Almost k-wise Independent Sample Spaces and Their Cryptologic Applications *

Kaoru Kurosawa
Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology
2–12–1 O-okayama, Meguro-ku, Tokyo 152-8552, Japan
kurosawa@ss.titech.ac.jp

Thomas Johansson
Department of Information Technology, Lund University,
PO Box 118, S-22100 Lund, Sweden
thomas@it.lth.se

Douglas Stinson
Department of Combinatorics and Optimization
University of Waterloo
Waterloo Ontario, N2L 3G1, Canada
dstinson@uwaterloo.ca

September 23, 1999

Abstract

An almost k-wise independent sample space is a small subset of m bit sequences in which any k bits are "almost independent". We show that this idea has close relationships with useful cryptologic notions such as multiple authentication codes (multiple A-codes), almost strongly universal hash families, almost k-resilient functions, almost correlation-immune functions, indistinguishable random variables, and k-wise decorrelation bias of block ciphers.

We use almost k-wise independent sample spaces to construct new efficient multiple A-codes such that the number of key bits grows linearly as a function of k (where k is the number of messages to be authenticated with a single key). This improves on the construction of Atici and Stinson [2], in which the number of key bits is $\Omega(k^2)$.

We introduce the concepts of ϵ-almost k-resilient functions and almost correlation-immune functions, and give a construction for almost k-resilient functions that has parameters superior to k-resilient functions. We also point out the connection between almost k-wise independent sample spaces and pseudorandom functions that can be distinguished from truly random functions, by a distinguisher limited to k oracle queries, with only a small probability. Vaudenay [31] has shown that such functions can be used to construct block ciphers with a small decorrelation bias.

*A preliminary version of this paper was presented at EUROCRYPT '97 and appeared in *Lecture Notes in Computer Science* **1233** (1997), 409–421.

Finally, new bounds (necessary conditions) are derived for almost k-wise independent sample spaces, multiple A-codes and balanced ϵ-almost k-resilient functions.

1 Introduction

An *almost k-wise independent sample space* is a probability space on m-bit sequences such that any k bits are almost independent. A *ϵ-biased sample space* is a space in which any (boolean) linear combination of the m bits takes the value 1 with probability close to 1/2. These notions were introduced by Naor and Naor [22] and further studied in [1] due to their applications to algorithms and complexity theory. However, there are also cryptographic applications: Krawczyk applied ϵ-biased sample spaces to the construction of authentication codes [18].

In this paper, we investigate several new relationships between almost k-wise independent sample spaces and useful cryptologic notions such as multiple authentication codes (multiple A-codes) [32, 20, 2]; k-resilient and correlation-immune functions [14, 3, 15, 29, 4]; and indistinguishable random variables and their application to k-wise decorrelation bias of block ciphers [31]. We begin our study with a summary of basic definitions and results on almost k-wise independent sample spaces in Section 1.1.

In Section 2, we study multiple A-codes. In a multiple A-code, $k \geq 2$ messages are authenticated with the same key. (In "usual" A-codes, just one message is authenticated with a given key.) Recently, Atici and Stinson [2] defined some new classes of almost strongly universal hash families which allowed the construction of multiple A-codes. Here, we prove that almost k-wise independent sample spaces are equivalent to multiple A-codes. This allows us to obtain a more efficient construction of multiple A-codes from the almost k-wise independent sample spaces of [1].

In Section 3, we present a lower bound on the size of the keyspace in a multiple A-code. Numerical examples show that the multiple A-codes we construct are quite close to this bound. Further, from the above equivalence, a lower bound on the size of almost k-wise independent sample spaces is obtained for free. (While a lower bound on the size of ϵ-biased sample spaces was given in [1], no lower bound was known for the size of almost k-wise independent sample spaces.)

In Section 4, we generalize the idea of resilient functions. A function $\phi : \{0,1\}^m \rightarrow \{0,1\}^l$ is called *k-resilient* if every possible output l-tuple is equally likely to occur when the values of k arbitrary inputs are fixed by an opponent and the remaining $m-k$ input bits are chosen at random. This is a useful tool for achieving key renewal: an m-bit secret key $(x_1, \cdots, x_m)$ can be renewed to a new l-bit secret key $\phi(x_1, \cdots, x_m)$ about which an opponent has no information if the opponent knows at most k bits of $(x_1, \cdots, x_m)$. We show that k can be made larger if the definition of resilient function is slightly relaxed. Thus, we define an ϵ-almost k-resilient function as a function ϕ such that every possible output l-tuple is almost equally likely to occur when the values of k arbitrary inputs are fixed by an opponent. (The statistical difference between the output distribution of a k-resilient function and an ϵ-almost k-resilient function is bounded above by ϵ.) We prove that a large set of almost k-wise independent sample spaces is equivalent to a balanced ϵ-almost k-resilient function, generalizing a result of [29]. (A similar result is shown for correlation-immune functions.) From this equivalence, we are able to obtain both efficient constructions and bounds for balanced ϵ-almost k-resilient functions.

Finally, in Section 5, we point out the connection between almost k-wise independent sample spaces and pseudorandom functions that can be distinguished from truly random functions, by a distinguisher limited to k oracle queries, with only a small probability. Vaudenay [31] has shown that such functions can be used to construct block ciphers with a small decorrelation bias. Thus almost k-wise independent sample spaces can potentially be used as round functions for Feistel type ciphers.

1.1 Almost k-wise independent sample spaces

In this paper, a *sample space* will be a set of binary m-tuples $S_m \subseteq \{0,1\}^m$. A sample space is *linear* if it is a subspace of the vector space $\{0,1\}^m$. If $S_m \subseteq \{0,1\}^m$ is a sample space, we will let $X = x_1 \cdots x_m$ be the random variable obtained by choosing each m-tuple from S_m with the same probability $1/|S_m|$.

Definition 1.1 [1] A sample space $S_m \subseteq \{0,1\}^m$ is (ϵ, k)-*independent* if, for any k positions $i_1 < i_2 < \cdots < i_k$ and any k-bit string α, it holds that

$$|\Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-k}| \leq \epsilon. \tag{1}$$

A $(0,k)$-independent sample space is said to be k-independent.

k-independent sample spaces are also known as *perfect local randomizers*. These objects were introduced by Schnorr [25] and further studied in [21, 23]. They are in fact equivalent to certain combinatorial structures which we define now.

Definition 1.2 An *orthogonal array* $\mathrm{OA}_\lambda(t, m, v)$ is a λv^t by m array on v symbols, say M, such that, within any t columns of M, every t-tuple occurs in exactly λ rows.

The following observation is due to Maurer and Massey [21].

Theorem 1.1 *A k-independent sample space $S_m \subseteq \{0,1\}^m$ is equivalent to an orthogonal array $OA_\lambda(k, m, 2)$, where $\lambda = |S_m|/2^k$.*

It is well-known that k-independent sample spaces are "large". For example, the classical Rao bound for orthogonal arrays, together with Theorem 1.1, shows that $|S_m|$ is $\Omega(m^{k/2})$ if $S_m \subseteq \{0,1\}^m$ is k-independent. Thus several researchers have studied (ϵ, k)-independent sample spaces, where $\epsilon > 0$ and is $|S_m|$ is "small". See, for example, [1, 7, 8]. The following efficient construction for (ϵ, k)-independent sample spaces is proved by Alon, Goldreich, Hastad and Peralta.

Proposition 1.2 [1] *There exists an (ϵ, k)-independent sample space $S_m \subseteq \{0,1\}^m$ such that*

$$\log_2 |S_m| = 2(\log_2 \log_2 m - \log_2 \epsilon + \log_2 k - 1).$$

For completeness, we briefly review the main construction method for (ϵ, k)-independent sample spaces. We require another definition.

Definition 1.3 [22] A sample space $S_m \subseteq \{0,1\}^m$ is ϵ-*biased* if, for any $\alpha \in \{0,1\}^m$, it holds that

$$|\Pr[x \cdot \alpha = 1] - \Pr[x \cdot \alpha = 0]| \leq \epsilon, \tag{2}$$

where "$\cdot$" denotes the inner product modulo 2.

The following result is due to Naor and Naor.

Proposition 1.3 [22] *Suppose the following exist:*

1. *An ϵ-biased sample space $S \subseteq \{0,1\}^n$ with $|S| = 2^m$, and*
2. *A linear k-independent sample space $L \subseteq \{0,1\}^N$ with $|L| = 2^n$.*

Then there exists an (ϵ, k)-independent sample space $R \subseteq \{0,1\}^N$ with $|R| = 2^m$.

Appropriate ingredients required for Proposition 1.3 can be obtained as follows. Three explicit constructions for ϵ-biased sample spaces are presented in [1]. One of these, the "powering construction", yields an $n/2^r$-biased sample space $S \subseteq \{0,1\}^n$ with $|S| = 2^{2r}$ for integers n and r. (Note: It is observed in [7, 8] that this construction can be viewed as an application of Reed-Solomon codes, and more efficient variations can be obtained using algebraic geometry codes.)

The second ingredient is a linear k-independent sample space $L \subseteq \{0,1\}^N$ with $|L| = 2^n$. Such a sample space is a linear $\mathrm{OA}_{2^{n-k}}(k, N, 2)$. The orthogonal complement of this sample space is therefore an $[N, N-n, k+1]$-code, and conversely, from a code with these parameters we can construct the desired sample space L (see, for example, [19]). It is suggested in [22] to use a sample space of this type constructed from a BCH code. Suppose $N = 2^t - 1$ for some integer t, and suppose k is odd. Then the sample space L obtained by this method has $|L| = 2^{(k+1)t/2}$.

If we construct our ingredients as described above, and apply Proposition 1.3, then the resulting sample space has parameters as stated in Proposition 1.2.

2 Multiple A-codes and ASU-k hash families

In this section, we prove that almost k-wise independent sample spaces are equivalent to multiple authentication codes (more precisely, almost strongly universal-k hash families, as defined in [2]). This allows us to obtain more efficient multiple A-codes than were previously known.

First, we briefly review basic concepts of (multiple) authentication codes. In the usual Simmons model of authentication codes (A-codes) [26, 27], there are three participants, a *transmitter*, a *receiver* and an *opponent*. In an *A-code without secrecy*, the transmitter sends a *message* (s, a) to the receiver, where s is a *source state* (plaintext) and a is an *authenticator*. The authenticator is computed as $a = e(s)$, where e is a secret *key* shared between the transmitter and the receiver. The key e is chosen according to a specified probability distribution.

In a *multiple* A-code, we suppose that an opponent observes $i \geq 2$ messages which are sent using the same key. Then the opponent places a new bogus message (s', a') into the channel, where s' is distinct from the i source states already sent. This attack is called a *spoofing attack of order i*. P_{d_i} denotes the success probability of a spoofing attack of order i, see [20].

Almost strongly universal hash families are a very useful way of constructing practical A-codes. This idea was introduced by Wegman and Carter [32], and further developed and refined in papers such as [28, 6, 18, 17]. Atici and Stinson [2] generalized the definitions so that they could be applied to multiple A-codes. We review these definitions now.

Definition 2.1 An $(N; m, n)$ *hash family* is a set F of N functions such that $f : A \to B$ for each $f \in F$, where $|A| = m$, $|B| = n$ and $m > n$.

Definition 2.2 [2] An $(N; m, n)$ hash family F of functions from A to B is *ϵ-almost strongly universal-k* (and denoted ϵ-ASU $(N; m, n, k)$) provided that, for all distinct elements $x_1, x_2, \cdots, x_k \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_k \in B$, we have

$$|\{f \in F : f(x_i) = y_i, 1 \leq i \leq k\}| \leq \epsilon \times |\{f \in F : f(x_i) = y_i, 1 \leq i \leq k-1\}|.$$

The following result of Atici and Stinson gives the connection between ϵ-ASU $(N; m, n, k)$ hash families and multiple A-codes.

Proposition 2.1 [2] *There exists an A-code without secrecy for m source states, having n authenticators and N equiprobable authentication rules and such that $P_{d_{k-1}} \leq \epsilon$, if and only if there exists an ϵ-ASU $(N; m, n, k)$ hash family F.*

2.1 Equivalence of hash families and sample spaces

We can can rephrase Definition 1.1 in terms of hash families, and generalize it to the non-binary case, as follows.

Definition 2.3 An $(N; m, n)$ hash family F of functions from A to B is (ϵ, k)-*independent* if for all distinct elements $x_1, x_2, \cdots, x_k \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_k \in B$, we have

$$|\Pr(f(x_i) = y_i, 1 \leq i \leq k) - n^{-k}| \leq \epsilon, \tag{3}$$

where $f \in F$ is chosen uniformly at random.

The following results are straightforward.

Proposition 2.2 *An (ϵ, k)-independent sample space $S_m \subseteq \{0,1\}^m$ is equivalent to an (ϵ, k)-independent $(|S_m|; m, 2)$ hash family.*

Proof. Let $S_m \subseteq \{0,1\}^m$ be an (ϵ, k)-independent sample space. Define an $(|S_m|; m, 2)$ hash family F as follows: For each $x = (x_1, \cdots, x_m) \in S_m$, define a function $f_x : \{1, \cdots, m\} \rightarrow \{0,1\}$ by the rule $f_x(i) = x_i$. Then define $F = \{f_x : x \in S_m\}$. F is easily seen to be an (ϵ, k)-independent hash family.

Conversely, suppose F is an (ϵ, k)-independent $(N; m, 2)$ hash family, where without loss of generality, $f : \{1, \cdots, m\} \rightarrow \{0,1\}$ for each $f \in F$. For each $f \in F$, define $x_f = (f(1), \cdots, f(m))$, and define $S_m = \{x_f : f \in F\}$. Then $S_m \subseteq \{0,1\}^m$ is an (ϵ, k)-independent sample space with $|S_m| = N$. □

Proposition 2.3 *If there exists an (ϵ, k)-independent sample space $S_m \subseteq \{0,1\}^m$, then there exists an $(\epsilon, k/t)$-independent $(|S_m|; m/t, 2^t)$ hash family.*

Proof. Suppose S_m is the hypothesized sample space. By Proposition 2.2, there exists an (ϵ, k)-independent $(|S_m|; m, 2)$ hash family, say F, where $f : \{1, \cdots, m\} \rightarrow \{0,1\}$ for each $f \in F$. Let $m' = m/t$. For each $f \in F$, define a function $f' : \{1, \cdots, m'\} \rightarrow \{0,1\}^t$ by the following rule:

$$\begin{aligned}
f'(1) &= (f(1), \cdots, f(t)) \\
f'(2) &= (f(t+1), \cdots, f(2t)) \\
\vdots \quad &\vdots \quad \vdots \\
f'(m') &= (f((m'-1)t+1), \cdots, f(m)).
\end{aligned}$$

Then define $F' = \{f' : f \in F\}$. It is straightforward to see that F' is an $(\epsilon, k/t)$-independent $(|S_m|; m/t, 2^t)$ hash family. □

Now we show the equivalence of (ϵ, k)-independent sample spaces and almost strongly universal-k hash families.

Theorem 2.4 *If F is an (ϵ, k)-independent $(N; m, n)$ hash family, then F is a δ-ASU $(N; m, n, k)$ hash family, where*

$$\delta = \frac{(n^{-k} + \epsilon)}{n(n^{-k} - \epsilon)}.$$

Proof. Suppose that Eq. (3) holds. Then for any $y_1, \cdots, y_k \in B$, we have

$$\begin{aligned}
\Pr[f(x_i) = y_i, 1 \leq i \leq k] &\geq n^{-k} - \epsilon, \\
\sum_{y_k \in B} \Pr[f(x_i) = y_i, 1 \leq i \leq k] &\geq \sum_{y_k \in B} (n^{-k} - \epsilon), \quad \text{and} \\
\Pr[f(x_i) = y_i, 1 \leq i \leq k-1] &\geq n(n^{-k} - \epsilon).
\end{aligned}$$

From the above inequality and Eq. (3), we have

$$\frac{\Pr[f(x_i) = y_i, 1 \le i \le k]}{\Pr[f(x_i) = y_i, 1 \le i \le k-1]} \le \frac{n^{-k} + \epsilon}{n(n^{-k} - \epsilon)}.$$

Let $\delta \triangleq (n^{-k} + \epsilon)/(n(n^{-k} - \epsilon))$. Then

$$|\{f \in F : f(x_i) = y_i, 1 \le i \le k\}| \le \delta \times |\{f \in F : f(x_i) = y_i, 1 \le i \le k-1\}|.$$

Hence, F is a δ-ASU $(N; m, n, k)$ hash family. ∎

Definition 2.4 An $(N; m, n)$ hash family F of functions from A to B is *strongly* (ϵ, k)*-independent* if for any t such that $1 \le t \le k$ and for all distinct elements $x_1, x_2, \cdots, x_t \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_t \in B$, we have

$$|\Pr(f(x_i) = y_i, 1 \le i \le t) - n^{-t}| \le \epsilon \tag{4}$$

where $f \in F$ is chosen uniformly at random.

Theorem 2.5 *If an $(N; m, n)$ hash family F is strongly (ϵ, k)-independent, then F is a δ-ASU $(N; m, n, k)$ hash family, where $\delta = (n^{-k} + \epsilon)/(n^{-(k-1)} - \epsilon)$.*

Proof. The proof is similar to the proof of Theorem 2.4. ∎

Lemma 2.6 [2] *Suppose that a hash family F of functions from A to B is ϵ-ASU $(N; m, n, k)$. Then for for all $1 \le j \le k$, for all distinct elements $x_1, x_2, \cdots, x_j \in A$, and for all (not necessary distinct) $y_1, y_2, \cdots, y_j \in B$, we have*

$$|\{f \in F : f(x_i) = y_i, 1 \le i \le j\}| \le \epsilon^j \times N. \tag{5}$$

Lemma 2.7 [2] *If a hash family F is ϵ-ASU $(N; m, n, k)$, then $\epsilon \ge 1/n$.*

Theorem 2.8 *If a hash family F is ϵ-ASU $(N; m, n, k)$, then F is (δ, k)-independent, where $\delta = (n^k - 1)(\epsilon^k - n^{-k})$.*

Proof. From Lemma 2.6, we have

$$\begin{aligned} \Pr[f(x_i) = y_i, 1 \le i \le k] &\le \epsilon^k \quad \text{and} \quad (6) \\ \Pr[f(x_i) = y_i, 1 \le i \le k] - n^{-k} &\le \epsilon^k - n^{-k}. \quad (7) \end{aligned}$$

On the other hand, from Eq. (6), we have

$$\sum_{(\hat{y}_1, \cdots, \hat{y}_k) \ne (y_1, \cdots, y_k)} \Pr[f(x_i) = \hat{y}_i, 1 \le i \le k] \le (n^k - 1)\epsilon^k.$$

Therefore, we have

$$\begin{aligned} \Pr[f(x_i) = y_i, 1 \le i \le k] &= 1 - \sum_{(\hat{y}_1, \cdots, \hat{y}_k) \ne (y_1, \cdots, y_k)} \Pr[f(x_i) = \hat{y}_i, 1 \le i \le k] \\ &\ge 1 - (n^k - 1)\epsilon^k. \end{aligned}$$

Hence,

$$\begin{aligned} \Pr[f(x_i) = \hat{y}_i, 1 \le i \le k] - n^{-k} &\ge 1 - (n^k - 1)\epsilon^k - n^{-k} \\ &= 1 - \epsilon^k n^k + \epsilon^k - n^{-k} \\ &= -(n^k - 1)(\epsilon^k - n^{-k}). \end{aligned}$$

From Lemma 2.7, we see that $\epsilon^k - n^{-k} \geq 0$. Hence,

$$-(n^k - 1)(\epsilon^k - n^{-k}) \leq \Pr[f(x_i) = \hat{y}_i, 1 \leq i \leq k] - n^{-k} \leq \epsilon^k - n^{-k}$$

Then the family is (δ, k)-independent, where

$$\delta = \max\{|\epsilon^k - n^{-k}|, |-(n^k - 1)(\epsilon^k - n^{-k})|\} = (n^k - 1)(\epsilon^k - n^{-k})$$

$\square$

2.2 New multiple A-codes

By combining Propositions 1.2 and 2.3 with Theorem 2.4 or Theorem 2.5, we can obtain new multiple A-codes (ASU-k hash families) from an (ϵ, k)-independent sample space. The (ϵ, k)-independent sample spaces from [1] mentioned in Proposition 1.2 can be shown to produce strong ASU-k hash families. Therefore we can apply Theorem 2.5, obtaining the following.

Theorem 2.9 *There exists a δ-ASU $(N; m, n, k)$ hash family, where*

$$\log_2 N = 2(\log_2 \log_2(m \log_2 n) + k \log_2 n - \log_2(n\delta - 1) + \log_2(k \log_2 n) - 1). \tag{8}$$

Proof. Define $l = k \log_2 n$, $u = m \log_2 n$, and

$$\epsilon = \frac{n^{-k}(\delta n - 1)}{\delta + 1} \approx n^{-k}(\delta n - 1).$$

Apply Proposition 1.2 and 2.3, constructing a strongly (ϵ, k)-independent (N, m, n) hash family, where $\log_2 N = 2(\log_2 \log_2 u - \log_2 \epsilon + \log_2 l - 1)$. Now apply Theorem 2.5, to obtain a δ-ASU $(N; m, n, k)$ hash family. We compute $\log_2 N$ as

$$\begin{aligned} \log_2 N &= 2(\log_2 \log_2(m \log_2 n) - \log_2(n^{-k}(\delta n - 1)) + \log_2(k \log_2 n) - 1) \\ &= 2(\log_2 \log_2(m \log_2 n) + k \log_2 n - \log_2(\delta n - 1) + \log_2(k \log_2 n) - 1). \end{aligned}$$

$\square$

3 A lower bound

In this section, we present a new lower bound on the size of ASU-k hash families and almost k-wise independent sample spaces.

Theorem 3.1 *If there exists an ϵ-ASU$(N; m, n, k)$ hash family such that*

$$\epsilon^k \leq 1/n, \tag{9}$$

then

$$N \geq \frac{1}{\epsilon^k} \left(\frac{\log\left(\frac{mn}{k-1}\right)}{\log\left(\frac{1-\epsilon^k}{\frac{1}{n}-\epsilon^k}\right)} - 1 \right).$$

Proof. Suppose F is an ϵ-$ASU(N; m, n, k)$ hash family from A to B, where $|A| = m$, $|B| = n$ and $k \geq 2$. Construct an $N \times mn$ binary matrix $G = (g_{ij})$, with rows indexed by the functions in F and columns indexed by $A \times B$, defined by the rule

$$g_{f,(x,y)} = \begin{cases} 1 & \text{if } f(x) = y \\ 0 & \text{if } f(x) \neq y. \end{cases}$$

Interpret the columns of G as incidence vectors of the N-set F. We obtain a set-system $(F, \mathcal{C} = \{C_{x,y} : x \in A, y \in B\})$, where

$$C_{x,y} = \{f \in F : f(x) = y\}$$

for all $x \in A$, $y \in B$. Let

$$t \triangleq \lfloor \epsilon^k N \rfloor + 1. \tag{10}$$

This set-system satisfies the following properties: (A) $|F| = N$, (B) $|\mathcal{C}| = mn$, (C) $\sum_{C \in \mathcal{C}} |C| = Nm$, (D) there does not exist a subset of t points that occurs as a subset of k different blocks (see Lemma 2.6).

Property (D) says that $(F, \mathcal{C})$ is a *t-packing of index* $\lambda = k - 1$ (i.e., no t-subset of points occurs in more than λ blocks). Hence we obtain the following:

$$\lambda \binom{N}{t} \geq \sum_{C \in \mathcal{C}} \binom{|C|}{t}. \tag{11}$$

Property (C) implies that the average block size is $Nm/mn = N/n$. Define a real-valued function $f(x)$ as

$$f(x) = \begin{cases} 0 & \text{if } x < t \\ x(x-1)\ldots(x-t+1) & \text{otherwise.} \end{cases}$$

Since $f(x)$ is convex, we have

$$\frac{\lambda}{mn} \binom{N}{t} \geq \frac{1}{mn} \sum_{C \in \mathcal{C}} \binom{|C|}{t} \geq \frac{f(N/n)}{t!} \tag{12}$$

from Jensen's inequality. We observe that $N/n \geq t - 1$ follows from Eq. (9) and Eq. (10). Then, we obtain

$$(k-1) \frac{N(N-1)\cdots(N-t+1)}{\frac{N}{n}\left(\frac{N}{n}-1\right)\cdots\left(\frac{N}{n}-t+1\right)} \geq mn, \tag{13}$$

and hence

$$(k-1) \left(\frac{N-t+1}{\frac{N}{n}-t+1} \right)^t \geq mn. \tag{14}$$

From Eq. (10), we have $t \leq \epsilon^k N + 1$. Then Eq. (14) can be simplified as follows.

$$\begin{aligned} (k-1) \left(\frac{1-\epsilon^k}{\frac{1}{n}-\epsilon^k} \right)^t &\geq mn, \quad \text{and hence} \\ (\epsilon^k N + 1) \log \left(\frac{1-\epsilon^k}{\frac{1}{n}-\epsilon^k} \right) &\geq \log \left(\frac{mn}{k-1} \right), \end{aligned}$$

from which our bound is obtained. $\square$

Corollary 3.2 *Suppose S_m is an (ϵ, k)-independent sample space. Denote $\delta = (2^{-k} + \epsilon)/(2(2^{-k} - \epsilon))$. If $\delta^k \leq 1/2$, then*

$$|S_m| \geq \frac{1}{\delta^k} \left(\frac{\log\left(\frac{2m}{k-1}\right)}{\log\left(\frac{1-\delta^k}{\frac{1}{2}-\delta^k}\right)} - 1 \right).$$

Proof. This follows from Theorem 2.4. $\square$

This technique also gives us a bound on orthogonal arrays, which appears to be new. An $\mathrm{OA}_\lambda(s,m,n)$ is equivalent to a $\frac{1}{n}$-$ASU(\lambda n^s; m, n, s)$ hash family. Certainly inequality (9) holds. Applying Theorem 3.1 with $N = \lambda n^s$ and $\epsilon = 1/n$, we obtain the following.

Theorem 3.3 *If there exists an $OA_\lambda(s,m,n)$ with $s \geq 2$, then*

$$N \geq n^s \left(\frac{\log\left(\frac{mn}{s-1}\right)}{\log\left(\frac{n^s-1}{n^{s-1}-1}\right)} - 1 \right).$$

3.1 Some numerical examples of multiple A-codes

We give some numerical examples to compare the multiple A-codes constructed by Atici and Stinson in [2]; our new multiple A-codes obtained from Theorem 2.9; and the lower bound of Theorem 3.1. Suppose we want an authentication code for $m = 2^{2^{128}}$ source states with deception probability $\delta = 2^{-40}$. (In other words, we are authenticating a bit string of length 2^{128}, which is truly enormous!) We tabulate the number of key bits (i.e., $\log_2 N$) for $k = 3, 4, 10$. Note that we take $n = 2/\delta = 2^{41}$ in Theorem 2.9 and Theorem 3.1 (whereas in [2], $n > 2/\delta$).

k	[2]	Theorem 2.9	Lower bound
3	657	518	243
4	1043	602	283
10	5376	1096	523

A counter-based multiple authentication scheme ([32]) would (of course) require fewer key bits than the proposed construction. For example, tabulated values from [2] show that the construction from [6] would for the parameters above and $k = 4$ require 447 key bits. Hence, the $602 - 447 = 155$ additional key bits we use can be thought of as the price payed for having a stateless multiple authentication scheme. An interesting property that can be verified through Theorem 2.9 is the following. When $k \to \infty$, the number of key bits required per message approaches $\log_2 n$, which is the same as for the counter-based multiple authentication scheme.

4 Almost resilient functions

We now turn our attention to the concept of resilient functions, and we will show how almost independent sample spaces can be used to construct functions that are "almost resilient".

In what follows, let $m \geq l \geq 1$ be integers and let $\phi : \{0,1\}^m \to \{0,1\}^l$ be a function mapping m bit vectors into l bit vectors.

Definition 4.1 The function ϕ is called an *(m,l,k)-resilient function* if

$$\Pr[\phi(x_1,\ldots,x_m) = (y_1,\ldots,y_l) \mid x_{i_1}x_{i_2}\cdots x_{i_k} = \alpha] = 2^{-l}$$

for any k positions $i_1 < \cdots < i_k$, for any k-bit string $\alpha \in \{0,1\}^k$ and for any $(y_1,\cdots,y_l) \in \{0,1\}^l$, where the values x_j ($j \notin \{i_1,\ldots,i_k\}$) are chosen independently at random.

Resilient functions have been studied in several papers, e.g., [14, 3, 15, 29, 4]. We now introduce a generalization, which we call ϵ-almost resilient functions. Here the the output distribution may deviate from the uniform distribution by a small amount ϵ.

Definition 4.2 Let the function ϕ be called an *ϵ-almost (m, l, k)-resilient function* if

$$|\Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l) \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-l}| \leq \epsilon$$

for any k positions $i_1 < \cdots < i_k$, for any k-bit string $\alpha \in \{0,1\}^k$ and for any $(y_1, \cdots, y_l) \in \{0,1\}^l$, where the values x_j $(j \notin \{i_1, \ldots, i_k\})$ are chosen independently at random.

As will be demonstrated, by allowing this small deviation from the uniform distribution, one can obtain a substantial improvement on the parameters.

4.1 Relation with (ϵ, k)-independent sample space

It is well known that a resilient function is equivalent to a large set of orthogonal arrays [29]. Here we prove a similar result for almost resilient functions that involves k-wise independent sample spaces.

Definition 4.3 A *large set of (ϵ, k)-independent sample spaces*, denoted $LS(\epsilon, k, m, t)$, is a set of 2^{m-t} (ϵ, k)-independent sample spaces, each of size 2^t, such that their union contains all 2^m binary vectors of length m.

Theorem 4.1 *If there exists an $LS(\epsilon, k, m, t)$, then there exists a δ-almost $(m, m-t, k)$-resilient function, where $\delta = \epsilon/2^{m-t-k}$.*

Proof. There are 2^{m-t} (ϵ, k)-independent sample spaces in $LS(\epsilon, k, m, t)$. Name the (ϵ, k)-independent sample spaces C_γ, $\gamma \in \{0,1\}^{m-t}$. Then define a function $\phi : \{0,1\}^m \to \{0,1\}^{m-t}$ by the rule

$$\phi(x_1, \ldots, x_m) = \gamma \text{ if and only if } (x_1, \ldots, x_m) \in C_\gamma.$$

Due to Definition 4.3, ϕ is well defined. For any k positions $i_1 < \cdots < i_k$, any k-bit string $\alpha \in \{0,1\}^k$ and any $\gamma \in \{0,1\}^{m-t}$, let

$$L \triangleq |\{(x_1, \ldots, x_m) : x_{i_1} \cdots x_{i_k} = \alpha, (x_1, \ldots, x_m) \in C_\gamma\}|.$$

Then

$$\Pr[\phi(x_1, \ldots, x_m) = \gamma \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = \frac{L}{2^{m-k}}. \tag{15}$$

From Definition 1.1, we have

$$2^{-k} - \epsilon \leq \frac{L}{2^t} \leq 2^{-k} + \epsilon. \tag{16}$$

Hence, from (15) and (16), we obtain

$$|\Pr[\phi(x_1, \ldots, x_m) = \gamma \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - 2^{-(m-t)}| \leq \frac{\epsilon}{2^{m-t-k}}.$$

□

Definition 4.4 The function $\phi : \{0,1\}^m \to \{0,1\}^l$ is called *balanced* if we have

$$\Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l)] = 2^{-l}$$

for all $(y_1, \cdots, y_l) \in \{0,1\}^l$.

For balanced functions, we can prove the converse of Theorem 4.1.

Theorem 4.2 *If there exists a balanced ϵ-almost (m, l, k)-resilient function, then there exists an $LS(\delta, k, m, m-l)$, where $\delta = \epsilon/2^{k-l}$.*

Proof. For $\gamma \in \{0,1\}^l$, let

$$C_\gamma \triangleq \{(x_1, \ldots, x_m) : \phi(x_1, \ldots, x_m) = \gamma\}.$$

Since ϕ is balanced, $|C_\gamma| = 2^{m-l}$. If each C_γ is an (ϵ, k)-independent sample space, then we automatically get a large set. For any k positions $i_1 < \cdots < i_k$, for any k-bit string α for and any $\gamma \in \{0,1\}^l$, let

$$L \triangleq |\{(x_1, \ldots, x_m) : x_{i_1} \cdots x_{i_k} = \alpha, (x_1, \ldots, x_m) \in C_\gamma\}|.$$

Then, within the sample space C_γ, we have

$$\Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = \frac{L}{|C_\gamma|} = \frac{L}{2^{m-l}}. \tag{17}$$

From Definition 4.2, we get

$$2^{-l} - \epsilon \le \frac{L}{2^{m-k}} \le 2^{-l} + \epsilon. \tag{18}$$

Hence, from (17) and (18), we obtain

$$|\Pr(x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha) - 2^{-k}| \le \frac{\epsilon}{2^{k-l}}.$$

□

Finally, we can use the bound from Theorem 4.2 to obtain the following bound on almost resilient functions.

Corollary 4.3 *Suppose that there exists a balanced ϵ-almost (m, l, k)-resilient function. Let*

$$\delta = \frac{1 + \epsilon 2^k}{2(1 - \epsilon 2^k)}.$$

If $\delta^k \le 1/2$, then

$$l \le m + k \log \delta - logZ,$$

where

$$Z = \left(\frac{\log\left(\frac{2m}{k-1}\right)}{\log\left(\frac{1-\delta^k}{\frac{1}{2}-\delta^k}\right)} - 1 \right)$$

Proof. From Corollary 3.2, we have $2^{m-l} \ge Z/\delta^k$. □

In summation, we have in this subsection established the relations between the notions of almost independent sample spaces, large sets of almost independent sample spaces, and almost resilient functions. This can be thought of as an "almost" version of the relations between orthogonal arrays, large sets of orthogonal arrays and resilient functions. The main motivation is that by considering "almost" versions we will be able to significantly improve certain parameters, as compared to the traditional case.

4.2 Constructions of ϵ-almost resilient functions

In order to construct almost resilient functions, we first exhibit a construction of almost independent sample spaces. It will then be extended to obtain a large set of almost independent sample spaces, i.e., an almost resilient function.

Definition 4.5 An (ϵ, k)-independent sample space S_m is called *t-systematic* if $|S_m| = 2^t$, and there exist t positions $i_1 < \cdots < i_t$ such that each t-bit string occurs in these positions for exactly one m-tuple in S_m.

A t-systematic (ϵ, k)-independent sample space can be transformed into an $LS(\epsilon, k, m, t)$ by using the same technique as [30, Theorem 3]. We have the following result.

Theorem 4.4 *If there exists a t-systematic (ϵ, k)-independent sample space S_m, then there exists a balanced δ-almost $(m, m-t, k)$-resilient function, where $\delta = \epsilon/2^{m-t-k}$.*

Proof. By using the same technique as [30, Theorem 3], we can obtain a large set of (ϵ, k, m, t)-independent sample spaces from a t-systematic (ϵ, k)-independent sample space S_m as follows. Without loss of generality, assume that the first t positions in S_m run through all possible t-bit strings. We then obtain 2^{m-t} sample spaces C_α indexed by $\alpha = (\alpha_1, \ldots, \alpha_{m-t}) \in \{0,1\}^{m-t}$ by

$$C_\alpha = S_m + (\underbrace{0, 0, \ldots, 0}_{t}, \alpha_1, \ldots, \alpha_{m-t}).$$

These sample spaces form a $LS(\epsilon, k, m, t)$.

Then from Theorem 4.1, a δ-almost $(m, m-t, k)$-resilient function is obtained, where $\delta = \epsilon/2^{m-t-k}$. □

We now present a summary of our construction of t-systematic (ϵ, k)-independent sample spaces. Our approach is similar to [17] (see also [23]), and depends on the Weil-Carlitz-Uchiyama bound. In what follows, let Tr denote the *trace* function from $GF(2^t)$ to $GF(2)$.

Proposition 4.5 (Weil-Carlitz-Uchiyama bound, [13]) *Let $f(x) = \sum_{i=1}^{D} f_i x^i \in GF(2^t)[x]$ be a polynomial that is not expressible in the form $f(x) = g(x)^2 - g(x) + \theta$ for any polynomial $g(x) \in GF(2^t)[x]$ and for any $\theta \in F_{2^t}$. Then*

$$\left| \sum_{\alpha \in GF(2^t)} (-1)^{Tr(f(\alpha))} \right| \leq (D-1)\sqrt{2^t}.$$

Definition 4.6 A polynomial $h(x) \in GF(2^t)[x]$ is called a $(2^t, D)$-*polynomial* if h has degree at most D and $a_i = 0$ for all even i, where $h = \sum_{i=0}^{D} a_i x^i$. Define $H(2^t, D, k)$ to be a set of $(2^t, D)$-polynomials such that any k polynomials in the set are independent over $GF(2)$.

Observe that the condition $a_{2i} = 0$ for all i guarantees that $h(x)$ is not expressible in the form $f(x) = g(x)^p - g(x) + \theta$ for any polynomial $g(x)$ over $GF(2^t)$ and $\theta \in F_{2^t}$. Hence, Proposition 4.5 can be applied.

For $h_{i_1}, h_{i_2}, \ldots, h_{i_k} \in H(2^t, D, k)$ and for any k elements $\alpha_1, \cdots, \alpha_k \in GF(2)$, define

$$N_{\alpha_1,\ldots,\alpha_k}(h_{i_1}, \ldots, h_{i_k}) \triangleq |\{x \in GF(2^t) : Tr(h_{i_1}(x)) = \alpha_1, \cdots, Tr(h_{i_k}(x)) = \alpha_k\}|.$$

Lemma 4.6 [17] $|N_{\alpha_1,\ldots,\alpha_k}(h_{i_1}, \ldots, h_{i_k}) - 2^{t-k}| \leq (D-1)\sqrt{2^t}$.

Proof. The proof is an application of Proposition 4.5. The case $k = 2$ can be found in [17] and the general case is proved similarly. □

Theorem 4.7 *Suppose that β is a primitive element of $GF(2^t)$, and $H(2^t, D, k)$ is chosen such that*

$$\{x, \beta x, \beta^2 x, \ldots, \beta^{t-1} x\} \subseteq H(2^t, D, k).$$

Then there exists a t-systematic (ϵ, k)-independent sample space S_m where $m = |H(2^t, D, k)|$ and $\epsilon = (D-1)/\sqrt{2^t}$.

Proof. Let $H(2^t, D, k) = \{h_1, \cdots, h_m\}$. Construct a sample space S_m as follows: A binary string $X_\gamma = x_1 x_2 \cdots x_m \in S_m$ is specified by any $\gamma \in GF(2^t)$, where the ith bit of X_γ is $x_i = Tr(h_i(\gamma))$.

Then from Lemma 4.6, for a k bit string α,

$$|\Pr(x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha) - 2^{-k}| = |N_\alpha(h_{i_1}, \ldots, h_{i_k})/2^t - 2^{-k}| \leq (D-1)/\sqrt{2^t}.$$

Therefore, S_m is an (ϵ, k)-independent sample space, where $\epsilon = (D-1)/\sqrt{2^t}$.

Let β be a primitive element of $GF(2^t)$. Then $x, \beta x, \beta^2 x, \ldots, \beta^{t-1} x$ are independent over $GF(2)$. Now, $H(2^t, D, k)$ was chosen such that $\{x, \beta x, \beta^2 x, \ldots, \beta^{t-1} x\} \subseteq H(2^t, D, k)$.

It is a well known fact that

$$Y_x = (Tr(x), Tr(\beta x), \ldots, Tr(\beta^{t-1} x))$$

runs through $\{0,1\}^t$ when x runs through $GF(2^t)$. Hence, the sample space is t-systematic. ∎

In our approach, using Theorem 4.7, we need to construct a set of polynomials $H(2^t, D, k)$ such that any k of them are linearly independent over $GF(2)$. For this we can use linear error-correcting codes (see [19]). For a fixed (odd) degree D, we can express each polynomial as a linear combination of polynomials in the set

$$\{x, \beta x, \ldots, \beta^{t-1} x, x^3, \beta x^3, \ldots, \beta^{t-1} x^3, \ldots, x^D, \beta x^D, \ldots, \beta^{t-1} x^D\}.$$

The polynomials in this set are clearly independent over $GF(2)$. Indexing the polynomials in $H(2^t, D, k)$ as $h_1, h_2, \ldots, h_m$ we obtain a binary $tD' \times m$ matrix, where $D' = (D+1)/2$,

$$\begin{pmatrix} h_{1,1} & h_{1,2} & \ldots & h_{1,m} \\ h_{2,1} & h_{2,2} & \ldots & h_{2,m} \\ \vdots & \ddots & \ddots & \vdots \\ h_{tD',1} & h_{tD',2} & \ldots & h_{tD',m} \end{pmatrix},$$

where $h_i(x) = h_{1,i} x + h_{2,i} \beta x + \ldots + h_{tD',i} \beta^{t-1} x^D$. Any k polynomials are independent over $GF(2)$ means that any k columns of the above matrix are linearly independent. Hence the matrix corresponds to a parity check matrix of an $[m, l, d]$ error correcting code, a code of length $m = |H(2^t, D, k)|$, dimension $m - l = tD'$ and minimum Hamming distance $d = k + 1$ [19].

In order to get a t-systematic sample space, we have already chosen the polynomials $h_1 = x, h_2 = \beta x, \ldots, h_t = \beta^{t-1} x$. But clearly, this is no restriction, since any parity check matrix can be rewritten into such a form without changing the code parameters. Conversely, given such a code, we obtain a t-systematic sample space, and hence a balanced ϵ-almost $(m, m-t, k)$-resilient function, as follows.

Theorem 4.8 *Suppose $D = 2D' - 1$ and there is a $[m, m - tD', k+1]$ code. Then there exists a balanced ϵ-almost $(m, m-t, k)$-resilient function such that*

$$\epsilon = \frac{(D-1)\sqrt{2^t}}{2^{m-k}}.$$

Proof. From Theorem 4.4 and Theorem 4.7. ∎

A suitable value of ϵ could be 2^{-m+t-1}. We obtain the following corollary of Theorem 4.8 by taking $D = 3$ and $k = (t/2) - 2$.

Corollary 4.9 *Suppose there is an $[m, m-4k-8, k+1]$ code. Then there exists a balanced $2^{-m+2k+3}$-almost $(m, m-2k-4, k)$-resilient function.*

input bits m	output bits l	maximum known resiliency for linear resilient function	resiliency for constructed ϵ-almost resilient function
80	60	7	8
80	40	15	18
80	20	24	28
120	80	9	13
120	60	19	28
120	30	33	43
160	120	11	18
160	80	22	38
160	40	41	58
200	150	13	23
200	100	27	48
200	50	49	73

Table 1: Maximum resiliency for (m, l, k) resilient functions and ϵ-almost resilient functions with $\epsilon = 2^{-l-1}$.

4.3 Examples and comparison

Example 4.1 (*Numerical comparison*) In the first example we do a numerical comparison in the following way. An (m, l, k)-resilient function has probability 2^{-l} on each output. We allow our ϵ-almost (m, l, k)-resilient function to have probability at most $3/2 \cdot 2^{-l}$ on each output, i.e. $\epsilon = 2^{-l-1}$. Furthermore, we set $D = 3$ in the construction from the previous subsection and can thus use Corollary 4.9. Some numerical results are given in the following Table 1.

We use tables of the best known binary linear codes to verify the existence of the required codes in Corollary 4.9. The best possible parameters for binary linear codes can be found in [9]. □

Example 4.2 (*Asymptotic results on resilient functions*) This example demonstrates a strictly better asymptotic behaviour for ϵ-almost (m, l, k)-resilient functions compared to resilient functions ($\epsilon = 0$).

Let us consider a family of $(m, m-t, k)$-resilient functions when $m \to \infty$. Introduce the notation $\tau = t/m$ and $\kappa = k/m$. We consider the maximum normalized resiliency κ as a function of τ.

For resilient functions ($\epsilon = 0$), the best known construction is through linear codes. Existence of a $(m, (1-\tau)m, \kappa m)$-resilient function is equivalent to the existence of a $[m, (1-\tau)m, \kappa m + 1]$ linear code. We use the asymptotic form of the Varshamov-Gilbert bound [19], which in this case states that there exist linear codes such that $(1-\tau) = 1 - h(\kappa)$, where $h()$ is the binary entropy function. This brings us to the conclusion that when $m \to \infty$, the maximum normalized resiliency that can be obtained (through the best known methods) is

$$\kappa = h^{-1}(\tau), \tag{19}$$

where $h^{-1}()$ is the inverse of the binary entropy function, under the constraint $\kappa < 0.5$.

Consider now the same problem, but for an ϵ-almost resilient function. The truly resilient function has probability $2^{-(m-t)}$ on each output. In order to have a fair comparison, we fix ϵ to be (arbitrarily) small compared to this value, e.g. $\epsilon < 2^{-(m-t)} \cdot 2^{-c}$ for some constant c. Using the proposed construction, let $D' = t^{-1}m$. The requirement of a $[m, (1-\tau D')m, \kappa m + 1]$ linear code is then trivially fullfilled for any κ. This leaves us with the condition for a small ϵ. For $\epsilon < 2^{-(m-t)} \cdot 2^{-c}$ we must have $2^{k+c}(2D' - 2) < 2^{t/2}$. Considering the asymptotic form of this expression we can get a maximum normalized resiliency of

$$\kappa = \tau/2, \tag{20}$$

with $\epsilon < 2^{-(m-t)} \cdot 2^{-c}$ for any fixed c. □

Comparing with (19), we have a strictly better asymptotic behaviour for all τ in $0 < \tau < 1$.

Example 4.3 (*Constructing multiple A-codes*) The constructive results of this section can also be used to construct multiple A-codes. Using the constructed almost independent sample space from the previous subsection, one can verify that there exists a ϵ-ASU $(q^{m+1}; q^{m(p-1)D/(k-\lfloor k/p \rfloor)} - 1, q, k)$ hash family such that

$$\epsilon = \frac{1}{q} + \frac{(D-1)(\sqrt{q^m} + \sqrt{q^m - 2})}{q^{m-k+2} - (D-1)\sqrt{q^m}}.$$

□

Example 4.4 (*Implementation aspects*) We make a remark on implementing an (ϵ, k)-almost resilient function. Again, let $H(2^t, D, k) = \{h_1, \ldots, h_m\}$. The proposed construction is very simple to implement. Following the construction, one should take the first t bits of the input x and solve a set of linear equations $Tr(z) = x_1, Tr(\beta z) = x_2, \ldots, Tr(\beta^{t-1} z) = x_t$ to obtain $z \in F_{2^t}$. However, it is easy to see that this can be simplified and that one can actually just put the first t bits of x to be z. Then generate the remaining sequence, call it S, $S = (Tr(h_{t+1}(z)), Tr(h_{t+2}(z)), \ldots, Tr(h_m(z))$. The output is finally $\phi(x_1, \ldots, x_m) = S \oplus (x_{t+1}, x_{t+2}, \ldots, x_m)$.

In conclusion, a compact description of the almost resilient function ϕ is as follows. Split the input x in two parts, $x = (z, w)$, where $z = (x_1, \ldots, x_t) \in GF(2^t)$ and $w = (x_{t+1}, \ldots, x_m)$. The function $\phi(z, w)$ is defined as

$$\phi(z, w) = (Tr(h_{t+1}(z)), Tr(h_{t+2}(z)), \ldots, Tr(h_m(z)) \oplus w.$$

□

4.4 Almost correlation immune functions

Our results on almost resilient functions can easily be generalized to almost correlation immune functions. We begin with a definition.

Definition 4.7 ϕ is called an (m, l, k)-*correlation immune function* if

$$\Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l) \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = \Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l)]$$

for any k positions $i_1 < \cdots < i_k$, for any k-bit string α, and for any $(y_1, \cdots, y_l) \in \{0,1\}^l$, where the values x_j $(j \notin \{i_1, \ldots, i_k\})$ are chosen independently at random.

We introduce a generalization, which we call ϵ-almost correlation immune functions.

Definition 4.8 We say that ϕ is an ϵ-*almost* (m, l, k)-*correlation immune function* if

$$|\Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l) \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - \Pr[\phi(x_1, \ldots, x_m) = (y_1, \ldots, y_l)]| \leq \epsilon$$

for any k positions $i_1 < \cdots < i_k$, for any k-bit string α, and for any $(y_1, \cdots, y_l) \in \{0,1\}^l$, where the values x_j $(j \notin \{i_1, \ldots, i_k\})$ are chosen independently at random.

Definition 4.9 A *nonuniform large set of* $(\epsilon, k, m, T_1, \cdots, T_{2^l})$-*independent sample spaces*, which we denote as $NLS(\epsilon, k, m, T_1, \cdots, T_{2^l})$, is a set of 2^l pairwise disjoint (ϵ, k)-independent sample spaces, of sizes $T_1, \cdots, T_{2^l}$ respectively, such that their union contains all 2^m binary vectors of length m.

Theorem 4.10 *If there exists an* $NLS(\epsilon, k, m, T_1, \cdots, T_{2^l})$, *then there exists a* δ-*almost* (m, l, k)-*correlation immune function, where*

$$\delta = \max_i \frac{\epsilon T_i}{2^{m-k}}.$$

Proof. There are 2^l (ϵ, k)-independent sample spaces in the set. Name the (ϵ, k)-independent sample spaces C_γ, $\gamma \in \{0,1\}^l$. Then define a function $\phi : \{0,1\}^m \to \{0,1\}^l$ by the rule

$$\phi(x_1, \ldots, x_m) = \gamma \text{ if and only if } (x_1, \ldots, x_m) \in C_\gamma.$$

Then

$$\Pr[\phi(x_1, \ldots, x_m) = \gamma] = \frac{T_\gamma}{2^m}.$$

For any k positions $i_1 < \cdots < i_k$, for any k-bit string α, and for any $\gamma \in \{0,1\}^l$, let

$$L \triangleq |\{(x_1, \ldots, x_m) : x_{i_1} \cdots x_{i_k} = \alpha, (x_1, \ldots, x_m) \in C_\gamma\}|.$$

Then

$$\Pr[\phi(x_1, \ldots, x_m) = \gamma \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = \frac{L}{2^{m-k}}. \tag{21}$$

From Definition 1.1, we have

$$2^{-k} - \epsilon \le \frac{L}{T_\gamma} \le 2^{-k} + \epsilon. \tag{22}$$

Hence, from (21) and (22), we obtain

$$|\Pr[\phi(x_1, \ldots, x_m) = \gamma \mid x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] - \Pr[\phi(x_1, \ldots, x_m) = \gamma]| \le \frac{\epsilon T_\gamma}{2^{m-k}}.$$

□

We now prove a converse to Theorem 4.10.

Theorem 4.11 *If there exists a ϵ-almost (m, l, k)-correlation immune function, ϕ, then there exist integers $T_1, \cdots, T_{2^l}$ and an $NLS(\delta, k, m, T_1, \cdots, T_{2^l})$ in which*

$$\delta = \max_i \frac{\epsilon 2^{m-k}}{T_i}.$$

Proof. For $\gamma \in \{0,1\}^l$, let

$$C_\gamma \triangleq \{(x_1, \ldots, x_m) : \phi(x_1, \ldots, x_m) = \gamma\}$$

and let

$$T_\gamma = |C_\gamma|.$$

If each C_γ is an (ϵ, k)-independent sample space, then we automatically get a (nonuniform) large set of sample spaces. For any k positions $i_1 < \cdots < i_k$, for any k-bit string α for and any $\gamma \in \{0,1\}^l$, let

$$L \triangleq |\{(x_1, \ldots, x_m) : x_{i_1} \cdots x_{i_k} = \alpha, (x_1, \ldots, x_m) \in C_\gamma\}|.$$

Then, within the sample space C_γ, we have

$$\Pr[x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha] = \frac{L}{|C_\gamma|} = \frac{L}{T_\gamma}. \tag{23}$$

From Definition 4.8, we get

$$T_\gamma / 2^m - \epsilon \le \frac{L}{2^{m-k}} \le T_\gamma / 2^m + \epsilon. \tag{24}$$

Hence, from (23) and (24), we obtain

$$|\Pr(x_{i_1} x_{i_2} \cdots x_{i_k} = \alpha) - 2^{-k}| \le \frac{\epsilon 2^{m-k}}{T_\gamma}.$$

□

5 Indistinguishability and almost k-wise independence

Indistinguishability of random variables plays an important role in cryptography. In this section, we study the indistinguishability of almost k-wise independent sample spaces from truly random sample spaces.

Let us consider a computationally unbounded distinguisher $\mathcal{D}$ which is limited to k queries to an oracle $\mathcal{O}$. Its aim is to distinguish if the oracle $\mathcal{O}$ implements a truly random function or a pseudorandom function. First, we consider an *adaptive* distinguisher, i.e., a distinguisher in which each query may depend on the answers to the previous queries.

Without loss of generality, we can represent a pseudorandom function as an $(N; m, n)$ hash family F_1 of functions from A to B. A truly random function corresponds to the $(n^m; m, n)$ hash family F_0 consisting of all functions from A to B. The *advantage* of the distinguisher $\mathcal{D}$ is defined to be

$$Adv^{\mathcal{D}}(F_1) = |\Pr[\mathcal{D}^{\mathcal{O}=F_1} = 1] - \Pr[\mathcal{D}^{\mathcal{O}=F_0} = 1]|.$$

We have the following theorem.

Theorem 5.1 *Let F_1 be an $(N; m, n)$ hash family. Suppose that*

$$\max Adv^{\mathcal{D}}(F_1) \leq \alpha,$$

where the maximum is taken over all adaptive distinguishers $\mathcal{D}$ which are limited to k oracle queries. Then F_1 is $(2\alpha, k)$-independent. Conversely, if F_1 is (ϵ, k)-independent, then

$$\max Adv^{\mathcal{D}}(F_1) \leq n^k \epsilon/2.$$

Proof. Vaudenay showed in [31] that the following formula holds:

$$\max Adv^{D}(F_1) = \frac{1}{2} \max_{x_1} \sum_{y_1} \max_{x_2} \sum_{y_2} \cdots \max_{x_k} \sum_{y_k} |\Pr(f(x_i) = y_i, 1 \leq i \leq k) - n^{-k}|. \tag{25}$$

Then our assumption is written as

$$\frac{1}{2} \max_{x_1} \sum_{y_1} \max_{x_2} \sum_{y_2} \cdots \max_{x_k} \sum_{y_k} |\Pr(f(x_i) = y_i, 1 \leq i \leq k) - n^{-k}| \leq \alpha.$$

From the above equation, we have, for all distinct $x_1, x_2, \cdots, x_k \in A$ and for all $y_1, \cdots, y_k \in B$, that

$$|\Pr(f(x_i) = y_i, 1 \leq i \leq k) - n^{-k}| \leq 2\alpha.$$

This implies that F_1 is $(2\alpha, k)$-independent.

Conversely, suppose that

$$\max_{x_1, \cdots, x_k} \max_{y_1, \cdots, y_k} |\Pr(f(x_i) = y_i, 1 \leq i \leq k) - n^{-k}| \leq \epsilon.$$

Then we obtain that

$$\max Adv^{\mathcal{D}}(F_1) \leq n^k \epsilon/2$$

from Eq. (25). □

A particular security property of block ciphers is also related to almost k-wise independence, as follows. Let F_1 denote the set of round functions of a Feistel type block cipher. A key K has the effect of selecting one of the functions $f \in F_1$. Vaudenay [31] defined the concept of *k-wise decorrelation bias* of F_1, which is denoted by $DecF^k(F_1)$. In our terminology, this quantity can be defined as

$$DecF^k(F_1) = \max_{x_1,\cdots,x_k} \sum_{y_1,\cdots,y_k} |\Pr(f(x_i) = y_i, 1 \leq i \leq k) - n^{-k}|.$$

Vaudenay considered several constructions of hash families F_1 with small values of $DecF^k(F_1)$ which are suitable for block ciphers.

We prove the following corollary of Theorem 5.1.

Corollary 5.2 *If F_1 is (ϵ, k)-independent, then*

$$DecF^k(F_1) \leq n^k \epsilon/2.$$

Proof. Note that

$$DecF^k(F_1) \leq \max Adv^D(F_1)$$

follows from Eq. (25). Then we see from Theorem 5.1 that, if F_1 is (ϵ, k)-independent, then

$$DecF^k(F_1) \leq n^k \epsilon/2.$$

□

Conversely, in a similar manner as in the proof of Theorem 5.1, it is striaghtforward to show that, if

$$DecF^k(F_1) \leq \alpha,$$

then F_1 is $(2\alpha, k)$-independent.

6 Conclusion

In this paper, we have presented several applications of almost k-wise independent sample spaces in cryptology. In particular, we have found significantly improved constructions for multiple authentication codes by this approach. We hope that almost k-wise independent sample spaces will find further cryptologic applications in the future.

Acknowledgements

This work was supported in part by the Swedish Research Council for Engineering Sciences under Grant 97-130 (Johansson), and by the Natural Sciences and Research Coucil of Canada under grants NSERC-IRC 216431-96 and NSERC-RGPIN 203114-98 (Stinson).

References

[1] N. Alon, O. Goldreich, J. Hastad, and R. Peralta. Simple constructions of almost k-wise independent random variables. *Random Structures and Algorithms* **3** (1992), 289–304.

[2] M. Atici and D.R. Stinson. Universal hashing and multiple authentication. *Lecture Notes in Computer Science* **1109** (1996), 16–30 (CRYPTO '96).

[3] C.H. Bennett, G. Brassard, and J.-M. Robert. Privacy amplification by public discussion. *SIAM Journal on Computing* **17** (1988), 210–229.

[4] J. Bierbrauer, K. Gopalakrishnan and D.R. Stinson. Bounds for resilient functions and orthogonal arrays. *Lecture Notes in Computer Science* **839** (1994), 247–257 (CRYPTO '94).

[5] J. Bierbrauer, K. Gopalakrishnan and D.R. Stinson. Orthogonal arrays, resilient functions, error-correcting codes and linear programming bounds. *SIAM Journal on Discrete Math.* **9** (1996), 424–452.

[6] J. Bierbrauer, T. Johansson, G. Kabatianskii and B. Smeets. On families of hash functions via geometric codes and concatenation. *Lecture Notes in Computer Science* **773** (1994), 331–342 (CRYPTO '93).

[7] J. Bierbrauer and H. Schellwat. Efficient constructions of ϵ-biased arrays, ϵ-dependent arrays and authentication codes. Preprint.

[8] J. Bierbrauer and H. Schellwat. Weakly biased arrays, weakly dependent arrays and error-correcting codes. Preprint.

[9] A.E. Brouwer. Bounds on the minimum distance of binary linear codes. `http://www.win.tue.nl/win/math/dw/voorlincod.html`

[10] P. Camion and A. Canteaut. Generalization of Siegenthaler inequality and Schnorr-Vaudenay multipermutations. *Lecture Notes in Computer Science* **1109** (1996), 372–386 (CRYPTO '96).

[11] P. Camion and A. Canteaut. Correlation-immune and resilient functions over a finite alphabet and their applications in cryptography. *Designs, Codes and Cryptography* **16** (1999), 121–149.

[12] P. Camion, C. Carlet, P. Charpin and N. Sendrier. On correlation-immune functions. *Lecture Notes in Computer Science* **576** (1992), 86–100 (CRYPTO '91).

[13] L. Carlitz and S. Uchiyama. Bounds for exponential sums. *Duke Math. Journal* **24** (1957), 37–41.

[14] B. Chor, O. Goldreich, J. Hastad, J. Friedman, S Rudich and R. Smolensky. The bit extraction problem or t-resilient functions. *26th IEEE symposium on Foundations of Computer Science*, pages 396–407, 1985.

[15] J. Friedman. On the bit extraction problem. *33rd IEEE symposium on Foundations of Computer Science*, pages 314–319, 1992.

[16] K. Gopalakrishnan and D.R. Stinson. Three characterizations of non-binary correlation-immune and resilient functions. *Designs, Codes and Cryptography* **5** (1995), 241–251.

[17] T. Helleseth and T. Johansson. Universal hash functions from exponential sums over finite fields and Galois rings. *Lecture Notes in Computer Science* **1109** (1996), 31–44 (CRYPTO '96).

[18] H. Krawczyk. New hash functions for message authentication. *Lecture Notes in Computer Science* **921** (1995), 301–310 (EUROCRYPT '95).

[19] F.J. MacWilliams and N.J.A. Sloane. *The Theory of Error-Correcting Codes.* North-Holland, 1977.

[20] J.L. Massey. Cryptography – A selective survey. *Digital Communications*, North-Holland (1986), 3–21.

[21] U.M. Maurer and J.L. Massey. Local randomness in pseudo-random sequences. *Journal of Cryptology* **4** (1991), 135–149.

[22] J. Naor and M. Naor. Small bias probability spaces: efficient constructions and applications. *SIAM Journal on Computing* **22** (1993), 838–856.

[23] H. Niederreiter and C.P. Schnorr. Local randomness in polynomial random number and random function generators. *SIAM Journal on Computing* **22** (1993), 684–694.

[24] T. Siegenthaler. Correlation-immunity of nonlinear combining functions for cryptographic applications. *IEEE Trans. Inform. Theory* **30** (1984), 776–780.

[25] C.P. Schnorr. On the construction of random number generators and random function generators. *Lecture Notes in Computer Science* **330** (1988), 225–232 (EUROCRYPT '88).

[26] G.J. Simmons. A game theory model of digital message authentication. *Congressus Numeratium* **34** (1982), 413–424.

[27] G.J. Simmons. Authentication theory/coding theory, *Lecture Notes in Computer Science.* **196** (1985), 411–431 (CRYPTO '84).

[28] D.R. Stinson. Universal hashing and authentication codes. *Designs, Codes and Cryptography* **4** (1994), 369–380.

[29] D. R. Stinson. Resilient functions and large set of orthogonal arrays. *Congressus Numerantium* **92** (1993), 105–110.

[30] D.R. Stinson and J.L. Massey. An infinite class of counterexamples to a conjecture concerning nonlinear resilient functions. *Journal of Cryptology* **8** (1995), 167–173.

[31] S. Vaudenay. Adaptive-attack norm for decorrelation and super-pseudorandomness. To appear in *Lecture Notes in Computer Science* (SAC '99).

[32] M.N. Wegman and J.L. Carter. New hash functions and their use in authentication and set equality. *Journal of Computer and System Sciences* **22** (1981), 265–279.