f(n_1, n_2, d_1, d_2)$

（2）BCH符号・RS符号など具体的符号パラメータ

（3）Gilbert下界式: $d/n \geq H^{-1}(1\text{-}R) \quad (n \to \infty)$

4.5 数値計算結果

- Pr(＊):二項分布

(1)均一誤り訂正符号C

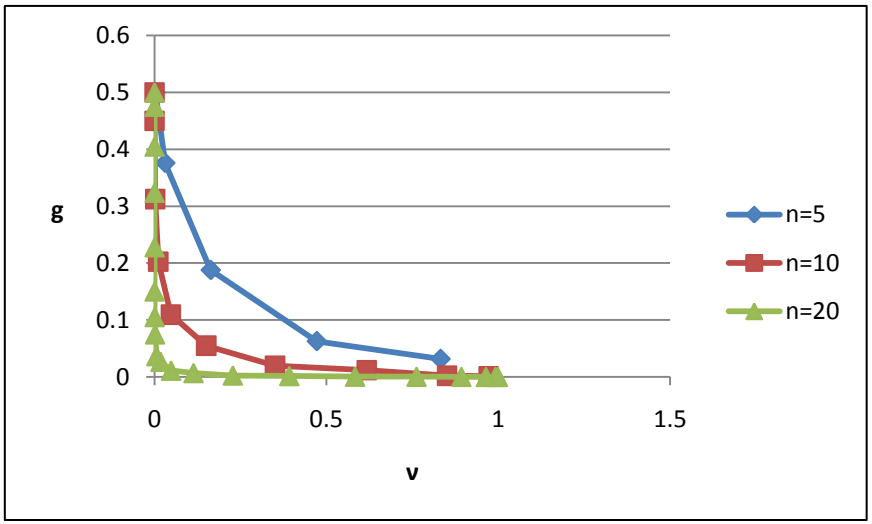


図2 LP上界式 $p = 0.3$

(エラスティック)

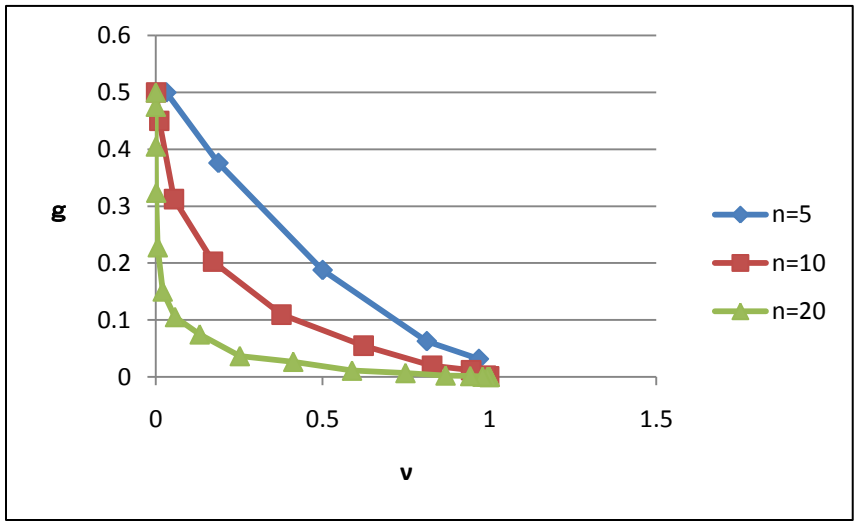


図3 LP上界式 $p = 0.5$

(エラスティック)

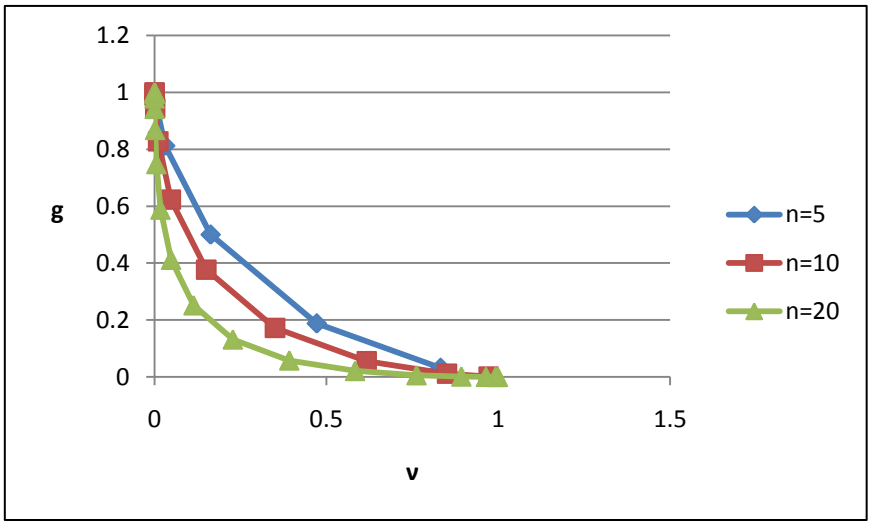


図4　Gilbert下界式　$p = 0.3$

（エラスティック）

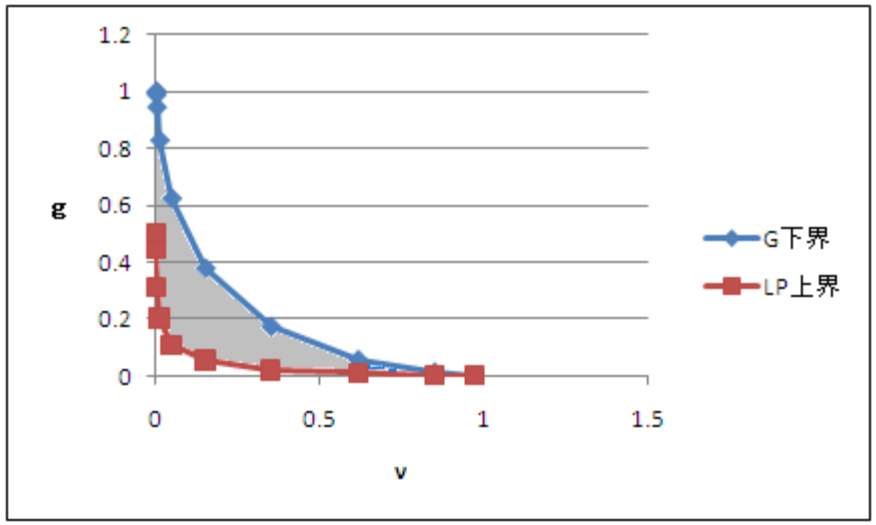


図5　LP上界式とGilbert下界式

$p = 0.3, n=10$

（符号の存在範囲）

(2)不均一誤り訂正符号C_u

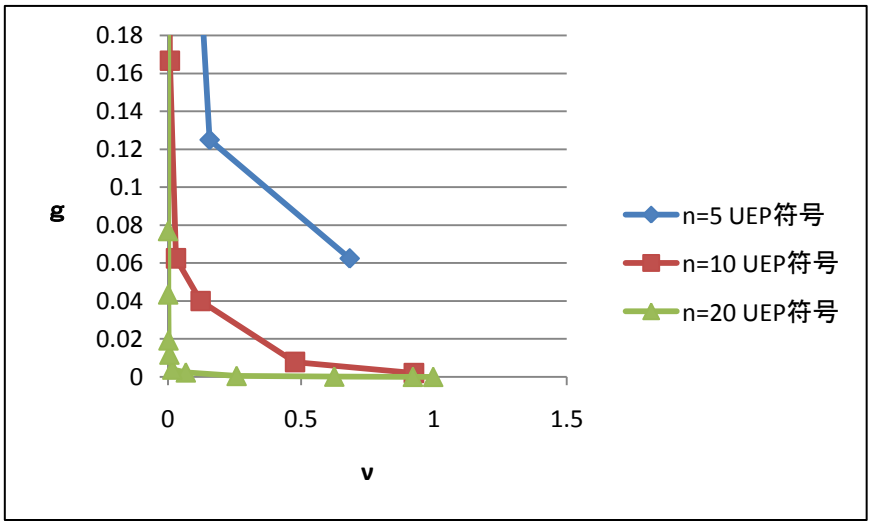


図6　LP上界式(UEP符号)
p_1 = 0.5, p_2 = 0.25
(エラスティック)

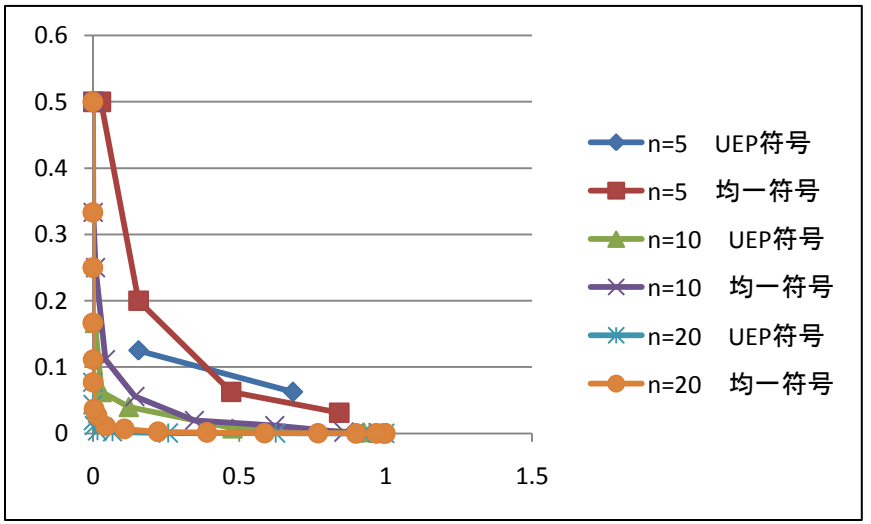


図7　LP上界式(UEP符号，均一符号)
p_1 = 0.5, p_2 = 0.25
(フレキシブル)

効果的エラスティック

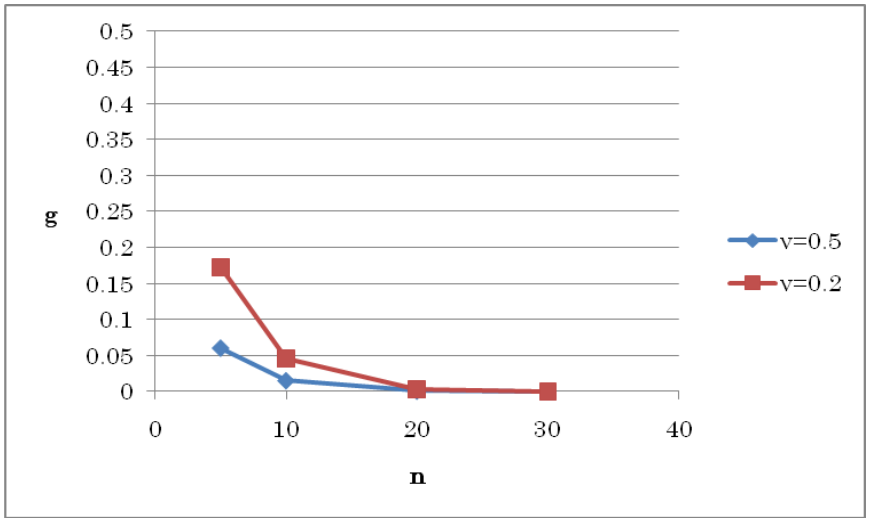


図8　効果的エラスティック(p=0.3)

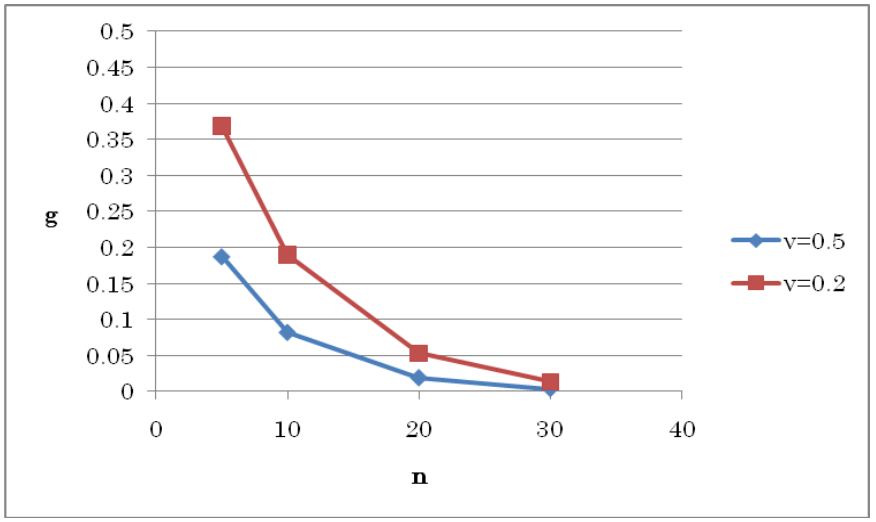


図9　効果的エラスティック(p=0.5)

5. 考察

5.1 均一誤り訂正符号

(1.1) (効果的)エラスティック(LP上界式)・・・図2，図3，(図8，図9)
(1.2) エラスティック　　(Gilbert下界式)・・・図4
(1.3) 符号Cの存在範囲・・・図5

5.2 不均一誤り訂正符号

(2.1) エラスティック(LP上界式)・・・図6
(2.2) 不均一誤り訂正符号が有効(フレキシブル)・・・図7

6. むすび

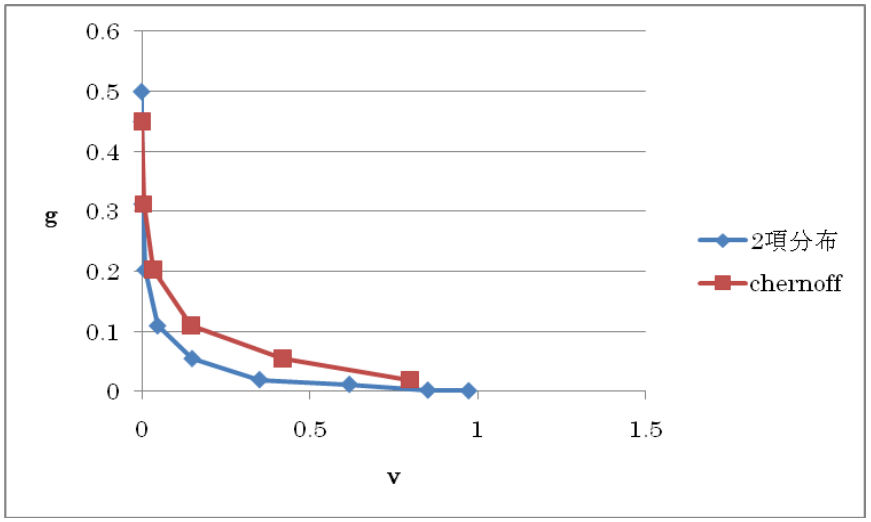
(1) 下に凸の関数である.

(2) システムの規模を示すパラメータ n が増加するにしたがい，効率が良くなる．すなわち，エラスティックである.

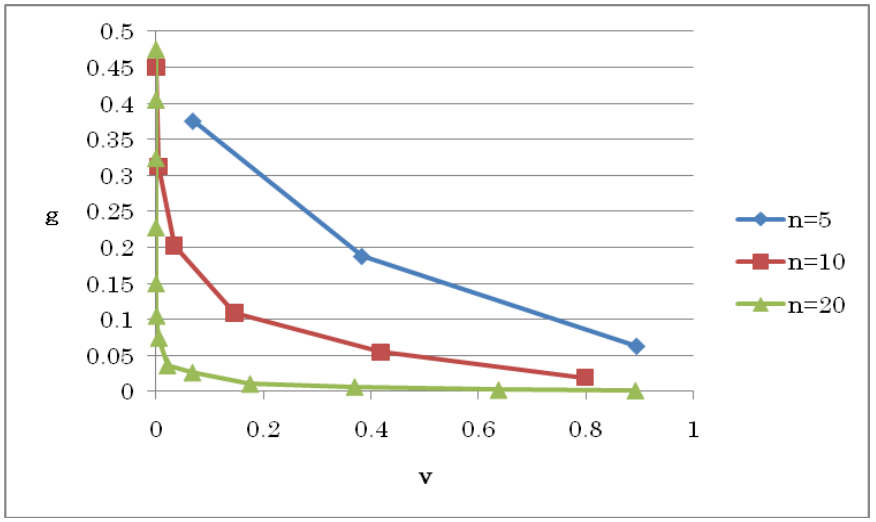
(3) ＊の生起が属性により不均一なときは，不均一誤り訂正符号を用いることが有効である．すなわち，フレキシブルである.

付録

サイバー大学
Cyber University



図A1　2項分布とChernoff限界
（n=10, p=0.3, LP上界式）



図A2　Chernoff限界
（p=0.3, LP上界式）