2022 국민대학교 정보보안암호수학과 학술동아리 주관 경진대회

1번 문제 - 난수발생기의 엔트로피 (출제: 동아리 RB)

[배경 지식1]

공개키 방식의 키 생성, 블록암호 운영모드, 양자내성암호 등에서 쓰이는 난수는 보안의 필수적인 요소 중 하나이다. 따라서, 난수발생기의 안전성을 평가하는 것은 필수적이며 이를 위한 측정도구로 Min Entropy(앞으로 최소 엔트로피라 표현)라는 개념을 사용한다. 난수발생기 X가 n개의 원소 x_i를 출력할 수 있고 각각 출력될 확률이 $p_i = p(x_i)$일 때 최소 엔트로피는 다음과 같이 정의된다.

$$\text{최소 엔트로피} \rightarrow H(X) := -\log_2 \max\{p_i\}.$$

난수발생기의 출력 가능 값이 n개일 때 최소 엔트로피는 $0 \le H(X) \le \log_2 n$ 사이의 값을 갖는다. 난수발생기 X가 n개의 원소 x_i를 출력할 때 $p(x_k) = 1$인 x_k가 존재할 경우(즉, x_k만 출력하는 난수 발생기) 최소 엔트로피는 0이 되고

모든 x_i에 대하여 $p(x_i) = \frac{1}{n}$ 일 경우(즉, 균등분포일 경우) 최소 엔트로피는 $\log_2 n$이 된다.

[예시1]

난수발생기 X의 출력이 $\{a, b, c, d\}$ 중 하나이고, X의 출력이 a, b, c, d일 확률이 각각 $p(a) = \frac{1}{8}$, $p(b) = \frac{1}{8}$, $p(c) = \frac{1}{4}$, $p(d) = \frac{1}{2}$일 때, X의 최소 엔트로피는

$$H(X) := -\log_2 \max\{p(a), p(b), p(c), p(d)\} = -\log_2 p(d) = -\log_2 \frac{1}{2} = 1$$

이다.

[예시2]

난수발생기 Y의 출력이 $\{a, b, c, d\}$ 중 하나이고, Y의 출력이 a, b, c, d일 확률이 각각 $p(a) = \frac{1}{4}$, $p(b) = \frac{1}{4}$, $p(c) = \frac{1}{4}$, $p(d) = \frac{1}{4}$일 때, Y의 최소 엔트로피는

$$H(Y) := -\log_2 \max\{p(a), p(b), p(c), p(d)\} = -\log_2 p(d) = -\log_2 \frac{1}{4} = 2$$

이다.

[예시3]

난수발생기 Z의 출력이 $\{a, b\}$ 중 하나이고, Z의 출력이 a, b일 확률이 각각 $p(a) = \frac{1}{2}$, $p(b) = \frac{1}{2}$일 때, Z의 최소 엔트로피는

$$H(Z) := -\log_2 \max\{p(a), p(b)\} = -\log_2 (p(b)) = -\log_2 (\frac{1}{2}) = 1$$

이다.

[예시 결과 해석]

X와 Y는 동일한 개수의 출력을 가지지만 Y의 출력이 균등하기 때문에 Y가 최소 엔트로피 관점에서 더 좋은 난수를 출력하는 난수발생기라 할 수 있다.

X가 Z보다 출력할 수 있는 난수값이 많은데도 불구하고 엔트로피 값이 같은 이유는 X의 분포가 균일하지 않고 이것이 보안강도를 떨어뜨리기 때문이다.

[배경 지식2]

폰노이만 교정자(von Neumann Corrector)란 "난수발생기 X가 0을 출력할 확률이 p_0, 1을 출력할 확률이 p_1일 때 연속된 두 비트가 "01"일 확률과 "10"일 확률은 p_0p_1으로 같은 점"을 이용하여 불균등한 분포를 가지는 이진난수발생기의 출력을 균등분포로 만드는 기법이다.

이 교정자는 난수발생기가 $2n$ 비트를 출력하면 n개의 연속된 두 비트 값을 본 뒤 "11"이거나 "00"인 경우는 버리고 "01"인 경우엔 0을 할당, "10"인 경우엔 1을 할당한다.

예를 들어 $p(0)=\frac{1}{4}$, $p(1)=\frac{3}{4}$ 인 난수발생기 X가 10비트 "1011011101"을 출력했을 경우, 연속된 두비트는 5개이고 앞에서부터 "10", "11", "01", "11", "01"이다. 폰 노이만 교정자를 사용할 경우 "11"은 제거되고 "01"에 0, "10"에 1을 할당하여 새로운 난수열 "100"을 얻게 된다. 결론적으로 불균등한 분포로부터 얻은 10비트를 사용하여 균등분포로 출력된 3비트를 얻은 것으로 볼 수 있다.

[배경 지식3]

엔트로피는 또한 정보량을 의미하기도 하는데 여기서 정보량이란 앞으로 새롭게 만들어 낼 수 있는 정보량을 뜻한다. 예를 들어 난수발생기에서 뽑은 10비트가 0과 1이 각각 5개라는 정보가 주어졌다고 가정하자. 이때 가능한 경우의 수는 ${}_{10}C_5$가지이다. 반면에 난수발생기로 뽑은 10비트 중 0이 1개고 나머진 1이라는 정보가 주어졌다면 가능한 경우의 수는 ${}_{10}C_1$로 10가지이다. 따라서 후자가 주어진 정보의 가치는 더 크지만 앞으로 새롭게 얻을 수 있는 정보는 더 적으므로 엔트로피가 낮은 것으로 생각할 수 있다. 마찬가지로 난수발생기의 분포가 균일할수록 비트는 예측하기 어려워지고 앞으로 얻어낼 수 있는 정보가 많은 것으로 생각할 수 있다. 따라서 난수발생기의 분포가 균등분포일 때 엔트로피는 최대가 되고 분포가 불균일할수록 엔트로피는 낮아진다.

[문제 1]

난수발생기 X의 출력 가능 값이 $a_i\ (1 \le i \le 10)$이고 $p(a_i) = \frac{i}{55}$일 때 $H(X)$를 구하시오. (25pt)

[문제 2]

이진난수발생기 Y가 난수열 "01010001101100"을 출력했을 때 폰 노이만 교정자에 의해 생긴 새로운 난수열을 구하시오. (25pt)

[문제 3]

$p(0) = p$, $p(1) = 1 - p$인 ($0 \le p \le 1$) 이진난수발생기 Y의 출력을 폰 노이만 교정자로 교정한 뒤 출력하는 난수발생기 Z가 있다. Y로 $2n$비트를 생성할 때 최종 출력 Z의 길이의 기댓값 E_n을 구하고 Y1비트당 Z의 출력 길이 $l(\lim_{n\to\infty} \frac{E_n}{2n})$을 구하시오. (25pt)

[문제 3]

부등식 $l \le H(Y)$을 증명하고 그 의미를 엔트로피와 연관지어 설명하시오.(l, Y 는 3번에서 쓰인 l, Y임) (25pt)

문의 : ofryuji@kookmin.ac.kr