Entropy에 의한 Randomness 검정법

최봉대*, 신양우**, 이경현***

* 한국과학기술원 수학과 ** 창원대학교 자연과학대학 통계학과

*** 한국 전자통신 연구소

A Randomness Test by the Entropy

Bong Dae Choi* , Yang Woo Shin**, Kyung Hyune Rhee***

* Department of Mathematics, Korea Advanced Institute of Science and Technology

** Department of Statistics, Changwon National University

*** Electronics and Telecommunications Research Instiute

초 록

본 논문에서는 임의의 이진 난수발생기의 source가 BMS_p이거나 M-memory를 갖는 마르코프연쇄로 모델화 되었을 경우에 비트당 entropy와 관련이 있는 새로운 randomness에 관한 통계적 검정법을 제안한다. 기존에 알려진 이진 난수발생기의 randomness검정법이 0또는 1의 분포의 편향성(bias)이나 연속된 비트간의 상관성(correlation)중의 한 종류만의 non-randomness를 추적해낼수 있는 반면에 새로운 검정법은 위의 두가지 검정을 통과하였을 때 암호학적으로 중요한 측도인 비트당 entropy 를 측정하여 암호학적인 약점을 검정할 수 있다. 또한 대칭(비밀키)암호시스템의 통계적 결점을 바탕으로 하여 키를 찾는 공격자의 최적 전략(optimal strategy) 문제를 분석하여 이 최적 전략이 이진 수열의 비트당 entropy와 밀접한 관계가 있음을 보이고 이 비트당 entropy와 관련이 있는 새로운 통계량을 도입하여 이진 난수 발생기의 source의 이진수열이 다음 3가지 경우, 즉, *i.i.d.* symmetric인 경우, BMS_p인 경우, M-memory를 갖는 마르코프연쇄인 경우의 각각에 대하여 특성을 조사하고 새로운 통계량의 평균과 분산을 구한다. 이때 구한 새로운 통계량은 잘 알려진 중심 극한 정리에 의하여 근사적으로 정규분포를 따르므로 위의 평균과 분산을 이용하여 스트림 암호시스템에서 구성요소로 많이 사용되는 몇몇 간단한 이진 난수 발생기에 적용하여 통계적 검정을 실시함으로써 entropy 관점의 검정법이 새로운 randomness검정법으로 타당함을 보인다.

한글 $\mathcal{AMS}$-TEX에 의한 전자조판

제 1 절 서론

본 논문에서는 난수발생기의 source가 BMS_p 이거나 M-memory 를 갖는 마르코프연쇄로 모형화 되었을 경우에 비트당 entropy와 관련이 있는 새로운 통계적 검정법을 제안한다. 기존에 알려진 여러가지 검정법은 0 혹은 1의 분포의 편향성, 연속된 비트간의 상관성 중의 한 종류만을 검정할수 있는 반면에, 이 새로운 검정법은 암호적으로 중요한 측도인 비트당 entropy를 측정하여 암호학적인 약점을 검정할 수 있다. 제 2절에서는 entropy의 개념을 도입하고 entropy를 불확실성의 정도 또는 한 실험에 의하여 얻어진 정보량으로 해석할 수 있음을 설명하고, 앞으로 사용될 entropy의 성질을 기술한다. 제 3절에서는 먼저 M-memory을 갖는 마르코프연쇄를 정의하고, $M = 1$ 인경우, 즉, 보통의 의미의 마르코프연쇄의 entropy를 정의한다. 다음에 비밀키의 통계적 결점을 바탕으로 하여 키를 찾는 적의 최적 전략 (optimal strategy)문제의 분석이 제시되고 이 최적 전략이 비트당 entropy와 밀접한 관계가 있음을 보인다. 마르코프연쇄에서 전이확률 (transition probability)의 추정, memory의 크기의 추정문제를 다룬다. 4절에서는 비트당 entropy 에 관련이 있는 새로운 통계량을 도입하고 이 통계량이 이진수열을 생성하는 난수발생기의 source가 *i.i.d.* symmetric 인 경우, BMS_p 인 경우, M-memory 를 갖는 마르코프 연쇄인 각각 경우의 특성을 조사하고 각 경우에 새로운 통계량의 평균과 분산을 구한다. 새로운 통계량은 중심극한 정리에 의하여 근사적으로 정규분포를 따르므로 위의 평균과 분산을 이용하여 통계적 검정을 시행 할 수 있다.

제 2절 Entropy

Entropy의 개념은 여러 종류의 정보(informations)의 전송을 위한 이론적 모형을 만드는 단계에서 C. Shannon (1948)에 의하여 도입되었다.

$\{A_1, A_2, \cdots, A_n\}$ 이 확률공간 $(\Omega, \mathcal{F}, P)$의 분할(partition)이란 $\{A_i\}$ 이 서로소이고 $\bigcup_{i=1}^{n} A_i = \Omega$ 일때를 말한다. 즉 한 실험에서 $\{A_i\}$ 들중에 단 한사건만 일어나는 것을 의미한다. 예를 들면 주사위를 던지는 실험에서 $A_i = \{i\}$ 라 하면 $\{A_i | 1 \le i \le 6\}$ 은 분할이 된다.

$\{A_i | 1 \le i \le n\}$ 가 분할이고 $P(A_i) = p_i$ 라 두면 $p_i \ge 0, \sum_{i=1}^{n} p_i = 1$ 이다. 이때

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$$

을 유한 체계(finite scheme)이라고 한다. 예를 들면 공정한 주사위를 던지는 실험인 경우에 다음과 같은 유한 체계를 갖는다

$$\begin{pmatrix} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}.$$

모든 유한 체계는 다음과 같은 의미에서 불확실성(uncertainty)의 상태를 나타낸다. 한 실험의 결과는 사건 $A_1, A_2, \cdots, A_n$ 중의 어느 하나로 나타나고 실험하기 이전에는 각 사건이 일어날 확률만을 알고 있기 때문이다. 불확실성의 정도는 유한 체계에 따라서 다르다.

예를 들면 다음 두개의 유한 체계에서

$$\begin{pmatrix} A_1 & A_2 \\ 0.5 & 0.5 \end{pmatrix}, \quad \begin{pmatrix} A_1 & A_2 \\ 0.99 & 0.01 \end{pmatrix},$$

첫번째의 체계는 두번째보다 훨씬 더 많은 불확실성을 나타낸다 ; 두번째 체계에서는 실험의 결과는 거의 확실시 ("almost surely") A_1 이 되지만, 첫번째 체계에서는 어떠한 예측도 하기가 힘들다. 다음 체계

$$\begin{pmatrix} A_1 & A_2 \\ 0.3 & 0.7 \end{pmatrix},$$

은 위의 두 체계의 중간 정도의 불확실성을 나타낸다.

주어진 유한 체계에 합리적인 방법으로 불확실성의 양을 측정하는 측도를 도입하는 것이 바람직할 것이며, 이것이 다음에 정의하는 entropy 이다.

정의: $A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}$ 가 유한 체계일때

$$H(p_1, p_2, \cdots, p_n) = -\sum_{k=1}^{n} p_k \log p_k$$

을 A의 entropy라고 한다.

주의: 지수의 기저는 임의로 택해도 좋으나 일반적으로 2를 취한다. 그리고 $p_k = 0$인 경우 $p_k \log p_k = 0$으로 정의한다.

불확실성의 합리적인 측도가 가져야 하는 성질들을 위에서 정의한 entropy $H(p_1, p_2, \cdots, p_n)$가 갖고 있음을 직관과 이론적으로 설명을 하고저 한다. 먼저 우리는 다음사실을 쉽게 알 수 있다.

(a). $H(p_1, p_2, \cdots, p_n) = 0$ 일 필요충분조건은 $p_1, p_2, \cdots, p_n$ 중에 어느 하나가 1이고 나머지 모두는 0이다.

위의 사실은 실험의 결과가 시행이전에 확실하게 예측될 수 있어서 그의 결과에는 불확실성이 전혀 없는 경우를 설명 하는 것이다.

(b). $\phi(x) = \begin{cases} 0, & x = 0 \\ x \log x, & x \neq 0 \end{cases}$

$\phi(x)$는 convex함수이다. 즉, $x, y \in [0, \infty)$, $\alpha + \beta = 1$, $a, \beta \geq 0$이면

$\phi(\alpha x + \beta y) \leq \alpha\phi(x) + \beta\phi(y)$이다.

수학적 귀납법에 의하여, $x_i \in [0, \infty)$, $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i = 1$이면

$$\phi(\sum_{i=1}^{n} \alpha_i x_i) \leq \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

이다.

(c). $H(p_1, p_2, \cdots, p_n) \leq \log n$

증명:

$\phi(x) = x \log x$라 하고 윗식에서 $x_k = p_k$, $a_k = \frac{1}{n}$ 라 두면,

$$\phi\left(\frac{1}{n}\right) = \frac{1}{n} \log \frac{1}{n} \leq \frac{1}{n} \sum_{k=1}^{n} p_k \log p_k = -\frac{1}{n} H(p_1, p_2, \cdots, p_n)$$

을 얻는다. 따라서

$$H(p_1, p_2, \cdots, p_n) \leq \log n = H(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})$$

이다.

위의 사실 (c)는 Ω의 모든 n개의 사건으로 이루어진 체계 가운데 가장 큰 entropy 가 모든 사건이 같은 확률을 갖는 체계임을 보여준다. 이것은 entropy 의 직관적인 해석과 일치한다.

(d) 두개의 유한 체계

$$A = \begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_1 & p_2 & \cdots & p_n \end{pmatrix}, \quad B = \begin{pmatrix} B_1 & B_2 & \cdots & B_n \\ q_1 & q_2 & \cdots & q_n \end{pmatrix},$$

이 주어졌을때, $\{A_k \cap B_l | 1 \leq k \leq n,\ 1 \leq l \leq m\}$과 $\pi_{kl} = P(A_k \cap B_l)$은 다른 유한 체계가 되고, 이것을 체계 A와 B의 곱(product)라고 부르고 AB로 적는다.

만약 체계 A와 B가 독립, 즉,

$\pi_{kl} = P(A_k \cap B_l) = P(A_k)P(B_l) = p_k q_l$ 이면

$H(AB) = H(A) + H(B)$ 이다. 단, $H(A), H(B), H(AB)$는 각 체계 A, B, AB의 entropy 이다.

왜냐하면,

$$-H(AB) = \sum_k \sum_l \pi_{kl} \log \pi_{kl} = \sum_k \sum_l p_k q_l (\log p_k + \log q_l)$$

$$= \sum_k p_k \log p_k + \sum_l q_l \log q_l = -H(A) - H(B)$$

가 성립하기 때문이다.

(e). 체계 A와 B가 독립이 아닌 종속인 경우를 생각하여 보자.

체계 A의 한 사건 A_k가 일어났다는 가정하에 체계 B의 원소 B_l의 조건부 확률을 q_{kl}로 두면, 즉,

$$q_{kl} = P(B_l|A_k) = \frac{\pi_{kl}}{p_k},$$

A_k상에 유도된 체계 $\{B_l \cap A_k | 1 \le l \le m\}$의 조건부 entropy는 $H_k(B) = -\sum_{l=1}^{m} q_{kl} \log q_{kl}$

이다. 그리고 난후 k에 관하여 평균을 취하면 $H(B|A) = -\sum_{k=1}^{m} p_k \sum_{l=1}^{m} q_{kl} \log q_{kl}$을 얻는다. 이것을 주어진 체계 A에 관하여 체계 B의 조건부 entropy 라고 한다. 이것을 다시 요약하면 다음과 같다.

정의: 체계 $A = (A_1, A_2, \cdots, A_n)$, $B = (B_1, B_2, \cdots, B_m)$에 대하여

$$H(B|A) = -\sum_{k=1}^{m} P(A_k) \sum_{l=1}^{m} \frac{P(A_k \cap B_l)}{P(A_k)} \log \frac{P(A_k \cap B_l)}{P(A_k)}$$

$$= -\sum_{k,l} P(A_k \cap B_l) \log \frac{P(A_k \cap B_l)}{P(A_k)} \ge 0$$

을 주어진 체계 A에 관한 체계 B의 조건부 entropy라고 한다. 조건부 entropy 에 관하여 다음 사실이 성립한다.

(e.1) $A = (\phi, \Omega) \Rightarrow H(B|A) = H(B)$

(e.2) $H(AB) = H(A) + H(B|A)$

(e.3) $H(B|A) \le H(B)$

주의: (1)의 사실은 A는 trivial 한 실험의 결과를 나타내므로 A로부터는 아무런 정보도 얻을 수 없다는 것을 나타낸다. (3)의 사실은 A의 결과를 아는 것은 평균적으로 B의 불확실성을 감소시킨다는 것이다.

증명: (1)은 조건부 정의로부터 쉽게 얻을 수 있고 (2)는 (e)에 주어진 독립인 경우와 마찬가지의 방법을 따라가면 얻을 수 있다.

(3)의 증명 ; $P(A_k) = p_k$, $P(B_l) = q_l$, $P(A_k \cap B_l) = \pi_{kl}$, $q_{kl} = \frac{\pi_{kl}}{p_k}$ 로 두면

$$H(B|A) = -\sum_k p_k \sum_l q_{kl} \log q_{kl}$$

이된다. $f(x) = x \log x$ 는 convex 함수이고, 부등식 $\sum_k \lambda_k f(x_k) \geq f(\sum_n \lambda_k x_k)$ 을 만족하므로.

$\lambda_k = p_k$, $x_k = q_{kl}$ 로 두면 임의의 l 에 대하여

$$\sum_k p_k q_{kl} \log q_{kl} \leq (\sum_k p_k q_{kl}) \log(\sum_k p_k q_{kl}) = q_l \log q_l$$

을 얻는다. 양변에 l에 관하여 합하면

$$-H(B|A) \geq -H(B)$$

을 얻을수 있다.

(f). 불확실성은 실험에 의하여 얻어지는 정보와 같다는 개념을 설명하고자 한다.

실험의 결과가 주어진 유한 체계 A에 의하여 기술되어지는 경우에, 그 실험을 시행한 후에 실제적으로 어느 사건이 일어났는가 하는 정보(information)을 얻게 되고 유한 체계의 불확실성이 완전히 제거되어 진다. 따라서 어떠한 실험을 시행하므로써 얻어지는 정보는 실험이전에 존재했던 불확실성을 제거하는 것과 같이 볼 수 있다. 불확실성이 크면 클수록 그것을 제거하므로써 얻어지는 정보의 양이 많다는 것으로 간주한다. Entropy 의 성질로부터 정보의 양이 불확실성의 측도인 entropy에 비례하는것으로 택하는 것이 편리함을 알수있다. 예를 들면 유한 체계 A와 B, 그리고 곱 AB를 생각하여 보자. AB가 일어난다 함은 A와 B가 독립이면, $H(AB) = H(A) + H(B)$ 가 되고 따라서 A와 B에 의하여 얻어진 정보의 양의 합이 된다는것은 자연스러운 것이다. 비례

상수를 1로 취하므로써, 한 유한 체계의 실현으로 얻어지는 정보의 양은 그 유한 체계의 entropy로 정의한다. 이러한 약정으로인하여 entropy의 개념이 정보이론에 매우 유용하게 이용되고 있다. 따라서 얻어진 정보의 양의 관점에서의 (e. 2)식의 해석은 A와 B의 실현에 의하여 얻어진 정보의 양은 A의 실현에 의하여 얻어진 정보의 양에 A의 실현이 이루어진 후에 B의 실현에 의하여 부가적인 정보의 양의 수학적 기대치를 합한 것과 같다. (e. 3)의 해석은 B의 실현에 의하여 얻어진 정보의 양은 그 이전에 다른 A의 실현이 이루어졌다면 감소한다.

(g). 우리가 증명한 entropy의 기본적인 성질을 열거하면 다음과 같다.

(g.1) $H(p_1, p_2, \cdots, p_n)$는 $p_k = \frac{1}{n}(1 \leq k \leq n)$인 경우에 최대치를 갖는다.

(g.2) $H(AB) = H(A) + B(B|A)$

(g.3) $H(p_1, p_2, \cdots, p_n, 0) = H(p_1, p_2, \cdots, p_n)$.

정리 2.1: $\Delta_n = \{(p_1, p_2, \cdots, p_n) \in R^n | p_k \geq 0, \sum_{k=1}^{n} p_k = 1\}$

함수 $H : \bigcup_{n=1}^{\infty} \Delta_n \to R$가 위의 (g.1) - (g.3)을 만족하고 또는 $H|_{\Delta_n}$가 연속이면, 적당한 $\lambda > 0$가 존재하여

$$H(p_1, p_2, \cdots, p_n) = -\lambda \sum_{k=1}^{n} p_k \log p_k$$

로 나타난다.

위의 정리의 증명은 Khinchine의 책[5, p9]에 있다.

위의 정리는 entropy의 정의가 실제적 의미의 불확실성의 측도 (또는 정보의 양)으로 해석하였을때 만족하여야 할 성질을 갖는 것으로는 유일하다는 것을 보인다.

제 3절 마르코프연쇄와 그의 entropy

일반적으로 난수 발생기 (혹은 source)에서 생성된 난수열이 가지고 있는 약점은 빈도의 편향성 (bias)과 각 항 사이의 상호관련성이다. 2진 난수발생기 S가 확률변수열 $U_1, U_2, \cdots$ 에 의하여 난수를 생성한다고 하자. n번째 항 U_n이 바로 앞의 M 비트 에만 의존한다면, 즉 $U_1, U_2, \cdots, U_{n-1}$ 이 주어졌을 때 U_n의 분포가 $U_{n-1}, \cdots U_{n-M}$ 에만 의존할때 $\{U_n\}$을 M-memory를 갖는 마르코프연쇄라고 부른다. 위의 사실을 수학적으로 나타내면, $P_{U_n|U_{n-1},\cdots,U_1}(u_n|u_{n-1},\cdots,u_1) = P(U_n = u_n|(U_{n-1},\cdots,U_1) = (u_{n-1},\cdots u_1))$을 조건부 확률이라 하면, 모든 $n > M$, $(u_1,\cdots,u_n)$ B^n에 대하여

$$P_{U_n|U_{n-1},\cdots,U_1}(u_n|u_{n-1},\cdots,u_1) = P_{U_n|U_{n-1},\cdots,U_{n-M}}(u_n|u_{n-1},\cdots,u_{n-M}) \tag{3.1}$$

여기서 $B = \{0,1\}$이다. (3.1)을 만족하는 최소의 정수 M을 S의 memory라고한다. M이 0인 경우, 즉, U_n이 $U_{n-1},\cdots,U_1$과 독립일때 S는 memoryless라고 한다. 특히 memoryless suource S에서 확률 p로 1 값을 갖는 비트를, 확률 $1-p$로 0 값을 갖는 비트를 생성할때 source S를 BMS_p (Biased Memoryless Source)라고 한다. $M = 1$인 경우 우리는 보통 의미의 마르코프연쇄을 얻는다. $\sum_1 = [U_0,\cdots,U_{n-M}]$은 회수 n이전의 M개의 비트의 상태를 나타낸다. $\sum_1 = [U_0,\cdots,U_{-M+1}]$은 초기상태를 나타내고 $U_0, U_{-1},\cdots U_{-M+1}$은 가상의 확률변수들이다. 특히 확률변수열 $\{U_n\}$이 (3.1)을 만족하고 또한 모든 $n > M$, $u \in B$, $\sigma \in B^n$에 대하여 $P_{U_n|\sum_n}(u|\sigma) = P_{U_1|\sum_1}(u|\sigma)$을 만족할때 source S를 stationary 하다고 한다. memory M을 갖는 stationary source S는 초기상태의 분포 $P_{\sum_1}$과 전이확률 $P_{\sum_2|\sum_1}(\sigma_2|\sigma_1)$에 의해서 완전히 결정되어 진다. 만일 random vector $\sum_n$의 한 state $(u_1,u_2,\cdots,u_M)$ $(\in B^n)$이 정수 j의 2진수일 때 $(u_1,u_2,\cdots,u_M)$을 j와 동일시하여 $\sum_n = j$라고 표시하기로한다. 그러면 $\{\sum_n\}$은 상태공간을 $[0, 2^M - 1]$로 갖는 보통의미의 Markov chain이 된다. 그러므로 M-memory를 갖는 stationary source S 는 상태공간 $[0, 2^M - 1]$을 갖는 homogeneous 1-memory Markov chain 으로 모형화 될 수 있다. 만일 $\{\sum_n\}$이 ergodic Markov chain이

면 $\sum_n$의 극한분포

$$\lim_{n\to\infty} P_{\sum_n}(j) = P_j,\ 0 \le j \le 2^M - 1$$

가 존재하고 다음의 관계식을 만족한다.

$$\sum_{j=0}^{2^M-1} P_j = 1$$

$$P_j = \sum_{k=0}^{2^M-1} P_{\sum_2|\sum_1}(j|k)P_k,\ \ 0 \le j \le 2^M - 1$$

보통 의미의 마르코프연쇄에 관한 entropy를 정의하기로 하자. 전이확률 $(p_{ik})(i,k = 1,2,\cdots,n)$와 상태공간 $\{1,2,\cdots,n\}$을 갖는 ergodic이고 stationary인 마르코프연쇄에서 극한분포를 P_j라하면

$$P_j = \sum_{k=1}^{n} P_k p_{kj}$$

를 만족한다. 만약 마르코프연쇄가 현재 state i에 있고, 다음에 다른 state로의 전이확률이 $(p_{i1}, p_{i2}, \cdots, p_{in})$일때, 이것의 entropy

$$H_i = -\sum_{k=1}^{n} p_{ik} \log p_{ik}$$

는 i에만 의존하고, H_i는 마르코프연쇄가 state i에서 출발하여 한 단계 앞으로 나가갔을때 얻어진 정보량의 측도로 간주 되어질수 있다. 모든 초기 state i에 관한 H_i의 평균

$$H = \sum_{i=1}^{n} P_i H_i = -\sum_{i=1}^{n}\sum_{k=1}^{n} P_i p_{ik} \log p_{ik}$$

는 마르코프연쇄가 한 단계 앞으로 움직였을때 얻어진 평균정보량의 측도로 볼 수 있다. 이것을 마르코프연쇄의 entropy라고 부른다.

좋은 암호계는 적의 공격에 대해서 안정되어야 한다. 즉, 좋은 암호계란 권한이 없는 암호 해독자가 암호방식을 해독하기 위하여 key를 찾을 때 어떠한 통계적인 공격방식도 key space 에 있는 모든 key 를 하나하나 적용하는 방식 (exhaustive key search) 보다 본질적으로 더 빠른 방식이 없도록 고안된 것이다. 만약 비밀 key가 truly random이 아니면 즉, 모든 key가 같은 확률을 가지지 않으면 적의 최적의 전략은 확률이 가장 높은 키값으로부터 시작하여 확률이 낮은 키 값으로 순서대로 적용할 것이다. Z를 비트의 길이가 n 인 비밀키라 하고 $P_Z(z)$를 Z의 값이 z 일 확률이라고 하자. 이때 $z_1, z_2, \cdots, z_{2^n}$ 은 $P_Z(z_i)$의 값이 큰 순서대로 key space의 원소들에 번호를 붙여 놓은것이라 하자. 즉, $P_Z(z_1) \geq P_Z(z_2) \geq \cdots \geq P_Z(z_{2^n})$.

주어진 source S로부터 n - bit key Z를 만들었들때 최소한 δ의 확률로 key를 찾는데 성공하기 위하여 암호해독자가 최적의 전략을 따라서 test 해야할 최소한의 시행회수를 $\mu_S(n,\delta)$ 라 하자. 즉,

$$\mu_S(n,\delta) = \min\{k : \sum_{i=1}^{k} P_Z(z_i) \geq \delta\}.$$

이때 $\log_2 \mu_S\left(n, \frac{1}{2}\right)$ 을 effective key size라 한다. 여기서 $\delta = \frac{1}{2}$ 을 선택한것은 임의적인것이며 일반적으로 key의 길이 n이 충분히 클때 δ가 0이나 1에 극단적으로 가깝지 않으면 $\log_2 \mu_s(n,\delta)$는 δ 에 거의 의존하지 않는다. 만약 S가 truly random source라면, $P_Z(z_i) = \frac{1}{2^n}, i = 1, 2, \cdots, 2^n$ 이므로 effective key size는 $\log_2 \mu_s\left(n, \frac{1}{2}\right) = n - 1$ 이다.

Effective key size와 entropy의 관련성을 아래의 정리 3.1 (M-memory를 갖는 source 의 경우) 와 정리 3.2 (BMS_p source의 경우)에서 알 수 있다.

정리 3.1 (Shannon(1948)): M - memory를 갖는 stationary source S로 부터 key Z가 만들어졌을 때 다음의 식이 성립한다.

$$\lim_{n\to\infty} \frac{\log_2 \mu_S(n,\delta)}{n} = H_S, \ 0 < \delta < 1 \qquad (3.2)$$

여기서 $H_S = -\sum_{j=0}^{2^M-1} P_j \sum_{k=0}^{2^M-1} P_{\Sigma_2|\Sigma_1}(k|j)\log_2 P_{\Sigma_2|\Sigma_1}(k|j)$는 M-memory을 갖는 마르코프연쇄인 source S의 단위 비트당 entropy이다.

증명은 Khinchine의 책[5, p20]에 있다.

길이 n인 모든 key의 개수는 2^n인 반면 정리 3.1로부터 $\mu_S(n,\delta)$는 약 2^{nH_s} 임을 나타낸다.

다음은 key source S가 BMS_p인 경우를 생각하자. 우리는 일반성을 잃지 않고 $0 < p \le \frac{1}{2}$ 을 가정할 수 있다. Z가 BMS_p로부터 만들어졌다면 Z의 확률분포는 $P_Z(z) = P^{w(z)}(1-p)^{n-w(z)}$로 주어진다. 여기서 $w(z)$는 z의 성분중에 있는 1의 개수이다. 약 $\frac{1}{2}$ 의 확률로 성공하기 위하여 해독자는 $w(z) \le pn$인 모든 key값 z에 대하여 조사를 하여야 한다. 이때 effective key size는

$$\log_2 \mu_{\mathrm{BMS}_p}\left(n, \frac{1}{2}\right) \approx \log_2 \sum_{i=0}^{pn} \binom{n}{i}. \tag{3.3}$$

다음의 부등식

$$\frac{1}{\sqrt{8t(n-t)/n}} 2^{nH(\frac{t}{n})} \le \binom{n}{i} \le 2^{nH(\frac{t}{n})},\ t \le \frac{n}{2} \tag{3.4}$$

에서 $t = np$ 를 대입한 다음 $\log_2$를 취한후 n으로 나누고 n을 ∞로 보내면 다음을 얻는다. 여기서 $H(x) = -x\log_2 x - (1-x)\log_2(1-x)$는 binary entropy function 이다.

정리 3.2:

$$\lim_{n\to\infty} \frac{\log_2 \mu_{\mathrm{BMS}_p}(n,\delta)}{n} = H(p),\ 0 < \delta < 1 \tag{3.5}$$

어떤 source (혹은 generator) 를 이용하여 암호계의 key를 만들었을때 source의 단위 비트당 entropy는 그 암호계의 안정성과 밀접한 관련이 있다는 것을 위의 결과들로부터 알 수 있다. 그러므로 암호키를 만드는데 사용된 source의 단위 비트당 entropy는 암호계의 안정성을 측정하는 하나의 척도가 될 수 있다.

다음에 이산 시각을 갖는 마르코프연쇄의 모수들의 추론 문제를 다룬다. 이산 시각을 갖는 마르코프연쇄에 대한 중요한 통계적 추론문제에는 다음과 같은것이 있다.

(1) 주어진 source가 Markovian 성질과 time homogeneous성질의 성립여부를 판정하는 문제

(2) 마르코프연쇄의 전이확률 (p_{ij})를 검정하는 문제

(3) M-memory를 갖는 마르코프연쇄에서 M을 검정하는 문제

(1) 전이확률의 추정

상태공간 $\{1,2,...m\}$ 를 갖는 유한 마르코프연쇄 $\{X_n, n \geq 0\}$이 유한시간 $[0,N]$동안 관측되었다고 가정하자. 이때 $X_0 = x_0, X_1 = x_1, \cdots X_N = x_N$을 실현치라 하자. 여기서 X_0는 random이 아니라고 가정하자. 주어진 data $x_1, \cdots, x_n$에 대한 우도함수 (likelihood function)는 다음과 같이 주어진다.

$$f(x_1, \cdots, x_N) = p_{x_0 x_1} p_{x_1 x_2} \cdots p_{x_{N-1} x_N}. \tag{3.6}$$

이때

$$I_{ijk} = \begin{cases} 1 & \text{if } x_k = i,\ x_{k+1} = j, \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

$$n_{ij} = \sum_{k=0}^{N-1} I_{ijk} \ : \text{state } i \text{ 에서 state } j \text{ 로의 one-step transition의 총개수} \tag{3.8}$$

$$n_i = \sum_{j=1}^{m} n_{ij} \ : \ 1\text{회에서 } N \text{ 회까지에서 state } i \text{를 방문하는 총개수} \tag{3.9}$$

그러면 (3.6)의 식은

$$f(p_{ij},\ 1 \leq i,j \leq m | x_1, \cdots x_n) = \prod_{i=1}^{m} \prod_{j=1}^{m} p_{ij}^{n_{ij}} \tag{3.10}$$

가 된다. 여기서 p_{ij}는 i에서 j로의 one step transition probability로 추정하고자 하는 모수(parameter)이다. (3.10)식으로 부터 로그우도함수(log-likelihood function)

$$\log f(p_{ij},\ 1 \le i,j \le m | x_1, \cdots x_n) = \sum_{i=1}^{m}\sum_{j=1}^{m} n_{ij} \log p_{ij}$$

를 얻는다. 이때 p_{ij}의 최우추정량은 위식의 우변을 최대로 하는 p_{ij}의 값이다. 이때 최우추정량(maximum likelihood estimator)는

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i},\ n_i > 0,\ i = 1, \cdots, m,\ j = 1, \cdots, m \qquad (3.11)$$

로 주어진다. 추정량 $\hat{p}_{ij}$는 다음과 같은 성질을 갖는다.

1. $\sqrt{n_i}(\hat{p}_{ij} - p_{ij}) \xrightarrow{w} N(0, p_{ij}(1-p_{ij}))$, $N \to \infty$. 여기서 $\xrightarrow{w}$는 약 수렴 (weak convergence)를 의미한다.
2. n_j가 주어졌다는 가정하에 p_{ij}에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\left(\sin(\sin^{-1}\sqrt{\hat{p}_{ij}} - z_{\alpha/2}\frac{2}{\sqrt{n_i}}),\ \sin(\sin^{-1}\sqrt{\hat{p}_{ij}} + z_{\alpha/2}\frac{2}{\sqrt{n_i}})\right)$$

으로 주어진다.

(2) 마르코프연쇄의 memory크기 M의 추정

보통 의미의 마르코프 성질은 현재의 상태가 주어졌을 때 미래의 상태와 과거의 상태가 독립인 것을 의미한다. 마르코프 성질의 일반화로서 r차 마르코프 성질이라는 것은 과거 $n-(r+1)$시각부터 현재 n시각까지 r시간 동안의 상태가 주어졌을 때 미래의 상태와 초기 부터 $n-r$시각까지의 상태가 독립인것을 뜻한다. r차 마르코프 성질을 가지는 확률과정을 r차 마르코프연쇄라 한다. 즉 상태공간 $\{1,2,3,\cdots,m\}$을 갖는 확률과정 $\{X_n, n \ge 0\}$이 모든 $n > r, i_0, \cdots, i_n, j$에 대하여

$$P(X_{n+1}=j|X_n=i_0,X_{n-1}=i_1,\cdots,X_{n-r+1}=i_{r-1},\cdots,X_0=i_n)$$
$$=P(X_{n+1}=j|X_n=i_0,X_n-1=i_1,\cdots,X_{n-r+1}=i_{r-1}) \quad (3.12)$$

을 만족할 때 (3.12)을 만족하는 최소의 정수를 M이라하면 M을 $\{X_n,n\geq 0\}$의 차수(order) 혹은 memory라고 한다. memory M을 갖는 확률과정 $\{X_n,n\geq 0\}$을 M차 마르코프연쇄 혹은 M-memory를 갖는 마르코프연쇄라 한다. 특히 $M=0$일때는 $\{X_n,n\geq 0\}$은 독립인 확률변수열이 되고 $M=1$ 일때는 $\{X_n,n\geq 0\}$은 보통의 의미의 마르코프연쇄가 된다. 이제 주어진 수열로 부터 source의 memory 크기를 추정하는 방법을 살펴보자.

$x_0,x_1,\cdots,x_N$ 을전이확율 (p_{ij})을 갖는 마르코프 연쇄로 부터 얻어진 데이타라 하자. 이때 전이확률 $P=(p_{ij})$는 알려지지 않은것으로 가정한다. 면저 우리는 관측된 data로 부터 source의 memory가 $M=0$ (독립된 수열) 혹은 $M=1$ (보통의 마르코프연쇄)인가를 검정한다. 즉

$H_0:M=0$

$H1:M=1$

만일 H_0가 참이면 $p_{ij}=p_j,\ i=1,2,\cdots,m$일 것이다. n_{ij} 와 n_i을 (3.8)과 같이 정의 하자. 그러면 p_{ij}의 최우추정량은 $\hat{p}_{ij}=\frac{n_{ij}}{n_i}$이므로 우도함수는

$$L_1=\prod_i\prod_j\left(\frac{n_{ij}}{n_{i\cdot}}\right)^{n_{ij}} \quad (3.13)$$

이다. $\sum_j p_{ij}=1,\ (j=1,2,\cdots,m)$이므로 parameter space Θ의 차원은 $m(m-1)$이다.

H_0 하에서는 $m-1$개의 parameter $\{p_j^0,j=1,2,\cdots,m\}$가 있고, p_j^0의 최우추정량은

$$\hat{p}_j=\frac{n_j}{N}=\frac{\sum_i\sum_k n_{ijk}}{\sum_i\sum_j\sum_k n_{ijk}} \quad (3.14)$$

로 주어진다. 여기서 n_{ijk}는 k step에서 i에서 j로의 전이가 일어나는 횟수이다. 즉

$$n_{ijk} = \begin{cases} 1 & x_k = i, x_{k+1} = 1 \text{ 일 경우} \\ 0 & \text{그밖의 경우} \end{cases}$$

그러므로 H_0 하에서의 우도함수는

$$L_0 = \prod_j \left(\frac{n_{\cdot j}}{N}\right)^{n_{\cdot j}} \tag{3.15}$$

이다.

일반화된 likelihood ratio 방법을 따라서 L_0 와 L_1의 비율 $\frac{L_0}{L_1}$을 계산한 다음 $\frac{L_0}{L_1}$이 작으면 H_0 를 기각한다. 정확하게 말하면 $-2\log(\frac{L_0}{L_1})$는 근사적으로 자유도 $m(m-1)-(m-1) = (m-1)^2$ 을 가지는 χ^2-분포를 따르므로, 만약

$$-2\log(\frac{L_0}{L_1}) > \chi^2_{(S-1)^2}(\alpha) \tag{3.16}$$

이면 유의 수준 α로 H_0 를 기각한다.

만일 $M=0$가 기각되었으면 다시 귀무가설 $M=1$에 대해 대립가설 $M=2$를 검정한다. 즉

H_0 : $M=1$

$H1$: $M=2$

H_0 : $M=1$ 이라는 가정하에서 우도함수는 (3.13)에서 주어졌다. H_1 하에서의 우도함수를 계산하기 위하여 2단계 전이확률 $p_{ijl} = P(X_k = l | X_{k-2} = i, X_{k-1} = j)$과 우도함수 $L = \prod_{k=3}^{N} p_{x_{k-2}x_{k-1}x_k}$ 를 정의한다.

$$I_{ijlk} = \begin{cases} 1 & \text{if } X_{k-2} = i, X_k = l \\ 0 & \text{otherwise} \end{cases}$$

$n_{ijl} = \sum_{k=3}^{N} I_{ijlk}$, $n_{ij\cdot} = \sum_{l} n_{ijl}$ 로 두면 우도함수는

$$L_2 = \prod_i \prod_j \prod_l \left(\frac{n_{ijl}}{n_{ij\cdot}}\right)^{n_{ijl}} \tag{3.17}$$

로 주어진다. $\sum_l p_{ijl} = 1,\ 1 \le i,j \le m$ 이므로 $H_1 : M = 2$하에서는 $m^2(m-1)$ 개의 parameters가 있다. N이 충분히 클때 $-2\log\left(\frac{L_1}{L_2}\right)$ 는 자유도 $S^2(S-1) - S(S-1) = S(S-1)^2$ 을 갖는 χ^2-분포를 따른다. 그러므로 만일

$-2\log\left(\frac{L_1}{L_2}\right) > \chi^2_{S(S-1)^2}(\alpha)$이면

$H_0 : M = 1$을 유의수준 α로 기각한다. 이와 같은 방법을 $H_0 : M = k$가 받아들여지는 순간까지 반복한다. 이때 source의 memory는 k가 된다.

제 4절 Entropy 관점에서의 Randomness 검정법

이 절에서 우리는 비트당 entropy와 밀접한 관련이 있는 통계량 T를 정의한다. $s^N = s_1 s_2 \cdots s_N$을 전체의 길이가 N인 이진 수열이라 하자. 먼저 s^N에서 길이가 L인 $\frac{N}{L}$개의 겨뇨치지 않는 blocks을 만든다. 처음 Q개의 blocks을 초기화를 위한 block으로 사용하고 나머지 $K = \frac{N}{L} - Q$개의 block은 test를 위해 사용한다. $b_n(s^N) = [s_{Ln}, s_{Ln+1}, \cdots, s_{Ln+L-1}]$, $0 \le n \le Q+K-1$을 n 번째 block이라하고 $Q \le n \le Q+K-1$ 에 대하여

$$A_n(s^N) = \begin{cases} \min\{i | 1 \le i \le n,\ b_n(s^N) = b_{n-i}(s^N)\} & \text{if } \{\cdots\} \neq \phi \\ n & \text{if } \{\cdots\} = \phi \end{cases}$$

라고 두자. 이때 통계량 T를 다음과 같이 정의한다.

$$T(s^N) = \frac{1}{K} \sum_{n=Q}^{Q+K-1} \log_2 A_n(s^N). \tag{4.1}$$

통계량 T의 parameters 로서 다음의 값을 추천한다. $8 \le L \le 16$, $Q \ge 30 \cdot 2^L$ 그리고 K는 클수록 좋다 (예 $K = 10^4$ 혹은 $K = 10^5$). 이러한 Q의 선택은 거의 확률 1로써 random 수열에서 모든 L-bit형태가 최소한 한번씩 나타나도록 한것이다. 통계량 $T(s^N)$ 을 계산하는 알고리즘은 PASCAL 형태의 기호로 표시하면 다음과 같다.

FOR $i := 0$ TO $2^L - 1$ DO Tab $[i] := 0$;

FOR $n := 0$ TO $Q - 1$ DO Tab $[b_n(s^N)] := n$;

sum = 0.0

FOR $n : Q$ TO $Q + K - 1$ DO BEGIN

sum := sum $+ \log_2(n -$ Tab $[b_n(s^N)])$;

Tab $[b_n(s^N)] := n$;

END

$T =$ sum $/K$

$L \to \infty$일 때 통계량 (4.1)과 단위 비트당 entropy와의 관계는 다음의 정리 4.1 (truly ramdom source의 경우), 정리 4.2 (BMS_p source의 경우), 정리 4.3 (M-memory를 갖는 source의경우) 에서와 같다.

정리 4.1: $R^N = R_1 R_2 \cdots R_N$이 *i.i.d.* symmetric binary sequence 이라 할 때 다음 식이 성립한다.

$$\lim_{L\to\infty} \left(E(T(R^N)) - L\right) = -0.832746$$

$$\lim_{L\to\infty} K \cdot \mathrm{Var}(T(R^N)) = 3.423715$$

정리 4.2: $U^N_{\mathrm{BMS}_p}$가 길이 N인 BMS_p의 output이라 할 때 다음 식이 성립한다.

$$\lim_{L\to\infty} \left(E(T(U^N_{\mathrm{BMS}_p})) - LH(p)\right) = -0.832746$$

정리 4.3: U^N_s가 M- memory 를 갖는 stationary source S의 output 이면

$\lim_{L\to\infty} \frac{E(T(U_s^N))}{L} = H_s$가 된다.

2진수열 U^N 의 randomness를 entropy 관점에서 검정할 때, 주어진 유의수준 ρ에 대한 기각역은 $\left|\frac{T(U^N) - E(T(U^N))}{\sigma}\right| \geq y$이다. 여기서 $\sigma = \sqrt{\mathrm{Var}(T(U_s^N))}$이고 $y = -\Phi^{-1}\left(\frac{\rho}{2}\right)$이다 (단 $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$).

2진수열 U^N이 주어졌을때 통계량 $T(U^N)$의 평균 $E(T(U^N))$와 분산 $\mathrm{Var}(T(U^N))$를 구하자.

(1) Source가 BMS_p인 경우

먼저 BMS_p source에 대하여 생각한다. BMS_p 의 block들인 $b_n(U^N_{\mathrm{BMS}_p})$은 서로 독립이므로

$$\begin{aligned}
&P_r(A_n(U^N_{\mathrm{BMS}_p}) = i) \\
&= \sum_{b\in B^N} P_r\Big(b_n(U^N_{\mathrm{BMS}_p}) = b,\ b_{n-1}(U^N_{\mathrm{BMS}_p}) \neq b, \cdots, \\
&\qquad b_{n-i+1}(U^N_{\mathrm{BMS}_p}) \neq b,\ b_{n-i}(U^N_{\mathrm{BMS}_p}) = b\Big) \\
&= \sum_{b\in B^N} \left[P(b_n(U^N_{\mathrm{BMS}_p}) = b)\right]^2 \left[1 - P(b_n(U^N_{\mathrm{BMS}_p}) = b)\right]^{i-1} \\
&= \sum_{k=0}^{L} \binom{L}{k} \left(p^k(1-p)^{L-k}\right)^2 \left(1 - p^k(1-p)^{L-k}\right)^{i-1}
\end{aligned}$$

이다. 그러므로

$$\begin{aligned}
E(T(U^N_{\mathrm{BMS}_p})) &= \frac{1}{K} \sum_{n=Q}^{Q+K-1} \sum_{i=1}^{n} P_r\left(A_n(U^N_{\mathrm{BMS}_p}) = i\right) \log_2 i \\
&= \frac{1}{K} \sum_{n=Q}^{Q+K-1} \sum_{k=0}^{L} \binom{L}{k} \left(p^k(1-p)^{L-k}\right)^2 \sum_{i=1}^{n} \left(1 - p^k(1-p)^{L-k}\right)^{i-1} \log_2 i
\end{aligned}$$

이다. $Q \to \infty$이면 $\sum_{i=1}^{n} \left(1 - p^k(1-p)^{L-k}\right)^{i-1} \to 0$이므로 $A_n(U^N_{\mathrm{BMS}_p})$은 n에 거의 의존하지 않게 된다. 그러므로 이 때는

$$E\left(T(U^N_{\mathrm{BMS}_p})\right) = \sum_{k=0}^{L} \binom{L}{k} \left(p^k(1-p)^{L-k}\right)^2 \sum_{i=1}^{\infty} \left(1 - p^k(1-p)^{L-k}\right)^{i-1} \log_2 i$$

이다. 또한 $Q \to \infty$일때

$$\begin{aligned} K \cdot \mathrm{Var}\left(T(U^N_{\mathrm{BMS}_p})\right) &= \mathrm{Var}\left(\log_2 A_n(U^N_{\mathrm{BMS}_p})\right) \\ &= E\left(\log_2 A_n(U^N_{\mathrm{BMS}_p})\right)^2 - \left(E(\log_2 A_n(U^N_{\mathrm{BMS}_p}))\right)^2 \\ &= \sum_{k=0}^{L} \binom{L}{k} \left(p^k - (1-p)^{L-k}\right)^2 \sum_{i=1}^{\infty} \left(1 - p^k(1-p)^{L-k}\right)^{i-1} (\log_2 i)^2 \\ &\quad - \left[E(\log_2 A_n(U^N_{\mathrm{BMS}_p}))\right]^2 \end{aligned}$$

이다. 위의 계산에서 $Q \to \infty$를 가정한것은 실제의 검정에서 Q를 충분히 크게 선택하므로 실제적으로 이용하는 데 차이가 없다.

(2) Source가 M-memory Markov Chain인 경우

이진 수열 U^N_s를 M-memory를 갖는 source S에서 출력된 길이 N인 2진 수열이라고 하자. $E(T(U^N_s))$와 $\mathrm{Var}(T(U^N_s))$를 구하기 위하여는 먼저 확률

$$P(A_n(U^N_s) = i) = \sum_{b \in B^n} P_r(b_n(U^N_s) = b, b_{n-1}(U^N_s) \neq b, \cdots, b_{n-i+1}(U^N_s) \neq b, b_{n-i}(U^N_s) = b)$$

를 구해야한다.

$$P(b_n(U_s^N) = b, b_{n-1}(U_s^N) \neq b, b_{n-i}(U_s^N) = b))$$

$$= P(b_n(U_s^N) = b, b_{n-i}(U_s^N) = b)$$

$$- \sum_{k=1}^{i-1} P(b_n(U_s^N) = b, b_{n-k}(U_s^N) = b, b_{n-i}(U_s^N) = b) + \cdots$$

$$+ (-1)^k \sum_{i_1,\cdots,i_k=1 i_1<i_2<\cdots i_k}^{i-1} P(b_n(U_s^N) = b, b_{n-i_1}(U_s^N), \cdots, b_{n-i_k}(U_s^N) = b, b_{n-i}(U_s^N) = b)$$

$$+ \cdots + (-1)^{i-1} P(b_n(U_s^N) = b, b_{n-1}(U_s^N), \cdots, b_{n-i+1}(U_s^N) = b, b_{n-i}(U_s^N) = b)$$

이므로 $Q \leq n_1 < n_2 < \cdots < n_k$ 일때, $P(b_n(U_s^N))$를 구하면 된다. 먼저 block 의 길이 L이 memory의 크기 M의 배수라고 가정하자. 즉 $L = Mm$, 여기서 m은 양의 정수이다. U_s^N를 계산하기 위하여 아래와 같이 $P\left(b_n(U_s^N) = b\right)$의 n번째 block $b_n(U_s^N)$를 $\{\sum_n\}$의 block 으로 나타내면 다음과 같이 된다.

$$b_n(U_s^N) = [U_{Ln}, U_{Ln+1}, \cdots, U_{Ln+L-1}]$$

$$= \left[\sum mnM, \sum (mn+1)M, \cdots, \sum (mn+m-1)M\right].$$

기호의 편의를 위하여 $\Gamma_n = \left[\sum_{mnM}, \sum_{(mn+1)M}, \cdots, \sum_{(mn+m-1)M}\right]$ 이라고 두자. $b \in B^N$ 일때 b의 n번째 block $(b_{Ln}, b_{Ln+1}, \cdots., b_{Ln+L-1})$을 다시 크기 M 인 m개의 block으로 나누자. 즉

$$(b_{Ln}, \cdots, b_{Ln+L-1}) = (B_{n0}, \cdots, B_{nM-1})$$

여기서 $B_{nk} = (b_{Ln+kM}, b_{Ln+kM+1}, \cdots, b_{Ln+(k+1)M-1})$이다. m비트 B_{nk}에 대응하는 십진수 (양의 정수)를 i_k 라 하면

$$\{b_n(U_s^N) = b\} = \{\Gamma_n(i_0, i_1, \cdots, i_{m-1})\}$$
$$= \{\sum mnM = i_0, \sum (mn+1)M = i_1, \cdots, \sum (mn+m-1)M = i_{m-1}\}$$

이다. 그러므로

$$P_r\left(b_n(U_s^N) = b\right) = P(\{\Gamma_n = (i_0, i_1, \cdots, i_{m-1})\})$$
$$= P(\sum mnM = i_0, \cdots, \sum (mn+m-1)M = i_{m-1})$$
$$= P(\sum mnM = i_0)P^M(i_1, i_2)\cdots P^M(i_{m-2}, i_{m-1})$$

을 얻는다. 여기서 $P^M(i,j)$는 Markov chain $\{\sum_n\}$의 M-step transition matrix의 (i,j) 성분이다. $\{\sum_n\}$의 stationary probability $\lim_{n\to\infty} P(\sum_n = i) = P(i)$ 일때, $Q \to \infty$이면 $P(\sum_{mnM} = i_0) \to P(i_0)$, $n \geq Q$ 이다. 그러므로 $Q \to \infty$ 일때

$$P_r(b_n(U_s^N) = b) = P(i_0)P^M(i_0, i_1)P^M(i_1, i_2)\cdots P^M(i_{m-2}, i_{m-1}). \qquad (4.2)$$

이다. 또한 $j > k \geq Q$ 일때, 마르코프의 성질에 의해서

$$P(b_j(U_s^N) = b | b_k(U_s^N) = b)$$
$$= P(\sum mLM = i_0, \sum (mj+1)M = i_0, \cdots,$$
$$\sum (mj+m-1)M = i_{m-1} | \sum (mj+m-1)M = i_{m-1}) \qquad (4.3)$$
$$= P^{M(m(j-k-1)+1)}(i_{m-1}, i_0)P^M(i_0, i_1)\cdots P^M(i_{m-L}, i_{m-1})$$
$$= P^{M(m(l-k-1)+1)}(i_{m-1}, i)P^M(i_0, i_1)\cdots P^M(i_{m-2}, i_{m-1})$$

을 얻는다. (4.3)식에서 $j-k$ 혹은 M 또는 m이 충분히 크면 $M(m(j-k-1)+1)$도 충분히 커지므로 $P^{M(m(j-k-1)+1)}(i_{m-1}, i_0) \approx P(i_0)$가 된다.

일반적으로 $Q \le n_1 < n_2 < \cdots < n_k$ 일때,

$$P(b_{n_1}(U_s^N) = b, b_{n_2}(U_s^N) = b, \cdots, b_{n_k}(U_s^N) = b)$$
$$= P(b_{n_k}(U_s^N) = b | b_{n_{k-1}}(U_s^N) = b) \cdots P(b_{n_2}(U_s^N) = b | b_{n_1}(U_s^N) = b) P(b_{n_1}(U_s^N) = b)$$

가 되며 이식의 오른편은 (4.2)와 (4.3)결과를 이용하여 계산할 수 있다. 또한

$$\sum_{b \in B^n} P(b_{n_1}(U_s^N) = b, b_{n_2}(U_s^N) = b, \cdots b_{n_k}(U_s^N) = b)$$
$$= \sum_{i_1=0}^{2^M-1} \sum_{i_2=0}^{2^M-1} \cdots \sum_{i_m=0}^{2^M-1} P(\Gamma_{n_1} = (i_1, i_2, \cdots, i_m), \cdots \Gamma_{n_k} = (i_1, i_2, \cdots i_m))$$

이므로 확률 $P_r(A_n(U_s^N) = i)$를 (4.3) - (4.7) 로부터 계산할 수 있다. $Q \to \infty$이면 $E(\log_2 A_n(U_s^N$는 n에 의존하지 않는다. 그러므로

$$E(A_n(U_s^N)) = E(\log_2 A_n(U_s^N)) = \sum_{i=1}^{\infty} P_r(\log_2 A_n(U_s^N) = i) \log_2 i$$

이고

$$K \cdot \mathrm{Var}(T(U_s^N)) = \mathrm{Var}(\log_2 A_n(U_s^N))$$
$$= \sum_{i=1}^{\infty} P_r(\log_2 A_n(U_s^N) = i)(\log_2 i)^2 - [E(T(U_s^N))]^2$$

이다. $\mathrm{Var}(T(U_s^N))$를 계산할 때 우리는 $A_n(U_s^N)$가 서로 독립이라고 가정하였다.

제 5 절 RANDOMNESS TEST SIMULATION 및 IMPLEMENTATION

이진 난수 발생기 (Random bit generator)에서 생성된 비트열의 randomness를 검정하는 데는 2가지의 전략이 있다. 하나는 완전한 난수 발생기가 가지고있는 어떤 특성에서의 일탈(예를들어 비트간의 독립성 결여, 혹은 0과1의 빈도의 편향성 등)을 발견하면 생성자를 기각하는것이고 다른 하나는 생성자의 암호학적인 약점이 어떤 기준을 넘게되면 생성자를 기각하는것이다. 이 절에서는 이진 난수 발생기 (Random bit generator) 에서 생성된 비트열에대하여 각항사이의 독립성을 검정하고 (생성자의 memory size를 계산, 계열검정이용) 동시에 빈도의 안정성을 검정 (도수검정이용) 한다. 위의 검정을 통과하였을때 암호학적인 약점을 검정하기위하여 entropy관점에서의 randomness를 검정한다.

(1) 난수 생성자

모의 실험을 위하여 우리는 다음과 같은 종류의 난수 생성자를 사용하였다.

(i) 선형합동법 (Linear congruential generator)

(ii) 선형귀환 쉬프트 레지스터 (Linear feedback shift register)

(iii) 비선형 알고리즘

(a) 승산 시스템 (Multiplication system)

(b) J-K 플립플롭 시스템 (J-K Flip Flop system)

(c) Geffe 시스템

(d) 상호대칭 시스템

(i) 선형 합동법

Linear congruential generator는 다음과 같이 정의된다.

$$x_i = (ax_{i-1} + c) \mod M$$

여기서 승수 a,쉬프트 (shift) c, 모듈러스 (modulus) M 은 정수이다. 우리는 $a = 7^5 = 16807$, $c = 0$, $M = 2^{31} - 1 = 2147483647$ 를 선택했다. 이 생성자의 주기는 $2^{31} - 2$이다. x_i가 $\frac{M}{2}$ 보다 크면 이진 비트열 $\{b_i\}$ 의 선택에 있어서 $b_i = 1$로 둔다.

(ii) 선형귀환 쉬프트 레지스터 (Linear feedback shift register)

쉬프트 레지스터에 의해 생성되는 비트열은

$$b_i = (a_1 b_{i-1} + \cdots + a_d b_{i-d}) \mod 2 \tag{5.1}$$

에 의해 생성된다.

(5.1)에 다항식 $f(x) = x^d + a_1 x^{d-1} + \cdots + a_d$ 이 대응된다. 우리는

$$f(x) = x^p + x^q + 1 \tag{5.2}$$

과 같은 형태의 다항식을 사용한다.

특성 다항식 (5.2)에 대응하는 쉬프트 레지스터를 이용하에 비트열을 생성하는 알고리즘은 다음과 같다.

Algorithm 5.1

1. $Y \leftarrow X$ (X 는 $b_{i+p-1}, b_{i+p-2} \cdots b_i$의 형태로 되어있다.)
2. Y를 q bit 만큼 오른쪽 으로 이동 시키고 빈자리는 0 으로 채운다 .
3. $Y \leftarrow X \leftarrow Y$ XOR X (여기서 XOR 은 exclusive OR 연산을 의미 한다.)
4. Y를 $p-q$ 비트만큼 왼쪽으로 이동시키고 빈자리는 0 으로 채운다.
5. $X \leftarrow Y$ XOR X (X는 다시 $b_{i+2p-1}\ b_{i+2p-2} \cdots b_{i+p}$로 구성된다.)

쉬프트 레지스터에 의해 생성되는 비트열의 주기는 $2^p - 1$이하이다. 최대주기 $2^p - 1$이 되도록 하는 (p, q)쌍의 예를 들어 보면 다음과 같다.

p	q	p	q
15	1, 4, 7, 8, 11, 14	31	3, 6, 7, 13, 18, 24, 25, 28
17	3, 5, 6, 11, 12, 14	33	13, 20
18	7, 11	35	2, 33
20	3, 17	36	11, 25
21	2, 19	89	38, 51
22	1, 21	98	27, 71
23	5, 9, 14, 18	521	32, 439
25	3, 7, 18, 22	607	273, 334
28	3, 9, 13, 15, 19, 25		
29	2, 27		

이진 수열 $\{b_i\}$의 주기가 최대주기인 $2^p - 1$이 되는 선형 쉬프트 레지스터를, 최대주기를 갖는 선형 쉬프트 레지스터라 하고 m-LFSR (maximum length Linear Feedback shift Register) 이라고 정의 한다.

(iii) 비선형 알고리즘

앞에서 언급한 m-LFSR한개로 구성된 생성자의 약점을 보완하여 몇 개의 m-LFSR을 비선형 논리구조로 결합하여 난수 생성자로 이용하는 방법이 많이 제안되었다.

(a) 승산(multiplication)시스템

m-LFSR 2개의 출력를 서로 곱하여 최종의 출력수열을 발생하는 시스템을 승산시스템이라고 한다. m-LFSR1의 출력수열이 $\{a_i\}$이고 m-LFSR2 의 출력수열이 $\{b_i\}$ 인 경우 승산시스템의 출력수열 $\{c_i\}$는

$$c_i = a_i \times b_i$$

가 된다.

(b) J-K 플립-플롭 (J-K Flip-Flop)

m-LFSR 2개의 출력을 J-K 플립-플롭에 의해 조합하여 출력수열을 발생 하는 시스템이다. m-LFSR 1의 출력수열이 $\{a_i\}$이고 m-LFSR 2의 출력수열이 $\{b_i\}$ 일 때 J-K 플립-플롭 시스템의 출력수열 $\{c_i\}$는

$$c_i = ((a_i + b_i + 1)c_{i-1} + a_i) \mod 2, \qquad c_0 = 0$$

가 된다.

(c) Geffe 시스템

Geffe 시스템은 3개의 m-LFSR로 구성된다. m-LFSR1, m-LFSR2, m-LFSR3의 출력수열을 각각 $\{a_i\}$, $\{b_i\}$, $\{c_i\}$라고 하면 Geffe 시스템의 출력수열 $\{g_i\}$는 다음과 같이 생성된다.

$$g_i = a_i b_i \oplus c_i b_i \oplus c_i$$

m-LFSR1, m-LFSR2, m-LFSR3의 차수가 m, n, k이고 각 쌍마다 서로소가 될때 Geffe 시스템에서 발생되는 출력수열의 주기는 $(2^m - 1)(2^n - 1)(2^k - 1)$이 된다.

(d)상호 대칭 시스템

Geffe 시스템과 마찬가지로 3개의 m-LFSR로 구성되며 m-LFSR1, m-LFSR2, m-LFSR3의 출력수열을 각 각 $\{a_i\}$, $\{b_i\}$, $\{c_i\}$라고 하면 시스템의 출력수열 $\{s_i\}$는

$$s_i = a_i b_i \oplus b_i c_i \oplus c_i a_i$$

가 된다. m-LFSR1, m-LFSR2, m-LFSR3의 차수가 각각 m, n, k이고 서로소일 때 출력되는 수열의 주기는 $(2^m-1)(2^n-1)(2^k-1)$이 된다.

(2) Randomness 검정법

본절에서 Random test시뮬레이션에 사용된 각종검정들을 간략히 언급한다.

(i) Memoryless 검정 (제 3 절의 (3.16))

(ii) Serial 검정

(a) 연속된 2비트 u_0u_1, u_1, u_2, $\cdots$, $u_{N-1}u_{N-1}$로 분할

(b) n_{00}: "00"의 수

n_{01}: "01"의 수

n_{10}: "10"의 수

n_{11}: "11"의 수

n_0: "0"의 수

n_1: "1"의 수

(iii) 도수검정

n 비트 이진수열 $< x_0, x_1, \cdots, x_{n-1} >$에서 x_n을 "0"인 항의 개수, n_1을 "1"인 항의 개수.

$$\chi^2 = \frac{(n_0 - n_1)}{n}$$

(3) 결과 및 논의

우리는 $BMS_{\frac{1}{2}}$로 모형화 할 수 있는 몇개의 2진 난수 발생기에 대하여 entropy에 의한 randomness를 검정하였다. 실험의 결과를 <표1 >에 나타내었다. Table에서 난수 생성자 1은 선

형합동법(Linear congruential generator)를, 2는 선형귀한 쉬프터 레지스터를, 3은 승산 시스템을, 4는 J-K 플립플롭을, 5는 Geffe시스템을, 6은 상호대칭 시스템을 나타낸다.생성자 2에 해당하는 다항식은 (5.2)이며 이 때 $p = 28$, $q = 9$를 사용하였다. 이 때 생성되는 비트열의 주기는 $2^{28} - 1 = 268435455$이다. 생성자 3과 4에서 사용된 선형귀환 쉬프트 레지스터는 $(p, q) = (31, 13)$과 $(p, q) = (33, 13)$인 경우이며 생성자 5와 6에 사용된 선형귀환 쉬프트 레지스터는 $(p, q) = (31, 13)$ $(p.q) = (33, 13)$ 그리고 $(p.q) = (28, 9)$인 경우이다. 검정의 결과 승산 시스템은 독립성 검정인 계열검정정과 빈도의 안정성검정인 도수검정을 통과하지 못하므로 entropy에 의한 검정에서 제외하였다. 그 밖의 난수 생성자는 1% 유의수준에서 연속되는 비트간의 독립성 검정인 Memoryless검정, 계열검정을 통과하고 빈도 의 안정성에 관한 검정인 도수검정을 통과하였으므로 $BMS_{\frac{1}{2}}$로 모형화하여 entropy에 의한 randomness검정을 하였다.

Block의 크기는 8부터 16까지에 대하여 적용하였으며초기화를 위한 block의 수 Q는 L이 8과 11 사이에 있을 때는 $Q = 30 \cdot 2^L$ 개를 L이 12와 16사이에 있을 때는 $Q = 5 \cdot 2^L$개를 사용하였다.검정을 위한 block의 수는 $K = 10000$개를 사용하였다.그러므로 각각의 경우 사용된 총 비트의 수는 $N = (Q + K)L$이다. 이 때 승산 시스템을 제외한 모든 난수발생기가 1% 유의수준에서 entropy에 의한 검정을 통과하였다. 기존의 전통적인 통계적 검정방법에 비하여 entropy에 의한 검정방법은 암호학적으로 중요한 측도인 비트당 entropy와 관련이 있다는 점에서 새로운 randomness의 검정법으로 타당하다.

참 고 문 헌

[1] H. Beker and F. Piper, Ciper Systems - The Protection of Communications, John Wily and Sons, 1982.

[2] P. Bratky, B. L. Fox and L. F. Schrage, A Guide to Simulation, Springer-Verlag, 1983.

[3] S. W. Golombo, Shift Register Sequences, Holden - Day,1967.

[4] I. J. Good, On the serial tests for random sequences, Ann. Math. Statist., 28, 262 -264, 1967.

[5] A. I. Khinchin, Mathematical Fundation of Information Theory, DoverPubulication, Inc, 1957.

[6] D. E. Knuth, The Art of Computer Programming, Vol. 2 Semi Numerical Algorithms, Addison - Wesley Publishing Company, 1981.

[7] A. N. Kolmogorv and V. A. Uspenskii, Algorithms and Randomness, Theory Prob. Appl. vol. 32, No. 3, 389 - 412.

[8] J. Lehoczky, Statistical Methods, Handbooks in Operations Research and Managenent Science, Vol. 2, Stochastic Models (ed) by D. P. Heyman & M. J. Sobel, pp 255–294, North-Holland, 1990.

[9] U. M. Maurer, A universal statistical test for random bit generators, Crypto '90, 401 - 413.

[10] A. M. Mood, The distribution theory of runs, Ann. Math. Statist., 11, 367 - 392, 1940.

[11] B. Riphey, Stochastic Simulation, John Wiley & Sons, 1987.

12] C. E. Shanon, A mathematical theroy of communication, Bell Syst. Tech. J. vol. 27, 379 -423, 623 - 656, 1948.

[13] J. M. Wozen Craft and B. Reiffen, Sequential Decoding, Cambridge, MA. Techn. Press of the MIT, 1960.

[14] 현대 암호학, 한국전자통신연구소편저, 1991.

감사의 글 : 본 논문은 과학 기술처의 출연금에 의하여 이루어 졌습니다.