擬似乱数の識別不可能性概念の相互関係について[1]

縫田　光司（NUIDA, Koji）[2]

k.nuida@aist.go.jp

二つの有限集合 X, Y の間の写像 $G\colon X \to Y$ を、一様ランダムな入力 $x \in X$ に応じて何らかの意味で「ランダム」な出力 $G(x) \in Y$ を生成する擬似乱数生成器と考えたい。暗号理論においてこの G に課される代表的な条件は G の出力分布の「識別不可能性」であり、例えば以下のように定式化される。

定義 1. T と ε を正の実数とする。時間 T 以内で停止する任意の 1 ビット出力（確率的）アルゴリズム $D\colon Y \to \{0,1\}$ に対して、不等式

$$\mathrm{Adv}_G(D) := Pr[D(G(x)) = 1] - Pr[D(y) = 1] \leq \varepsilon \qquad (1)$$

（ただし、$x \in X$ と $y \in Y$ はともに一様分布に従って選ばれるとする）が常に成り立つとき、G は (T, ε)-識別不可能であると定義する。

上のアルゴリズム D（擬似乱数生成器 G に対する識別者と称される）の出力 0 を「真の乱数」、1 を「擬似乱数」と対応付けて、G の出力もしくは Y 上の一様分布に従って選ばれた元のいずれかを D に渡したとき、それが前者（擬似乱数）であるか後者（真の乱数）であるかを正しく判定する確率が、でたらめに半々の確率で判定するのと比べて殆ど改善できない（つまり、真の乱数と見分けることすらできないほど G の出力分布はランダムである）というのが G の識別不可能性の直感的な解釈の一つである。さて、定義 1 では識別者 D として 1 ビット出力のものだけを考慮したが、より木目細かく多ビット出力を行う識別者 $D\colon Y \to \{0,1\}^n$ を考慮することもできる。そうして得られる識別不可能性の定義は以下のようになる [1][3]。

定義 2. T と ε を正の実数、n を正の整数とする。時間 T 以内で停止し、出力が n ビット以下であるような任意の（確率的）アルゴリズム $D\colon Y \to \{0,1\}^m$ $(1 \leq m \leq n)$ に対して、不等式

$$\mathrm{Adv}_G(D) := \frac{1}{2} \sum_{z \in \{0,1\}^m} |Pr[D(G(x)) = z] - Pr[D(y) = z]| \leq \varepsilon \qquad (2)$$

（ただし、$x \in X$ と $y \in Y$ はともに一様分布に従って選ばれるとする）が常に成り立つとき、G は (T, n, ε)-識別不可能であると定義する。

[1]本発表は、京都大学の阿部拓郎・前野俊昭両氏、山口大学の鍛冶静雄氏、東京大学/JST CREST の沼田泰英氏との共同研究に基づく。

[2]産業技術総合研究所 (AIST) 情報セキュリティ研究センター (RCIS)

[3]原論文では、個々の G ではなく G のパラメータを変化させて得られる擬似乱数生成器の族に対して識別不可能性を定義している（し、実はそういう定義の方が現在の暗号理論では一般的である）が、本発表では話を簡単にするために個々の G に着目した定義を採用する。

$m = 1$ のときは定義 1 と定義 2 に現れる量 $\mathrm{Adv}_G(D)$（の絶対値）が一致するため、G が (T, n, ε)-識別不可能であれば (T, ε)-識別不可能でもある。逆に、G が (T, ε)-識別不可能であれば、T と ε と n から定まる何らかのパラメータ T' と ε' について (T', n, ε')-識別不可能でもある、という方向の関係は存在するだろうか。この問題について、話者らの研究 [2] で以下の結果を得た。

$\mathcal{C}$ を $\{0,1\}^n$ の冪集合、$\mathcal{C}' \subset \mathcal{C}$ とする。$Z_1, Z_2 \in \mathcal{C}$ について、集合 Z_1 と Z_2 の対称差の濃度を $d(Z_1, Z_2)$ と定めると、d は $\mathcal{C}$ 上の距離となる。ここで、「$\mathcal{C}'$ を中心とする $\mathcal{C}$ の半径」を r と置く：

$$r = r(\mathcal{C}; \mathcal{C}') := \max_{Z \in \mathcal{C}} \min_{Z' \in \mathcal{C}'} d(Z, Z') \ . \tag{3}$$

一方、$z \in \{0,1\}^n$ について、$\{z\}$ の特性関数 $\chi_z = \chi_{\{z\}} \colon \{0,1\}^n \to \{0,1\}$ を識別者 $D \colon Y \to \{0,1\}^n$ と合成する際の計算量の増分（直感的には、大体 χ_z 自体の計算量）の上界を δ と置く。同様に、$Z \in \mathcal{C}'$ の特性関数 $\chi_Z \colon \{0,1\}^n \to \{0,1\}$ を識別者 D と合成する際の計算量の増分（直感的には、大体 χ_Z 自体の計算量）の上界を δ' と置く。この状況で以下の関係が成り立つ。

定理 1 ([2])**.** 上の状況で、G が $(T + \delta, \varepsilon)$-識別不可能かつ $(T + \delta', \varepsilon')$-識別不可能ならば、$G$ は $(T, n, r\varepsilon + \varepsilon')$-識別不可能である。

$\delta \approx 0$ と考えられるので、あとは δ' が充分小さくなるように $\mathcal{C}' \subset \mathcal{C}$ を選んだ場合に、「半径」r があまり大きくならなければ、定義 1 の識別不可能性から定義 2 の識別不可能性を導く際のパラメータに関する損失があまり大きくならずに済むことになる。従って、この値 r を見積もることが重要である。ここで、$\mathcal{C}$ の元 Z をその特性関数 χ_Z と同一視すると、値 r を見積もる問題は以下のように一般化することができる：

> 関数の集合 $\mathcal{C}$ とその部分集合 $\mathcal{C}'$（および二つの関数 $f, g \in \mathcal{C}$ の「距離」$d(f, g)$ の定義）が与えられたとき、(3) 式で定義される「$\mathcal{C}'$ を中心とする $\mathcal{C}$ の半径」の見積もりを与えよ。

この一般化された問題（「関数密度問題」と呼ぶことにする）には、上記の擬似乱数の理論以外にも、ハッシュ関数の安全性評価など情報セキュリティ分野への応用が考えられる [2] ため、今後の数学的な議論の展開を期待する。

参考文献

[1] B. Dubrov, Y. Ishai, “On the randomness complexity of efficient sampling,” Proceedings of STOC 2006, pp.711–720 (2006)

[2] K. Nuida, T. Abe, S. Kaji, T. Maeno, Y. Numata, “A mathematical problem for security analysis of hash functions and pseudorandom generators,” to appear in IWSEC 2011, November 9, 2011