A mathematical problem for security analysis of hash functions and pseudorandom generators

Koji Nuida*, Takuro Abe, Shizuo Kaji, Toshiaki Maeno, Yasuhide Numata

Abstract

In this paper, we specify a class of mathematical problems, which we refer to as "Function Density Problems" (FDPs, in short), and point out novel connections of FDPs to the following two cryptographic topics; theoretical security evaluations of keyless hash functions (such as SHA-1), and constructions of provably secure pseudorandom generators (PRGs) with some enhanced security property introduced by Dubrov and Ishai [STOC 2006]. Our argument aims at proposing new theoretical frameworks for these topics (especially for the former) based on FDPs, rather than providing some concrete and practical results on the topics. We also give some examples of mathematical discussions on FDPs, which would be of independent interest from mathematical viewpoints. Finally, we discuss possible directions of future research on other cryptographic applications of FDPs and on mathematical studies on FDPs themselves.

Keywords: Function Density Problem, hash function, pseudorandom generator, security evaluation

1 Introduction

1.1 Background and related works

It is widely understood that some mathematical problems have been playing indispensable roles in research on cryptography and information security. For instance, the (expected) computational difficulty of integer factorization is the source of security of RSA cryptosystem [8], while the problem of solving multivariate quadratic (MQ) equations has attracted several studies after the development of Matsumoto-Imai cryptosystem [5] and its variants, whose constructions are closely related to MQ equations. Hence, posing and studying an interesting mathematical problem which arises in certain cryptographic settings can contribute to the progress of cryptography and information security.

The aim of this paper is to emphasize the significance of a certain mathematical problem, which has connections to the following two major topics in information security; security analysis of keyless hash functions in the real world (such as MD5 and SHA-1), and construction of pseudorandom generators (PRGs) with some enhanced security property. First, we give some descriptions of these two topics.

Security analysis of keyless hash functions. Intuitively, a hash function is a function $H\colon X \to Y$ from some (finite) set X to another (finite) set Y that possesses a certain desirable security property. When we concern efficiency or computability of H, we consider an algorithm that computes H (also denoted by H) and call it a hash algorithm. One of the standard security requirements for hash functions is *collision resistance*, which informally means that it is difficult to find a collision pair (x_1, x_2) for H, i.e., $x_1 \neq x_2 \in X$ satisfying $H(x_1) = H(x_2)$. Hash functions have been playing central roles in various information security applications, and secure hash functions for real-life applications are usually expected to possess the collision resistance property.

However, most of the preceding successful studies that show security of hash functions actually dealt with *keyed* hash functions (or hash *families*); intuitively, a family of hash functions H_k parameterized by

*Corresponding author

a key k is called collision resistant if, for any (efficient) adversary, the attack to find a collision pair of H_k fails with high probability for a randomly chosen key k. Several constructions of keyed hash functions have been proposed so far (e.g., [2]). The above security notion of keyed hash functions can be interpreted as allowing one to (randomly) choose a concrete instance H_k of the hash family *after* an adversary is given. In contrast, in most of real-life applications, the concrete instance of hash algorithms is specified first (for example, by a standardization), and then an adversary can try to attack the fixed hash algorithm. This reversal of order causes a crucial difficulty in guaranteeing (or even formalizing in a reasonable manner) security of a keyless hash algorithm H, as (unless the trivial situation where the domain of H is not larger than the image of H) there *does* always *exist* a collision pair (x_1, x_2) for H and any adversary (existing in theory) who innately knows the pair (x_1, x_2) is obviously able to efficiently attack the fixed hash algorithm H. In fact, even an instance of standardized (or de facto standard) hash algorithms, whose security must be evaluated well before the standardization, has been suffered from feasible attacks (e.g., [10]). In this paper, we try to propose a theoretical and unified way to say something, preferably affirmative, about security of a concrete (keyless) instance of hash algorithms.

For related works, Rogaway [9] gave a detailed observation about the difference between "inexistence of effective attack algorithms" and "lack of knowledge on construction of effective attack algorithms" for keyless hash algorithms. He emphasized the difference of the two situations (by the term "human ignorance"), and discussed how to prove security of a cryptographic protocol by reducing the security into "lack of knowledge on concrete attacks" on the hash algorithm internally used by the protocol. However, he did not discuss how to theoretically evaluate security of keyless hash algorithms themselves, which we study in this paper. On the other hand, in this paper we adopt concrete security formulation rather than asymptotic one; while some observation for security of keyless hash algorithms in asymptotic security formulation is also given in Rogaway's paper.

Construction of enhanced PRGs. A PRG is an algorithm $G\colon S \to X$ with (finite) set S of inputs (*seeds*) and (finite) output set X with the property that, when a seed $s \in S$ is chosen uniformly at random, the output $G(s) \in X$ of G is also "random" in some sense. Conventionally, the meaning of "randomness" here is formulated by using the notion of *distinguisher*, which is an algorithm $D\colon X \to \{0,1\}$ with 1-bit output and the input set being the output set X of G. In this paper we adopt concrete security formulation rather than asymptotic one, in which case the security requirement for PRGs can be formulated as (T,ε)-security; namely, G is called (T,ε)-*secure* [4] if, for any distinguisher D for G with (time) complexity bounded by T, the statistical distance between the output distribution $D(G(U_S))$ of D with input given by G with uniformly random seed $s \in S$ (referred to as "pseudorandom input") and the output distribution $D(U_X)$ of D with uniformly random input $x \in X$ (referred to as "random input") is bounded by ε. (Intuitively, any such D cannot distinguish the random element x and the pseudorandom element $G(s)$ in X with significant advantage.) There are a large number of constructions of PRGs, most of which are provably secure (possibly in asymptotic security formulation) under standard computational assumptions (e.g., [1, 4]).

On the other hand, in a preceding work of Dubrov and Ishai [3], an enhanced notion for PRGs, called *pseudorandom generators that fool non-boolean distinguishers* (*nb-PRGs*, in short), was proposed. This notion is obtained by allowing the distinguishers D in the above security notion to have larger output sets; namely, G is called (T,n,ε)-*secure* if, for any "non-boolean" distinguisher $D\colon X \to Y$ for G with (time) complexity bounded by T *and output set* Y *of size at most* n, the statistical distance between the output distributions of D with random and pseudorandom inputs is bounded by ε. Dubrov and Ishai showed interesting applications of nb-PRGs, e.g., secure pseudorandomization of a certain kind of information-theoretically secure protocols without any restriction on computational complexity of the adversary's attack algorithm.

However, constructing secure nb-PRGs seems much more difficult than the case of the usual PRGs. Indeed, to the authors' best knowledge, the only constructions of nb-PRGs proposed so far are ones in the original paper [3], which are based on certain less standard computational assumptions. Hence it will be fruitful if we can give some results implying that *any* usual PRG (with some parameter) is also an nb-PRG (with a possibly different parameter). In fact, a straightforward implication has been mentioned in [3], but

this is far from being efficient (i.e., to obtain nb-PRGs with reasonable security parameters, the original PRGs are required to have somewhat impractical security parameters). In this paper, we try to establish a more efficient implication result.

1.2 Our contributions, and organization of this paper

In Section 2, we propose a class of mathematical problems, which we refer to as "Function Density Problems". Intuitively, this problem is to evaluate the possibility of close approximations of arbitrary functions by using some "easily describable (or analyzable)" functions.

Then we introduce motivating applications of Function Density Problems to two topics in information security. First, in Section 3, we discuss theoretical analysis of collision resistance of keyless hash algorithms. We give an abstract framework for attacking a given hash algorithm by using known attacks on some other "easily breakable" hash algorithms. In the framework, it is essential to evaluate how closely a target hash algorithm can be approximated by "easily breakable" hash algorithms; thus Function Density Problems play a significant role in the security evaluation of hash algorithms.

Secondly, in Section 4, we study an enhanced security notion for PRGs (called nb-PRG) introduced by Dubrov and Ishai [3]. We give some implication results showing that any secure PRG with some parameter is also a secure nb-PRG with somewhat modified security parameter. In the results, the overheads in the bounds of (time) complexity and of advantages for the distinguishers are in trade-off relations, and Function Density Problems can be applied to evaluate to what extent the trade-off will be improved by our proposed result.

Then, in order to arise some image or intuition of how Function Density Problems can be mathematically studied, in Section 5 we give some concrete examples of mathematical discussions on Function Density Problems themselves, using combinatorial and geometric arguments and techniques in Gröbner bases. In particular, we deal with special cases where the set of "easily describable (or analyzable)" functions forms a linear subspace (related to low-degree boolean functions, perfect linear codes and Reed–Solomon codes), which would be of independent interest from mathematical viewpoints.

Finally, in Section 6 we give a concluding remark, which includes discussions on further possible applications of Function Density Problems in information security, and on possible directions of future research on Function Density Problems themselves.

2 Function Density Problems

In this section, we specify a class of mathematical problems, which we call *Function Density Problems* (FDPs) in this paper. As the class of FDPs in a most general form will include too various problems to obtain meaningful insights for their properties, it is significant to restrict the class suitably according to each situation under consideration. Relations of FDPs to some concrete topics in cryptography will be shown in the following sections.

We give a general description of our problem:

Definition 1 (Function Density Problems)**.** Let $\mathcal{C}$ be a set of some functions, and let $\mathcal{C}'$ be a subset of $\mathcal{C}$. Let $d(\cdot,\cdot)$ be a distance function for the pairs of functions in $\mathcal{C}$. In this setting, we define a *Function Density Problem* to be a problem of estimating the following quantity:

$$r(\mathcal{C},\mathcal{C}') := \sup\{d(f,\mathcal{C}') \mid f \in \mathcal{C}\} \ , \tag{1}$$

where, for each $f \in \mathcal{C}$, $d(f,\mathcal{C}') := \inf\{d(f,g) \mid g \in \mathcal{C}'\}$ is the distance from f to $\mathcal{C}'$. (The symbol 'r' stands for "radius", by an analogy as if $\mathcal{C}'$ is a single central point in the figure $\mathcal{C}$, in which case the r is the radius of $\mathcal{C}$ in usual sense.)

Among very various situations covered by Definition 1 (where $\mathcal{C}$ in fact need *not* even to be a set of functions!), in the applications of FDPs discussed in this paper we will focus on the following typical cases:

Definition 2 (Function Density Problems – typical cases)**.** Let $\mathcal{C}$ be the set of all functions $f\colon X \to Y$ from a given finite set X to a given finite set Y. Let $\mathcal{C}' \subset \mathcal{C}$. For any $f, g \in \mathcal{C}$, we define the distance between f and g by

$$d_{\mathrm{H}}(f, g) := |\{x \in X \mid f(x) \neq g(x)\}| \ . \tag{2}$$

In this setting, a *Function Density Problem* is a problem of estimating the quantity $r(\mathcal{C}, \mathcal{C}')$ defined by (1) with $d(\cdot, \cdot) = d_{\mathrm{H}}(\cdot, \cdot)$.

In the case of Definition 2, the "sup" and "inf" in Definition 1 can be simply replaced with "max" and "min", respectively. Moreover, the distance defined by (2) coincides with the (generalized) Hamming distance when members of $\mathcal{C}$ are identified with sequences of length $|X|$ over the alphabet Y in a natural manner. Note that the quantity $r(\mathcal{C}, \mathcal{C}')$ can be regarded as a special case of so-called Hausdorff distance for two subsets of a metric space, which would support that it is reasonable to consider $r(\mathcal{C}, \mathcal{C}')$.

An intuitive explanation of a motivation for the above definition is as follows. Given a set $\mathcal{C}$ of functions, a subset $\mathcal{C}'$ consists of members of $\mathcal{C}$ which are in some sense "easily analyzable" or "with simple descriptions". The distance $d(f, g)$ measures how two functions f and g are similar. Then the quantity $d(f, \mathcal{C}')$ evaluates how accurately a function $f \in \mathcal{C}$ can be approximated by an "easy" function in $\mathcal{C}'$, and the quantity $r(\mathcal{C}, \mathcal{C}')$ evaluates how densely the "easy" functions distribute among the entire set $\mathcal{C}$. In other words, when $r(\mathcal{C}, \mathcal{C}')$ is revealed to be small, it shows potential availability of a close approximation of any member of $\mathcal{C}$ by an "easy" function in $\mathcal{C}'$. For example, in the case of Definition 2, any function $f \in \mathcal{C}$ can in principle be converted into some function $g \in \mathcal{C}'$ by changing the values $f(x)$ for at most $r(\mathcal{C}, \mathcal{C}')$ points $x \in X$. (We emphasize that it does *not* mean that a close approximation of f by a function in $\mathcal{C}'$ can be *efficiently computable*. Such a difference between existence and efficient computability is also relevant to a preceding observation for "human ignorance" by Rogaway [9].)

3 Hash functions and FDPs

In this section, we point out a relation of FDPs introduced in Section 2 to security analysis of keyless hash functions. Here we propose a new framework for theoretical security evaluation of keyless hash functions based on FDPs. Although theoretical security evaluation of keyless hash functions is evidently an extremely difficult problem and our proposed framework is unfortunately not yet practical, we hope that our framework can be a clue to this problem.

We consider a keyless hash function $H\colon X \to Y$ with possibly large but finite domain X and relatively small (finite) range Y. Among the major security requirements for hash functions, we focus on the collision resistance of H; we discuss how it is difficult to find a collision pair (x_1, x_2) for H (recall that (x_1, x_2) is called a collision pair for H if we have $x_1, x_2 \in X$, $x_1 \neq x_2$ and $H(x_1) = H(x_2)$). To show the relevance of FDPs to this problem, first we give a somewhat informal description of an abstract "typical" strategy for finding a collision pair:

1. Construct a close approximation $H'\colon X \to Y$ of H in such a way that collision pairs for H' can be found with reasonable computational time.
2. Find randomly a collision pair (x'_1, x'_2) for H'.
3. Construct from (x'_1, x'_2) a candidate (x_1, x_2) of a collision pair for H (in the simplest case, we just set $(x_1, x_2) = (x'_1, x'_2)$).
4. Check if (x_1, x_2) is a collision pair of H; if it is indeed a collision pair of H, then output (x_1, x_2) and stop the process.
5. If (x_1, x_2) is not a collision pair of H, go back to Step (2) and repeat the process.

Intuitively, the number of iterations in the above strategy before finding a collision pair for H would be expected to be small if the approximation H' is sufficiently close to H (see Lemma 1 below for a quantitative

expression of this expected tendency). Hence security of a hash algorithm H against such an attack strategy is related to the possibility of finding its close approximation.

More precisely, we set $(x_1, x_2) = (x'_1, x'_2)$ in the above strategy for simplicity. We consider the case of Definition 2, and let $\mathcal{C}'$ be a subset of $\mathcal{C}$ with the property that any hash function H' in $\mathcal{C}'$ admits an efficient attack (finding a collision pair) by a certain known attack strategy. In the above attack strategy, the approximation H' for H specified in Step (1) is supposed to be chosen from $\mathcal{C}'$. Now we have the following lemma:

Lemma 1. *Suppose that H and H' are functions $X \to Y$ with $|Y| = n \geq 2$, and $d_{\mathrm{H}}(H, H') = d$, $0 < d < |X|$. Then the probability that a collision pair for H', which is chosen uniformly at random from the set of all collision pairs for H', is also a collision pair for H is not lower than*

$$\frac{2\alpha_0|X| - n(\alpha_0+1)\alpha_0 - 2d\alpha_0}{2\alpha_0|X| + 2d|X| - n(\alpha_0+1)\alpha_0 - 2d\alpha_0 - d^2 - d} \ , \tag{3}$$

where $\alpha_0 = \lfloor (|X| - d - 1)/n \rfloor$. Moreover, when $|X| \geq d + (n-1)^2$, the value in (3) is getting larger as d becomes smaller.

A proof of Lemma 1 will be provided in the last of this section. Now let us imagine the following situation. Two candidate sets $\mathcal{C}_1, \mathcal{C}_2$ for a new standard hash function are given, and we can specify subsets $\mathcal{C}'_1 \subset \mathcal{C}_1$ and $\mathcal{C}'_2 \subset \mathcal{C}_2$ in such a way that each $\mathcal{C}'_i$ $(i = 1, 2)$ consists of some hash functions for which collision pairs can be found in reasonable computational time by using some known techniques. We suppose that $r(\mathcal{C}_1, \mathcal{C}'_1)$ is significantly small and $r(\mathcal{C}_2, \mathcal{C}'_2)$ is significantly large. Then *any* hash function H chosen from $\mathcal{C}_1$ can be *potentially* attacked by just finding a close approximation $H' \in \mathcal{C}'_1$ of H (using some expert's sixth sense, for example) and applying the above attack strategy combined with known collision finding techniques. On the other hand, $\mathcal{C}_2$ contains at least one hash function H for which the above attack strategy combined with any known collision finding technique will not succeed. This would suggest that it can be potentially safer to choose a new hash function from $\mathcal{C}_2$ rather than $\mathcal{C}_1$, as we already know the potential attack on any hash function in $\mathcal{C}_1$ but not the same for $\mathcal{C}_2$.

The authors hope that studies of FDPs can contribute to security analysis of keyless hash functions in the above manner, though how to specify the subset $\mathcal{C}'$ in practical cases is of course a big problem to be concerned. One may also feel that it seems infeasible to compute the quantity $r(\mathcal{C}, \mathcal{C}')$ for practical classes of hash functions; even if so, some estimate of a bound or tendency of $r(\mathcal{C}, \mathcal{C}')$ would still give us an insight into the security level of those hash functions.

Remark 1. Here we notice that, although we have focused on the collision resistance in the above argument, a similar idea would also be applicable to other security notions for keyless hash functions, such as the (second) preimage resistance.

To conclude this section, we give a proof of Lemma 1.

Proof of Lemma 1. We write $(m)_2 := m(m-1)$ for any integer m. Put $Y := \{y_1, \ldots, y_n\}$, and for each $1 \leq i \leq n$, put

$$a_i := |\{x \in X \mid H'(x) = y_i\}| \ , \ b_i := |\{x \in X \mid H(x) \neq H'(x) = y_i\}| \ . \tag{4}$$

Moreover, put

$$\varphi_1(\vec{a}; \vec{b}) := \sum_{i=1}^{n} (a_i)_2 \ , \ \varphi_2(\vec{a}; \vec{b}) := \sum_{i=1}^{n} (a_i - b_i)_2 \ , \tag{5}$$

where $\vec{a} := (a_1, \ldots, a_n)$ and $\vec{b} := (b_1, \ldots, b_n)$. Then the number of collision pairs for H' is $\varphi_1(\vec{a}; \vec{b})$, while the number of collision pairs for H is at least $\varphi_2(\vec{a}; \vec{b})$. Therefore the probability specified in the statement of Lemma 1 is at least

$$\varphi(\vec{a}; \vec{b}) := \frac{\varphi_2(\vec{a}; \vec{b})}{\varphi_1(\vec{a}; \vec{b})} \ . \tag{6}$$

From now, we give a lower bound for the values of φ under the following conditions implied by the definitions: $0 \leq b_i \leq a_i$ for each i, $\sum_{i=1}^{n} a_i = |X|$, and $\sum_{i=1}^{n} b_i = d$. For the purpose, we show the following two lemmas:

Lemma 2. *In the above setting, if the minimum value of the function φ is attained by $\vec{a}$ and $\vec{b}$, then we have $b_i > 0$ for a unique index i, and $a_i - b_i \geq a_j$ for every index $j \neq i$.*

Proof. If we have $i \neq j$ and $b_i, b_j > 0$, and we suppose $a_i \leq a_j$ by symmetry, then we have

$$\big((a_i - 1)_2 + (a_j + 1)_2\big) - \big((a_i)_2 + (a_j)_2\big) = 2(a_j - a_i + 1) > 0 \ , \tag{7}$$

therefore the value of φ_1 increases when a_i, a_j, b_i and b_j are replaced with $a_i - 1$, $a_j + 1$, $b_i - 1$ and $b_j + 1$, respectively. On the other hand, the value of φ_2 is not changed by this replacement. Therefore the value of φ is decreased by this replacement, contradicting the assumption on the choice of $\vec{a}$ and $\vec{b}$. Hence an index i with $b_i > 0$ is unique, therefore $b_i = d$. Similarly, if $j \neq i$ and $a_i - b_i < a_j$, then we have

$$\big((a_i - b_i + 1)_2 + (a_j - 1)_2\big) - \big((a_i - b_i)_2 + (a_j)_2\big) = 2(a_i - b_i - a_j + 1) \leq 0 \ , \tag{8}$$

with equality holding when and only when $a_i - b_i = a_j - 1$. This implies that the value of φ at the $\vec{a}$ and $\vec{b}$ is larger than or equal to the value of φ with b_i and b_j $(= 0)$ being replaced with $b_i - 1$ and 1, respectively, where the equality holds if and only if $a_i - b_i = a_j - 1$. As the former value is assumed to be the minimum, the equality condition $a_i - b_i = a_j - 1$ should hold. Moreover, if $b_i - 1 > 0$, then the latter value of φ (which is now equal to the former) cannot be the minimum by the above argument, which also leads to a contradiction. Hence we have $b_i = 1$ (therefore $d = 1$) and $a_i = a_j$. Now we have

$$\big((a_i + 1)_2 + (a_j - 1)_2\big) - \big((a_i)_2 + (a_j)_2\big) = 2(a_i - a_j + 1) > 0 \ . \tag{9}$$

This implies that the value of φ will decrease when a_i and a_j are replaced with $a_i + 1$ and $a_j - 1$, respectively, contradicting the assumption that the former value is the minimum. Hence we have $a_i - b_i \geq a_j$ for every $j \neq i$, concluding the proof of Lemma 2. □

Lemma 3. *In the above setting, if the minimum of the function φ is attained by $\vec{a}$ and $\vec{b}$, then we have $|a_i - a_j| \leq 1$ for any pair of indices $i \neq j$ satisfying $b_i = b_j = 0$.*

Proof. Assume contrary that $a_i - a_j \geq 2$ for such a pair of indices $i \neq j$. For $\ell \in \{1, 2\}$, let α_ℓ denote the value of φ_ℓ at the $\vec{a}$ and $\vec{b}$, and let β_ℓ denote the value of φ_ℓ with a_i and a_j being replaced with $a_i - 1$ and $a_j + 1$, respectively. Then we have $\beta_1 - \alpha_1 = \beta_2 - \alpha_2 = 2(a_j - a_i + 1) < 0$. On the other hand, for the unique index i' with $b_{i'} > 0$ (see Lemma 2), we have $a_{i'} \geq b_{i'} + a_i \geq b_{i'} + a_j + 2 \geq 2$ by the assumption and Lemma 2, therefore $\alpha_1 > \alpha_2$. Now we present the following lemma, which is proven by an easy calculation:

Lemma 4. *If $p > q \geq 0$ and $r > 0$, then $q/p < (q + r)/(p + r)$.*

By using this lemma, we have

$$\frac{\alpha_2}{\alpha_1} = \frac{\beta_2 - 2(a_j - a_i + 1)}{\beta_1 - 2(a_j - a_i + 1)} > \frac{\beta_2}{\beta_1} \ , \tag{10}$$

contradicting the assumption that α_2/α_1 is the minimum of the value of φ. Hence Lemma 3 holds. □

By Lemma 2 and Lemma 3, the points $\vec{a}$ and $\vec{b}$ that attain the minimum of φ satisfy the following conditions: $b_i > 0$ for a unique i, and there is an integer α satisfying that $a_i - b_i \geq \alpha + 1$ and $a_j \in \{\alpha, \alpha + 1\}$ for every $j \neq i$. Note that this α can be taken as $\alpha \geq 0$; indeed, this is obvious if some a_j with $j \neq i$ is positive, while the remaining possibility that $a_j = 0$ for every $j \neq i$ allows us to choose $\alpha = 0$ as $a_i = |X| > d = b_i$ and $a_i - b_i \geq 1$. Let k be the number of indices $j \neq i$ with $a_j = \alpha + 1$, therefore $0 \leq k \leq n - 1$. Then we have $a_i = |X| - (n-1)\alpha - k$, while $b_i = d$, therefore the condition $a_i - b_i \geq \alpha + 1$ implies that $k \leq |X| - n\alpha - d - 1$. Now we write the values of φ_1 and φ_2 in this case as $\varphi_1(\alpha, k)$ and $\varphi_2(\alpha, k)$, respectively. Then we have

$$\begin{aligned} \varphi_1(\alpha, k) &= k(\alpha + 1)_2 + (n - 1 - k)(\alpha)_2 + (a_i)_2 \ , \\ \varphi_2(\alpha, k) &= k(\alpha + 1)_2 + (n - 1 - k)(\alpha)_2 + (a_i - d)_2 \ , \end{aligned} \tag{11}$$

therefore $\varphi_1(\alpha,k)-\varphi_2(\alpha,k)=2da_i-d^2-d$. Now by Lemma 4, we have

$$\begin{aligned}1-\frac{\varphi_2(\alpha,k)}{\varphi_1(\alpha,k)}&=\frac{2da_i-d^2-d}{\varphi_1(\alpha,k)}\le\frac{2da_i-d^2-d+2d((n-1)\alpha+k)}{\varphi_1(\alpha,k)+2d((n-1)\alpha+k)}\\&=\frac{2d|X|-d^2-d}{k(\alpha+1)_2+(n-1-k)(\alpha)_2+(a_i)_2+2d(n-1)\alpha+2dk}\end{aligned}\tag{12}$$

(note that $2d((n-1)\alpha+k)\ge 0$ as $\alpha\ge 0$). Let $\psi(\alpha,k)$ denote the denominator of the right-hand side. Then, by virtue of the property $\frac{\partial}{\partial k}a_i=-1$, we have

$$\frac{\partial}{\partial k}\psi(\alpha,k)=(\alpha+1)_2-(\alpha)_2-(2a_i-1)+2d=2\alpha-2a_i+1+2d<0\tag{13}$$

(note that $a_i-d\ge\alpha+1$), therefore $\psi(\alpha,k)$ is decreasing as k is increasing. On the other hand, we have $\psi(\alpha,n-1)=\psi(\alpha+1,0)$. Now note that $\alpha\le(|X|-d-1)/n$ as $0\le k\le|X|-n\alpha-d-1$. This implies that $\psi(\alpha,k)$ takes the minimum value at $\alpha=\lfloor(|X|-d-1)/n\rfloor=\alpha_0$ and $k=k_0:=|X|-n\alpha_0-d-1$ (note that $k_0\le n-1$). Moreover, we have $a_i=\alpha_0+d+1$ if $\alpha=\alpha_0$ and $k=k_0$. Hence a straightforward calculation shows that

$$\begin{aligned}1-\frac{\varphi_2(\alpha,k)}{\varphi_1(\alpha,k)}&\le\frac{2d|X|-d^2-d}{\psi(\alpha_0,k_0)}\\&=\frac{2d|X|-d^2-d}{2\alpha_0|X|+2d|X|-n(\alpha_0+1)\alpha_0-2d\alpha_0-d^2-d}\ ,\end{aligned}\tag{14}$$

therefore

$$\begin{aligned}\frac{\varphi_2(\alpha,k)}{\varphi_1(\alpha,k)}&\ge 1-\frac{2d|X|-d^2-d}{2\alpha_0|X|+2d|X|-n(\alpha_0+1)\alpha_0-2d\alpha_0-d^2-d}\\&=\frac{2\alpha_0|X|-n(\alpha_0+1)\alpha_0-2d\alpha_0}{2\alpha_0|X|+2d|X|-n(\alpha_0+1)\alpha_0-2d\alpha_0-d^2-d}\ ,\end{aligned}\tag{15}$$

which proves the lower bound (3) in the statement of Lemma 1.

Finally, suppose that $d\ge 2$, and let $\eta_1(d)$ and $\eta_2(d)$ denote the denominator and the numerator in (3), respectively. For any value x depending on d, let $\Delta[x]$ temporarily denote the value of x at $d-1$ minus the value of x at d. Then we have $\Delta(-d^2-d)=2d$, therefore

$$\begin{aligned}&\Delta[\eta_2(d)]=\Delta[2\alpha_0|X|-n(\alpha_0+1)\alpha_0-2d\alpha_0]\ ,\\&\Delta[\eta_1(d)]=\Delta[2\alpha_0|X|-n(\alpha_0+1)\alpha_0-2d\alpha_0]-2|X|+2d<\Delta[\eta_2(d)]\ .\end{aligned}\tag{16}$$

Moreover, we have $\Delta[\alpha_0]\in\{0,1\}$, and if $\Delta[\alpha_0]=0$, then $\Delta[\eta_2(d)]=2d\alpha_0>0$. On the other hand, if $\Delta[\alpha_0]=1$, then we have

$$\begin{aligned}\Delta[(\alpha_0+1)\alpha_0]&=(\alpha_0+2)(\alpha_0+1)-(\alpha_0+1)\alpha_0=2(\alpha_0+1)\ ,\\\Delta[2d\alpha_0]&=2(d-1)(\alpha_0+1)-2d\alpha_0=2d-2\alpha_0-2\ ,\end{aligned}\tag{17}$$

therefore

$$\begin{aligned}\Delta[\eta_2(d)]&=2|X|-2n(\alpha_0+1)-2d+2\alpha_0+2\\&=2|X|-2(n-1)\alpha_0-2n-2d+2\\&\ge 2|X|-2(n-1)\frac{|X|-d-1}{n}-2n-2d+2\\&=\frac{2}{n}\big(|X|-d+2n-1-n^2\big)\ge 0\end{aligned}\tag{18}$$

(where we used the assumption $|X|\ge d+(n-1)^2$). Now by Lemma 4, we have

$$\frac{\eta_2(d-1)}{\eta_1(d-1)}=\frac{\eta_2(d)+\Delta[\eta_2(d)]}{\eta_1(d)+\Delta[\eta_1(d)]}\ge\frac{\eta_2(d)}{\eta_1(d)+\Delta[\eta_1(d)]-\Delta[\eta_2(d)]}>\frac{\eta_2(d)}{\eta_1(d)}\ .\tag{19}$$

Hence the proof of Lemma 1 is concluded. □

4 PRGs and FDPs

As our second application of FDPs, in this section we present some results which prove that any (computationally indistinguishable) PRG with some parameter is also an nb-PRG with a (possibly different) specified parameter. The concrete relations between parameters for an algorithm as a PRG and as an nb-PRG, respectively, will be determined by applying FDPs.

First we recall the security notion for PRGs. We emphasize that, for the sake of simplicity, here we adopt definitions in forms of concrete security rather than asymptotic security. Let U_X denote the uniform probability distribution over a finite set X.

Definition 3 (see e.g., [4])**.** Let $G\colon S \to X$ be an algorithm with finite input set S and finite output set X. Given parameters $T \geq 0$ and $\varepsilon \geq 0$, G is called a (T, ε)-*secure pseudorandom generator* (*PRG*) if, for any algorithm (called a *distinguisher*) $D\colon X \to \{0, 1\}$ with time complexity bounded by T, we have $\mathsf{Adv}_D(G) \leq \varepsilon$ where $\mathsf{Adv}_D(G)$ denotes the *advantage* of D defined by

$$\mathsf{Adv}_D(G) := |Pr[D(U_X) = 1] - Pr[D(G(U_S)) = 1]| \ . \tag{20}$$

Let $\Delta(P_1, P_2)$ denote the statistical distance of two probability distributions P_1, P_2 over the same finite set Z defined by

$$\begin{aligned} \Delta(P_1, P_2) &:= \frac{1}{2} \sum_{z \in Z} |Pr[P_1 = z] - Pr[P_2 = z]| & (21) \\ &= \max_{E \subset Z} |Pr[P_1 \in E] - Pr[P_2 \in E]| \ . & (22) \end{aligned}$$

Then the advantage $\mathsf{Adv}_D(G)$ of a distinguisher D defined above is equal to $\Delta(D(U_X), D(G(U_S)))$, as both $D(U_X)$ and $D(G(U_S))$ are probability distributions over $\{0, 1\}$. This interpretation of the advantage gives us a motivation to enhance the above security notion of PRGs, as in the following definition introduced by Dubrov and Ishai [3] (with slightly different formulation):

Definition 4 ([3])**.** Let $G\colon S \to X$ be an algorithm with finite input set S and finite output set X. Given parameters $T \geq 0$, $\varepsilon \geq 0$ and an integer $n \geq 2$, G is called (T, n, ε)-*secure* if, for any algorithm (distinguisher) $D\colon X \to \{0, 1, \ldots, n-1\}$ with time complexity bounded by T, we have $\mathsf{Adv}_D(G) \leq \varepsilon$ where we put $\mathsf{Adv}_D(G) := \Delta(D(U_X), D(G(U_S)))$. Such an algorithm G is called a *PRG that fools non-boolean distinguishers* (*nb-PRG*, in short).

Note that $(T, 2, \varepsilon)$-security is equivalent to (T, ε)-security in Definition 3. Several applications of nb-PRGs are discussed in [3]. For example, it was shown that randomness used in some kinds of *information-theoretically secure* protocols (such as multi-party computation of certain types) can be replaced with outputs of nb-PRGs, without any restriction on computational complexity of the adversary against the protocol. However, despite the significance of nb-PRGs mentioned above, it seems much more difficult to construct secure nb-PRGs than the case of usual PRGs against 1-bit output distinguishers. Indeed, to the authors' best knowledge, the only constructions of nb-PRGs in the literature so far are the ones by Dubrov and Ishai themselves in the original paper [3], and their construction is based on certain computational assumption which is less standard than those used in constructions of usual PRGs. Hence, it is worthy to investigate a method to construct nb-PRGs (under standard computational assumptions).

Our proposal here is to establish a general theorem of the following form: Any (T', ε')-secure PRG is also a (T, n, ε)-secure nb-PRG, where the parameters T' and ε' as a usual PRG are determined by T, n and ε in a certain manner. Such an implication result is evidently meaningful, as it enables us to convert a large number of existing PRGs under standard assumptions into nb-PRGs. In fact, an implication relation as above has been mentioned (without proof) in [3]. Our aim here is to improve the preceding relation by introducing the idea of FDPs.

The above-mentioned relation is derived from the first expression (21) of statistical distance, in the following manner (which refers to a description in [7]). We introduce some notations. Put $Y := \{0, 1, \ldots, n-$

1} for simplicity. For any subset $Z \subset Y$, let $\chi_Z \colon Y \to \{0,1\}$ denote the characteristic function of Z defined by $\chi_Z(x) = 1$ if $x \in Z$ and $\chi_Z(x) = 0$ if $x \in Y \setminus Z$. We write $\chi_z = \chi_{\{z\}}$ for simplicity when $Z = \{z\}$. In this setting, for any PRG $G \colon S \to X$ and any non-boolean distinguisher $D \colon X \to Y$, the statistical distance $\Delta(D(U_X), D(G(U_S)))$ is equal to

$$\begin{aligned} &\frac{1}{2} \sum_{y \in Y} |Pr[D(U_X) = y] - Pr[D(G(U_S)) = y]| \\ &= \frac{1}{2} \sum_{y \in Y} |Pr[\chi_y \circ D(U_X) = 1] - Pr[\chi_y \circ D(G(U_S)) = 1]| \\ &= \frac{1}{2} \sum_{y \in Y} \mathsf{Adv}_{\chi_y \circ D}(G) \ , \end{aligned} \tag{23}$$

where $\chi_y \circ D$ denotes an algorithm performed by first executing the distinguisher D and then evaluating the output of D by the function χ_y. An important property is that $\chi_y \circ D$ is a 1-bit output algorithm, therefore it can be regarded as a distinguisher for the PRG G. This implies that, to show that a (T', ε')-secure PRG G is also a (T, n, ε)-secure nb-PRG, it suffices to choose the parameters as $T' = T + \delta_1$ and $\varepsilon' = 2\varepsilon/n$, where δ_1 is the maximum of the overhead in computational complexity of composing some χ_y $(y \in Y)$ to D (usually, δ_1 can be set to be almost zero in practical situations). In other words, we have the following proposition (which has been mentioned in [3]):

Proposition 1. *In this setting, any $(T + \delta_1, 2\varepsilon/n)$-secure PRG is also (T, n, ε)-secure, where the quantity δ_1 is defined in the above manner.*

A drawback of this result is that, in practical applications the parameter n (which is relevant to the allowable input size for an adversary against a protocol under consideration) should frequently be large, which makes the overhead in a bound of advantage in Proposition 1 too heavy. We try to resolve the drawback by improving or modifying the above result.

Our first idea is to use the second expression (22) of statistical distance instead of the first one (21) used in the preceding argument. Namely, in the same setting as above, the statistical distance $\Delta(D(U_X), D(G(U_S)))$ is equal to

$$\begin{aligned} &\max_{Z \subset Y} |Pr[D(U_X) \in Z] - Pr[D(G(U_S)) \in Z]| \\ &= \max_{Z \subset Y} |Pr[\chi_Z \circ D(U_X) = 1] - Pr[\chi_Z \circ D(G(U_S)) = 1]| \\ &= \max_{Z \subset Y} \mathsf{Adv}_{\chi_Z \circ D}(G) \ . \end{aligned} \tag{24}$$

In the same way as Proposition 1, the above argument implies the following result:

Proposition 2. *In this setting, any $(T + \delta_2, \varepsilon)$-secure PRG is also (T, n, ε)-secure, where δ_2 is the maximum of the overhead in computational complexity of composing some χ_Z with $Z \subset Y := \{0, 1, \ldots, n-1\}$ to D.*

In contrast to Proposition 1, there exists no overhead for a bound of advantage ε in Proposition 2. However, instead, the overhead δ_2 for a bound of time complexity of distinguishers is expected to be too heavy, as the set Y (of somewhat large size) may contain an extremely complicated subset Z, for which the computation of χ_Z would be inefficient.

From now, we try to improve the above-mentioned trade-off between overheads for bounds of advantage and of computational complexity, by applying the idea of FDPs. Put $Y := \{0, 1, \ldots, n-1\}$ as above, and let $\mathcal{C}$ be the set of characteristic functions $\chi_Z \colon Y \to \{0,1\}$ for subsets $Z \subset Y$, and let $d = d_{\mathrm{H}}$ (see (2)). Then for $\chi_{Y_1}, \chi_{Y_2} \in \mathcal{C}$, $d_{\mathrm{H}}(\chi_{Y_1}, \chi_{Y_2})$ is equal to the size of the symmetric difference $Y_1 \ominus Y_2 := (Y_1 \setminus Y_2) \cup (Y_2 \setminus Y_1)$ of two subsets Y_1 and Y_2. Now we fix a subset $\mathcal{C}'$ of $\mathcal{C}$. Let δ_3 be the maximum of the overhead in computational complexity of composing some $\chi_Z \in \mathcal{C}'$ to D. Moreover, we put $r := r(\mathcal{C}, \mathcal{C}')$ for simplicity. Then we have the following result (we notice that, when $\mathcal{C}' = \{\chi_\emptyset\}$, the theorem gives almost the same result as Proposition 1):

Theorem 1. *In the above situation, let δ_1 be as specified in Proposition 1. If $G\colon S \to X$ is $(T+\delta_1, \varepsilon_1)$-secure and $(T+\delta_3, \varepsilon_3)$-secure, then G is also $(T, n, r\varepsilon_1+\varepsilon_3)$-secure.*

Proof. For each distinguisher $D\colon X \to Y := \{0, 1, \ldots, n-1\}$, we write $\mu(Z) := Pr[D(U_X) \in Z]$ and $\mu'(Z) := Pr[D(G(U_S)) \in Z]$ for a subset $Z \subset Y$. Let Y_0 be a subset of Y that attains the maximum of the second expression (22) of the statistical distance;

$$\Delta(D(U_X), D(G(U_S))) = |\mu(Y_0) - \mu'(Y_0)| \ . \tag{25}$$

Note that Y_0 can be chosen in such a way that $\mu(Y_0) - \mu'(Y_0) \geq 0$ (if this inequality fails, use $Y \setminus Y_0$ instead of Y_0), therefore

$$\Delta(D(U_X), D(G(U_S))) = \mu(Y_0) - \mu'(Y_0) \ . \tag{26}$$

Moreover, by the definition of r, there is a subset $Y_1 \subset Y$ satisfying that $\chi_{Y_1} \in \mathcal{C}'$ and $d_{\mathrm{H}}(\chi_{Y_0}, \chi_{Y_1}) = |Y_0 \ominus Y_1| \leq r$. Now we have

$$\nu(Y_0) - \nu(Y_1) = \nu(Y_0 \setminus Y_1) - \nu(Y_1 \setminus Y_0) \text{ for each } \nu \in \{\mu, \mu'\} \ , \tag{27}$$

therefore we have

$$\begin{aligned} &(\mu(Y_0) - \mu'(Y_0)) - (\mu(Y_1) - \mu'(Y_1)) \\ &= (\mu(Y_0) - \mu(Y_1)) - (\mu'(Y_0) - \mu'(Y_1)) \\ &= (\mu(Y_0 \setminus Y_1) - \mu'(Y_0 \setminus Y_1)) - (\mu(Y_1 \setminus Y_0) - \mu'(Y_1 \setminus Y_0)) \ . \end{aligned} \tag{28}$$

Moreover, the right-hand side is equal to

$$\begin{aligned} &\sum_{y \in Y_0 \setminus Y_1} (\mu(\{y\}) - \mu'(\{y\})) - \sum_{y \in Y_1 \setminus Y_0} (\mu(\{y\}) - \mu'(\{y\})) \\ &\leq \sum_{y \in Y_0 \ominus Y_1} |\mu(\{y\}) - \mu'(\{y\})| \\ &= \sum_{y \in Y_0 \ominus Y_1} |Pr[\chi_y \circ D(U_X) = 1] - Pr[\chi_y \circ D(G(U_S)) = 1]| \\ &= \sum_{y \in Y_0 \ominus Y_1} \mathsf{Adv}_{\chi_y \circ D}(G) \ . \end{aligned} \tag{29}$$

Now if D has computational complexity bounded by T, then the assumption on G and the definition of δ_1 imply that

$$\sum_{y \in Y_0 \ominus Y_1} \mathsf{Adv}_{\chi_y \circ D}(G) \leq \sum_{y \in Y_0 \ominus Y_1} \varepsilon_1 = |Y_0 \ominus Y_1| \cdot \varepsilon_1 \leq r\varepsilon_1 \ . \tag{30}$$

Summarizing, we have

$$(\mu(Y_0) - \mu'(Y_0)) - (\mu(Y_1) - \mu'(Y_1)) \leq r\varepsilon_1 \ . \tag{31}$$

This and (26) implies that

$$\begin{aligned} &\Delta(D(U_X), D(G(U_S))) \\ &= (\mu(Y_0) - \mu'(Y_0)) - (\mu(Y_1) - \mu'(Y_1)) + (\mu(Y_1) - \mu'(Y_1)) \\ &\leq r\varepsilon_1 + (Pr[D(U_X) \in Y_1] - Pr[D(G(U_S)) \in Y_1]) \\ &\leq r\varepsilon_1 + |Pr[\chi_{Y_1} \circ D(U_X) = 1] - Pr[\chi_{Y_1} \circ D(G(U_S)) = 1]| \\ &= r\varepsilon_1 + \mathsf{Adv}_{\chi_{Y_1} \circ D}(G) \leq r\varepsilon_1 + \varepsilon_3 \ , \end{aligned} \tag{32}$$

concluding the proof of Theorem 1. □

Regarding the relation between parameters in Theorem 1, first note that it is natural by the definitions to expect that $\delta_1 \leq \delta_3$, which allows us to suppose that $\varepsilon_1 \leq \varepsilon_3$. Now let us imagine the following situation: We can find an appropriate subset $\mathcal{C}' \subset \mathcal{C}$ in such a way that every characteristic function $\chi_Z \in \mathcal{C}'$ has low computational complexity and the quantity $r := r(\mathcal{C}, \mathcal{C}')$ is small. In this case, δ_3 can be small as well as r, and it would make the implication relation given by Theorem 1 more efficient than those in Propositions 1 and 2, therefore the above-mentioned trade-off is improved. Hence a study of FDPs (in particular, those for functions with 1-bit output sets) will contribute to establish a better relation between PRGs and nb-PRGs.

Remark 2. We mention that, for the two applications of FDPs discussed in the last two sections, a kind of "risk-hedging" relation exists as follows. Namely, if we find that the quantity $r(\mathcal{C}, \mathcal{C}')$ tends to be large in general, then it would support the argument in Section 3 to show that keyless hash functions under consideration would have better security. On the other hand, if we find that the quantity $r(\mathcal{C}, \mathcal{C}')$ tends to be small in general, then it would support the argument in Section 4 to show that overheads in parameters for nb-PRGs compared to PRGs would be practically small.

5 Mathematical examples for FDPs

This section is devoted to describe some examples for mathematical studies of FDPs themselves, rather than their cryptographic applications such as ones discussed in Sections 3 and 4. The authors hope that one would feel that FDPs themselves are of independent interest as mathematical problems and mathematical studies of FDPs will be promoted.

5.1 Vector spaces and their subspaces: A general bound

The examples of FDPs discussed below can be interpreted in the following manner. The set $\mathcal{C}$ forms a finite-dimensional vector space over a finite field $\mathbb{F}$, with a distinguished basis $v_1, \ldots, v_d$ where $d := \dim(\mathcal{C})$, hence each element of $\mathcal{C}$ admits a vector expression. A subset $\mathcal{C}'$ is a linear subspace of $\mathcal{C}$, and the distance $d(f, g)$ is defined to be the (generalized) Hamming distance with respect to the vector expressions of $f, g \in \mathcal{C}$. In this subsection, we show a general upper and lower bounds of the quantity $r(\mathcal{C}, \mathcal{C}')$ in this case. Namely, we have the following:

Proposition 3. *In the above setting, let ℓ denote the minimal integer ℓ' satisfying that $\sum_{i=0}^{\ell'} \binom{d}{i}(|\mathbb{F}|-1)^i \geq |\mathbb{F}|^{\mathrm{codim}_{\mathcal{C}}(\mathcal{C}')}$, where $\mathrm{codim}_{\mathcal{C}}(\mathcal{C}')$ denotes the codimension $d - \dim(\mathcal{C}')$ of $\mathcal{C}'$ in $\mathcal{C}$. Then we have $\ell \leq r(\mathcal{C}, \mathcal{C}') \leq \mathrm{codim}_{\mathcal{C}}(\mathcal{C}')$.*

Proof. Put $d' := \dim(\mathcal{C}')$, therefore $\mathrm{codim}_{\mathcal{C}}(\mathcal{C}') = d - d'$. First we prove the lower bound. For each $w \in \mathcal{C}$ and $k \geq 0$, put $B(w, k) := \{w' \in \mathcal{C} \mid d(w, w') \leq k\}$. Then we have $|B(w, k)| = \sum_{i=0}^{k} \binom{d}{i}(|\mathbb{F}|-1)^i$. On the other hand, by the definition of $r(\mathcal{C}, \mathcal{C}')$, we have $\mathcal{C} \subset \bigcup_{w \in \mathcal{C}'} B(w, r(\mathcal{C}, \mathcal{C}'))$. This implies that

$$|\mathcal{C}| \leq |\mathcal{C}'| \cdot \sum_{i=0}^{r(\mathcal{C},\mathcal{C}')} \binom{d}{i}(|\mathbb{F}|-1)^i \ , \tag{33}$$

or equivalently $|\mathbb{F}|^{d-d'} = |\mathcal{C}|/|\mathcal{C}'| \leq \sum_{i=0}^{r(\mathcal{C},\mathcal{C}')} \binom{d}{i}(|\mathbb{F}|-1)^i$. Hence we have $\ell \leq r(\mathcal{C}, \mathcal{C}')$ by the choice of ℓ.

Secondly, we prove the upper bound. By applying Gaussian elimination to any basis of $\mathcal{C}'$, it follows that there exist a basis $u_1, \ldots, u_{d'}$ of $\mathcal{C}'$ and distinct indices $i_1, \ldots, i_{d'} \in \{1, 2, \ldots, d\}$ with the property that, for each $1 \leq j \leq d'$, the coefficient of a basis element v_{i_j} of $\mathcal{C}$ in u_j is 1 and the coefficient of v_{i_j} in any other u_k $(k \neq j)$ is 0. Now for an arbitrary element $w = \sum_{i=1}^{d} c_i v_i \in \mathcal{C}$ $(c_i \in \mathbb{F})$, the above property of $u_1, \ldots, u_{d'}$ implies that the distance between w and $w' := \sum_{j=1}^{d'} c_{i_j} u_j \in \mathcal{C}'$ is at most $d - d'$, therefore $d(w, \mathcal{C}') \leq d - d'$. Hence we have $r(\mathcal{C}, \mathcal{C}') \leq d - d'$, concluding the proof of Proposition 3. □

The next result shows how the lower and upper bounds in Proposition 3 are close to each other:

Proposition 4. *In the setting of Proposition 3, we have*

$$\begin{aligned}\ell \le \mathrm{codim}_{\mathcal{C}}(\mathcal{C}') &\le \log_{|\mathbb{F}|}\left(c_\ell(|\mathbb{F}|-1)^\ell d^\ell/\ell!\right)\\ &= \ell(\log_{|\mathbb{F}|}(|\mathbb{F}|-1)+\log_{|\mathbb{F}|} d)+\log_{|\mathbb{F}|} c_\ell - \log_{|\mathbb{F}|}\ell!\ ,\end{aligned} \tag{34}$$

where $c_\ell = \ell+1$ if $\mathbb{F}$ is the two-element field $\mathbb{F}_2$, and $c_\ell = (|\mathbb{F}|-1)/(|\mathbb{F}|-2)$ otherwise.

Proof. It suffices to prove the second inequality. As $|\mathbb{F}|^{\mathrm{codim}_{\mathcal{C}}(\mathcal{C}')} \le \sum_{i=0}^{\ell}\binom{d}{i}(|\mathbb{F}|-1)^i$ by the definition of ℓ, it suffices to show that $\sum_{i=0}^{\ell}\binom{d}{i}(|\mathbb{F}|-1)^i \le c_\ell(|\mathbb{F}|-1)^\ell d^\ell/\ell!$, or more generally, $\sum_{i=0}^{m}\binom{N}{i}(q-1)^i \le c'_m(q-1)^m N^m/m!$ for all integers $N \ge m \ge 0$ and $q \ge 2$, where we put $c'_m := (q-1)/(q-2)$ if $q \ge 3$ and $c'_m := m+1$ if $q = 2$, and we set $0^0 = 1$ (note that $\ell \le \mathrm{codim}_{\mathcal{C}}(\mathcal{C}') \le d$). We use induction on m. The case $m = 0$ is trivial. For the case $m \ge 1$, we have

$$\begin{aligned}\sum_{i=0}^{m}\binom{N}{i}(q-1)^i &= \sum_{i=0}^{m-1}\binom{N}{i}(q-1)^i + \binom{N}{m}(q-1)^m\\ &\le \frac{c'_{m-1}(q-1)^{m-1}N^{m-1}}{(m-1)!} + \binom{N}{m}(q-1)^m\\ &\le \frac{c'_{m-1}(q-1)^{m-1}N^{m-1}}{(m-1)!} + \frac{(q-1)^m N^m}{m!}\\ &= \frac{(q-1)^m N^m}{m!}\left(\frac{c'_{m-1}m}{(q-1)N}+1\right)\ .\end{aligned} \tag{35}$$

By the relation $m \le N$ and the definition of c'_m, we have

$$\frac{c'_{m-1}m}{(q-1)N}+1 \le \frac{c'_{m-1}}{q-1}+1 = c'_m\ , \tag{36}$$

therefore the desired inequality holds for this m as well. Hence the claim of Proposition 4 holds. □

5.2 Boolean functions of low degrees

As a first concrete example, here we deal with the set $\mathcal{C}$ of the functions $X \to Y$ with n-bit inputs and 1-bit outputs, i.e., we set $X := \{0,1\}^n$ and $Y := \{0,1\}$ (which is relevant to the situation of Section 4). First note that, when we identify $\{0,1\}$ naturally with $\mathbb{F}_2$, each function $f\colon X \to Y$ can be expressed as an n-variable square-free polynomial;

$$f(x_1,\dots,x_n) = \sum_{\vec{a}=(a_1,\dots,a_n)\in\{0,1\}^n} f(\vec{a})\chi_{\vec{a}}(x_1,\dots,x_n) \tag{37}$$

where we put

$$\chi_{\vec{a}}(x_1,\dots,x_n) := \prod_{i;a_i=0}(1-x_i)\prod_{i;a_i=1}x_i \text{ for } \vec{a}=(a_1,\dots,a_n) \tag{38}$$

(note that $\chi_{\vec{a}}(x_1,\dots,x_n) = 1$ if $x_i = a_i$ for every i and $\chi_{\vec{a}}(x_1,\dots,x_n) = 0$ otherwise, therefore $\chi_{\vec{a}}$ is indeed the characteristic function of $\vec{a} \in \{0,1\}^n$). For example, when $n = 2$ we have

$$\begin{aligned}f(x_1,x_2) = f(0,0)(1-x_1)(1-x_2) + f(0,1)(1-x_1)x_2\\ + f(1,0)x_1(1-x_2) + f(1,1)x_1x_2\ .\end{aligned} \tag{39}$$

Now for each $0 \le k \le n$, we set $\mathcal{C}' = \mathcal{C}'_k$ to be the subset of $\mathcal{C}$ consisting of functions that can be expressed as a square-free polynomial of degree $\le k$. For example, $\mathcal{C}'_0$ is the set of constant functions, and $\mathcal{C}'_1$ is the set of affine functions. The distance $d(f,g) = d_{\mathrm{H}}(f,g)$ is defined as in (2). Note that changing the value of $f \in \mathcal{C}$ at a point $\vec{a} \in \{0,1\}^n$ is equivalent to adding the function $\chi_{\vec{a}}$ to the f. In this situation, we have the following upper and lower bounds for the quantity $r(\mathcal{C},\mathcal{C}'_k)$:

Proposition 5. *In the above setting, put* $u_{n,k} := \sum_{i=k+1}^{n} \binom{n}{i}$*, and let* $\ell_{n,k}$ *be the minimum integer* ℓ *satisfying that* $2^{u_{n,k}} \leq \sum_{i=0}^{\ell} \binom{2^n}{i}$*. Then we have*

$$\ell_{n,k} \leq r(\mathcal{C}, \mathcal{C}'_k) \leq \min\{u_{n,k}, 2^{n-1}\} \ . \tag{40}$$

Proof. For the upper bound, note that $r(\mathcal{C}, \mathcal{C}'_k) \leq 2^{n-1}$, as any function $f \in \mathcal{C}$ can be converted into a constant function by changing the value $f(x)$ at every point $x \in \{0,1\}^n$ with the property that $f(x)$ is in the minority among the 2^n values of f (the number of such points is at most 2^{n-1}). Then the upper bound follows from Proposition 3, as $\mathcal{C}$ is an $\mathbb{F}_2$-vector space of dimension 2^n and $\mathcal{C}'_k$ is its subspace of codimension $u_{n,k}$. The lower bound also follows from Proposition 3. □

By Proposition 4, the quantities $\ell_{n,k}$ and $u_{n,k}$ in Proposition 5 satisfy the relation $\ell_{n,k} \leq u_{n,k} \leq n\ell_{n,k} + \log_2(\ell_{n,k}+1) - \log_2 \ell_{n,k}!$. Table 1 gives the precise values of $\ell_{n,k}$ for some smaller cases.

Table 1: The values of $\ell_{n,k}$ for some small parameters

		$n-k$							
		1	2	3	4	5	6	7	8
	2	1	2						
	3	1	2	4					
	4	1	2	4	8				
n	5	1	2	5	10	16			
	6	1	2	5	13	22	32		
	7	1	2	6	16	31	49	64	
	8	1	2	6	19	43	75	105	128

Here we introduce a geometric point of view to the above problem. We introduce some notations. For a subset $I \subset [n] := \{1, 2, \ldots, n\}$, put $x_I := \prod_{i \in I} x_i$, and let a_I be the element $(a_1, \ldots, a_n)$ of $\{0,1\}^n$ determined by $a_i = 1$ when and only when $i \in I$. We write $\delta_I := \chi_{a_I}$ for simplicity. Let Δ_+^{n-1} be the disjoint union of an isolated point P and the standard $(n-1)$-simplex Δ^{n-1} on the vertex set $[n]$; we regard P as "the (-1)-dimensional face" of Δ_+^{n-1}. For each $\emptyset \neq I \subset [n]$, let $\langle I \rangle$ denote the $(|I|-1)$-dimensional sub-simplex of Δ^{n-1} spanned by I, and let $\langle I \rangle^o$ be its relative interior (note that $\langle \{i\} \rangle^o = \langle \{i\} \rangle = \{i\}$ for each $i \in [n]$). On the other hand, we put $\langle \emptyset \rangle = \langle \emptyset \rangle^o := P$. Now for each function $f(x) = \sum_{I \subset [n]} c_I x_I$ ($c_I \in \mathbb{F}_2$), we define its *geometric realization* G_f by

$$G_f := \bigcup_{I; c_I = 1} \langle I^c \rangle^o \quad \text{(disjoint union)}, \tag{41}$$

where I^c denotes the complement $[n] \setminus I$ of I in $[n]$. For each $I \subset [n]$, by the definition and the fact that $\delta_I = \sum_{J \supset I} x_J$ (recall that now the values of functions are in $\mathbb{F}_2$), G_{δ_I} is the (disjoint) union of P and $\langle J \rangle^o$ for all $\emptyset \neq J \subset I^c$, therefore we have $G_{\delta_I} = P \cup \langle I^c \rangle$. Moreover, for any $0 \leq k \leq n$ and $I \subset [n]$, we have $|I| \geq k+1$ if and only if $\langle I^c \rangle$ is at most $(n-k-2)$-dimensional. This implies that a function $f \in \mathcal{C}$ belongs to $\mathcal{C}'_k$ if and only if G_f does not intersect with the $(n-k-1)$-dimensional skeleton Δ^n_{n-k-1} of Δ_+^{n-1}, which consists of the faces of Δ_+^{n-1} of dimension up to $n-k-1$.

Based on the above observation, we consider the following puzzle. We imagine a situation that a lamp is associated to each face of Δ_+^{n-1}. A *state* of Δ_+^{n-1} is a collection of light/dark properties of all the lamps. Given a function f, the corresponding *initial state* $\mathcal{I}_f$ is defined in such a way that a lamp at a face is light if and only if the relative interior of the face is contained in G_f. At any state, the player of the puzzle is allowed to indicate a face F of Δ_+^{n-1} (we call it "*push the face* F"), then the light/dark properties of lamps at P and every sub-face of F are flipped; such a process is regarded as a *move* of the puzzle. An initial state $\mathcal{I}_f$ is said to be *solved* when the lamps of all faces of Δ^n_{n-k-1} are switched off by a sequence of moves started from $\mathcal{I}_f$. With this interpretation, the distance $d(f, \mathcal{C}'_k)$ from $f \in \mathcal{C}$ to $\mathcal{C}'_k$ is the minimum of the number of moves to solve $\mathcal{I}_f$, and the quantity $r(\mathcal{C}, \mathcal{C}'_k)$ is the minimal necessary number of moves to solve *any* initial state.

Moreover, we also introduce a simplified puzzle on Δ^{n-1} instead of Δ_+^{n-1} by ignoring the isolated point P in the above puzzle. Let $r'_{n,k}$ denote the minimal necessary number of moves to solve (for the simplified puzzle) any initial state. Then we have $r(\mathcal{C}, \mathcal{C}'_k) = r'_{n,k} + 1$, as for an initial state $\mathcal{I}$ of the simplified puzzle for which solving $\mathcal{I}$ requires precisely $r'_{n,k}$ moves, one of the two initial states of the original puzzle obtained by adding a lamp at P which is light and dark, respectively, requires $r'_{n,k} + 1$ moves. Hence it suffices to consider the simplified puzzle on Δ^{n-1} for determining the quantity $r(\mathcal{C}, \mathcal{C}'_k)$.

Example 1. We set $n = 4$ and show that $r(\mathcal{C}, \mathcal{C}'_1) = 6$, or equivalently $r'_{4,1} = 5$. Note that the general bounds in Proposition 5 only guarantee that $4 \leq r(\mathcal{C}, \mathcal{C}'_1) \leq 8$ (note that $u_{4,1} = 11 > 8 = 2^{4-1}$). We identify naturally each state in the puzzle on $\Delta^{n-1} = \Delta^3$ with each family of non-empty subsets of $[n] = [4]$, and we write $\{i_1, i_2, \ldots, i_\ell\}$ as $i_1 i_2 \cdots i_\ell$ for simplicity. Moreover, to express each state we omit the subsets of [4] of size larger than 2, as the lamps at faces of dimension at least $n - k - 1 = 2$ are not relevant to determine whether the puzzle has been solved or not. In other words, in the present situation, we can regard each state as edge and vertex coloring of the complete graph K_4.

First, we show that the initial state $\mathcal{I} = \{13, 24\}$ requires more than 4 moves to solve. Assume contrary that $\mathcal{I}$ can be solved by at most 4 moves. If the player pushes the face 1234, then a state $\{1, 2, 3, 4, 12, 23, 34, 41\}$ is obtained. To solve the state by at most 3 remaining moves, the player has to push at least one 2-dimensional face; we may assume by symmetry that the face is 123. Then the resulting state is $\{4, 13, 34, 41\}$; however, a case-by-case analysis shows that to solve the state by at most 2 remaining moves is impossible. Therefore the player does not push the face 1234. On the other hand, if the player pushes a 2-dimensional face, then we may assume by symmetry that the face is 123, resulting in a state $\{1, 2, 3, 12, 23, 24\}$. To solve the state by at most 3 remaining moves, the player has to push at least one more 2-dimensional face. If it is 124, then we obtain a state $\{3, 4, 23, 41\}$, but a case-by-case analysis shows that to solve the state by at most 2 remaining moves is impossible (the case of 234 is similar by symmetry). If it is 134, then we obtain a state $\{2, 4, 12, 13, 14, 23, 24, 34\}$, but a case-by-case analysis shows that to solve the state by at most 2 remaining moves is impossible as well. Therefore the player does not push a 2-dimensional face. This implies that the player should push 13 and 24, resulting in a state $\{1, 2, 3, 4\}$, from which to solve the state by at most 2 remaining moves is impossible. Hence we have a contradiction, therefore the initial state $S = \{13, 24\}$ indeed requires more than 4 moves to solve.

Secondly, we show that any initial state $\mathcal{I}$ can be solved by at most 5 moves. The player can solve $\mathcal{I}$ by at most 4 moves when no lamps in $\mathcal{I}$ at 1-dimensional faces are light, therefore $\mathcal{I}$ can be solved by at most 5 moves when at most 1 lamp in $\mathcal{I}$ at 1-dimensional face is light. When 2 lamps in $\mathcal{I}$ at 1-dimensional faces are light, a case-by-case analysis shows that $\mathcal{I}$ can be solved by at most 4 moves unless $\mathcal{I}$ is of the form $\{i_1 i_2, i_3 i_4\}$ with $\{i_1, i_2\} \cap \{i_3, i_4\} = \emptyset$, and for any $\mathcal{I}$ of the latter form, $\mathcal{I}$ can be solved by pushing the faces 1234, $i_1 i_2 i_3$, $i_1 i_2 i_4$, i_1, and i_2. When 3 lamps in $\mathcal{I}$ at 1-dimensional faces are light, the problem can be reduced to the case of 2 light lamps at 1-dimensional faces by pushing one of the 3 light lamps at 1-dimensional faces. When 4 lamps in $\mathcal{I}$ at 1-dimensional faces are light, the problem can be reduced to the case of 2 light lamps at 1-dimensional faces by pushing the face 1234 unless $\mathcal{I}$ is of the form $\{1, 2, 3, 4, i_1 i_3, i_1 i_4, i_2 i_3, i_2 i_4\}$ with $\{i_1, i_2\} \cap \{i_3, i_4\} = \emptyset$, and for any $\mathcal{I}$ of the latter form, $\mathcal{I}$ can be solved by pushing the faces $i_1 i_2 i_3$, $i_1 i_2 i_4$, i_1, and i_2. When 5 lamps in $\mathcal{I}$ at 1-dimensional faces are light, the problem can be reduced to the case of 2 light lamps at 1-dimensional faces by pushing an appropriate 2-dimensional face. Finally, when 6 lamps in $\mathcal{I}$ at 1-dimensional faces are light, the problem can be reduced to the case of no light lamps at 1-dimensional faces by pushing the face 1234. Hence any initial state $\mathcal{I}$ can be solved by at most 5 moves, therefore we have $r'_{4,1} = 5$ as desired.

From now, we investigate FDPs in the above setting by using Gröbner bases. Recall that $X = \{0, 1\}^n$. Let $R := K[z_v \mid v \in X]$ be a polynomial ring in 2^n variables over a field K of characteristic 0. We define the following ideal of R:

$$I_0 := ({z_v}^2 - 1 \mid v \in X) \subset R \ . \tag{42}$$

For each $f \in \mathcal{C}$, put

$$z^f := \prod_{v \in X} {z_v}^{f(v)} \ . \tag{43}$$

Then the set $\{z^f \mid f \in \mathcal{C}\}$ of all square-free monomials in R forms a linear basis of the quotient ring $A_0 := R/I_0$. Note that $z^f z^g = z^{f+g} \pmod{I_0}$ and the degree $\deg(z^f)$ of z^f in R is equal to $d(f, \underline{0})$ for any $f, g \in \mathcal{C}$, where $\underline{0}$ denotes the function in $\mathcal{C}$ taking constant value 0.

Let $\mathcal{C}'$ be a subset of $\mathcal{C}$, which need not be a linear subspace of $\mathcal{C}$ unless otherwise specified. We define the following ideal of R:

$$\check{I}_{\mathcal{C}'} := (z^f - z^g \mid f, g \in \mathcal{C}') \subset R \ , \tag{44}$$

and consider the ideal $I_{\mathcal{C}'} := I_0 + \check{I}_{\mathcal{C}'}$ of R. We identify the quotient ring $A_{\mathcal{C}'} := R/I_{\mathcal{C}'}$ with the quotient ring of A_0 by the image of $\check{I}_{\mathcal{C}'}$. Now fix a graded monomial order, i.e., a monomial order $\prec$ satisfying that $\prod_{v\in X} z_v{}^{\alpha_v} \prec \prod_{v\in X} z_v{}^{\beta_v}$ for any exponents $(\alpha_v)_{v\in X}$ and $(\beta_v)_{v\in X}$ with $\sum_{v\in X} \alpha_v < \sum_{v\in X} \beta_v$. Let G be a Gröbner basis for the ideal $I_{\mathcal{C}'}$, and consider the reduction process with respect to the Gröbner basis G. As each generator of $I_{\mathcal{C}'}$ is of the form "(monic monomial) − (monic monomial)", G can be chosen in such a way that every element of G is of the same form, and the linear basis of A_0 consisting of the square-free monic monomials can be partitioned into equivalence classes when projected onto the quotient ring $A_{\mathcal{C}'}$. This also implies that the normal form $\mathrm{nf}(z^f)$ of each $f \in \mathcal{C}$ with respect to G is a square-free monic monomial, i.e., of the form z^g with $g \in \mathcal{C}$, and we have

$$\begin{aligned} \deg(\mathrm{nf}(z^f)) &= \min\{\deg(z^{f'}) \mid z^{f'} = z^f \pmod{I_{\mathcal{C}'}}\} \\ &= \min\{\deg(z^{f'}) \mid z^{f'} - z^f \in I_{\mathcal{C}'}\} \ . \end{aligned} \tag{45}$$

Now we consider the case that $\underline{0} \in \mathcal{C}'$. Note that $z^f z^f = 1 = z^{\underline{0}} \pmod{I_0}$ for any $f \in \mathcal{C}$. Now if $f, g \in \mathcal{C}$ and $z^f = z^g \pmod{I_{\mathcal{C}'}}$, then we have $z^{f+g} = z^f z^g = z^f z^f = z^{\underline{0}} \pmod{I_{\mathcal{C}'}}$, therefore $f + g \in \mathcal{C}'$. Conversely, if $f + g \in \mathcal{C}'$, then we have $z^f z^g = z^{f+g} = z^{\underline{0}} = 1 \pmod{I_{\mathcal{C}'}}$, therefore $z^f = z^f z^g z^g = z^g \pmod{I_{\mathcal{C}'}}$. Hence $\deg(\mathrm{nf}(z^f))$ is equal to the minimal degree of z^g with $g \in \mathcal{C}$ satisfying that $f + g \in \mathcal{C}'$, therefore $d(f, \mathcal{C}') = \deg(\mathrm{nf}(z^f))$. This argument reduces the FDP in this setting to the problem of computing (the degrees of) the normal forms of square-free monomials. More precisely, let h_i denote the number of monic monomials in $A_{\mathcal{C}'}$ whose normal forms have degree i, and put $s := \max\{i \mid h_i > 0\}$. (If the ideal is homogeneous, then $(h_i)_i$ is called the Hilbert function and it does not depend on the choice of a monomial order.) Now the above argument implies that $r(\mathcal{C}, \mathcal{C}') = s$. Moreover, if $\mathcal{C}'$ is a linear subspace of $\mathcal{C}$, then we have $h_i \cdot |\mathcal{C}'| = |\{f \in \mathcal{C} \mid d(f, \mathcal{C}') = i\}|$, therefore the data $(h_i)_i$ express the distributions of the distances $d(f, \mathcal{C}')$ over the functions $f \in \mathcal{C}$.

Based on the above argument, Proposition 3 can be restated for the present case as follows:

Proposition 6. *In the above setting, suppose that $\mathcal{C}'$ is a linear subspace of $\mathcal{C}$. Then we have*

$$\min\{\ell \mid \sum_{i=0}^{\ell} \binom{n}{i} \geq \frac{2^n}{|\mathcal{C}'|}\} \leq r(\mathcal{C}, \mathcal{C}') \leq \frac{2^n}{|\mathcal{C}'|} \ . \tag{46}$$

Proof. Note that the number of monic monomials in $A_{\mathcal{C}'}$ is $2^n/|\mathcal{C}'|$. Then the lower bound follows from the fact that the normal form of each monic monomial is also a monic monomial and that there exist $\binom{n}{i}$ square-free monic monomials of degree i, hence $h_i \leq \binom{n}{i}$. On the other hand, the upper bound is deduced from the fact that each divisor of a monic monomial of normal form is also of normal form, hence $h_i = 0$ if $h_j = 0$ and $j < i$. This concludes the proof. □

For the case $\mathcal{C}' = \mathcal{C}'_k$ as discussed above, Table 2 shows a calculation result of $r(\mathcal{C}, \mathcal{C}'_k)$ and $(h_i)_i$ for small parameters n and k, which is obtained by using computer algebra software `Singular/Sage`. By the table, we have $r(\mathcal{C}, \mathcal{C}'_k) = 6$ when $(n, k) = (4, 1)$, as explained in Example 1. Note that the values of $r(\mathcal{C}, \mathcal{C}'_k)$ in Table 2 are consistent with the lower bounds shown in Table 1.

5.3 Perfect codes and Reed–Solomon codes

In this subsection, we consider the case that $\mathcal{C}$ is an n-dimensional vector space over the q-element field $\mathbb{F}_q$, hence $\mathcal{C}$ is identified with $\mathbb{F}_q{}^n$, the distance $d(\cdot, \cdot)$ is defined to be the (generalized) Hamming distance

Table 2: Computer calculation result for some small parameters

n	k	$r(\mathcal{C},\mathcal{C}'_k)$	$(h_i)_{i\geq 0}$
2	1	1	$(1,1)$
3	1	2	$(1,8,7)$
3	2	1	$(1,1)$
4	1	6	$(1,16,120,560,875,448,28)$
4	2	2	$(1,16,15)$
4	3	1	$(1,1)$

(with respect to the vector expressions of elements), and $\mathcal{C}'$ is a linear subspace of $\mathcal{C}$ coming from the coding theory. Let the subspace $\mathcal{C}'$ be an (n,m,d)-code, i.e., $\dim(\mathcal{C}')=m$ and the minimum distance of $\mathcal{C}'$ is d. By the definition of minimum distance, we have the following well-known relation

$$q^m \sum_{i=0}^{\lfloor d/2\rfloor} \binom{n}{i}(q-1)^i \leq q^n \ . \tag{47}$$

This and the argument in Proposition 3 implies that $r(\mathcal{C},\mathcal{C}')\geq \lfloor d/2\rfloor$.

We say that $\mathcal{C}'$ is a *perfect code*, if the equality holds in (47). For a perfect code $\mathcal{C}'$, the above argument and Proposition 3 implies that $r(\mathcal{C},\mathcal{C}')=\lfloor d/2\rfloor$. For example, if $n=2^k-1$, $q=2$ and $\mathcal{C}'$ is the Hamming code H_k which is a $(2^k-1,2^k-k-1,3)$-code, then we have $r(\mathcal{C},\mathcal{C}')=1$. On the other hand, if $n=23$, $q=2$ and $\mathcal{C}'$ is the binary Golay code G_{23} which is a perfect $(23,12,7)$-code, then we have $r(\mathcal{C},\mathcal{C}')=3$. (In the case of the extended Golay code $\mathcal{C}'=G_{24}$ which is a nearly perfect $(24,12,8)$-code, where we set $n=24$ and $q=2$, we also have $r(\mathcal{C},\mathcal{C}')=4$ in a similar manner.)

As another concrete class of $\mathcal{C}'$ for which the quantity $r(\mathcal{C},\mathcal{C}')$ can be explicitly determined, from now we study the case of *Reed–Solomon codes*, which is also an important class of linear codes. We write $q=p^e$ with a prime number p and an integer $e\geq 1$, and choose an integer k with $1\leq k<n$. Take a primitive element α of $\mathbb{F}_q$, i.e., ${\mathbb{F}_q}^\times=\langle\alpha\rangle$. Define a polynomial $G(x)\in\mathbb{F}_q[x]$ of degree $n-k$ by

$$G(x):=(x-1)(x-\alpha)(x-\alpha^2)\cdots(x-\alpha^{n-k-1}) \ . \tag{48}$$

For any integer $j\geq 0$, let P_j denote the set of polynomials in $\mathbb{F}_q[x]$ of degrees up to j, which is a $(j+1)$-dimensional $\mathbb{F}_q$-linear subspace of $\mathbb{F}_q[x]$. We identify P_{n-1} with $\mathcal{C}$ via the correspondence $\sum_{i=0}^{n-1}a_ix^i \mapsto \sum_{i=0}^{n-1}a_iv_i$, where $(v_0,\ldots,v_{n-1})$ is a distinguished linear basis of $\mathcal{C}$. Now we introduce the following two linear maps:

$$\varphi_{n,k}\colon P_{k-1}\to P_{n-1}\,,\ f(x)\mapsto G(x)f(x) \ , \tag{49}$$

$$\psi_{n,k}\colon P_{n-1}\to {\mathbb{F}_q}^{n-k}\,,\ f(x)\mapsto (f(1),f(\alpha),f(\alpha^2),\ldots,f(\alpha^{n-k-1})) \ . \tag{50}$$

Let $\mathcal{C}'$ be the image of $\varphi_{n,k}$, which is a subspace of $\mathcal{C}$ (via the above identification $\mathcal{C}\simeq P_{n-1}$). This $\mathcal{C}'$ is a Reed–Solomon code. Note that $\mathcal{C}'$ coincides with the kernel of $\psi_{n,k}$. Now we have the following result:

Proposition 7. *In the above setting of Reed–Solomon code, we have $r(\mathcal{C},\mathcal{C}')=n-k$.*

Proof. As $\dim(\mathcal{C}')=k$, the inequality $r(\mathcal{C},\mathcal{C}')\leq n-k$ follows from Proposition 3. From now, we show that $r(\mathcal{C},\mathcal{C}')\geq n-k$, or equivalently, there exists an element $u\in\mathcal{C}$ satisfying that $d(u,\mathcal{C}')\geq n-k$.

For each polynomial $f(x)\in P_{n-1}$, the condition $d(f(x),\mathcal{C}')\leq n-k-1$ is equivalent to the following: There exist indices $0\leq\nu_1<\nu_2<\cdots<\nu_{n-k-1}\leq n-1$ and coefficients $c_j\in\mathbb{F}_q$ $(1\leq j\leq n-k-1)$ for which we have $f(x)-\sum_{j=1}^{n-k-1}c_jx^{\nu_j}\in\mathcal{C}'=\ker\psi_{n,k}$, or equivalently,

$$f(\alpha^i)=\sum_{j=1}^{n-k-1}c_j{\beta_{\nu_j}}^i \text{ for every } 0\leq i\leq n-k-1 \ , \tag{51}$$

where we put $\beta_{\nu_j} := \alpha^{\nu_j}$. The condition (51) can be expressed as

$$\begin{pmatrix} f(\alpha^0) \\ f(\alpha^1) \\ \vdots \\ f(\alpha^{n-k-1}) \end{pmatrix} = \begin{pmatrix} {\beta_{\nu_1}}^0 & {\beta_{\nu_2}}^0 & \cdots & {\beta_{\nu_{N-K-1}}}^0 \\ {\beta_{\nu_1}}^1 & {\beta_{\nu_2}}^1 & \cdots & {\beta_{\nu_{N-K-1}}}^1 \\ \vdots & \vdots & \vdots & \vdots \\ {\beta_{\nu_1}}^{n-k-1} & {\beta_{\nu_2}}^{n-k-1} & \cdots & {\beta_{\nu_{n-k-1}}}^{n-k-1} \end{pmatrix} \vec{c} \ , \tag{52}$$

where $\vec{c}$ denotes the column vector ${}^t(c_1, c_2, \ldots, c_{n-k-1})$. For simplicity, let B and $\vec{b}$ denote, respectively, the first $n-k-1$ rows and the last row of the above matrix; i.e., the above condition is written as

$${}^t(f(\alpha^0), f(\alpha^1), \ldots, f(\alpha^{n-k-1})) = \begin{pmatrix} B \\ \vec{b} \end{pmatrix} \vec{c} \ . \tag{53}$$

Now, as α is a primitive element of $\mathbb{F}_q$, all β_{ν_j} are distinct with each other and hence B is a Vandermonde matrix which is invertible. Therefore, the condition (51) implies that $\vec{c} = B^{-1} \cdot {}^t(f(\alpha^0), f(\alpha^1), \ldots, f(\alpha^{n-k-2}))$ and $f(\alpha^{n-k-1}) = \vec{b}\vec{c}$. On the other hand, the latter condition is not satisfied when $f(\alpha^i) = 0$ for every $0 \leq i \leq n-k-2$ and $f(\alpha^{n-k-1}) \neq 0$, e.g., $f(x) = \prod_{i=0}^{n-k-2}(x - \alpha^i)$. Hence this element $f(x) \in \mathcal{C}$ satisfies that $d(f(x), \mathcal{C}') \geq n-k$, as desired. This concludes the proof of Proposition 7. □

6 Concluding remarks

In this paper, we first specified a class of mathematical problems, which we call Function Density Problems. Then we pointed out novel connections of Function Density Problems to theoretical security evaluations of keyless hash functions and to constructions of provably secure pseudorandom generators with some enhanced security property. Our argument aimed at proposing new theoretical frameworks for these topics (especially for the former) based on Function Density Problems, rather than providing some concrete and practical results on the topics. We also gave some examples of mathematical discussions on the problems, which would be of independent interest from mathematical viewpoints.

To conclude this paper, we discuss some possible directions of future works. First, there exist some cryptographic protocols for which the constructions are motivated by some NP-complete/NP-hard problems, but actually the distributions of the problem instances in the protocols are somewhat biased, therefore it has not succeeded to prove the security of the protocols directly from the hardness of the underlying problems (e.g., McEliece cryptosystem and other code-based protocols relevant to decoding problem for random linear codes; knapsack cryptosystem relevant to Subset Sum Problem; etc.). We hope that the idea of Function Density Problems can be applied to measure the closeness of the approximations of the underlying hard problems in those protocols. Secondly, for the mathematical characteristics of Function Density Problems, it would be interesting to evaluate the computational difficulty of Function Density Problems (e.g., to prove, if possible, that Function Density Problems are NP-hard). Moreover, as the examples of Function Density Problems in this paper are for the case that the subset $\mathcal{C}'$ of $\mathcal{C}$ forms a linear subspace, it would be also significant to study the other cases that $\mathcal{C}'$ is not a linear subspace of $\mathcal{C}$.

Acknowledgments. A preliminary version of this paper was presented at The 6th International Workshop on Security (IWSEC 2011), November 8–10, 2011 [6]. The authors would like to thank the anonymous referees of IWSEC 2011 for their kind reviews and comments. The authors would also like to thank Dr. Goichiro Hanaoka for his precious comment on potential applications of the subject of this paper discussed in Section 6.

References

[1] Blum, L., Blum, M., Shub, M.: A simple unpredictable pseudo-random number generator. SIAM J. Comput. 15, 364–383 (1986)

[2] Carter, J.L., Wegman, M.N.: Universal classes of hash functions (extended abstract). In: Proc. STOC 1977, pp. 106–112 (1977)

[3] Dubrov, B., Ishai, Y.: On the randomness complexity of efficient sampling. In: Proc. STOC 2006, pp. 711–720 (2006)

[4] Farashahi, R.R., Schoenmakers, B., Sidorenko, A.: Efficient pseudorandom generators based on the DDH assumption. In: Proc. PKC 2007, pp. 426–441 (2007)

[5] Matsumoto, T., Imai, H.: Public quadratic polynomial-tuples for efficient signature-verification and message-encryption. In: Proc. EUROCRYPT 1988, pp. 419–453 (1988)

[6] Nuida, K., Abe, T., Kaji, S., Maeno, T., Numata, Y.: A mathematical problem for security analysis of hash functions and pseudorandom generators. In: Proc. IWSEC 2011, pp. 144–160 (2011)

[7] Nuida, K., Hanaoka, G.: On the security of pseudorandomized information-theoretically secure schemes. In: Proc. ICITS 2009, pp. 56–73 (2009)

[8] Rivest, R., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM 21(2), 120–126 (1978)

[9] Rogaway, P.: Formalizing human ignorance – collision-resistant hashing without the keys. In: Proc. VIETCRYPT 2006, pp. 211–228 (2006)

[10] Wang, X., Yin, Y.L., Yu, H.: Finding collisions in the full SHA-1: In: Proc. CRYPTO 2005, pp. 17–36 (2005)

Koji Nuida
Research Institute for Secure Systems (RISEC), National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan
k.nuida@aist.go.jp

Takuro Abe
Department of Mechanical Engineering and Science, Kyoto University, Kyoto, Japan
abe.takuro.4c@kyoto-u.ac.jp

Shizuo Kaji
Department of Mathematical Sciences, Faculty of Science, Yamaguchi University, Yamaguchi, Japan
skaji@yamaguchi-u.ac.jp

Toshiaki Maeno
Department of Mathematics, Meijo University, Aichi, Japan
tmaeno@meijo-u.ac.jp

Yasuhide Numata
Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan / Japan Science and Technology Agency (JST), CREST
numata@stat.t.u-tokyo.ac.jp