Bounds for the average degree-k monomial density of Boolean functions

Ana Sălăgean
Department of Computer Science
Loughborough University
Loughborough, UK
A.M.Salagean@lboro.ac.uk

Percy Reyes-Paredes
Department of Computer Science
Loughborough University
Loughborough, UK
A.P.Reyes-Paredes@lboro.ac.uk

Abstract

For a Boolean function f represented in algebraic normal form (i.e. as a multivariate polynomial function over $\mathbb{F}_2$) we consider the density of monomials of degree k in f, for each degree k, i.e. the number of monomials of degree k that appear in f, normalized by the total number of possible monomials of degree k. We then average this number over all functions which are affine equivalent to f; we call the resulting quantity, denoted by $\mathrm{add}_k(f)$, the average degree-k monomial density of f. We defined this quantity in our previous work, and showed it is closely related to a probabilistic test we introduced for deciding whether $\deg(f) < k$.

In this paper we give lower and upper bounds for $\mathrm{add}_k(f)$ for polynomials of any degree d (only the particular case $d = k$ having been dealt with in our previous work). There are several consequences of these bounds. Firstly, the $\deg(f) < k$ probabilistic test is guaranteed to have high accuracy when the actual degree of f is not much higher than k. Secondly, it answers negatively the question: does there exist a function f which has no monomials of a particular degree k (with $k < \deg(f)$) and, moreover, it still has no monomials of degree k after applying any affine invertible change of coordinates to f. Thirdly, while the average of $\mathrm{add}_k(f)$ over all n-variable functions f of a fixed degree $d > k$ is equal to 0.5, the distribution of the values is somewhat surprising; when $k \leq n - 10$ and $n \geq 20$, low values of $\mathrm{add}_k(f)$ exist (reaching approximately $\frac{1}{2^{d-k}}$), but there are no values higher than around 0.5005.

Keywords: Algebraic degree, Moebius transform, probabilistic testing, algebraic thickness

1 Introduction and motivation

A Boolean function f in n variables can be uniquely represented in ANF (algebraic normal form), i.e. as a polynomial in n variables over $\mathbb{F}_2$ (the finite field with 2 elements) of degree at most one in each variable. The degree of this polynomial is called the algebraic degree of f.

The algebraic degree is one of the parameters that measures the nonlinearity of Boolean functions used in cryptography. These Boolean functions must have high algebraic degree, otherwise some attacks can be effective; for example, the higher order derivative attacks, algebraic attacks, cube attacks, and integral attacks. However, the Boolean functions used in cryptography do not reach the highest degree because trade-offs with other parameters need to be considered.

We consider, for each degree k, the density of monomials of degree k in f, i.e. the number of monomials of degree k that appear in f, normalized by the total number of possible monomials of degree k. We then average this number over all functions which are affine equivalent to f; the resulting quantity, denoted by $\mathrm{add}_k(f)$, will be called the average degree-k monomial density of f. While the study of this parameter is interesting in itself, our original motivation comes from its connection to a probabilistic test that we introduced in previous work. Namely, when a cryptographic Boolean function f on $\mathbb{F}_2^n$ with a large number of variables is not given explicitly in ANF (e.g. it is given as a composition of functions, or even as a black box), it may not be feasible to compute its algebraic degree. The existence of a particular monomial $x_{i_1} \cdots x_{i_k}$ of degree k in the ANF of f can be decided by summing the values of f over a vector space generated by the k vectors of the canonical basis $e_{i_1}, \ldots, e_{i_k}$ (this method is also known as the Moebius transform). In [5, 6] we proposed the "$\deg(f) < k$" probabilistic test which generalizes this idea. One sums the values of f over a linear combination of k vectors, and if the result is zero, we say that f passes this instance of the test, otherwise it fails (we recall the full details in Section 2); when the k vectors are linearly independent, this is equivalent to testing the existence of a particular monomial of degree k after applying a random affine invertible change of variables to f. The number of monomials of degree k is likely to be high after the change of variables and therefore it would be easier to probabilistically detect their existence. The probability of failing the test is denoted by $\mathrm{dt}_k(f)$. In [6], we proved lower and upper bounds for $\mathrm{dt}_k(f)$ for the case when the actual degree of f (which is not known apriori) turns out to be k.

The main result of the present paper is Theorem 5 and its Corollary 6. They give lower and upper bounds on $\mathrm{dt}_k(f)$ and $\mathrm{add}_k(f)$ for a function f of any degree, generalising thus the existing result from [6] which only covers the case when f has degree k. These bounds have several consequences. Firstly, when the actual degree of f is not much higher than k, the $\deg(f) < k$ probabilistic test is guaranteed to have high accuracy (in the sense that it has a high probability of reaching the correct conclusion after a small number of tests). Secondly, it answers negatively the question: does there exist a function f which has no monomials of a particular degree k (with $k < \deg(f)$) and, moreover, it still has no monomials of degree k after applying any affine invertible change of coordinates to f. Thirdly, while the average of $\mathrm{add}_k(f)$ over all n-variable functions f of a fixed degree $d > k$ is equal to 0.5, the distribution of the values is somewhat surprising; when $k \leq n - 10$ and $n \geq 20$, low values of $\mathrm{add}_k(f)$ exist (reaching approximately $\frac{1}{2^{d-k}}$), but there are no values higher than around 0.5005.

2 Definitions and existing results

We denote by $\mathbb{F}_2$ the finite field with two elements, represented as $\{0, 1\}$, and we denote by $\oplus$ addition in $\mathbb{F}_2$ as well as in the vector space $\mathbb{F}_2^n$. Any function $f : \mathbb{F}_2^n \to \mathbb{F}_2$ can be represented in its algebraic normal form (ANF), i.e. as a polynomial function given by a polynomial of degree at most 1 in each variable:

$$f(x_1, \ldots, x_n) = \bigoplus_{a_1,\ldots,a_n \in \mathbb{F}_2} c_{a_1,\ldots,a_n} x_1^{a_1} \cdots x_n^{a_n},$$

with $c_{a_1,\ldots,a_n} \in \mathbb{F}_2$. The degree of this polynomial is called the algebraic degree of f, and here we will call it simply the degree of f and denote it by $\deg(f)$. The coefficients of the ANF of f can be computed by the following formula (see, for example, [4, Chapter 13, Theorem 1]) which is sometimes called the Moebius transform:

$$c_{a_1,\ldots,a_n} = \bigoplus_{x_1 \le a_1,\ldots,x_n \le a_n} f(x_1, \ldots, x_n). \tag{1}$$

Two n-variable Boolean functions f and g are *affine equivalent*, denoted by $f \sim g$, if $g = f \circ \varphi_{M,v}$ for some invertible affine function $\varphi_{M,v} : \mathbb{F}_2^n \to \mathbb{F}_2^n$, $\varphi_{M,v}(x) = Mx \oplus v$, where M is an $n \times n$ nonsingular matrix over $\mathbb{F}_2$ and $v \in \mathbb{F}_2^n$ is a vector. If f and g are affine equivalent, then $\deg(f) = \deg(g)$. We therefore say that the algebraic degree is an *affine invariant*.

Definition 1. [6] Let $0 \le k \le n$ be integers and let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a function. The degree-k monomial density of f, denoted by $\mathrm{dd}_k(f)$, is defined as the number of monomials of degree k in the ANF of f, divided by $\binom{n}{k}$ (the total number of monomials of degree k in n variables). In other words, if the ANF of f is $f(x) = \bigoplus_m c_m m$, with m ranging over all monomials in n variables and $c_m \in \mathbb{F}_2$, then

$$\mathrm{dd}_k(f) = \frac{\mid \{m : m \text{ monomial of degree } k \text{ and } c_m \neq 0\} \mid}{\binom{n}{k}}. \tag{2}$$

The average degree-k monomial density of f, denoted by $\mathrm{add}_k(f)$, is the average (arithmetic mean) of $\mathrm{dd}_k(g)$ over all the functions g such that $f \sim g$, i.e.

$$\mathrm{add}_k(f) = \frac{\sum_{g \sim f} \mathrm{dd}_k(g)}{\mid \{g : g \sim f\} \mid} = \frac{\sum_{M \in GL(n,\mathbb{F}_2), v \in \mathbb{F}_2^n} \mathrm{dd}_k(f \circ \varphi_{M,v})}{2^n(2^n-1)(2^n-2)\cdots(2^n-2^{n-1})}. \tag{3}$$

It was shown in [6, Remark 5] that the two ways of defining add_k in equation (3) are indeed equal.

We recall the proposed $\deg(f) < k$ probabilistictest [5, 6]: pick $u_0, u_1, \ldots, u_k \in \mathbb{F}_2^n$. If the equation

$$\bigoplus_{b_1,\ldots,b_k \in \mathbb{F}_2} f\left(\left(\bigoplus_{i=1}^{k} b_i u_i\right) \oplus u_0\right) = 0 \tag{4}$$

holds, we say that f passes this instance of the test, otherwise it fails. We denoted by $\mathrm{dt}_k(f)$ the probability of f failing the $\deg(f) < k$ test, taken over all all possible choices $u_0, u_1, u_2, \ldots, u_k \in \mathbb{F}_2^n$, i.e.

$$\mathrm{dt}_k(f) = \frac{|\{(u_0, u_1, u_2, \ldots, u_k) \in (\mathbb{F}_2^n)^{k+1} : \bigoplus_{b_1,\ldots,b_k \in \mathbb{F}_2} f\left(\left(\bigoplus_{i=1}^{k} b_i u_i\right) \oplus u_0\right) \neq 0\}|}{2^{(k+1)n}}. \tag{5}$$

If the degree of f is indeed less than k, then f always passes the $\deg(f) < k$ test, i.e. $\mathrm{dt}_k(f) = 0$. We are therefore interested in the values of $\mathrm{dt}_k(f)$ for the case when f has degree at least k. A value of $\mathrm{dt}_k(f)$ which is not very low would mean that after running the test a reasonably small number of times, we have a good chance of having at least one fail (namely, a probability of $1 - (1 - \mathrm{dt}_k(f))^t$ of at least one fail after t tests), and therefore decide, correctly, that $\deg(f) \geq k$.

One can easily check that $\mathrm{dt}_k(f)$ and $\mathrm{add}_k(f)$ are affine invariants. Moreover, it is easy to verify that they do not depend on the terms of f of degree strictly less than k, i.e. if $g = f \oplus h$ with $\deg(h) < k$ then $\mathrm{dt}_k(f) = \mathrm{dt}_k(g)$. Therefore we are working with the equivalence relation $\sim_{k-1}$ induced by affine equivalence on the quotient $RM(n,n)/RM(k-1,n)$, where $RM(k,n)$ denotes the set of polynomials of degree at most k in n variables (also known as the k-th order Reed-Muller code of length 2^n, see [4]). Namely, $f \sim_{k-1} g$ if there is a function h such that $f \sim h$ and $\deg(g - h) \leq k - 1$ (i.e. g and h coincide if we ignore all monomials of degree less than k).

It is noted in [6, Remark 3] that if the vectors $u_1, u_2, \ldots, u_k$ are linearly dependent, then any function f passes that particular instance of the $\deg(f) < k$ test. Therefore, in practice there is no need to run the test when they are linearly dependent. However, there are advantages in defining the probability for arbitrary vectors (one reason being the similar definition for the BLR linearity test; another reason being that Proposition 3 would not hold otherwise; see [6, Remark 3] for further discussion); the probability of failing the $\deg(f) < k$ test, taken over linearly independent vectors, equals $\mathrm{add}_k(f)$ and can be obtained by dividing $\mathrm{dt}_k(f)$ by the probability of k arbitrary vectors being linearly independent, see [6, Theorem 8]:

$$\mathrm{dt}_k(f) \quad = \quad \mathrm{add}_k(f) \prod_{i=n-k+1}^{n} \left(1 - \frac{1}{2^i}\right). \tag{6}$$

Recall that the discrete derivative of f in a non-zero direction $u \in \mathbb{F}_2^n$ is defined as $\mathrm{D}_u f(x) = f(x \oplus u) \oplus f(x)$. The derivative of order k in directions $u_1, \ldots, u_k$ is defined as $\mathrm{D}^{(k)}_{u_1,\ldots,u_k} f = \mathrm{D}_{u_1}(\mathrm{D}_{u_2}(\ldots \mathrm{D}_{u_k} f))$. For the $\deg(f) < k$ test, the equation (4) can be rewritten as $\mathrm{D}^{(k)}_{u_1,\ldots,u_k} f(u_0) = 0$.

In [6], we proved that if f has actually degree k, the following bounds on $\mathrm{dt}_k(f)$ hold:

Theorem 2. *[6, Theorem 14] Let f be a function of degree k in n variables. Then*

$$0.288788... < \prod_{i=1}^{k} \left(1 - \frac{1}{2^i}\right) \leq \mathrm{dt}_k(f) \leq \frac{1}{2}\left(1 - \frac{1}{2^n}\right)^{k-1} \leq 0.5, \tag{7}$$

where $0.288788...$ is the q-Pochhammer symbol at $(0.5, 0.5, \infty)$. The lower bound is achieved if and only if $f(x_1, \ldots, x_n)$ is affine equiavalent to $x_1 \ldots x_k \oplus h(x_1, \ldots, x_n)$ for a polynomial h of degree at most $k - 1$.

We also recall the following basic properties:

Proposition 3. *[6, Proposition 10] Let $f, g_1 : \mathbb{F}_2^n \to \mathbb{F}_2$ and $g_2 : \mathbb{F}_2^m \to \mathbb{F}_2$.*
(i) If $g(x_1, \ldots, x_n, x_{n+1}) = f(x_1, \ldots, x_n)$, then $\mathrm{dt}_k(g) = \mathrm{dt}_k(f)$.
(ii) If $g(x_1, \ldots, x_{n+m}) = g_1(x_1, \ldots, x_n) \oplus g_2(x_{n+1}, \ldots, x_{n+m})$, then $\mathrm{dt}_k(g) = \mathrm{dt}_k(g_1) + \mathrm{dt}_k(g_2) - 2\mathrm{dt}_k(g_1)\mathrm{dt}_k(g_2)$.

3 Bounds on $\mathrm{dt}_k(f)$ and $\mathrm{add}_k(f)$

We compute first $\mathrm{dt}_k(f)$ for the case when the ANF of f has just one monomial:

Proposition 4. *Let $f(x_1, \ldots, x_n) = x_1 x_2 \cdots x_d$. For any k with $1 \leq k \leq d$ we have:*

$$\mathrm{dt}_k(f) = \frac{1}{2^{d-k}} \prod_{i=d-k+1}^{d} \left(1 - \frac{1}{2^i}\right).$$

Proof. We can assume that the number of variables n equals d, see Proposition 3(i). Note that $f(x_1, \ldots, x_d) = 1$ if and only if $(x_1, \ldots, x_d) = \mathbf{1}$ where we denote $\mathbf{1} = (1, 1, \ldots, 1)$. Consider the $\deg(f) < k$ test on f at $(u_0, u_1, \ldots, u_k) \in (\mathbb{F}_2^n)^{k+1}$, which checks whether the following equation holds:

$$\bigoplus_{b_1,\ldots,b_k \in \mathbb{F}_2} f(u_0 \oplus \bigoplus_{i=1}^{k} b_i u_i) = 0.$$

This test fails if and only if $u_1, \ldots, u_k$ are linearly independent and $\mathbf{1} \in u_0 \oplus \langle u_1, \ldots, u_k \rangle$. In other words, there are constants $b_i \in \mathbb{F}_2$ such that $\mathbf{1} = u_0 \oplus \bigoplus_{i=1}^{k} b_i u_i$. The number of ways to choose k linearly independent vectors $(u_1, \ldots, u_k) \in (\mathbb{F}_2^n)^k$ is $(2^n - 1)(2^n - 2) \cdots (2^n - 2^{k-1})$. For each of these choices, there are 2^k ways to choose u_0

such that $\mathbf{1} \in u_0 \oplus \langle u_1, \ldots, u_k \rangle$; namely, for each of the 2^k elements $u \in \langle u_1, \ldots, u_k \rangle$, we choose $u_0 = \mathbf{1} \oplus u$.

Therefore, by using the definition of dt_k, we have:

$$\begin{aligned}
\mathrm{dt}_k(f) &= \frac{2^k(2^d-1)(2^d-2)\cdots(2^d-2^{k-1})}{2^{(k+1)d}} \\
&= \frac{1}{2^{d-k}} \prod_{i=d-k+1}^{d} \left(1 - \frac{1}{2^i}\right).
\end{aligned}$$

□

We now give lower and upper bounds for $\mathrm{dt}_k(f)$:

Theorem 5. *Let f be a polynomial of degree d in n variables. For any k with $1 \le k \le d$ we have:*

$$\frac{1}{2^{d-k}} \prod_{i=d-k+1}^{d} \left(1 - \frac{1}{2^i}\right) \le \mathrm{dt}_k(f) \le \frac{1}{2}\left(1 - \frac{1}{2^n}\right)^{k-1}.$$

The lower bound is tight; if f is affine equivalent to $x_1 x_2 \cdots x_d + g(x_1, \ldots, x_n)$ for a polynomial g of degree at most $k-1$, then $\mathrm{dt}_k(f)$ equals the lower bound.

Proof. The proof is by induction on d. For $d = 1$ we have $k = 1$, so Theorem 2 completes the proof of this case.

For the inductive step, we assume the statement is true for any degree less than d and prove it for degree d.

The $\deg(f) < k$ test on f at $(u_0, u_1, \ldots, u_k) \in (\mathbb{F}_2^n)^{k+1}$ checks whether the following equation holds:

$$\bigoplus_{b_1, \ldots, b_k \in \mathbb{F}_2} f(u_0 \oplus \bigoplus_{i=1}^{t} b_k u_i) = 0.$$

When $u_1 = \mathbf{0}$ this equation always holds, regardless of f; when $u_1 \ne \mathbf{0}$ the equation above can be rewritten as

$$\bigoplus_{b_2, \ldots, b_k \in \mathbb{F}_2} \mathrm{D}_{u_1} f(u_0 \oplus \bigoplus_{i=2}^{k} b_i u_i) = 0,$$

which is the $\deg(\mathrm{D}_{u_1} f) < k - 1$ test at $(u_0, u_2, \ldots, u_k)$. We have therefore

$$\mathrm{dt}_k(f) = \frac{1}{2^n} \sum_{u_1 \in \mathbb{F}_2^n \setminus \{\mathbf{0}\}} \mathrm{dt}_{k-1}(\mathrm{D}_{u_1} f). \tag{8}$$

For the lower bound, recall first that for any $u \in \mathbb{F}_2^n \setminus \{\mathbf{0}\}$ we have $\deg(\mathrm{D}_u f) \le \deg(f) - 1$ (see [2]); u is called a *fast point* for f if $\deg(\mathrm{D}_u f) < \deg(f) - 1$ ([1]). In [1, Theorem 3.2], it was shown that, for a function f of degree d in n variables, the vector $\mathbf{0}$ together with the fast points of f forms a vector space of dimension at most $n-d$.

By denoting by S the set of non-zero vectors in $\mathbb{F}_2^n$ which are not fast points for f, we have therefore $|S| \geq 2^n - 2^{n-d}$. Since $\deg(\mathrm{D}_{u_1} f) = d-1$ for all $u_1 \in S$, we can apply the induction hypothesis to $\mathrm{D}_{u_1} f$, obtaining $\mathrm{dt}_{k-1}(\mathrm{D}_{u_1} f)) \geq \frac{1}{2^{d-k}} \prod_{i=d-k+1}^{d-1} \left(1 - \frac{1}{2^i}\right)$. By using these results in (8), we have

$$\begin{aligned}
\mathrm{dt}_k(f) &= \frac{1}{2^n} \sum_{u_1 \in \mathbb{F}_2^n \setminus \{\mathbf{0}\}} \mathrm{dt}_{k-1}(\mathrm{D}_{u_1} f) \\
&\geq \frac{1}{2^n} \sum_{u_1 \in S} \mathrm{dt}_{k-1}(\mathrm{D}_{u_1} f) \\
&\geq \frac{|S|}{2^n} \frac{1}{2^{d-k}} \prod_{i=d-k+1}^{d-1} \left(1 - \frac{1}{2^i}\right) \\
&\geq \frac{2^n - 2^{n-d}}{2^n} \frac{1}{2^{d-k}} \prod_{i=d-k+1}^{d-1} \left(1 - \frac{1}{2^i}\right) \\
&= \frac{1}{2^{d-k}} \prod_{i=d-k+1}^{d} \left(1 - \frac{1}{2^i}\right)
\end{aligned}$$

as required. The fact that the lower bound is achieved with equality when $f \sim_{k-1} x_1 \cdots x_d$ is immediate from Proposition 4.

For the upper bound, since $\mathrm{D}_{u_1} f$ has degree strictly less than d, we know by the induction hypothesis that if it has degree at least $k-1$, then $\mathrm{dt}_{k-1}(\mathrm{D}_{u_1} f) \leq \frac{1}{2}\left(1 - \frac{1}{2^n}\right)^{k-2}$. If it has degree less than $k-1$, then $\mathrm{dt}_{k-1}(\mathrm{D}_{u_1} f) = 0 < \frac{1}{2}\left(1 - \frac{1}{2^n}\right)^{k-2}$. Hence, we have $\mathrm{dt}_{k-1}(\mathrm{D}_{u_1} f) \leq \frac{1}{2}\left(1 - \frac{1}{2^n}\right)^{k-2}$ for each of the $2^n - 1$ values of $u_1 \in \mathbb{F}_2^n \setminus \{\mathbf{0}\}$. Therefore, by using (8), we obtain $\mathrm{dt}_k(f) \leq \frac{1}{2}\left(\frac{2^n-1}{2^n}\right)\left(1 - \frac{1}{2^n}\right)^{k-2} = \frac{1}{2}\left(1 - \frac{1}{2^n}\right)^{k-1}$ so the upper bound holds. □

Theorem 5 and (6) also yield bounds for $\mathrm{add}_k(f)$:

Corollary 6. *Let f be a polynomial of degree d in n variables. For any k with $1 \leq k \leq d$ we have:*

$$\frac{1}{2^{d-k}} \left(\frac{\prod_{i=d-k+1}^{d} \left(1 - \frac{1}{2^i}\right)}{\prod_{i=n-k+1}^{n} \left(1 - \frac{1}{2^i}\right)} \right) \leq \mathrm{add}_k(f) \leq \frac{1}{2} \left(\frac{\left(1 - \frac{1}{2^n}\right)^{k-1}}{\prod_{i=n-k+1}^{n} \left(1 - \frac{1}{2^i}\right)} \right). \qquad (9)$$

The lower bound is tight; if f is affine equivalent to $x_1 x_2 \cdots x_d + g(x_1, \ldots, x_n)$ for some polynomial g of degree at most $k-1$, then $\mathrm{add}_k(f)$ *equals the lower bound.*

We will now examine a number of consequences of Theorem 5 and Corollary 6. Firstly, note that the lower bounds in both cases are non-zero. Therefore:

Corollary 7. *Let f be a Boolean function and let $k < \deg(f)$. There is at least one function g which is affine equivalent to f and has at least one monomial of degree k which appears with non-zero coefficient in the ANF of g.*

In other words, Corollary 7 says that even if a function f has no monomials of a certain degree $k < \deg(f)$, it is not possible that all the functions in its affine equivalence class also have this property.

Next, we estimate the numerical values of the bounds. We are particularly interested in functions in at least 20 variables, as for functions in fewer variables the ANF can be explicitly computed even if the function is given as a black box. For k, we are interested in values of at most 40, as the values of f at 2^k points need to be summed for one $\deg(g) < k$ test, which becomes unfeasible for $k > 40$.

When $n \geq 20$ and $2 \leq k \leq 40$ we have

$$0.49998 < \frac{1}{2}\left(1 - \frac{1}{2^n}\right)^{k-1} \leq 0.5.$$

so we will approximate the upper bound on $\mathrm{dt}_k(f)$ in Theorem 5 by 0.5.

For the other bounds, estimates for the following quantity will be particularly useful:

$$P(a, b) = \prod_{i=a}^{b} \left(1 - \frac{1}{2^i}\right)$$

for integers $1 \leq a \leq b$. (Recall that the q-Pochhammer symbol is defined as $(c; q)_n = \prod_{i=0}^{n-1}(1 - cq^i)$ with n a positive integer or ∞, so $P(a, b) = (\frac{1}{2^a}; \frac{1}{2})_{b-a})$. It is obvious that for a fixed a, $P(a, b)$ decreases as b increases. When $a = 1$ as b tends to infinity, this quantity converges to a limit (namely $(\frac{1}{2^a}; \frac{1}{2})_\infty$) which is in the interval $(0.288788, 0.288789)$; it converges fast, for example for $b = 20$, we are already within this interval and therefore closer than 10^{-6} to the limit. Similarly, the values of $P(a, b)$ for small values of $a = 1, \ldots, 10$ and $b \geq 20$ are, to 6 decimal places: 0.288788, 0.577576, 0.770102, 0.880116, 0.938791, 0.969074, 0.984456, 0.992208, 0.996099, 0.998048. For larger values of a, the quantity $P(a, b)$ is close to 1. For example, when $a \geq 11$ we have $P(a, b) \in (0.999, 1)$, and when $a \geq 21$ we have $P(a, b) \in (0.999999, 1)$, for any $b \geq a$.

Using the estimates above for $P(a, b)$, the values of the lower bound on $\mathrm{dt}_k(f)$ from Theorem 5 are tabulated (to 6 decimal places) in Table 1 for $d - k = 0, 1, \ldots, 8$, with the assumption $d \geq 20$. We also computed, based on this lower bound, the number of tests t that we would need to run in order to guarantee a probability of at least 0.95 that a correct decision is reached (i.e. $1 - (1 - \mathrm{dt}_k(f))^t \geq 0.95$). These numerical values show that by running the $\deg(f) < k$ test 769 times we can say with 0.95 confidence that if the actual degree of f (which is unknown) is at most $k + 8$ then the correct conclusion will be reached.

In Table 1, we also computed the lower and upper bounds for $\mathrm{add}_k(f)$, again under the assumption $n \geq d \geq 20$. These bounds were obtained by dividing the lower and upper bounds of $\mathrm{dt}_k(f)$ (with the upper bound approximated as 0.5) by

$P(n-k,n) = P(n-d-(d-k))$. We tabulated the values for $d-k = 0,\ldots,8$, considering two cases: $n-d=2$ and $n-d=10$ (any value of $n-d$ higher than 10 giving results very close to the ones for $n-d=10$).

Let us fix k and d with $k < d$. When f ranges over all n-variable functions of degree d, the number of monomails of degree k in the ANF of f has a binomial distribution with parameters $\binom{n}{k}$ and 0.5; each value $i = 0, 1, \ldots, \binom{n}{k}$ appears with probability $\binom{\binom{n}{k}}{i}\frac{1}{2^{\binom{n}{k}}}$. The number of degree-$k$ monomials in f, averaged over all functions f of degree d, equals $\frac{1}{2}\binom{n}{k}$ and the standard deviation equals $\frac{1}{2}\sqrt{\binom{n}{k}}$. Since we are interested in the case $n \geq 20$, as long as $k > 0$, this binomial distribution can be approximated by the normal distribution, so, for example, 95% of the values of the degree-k density, $\mathrm{dd}_k(f)$ (which is the number of degree-k monomials divided by $\binom{n}{k}$, as defined in Definition 1), will fall within the interval $[0.5 - \frac{1}{\sqrt{\binom{n}{k}}}, 0.5 + \frac{1}{\sqrt{\binom{n}{k}}}]$, i.e. will be very close to 0.5.

If we average $\mathrm{dd}_k(f)$ within each affine equivalence class (note the classes are not all of the same size), i.e we compute $\mathrm{add}_k(f)$, we would intuitively expect that most of the $\mathrm{add}_k(f)$ values would be close to 0.5, with possibly a small number of them being much lower or much higher.

However, what is surprising is that, for most values of $k < d < n$, Corollary 6 can be used to show that there exist at least one class where $\mathrm{add}_k(f)$ is much lower than 0.5, but there are no classes with $\mathrm{add}_k(f)$ much higher than 0.5. More precisely, from Table 1 we can see that for functions f in at least $n = 30$ variables, if the degree d of f is at most $n-10$, then $\mathrm{add}_k(f)$ can be quite low (keeping in mind that the lower bound is tight and, for example, for $d-k \leq 6$ the lower bound is just under $\frac{1}{2^{d-k}}$), but, surprisingly, $\mathrm{add}_k(f)$ cannot be any higher than about 0.5005. Namely, by using the previous estimates for $P(a,b)$, we have that $\mathrm{add}_k(f) \leq 0.5/P(n-k,n) < 0.5005$ when $n-k \geq 10$ and $\mathrm{add}_k(f) < 0.5000005$ when $n-k \geq 20$. It is only for functions f where n, d and k are very close to each other that $\mathrm{add}_k(f)$ can be significantly higher than 0.5 (with the extreme case of $n = d = k$, where the density of monomials of degree n for a function of degree n in n variables is obviously equal to 1 as there is only one monomial of degree n and the degree is an affine invariant).

We computed the exact values of $\mathrm{dt}_k(f)$, $k = 3, 4$, for all the 68431 classes of functions of degree 4 in 7 variables, under the equivalence $\sim_2$. A representative for each class was determined by Langevin in [3]. The values of $dt_4(f)$ range between 0.307617 and 0.451813. The lower and upper bounds given by Theorem 5 would be 0.307617 and 0.488373 respectively; so, while the lower bound is achieved, the upper bound is not tight in this case. Likewise, the values of $dt_3(f)$ range between 0.307617 and 0.481934. The lower and upper bounds given by Theorem 5 would be 0.307617 and 0.492218 respectively; so, while the lower bound is achieved, the upper bound is not tight in this case.

Using (6), for polynomials of degree 4 in 7 variables, $\mathrm{add}_4(f) \in [0.346795, 0.509356]$ and $\mathrm{add}_3(f) \in [0.325120, 0.516162]$.

We also considered functions that describe the whole cipher Trivium, Grain-128a,

$d-k$	$\mathrm{dt}_k(f)$	t	$\mathrm{add}_k(f)$	$\mathrm{add}_k(f)$	$\mathrm{add}_k(f)$	$\mathrm{add}_k(f)$
	lower bound		lower bound	upper bound	lower bound	upper bound
			$n-d=2$	$n-d=2$	$n-d=10$	$n-d=10$
0	0.288788	9	0.375000	0.649265	0.289070	0.500489
1	0.288788	9	0.328125	0.568107	0.288929	0.500244
2	0.192525	15	0.205078	0.532600	0.192572	0.500122
3	0.110015	26	0.113525	0.515957	0.110028	0.500061
4	0.058674	50	0.059601	0.507895	0.058678	0.500031
5	0.030284	98	0.030521	0.503927	0.030284	0.500015
6	0.015382	194	0.015442	0.501958	0.015382	0.500008
7	0.007752	385	0.007767	0.500978	0.007752	0.500004
8	0.003891	769	0.003895	0.500489	0.003891	0.500002

Table 1: Bounds for $\mathrm{dt}_k(f)$ and $\mathrm{add}_k(f)$ for $n \geq d \geq 20$, and number of tests t for a 0.95 probability of reaching a correct decision

and SNOW-V. Namely, for each cipher, the inputs are the key and initialisation vector and the output is the first bit of the key stream. In each case, we ran the $\deg(f) < k$ test, for $k = 1, 2, \ldots, 10$ at least 20 times. Not surprisingly, the test confirmed that the functions have degree at least 10. The experimental probability of failing the test was within the interval (0.47-0.52) in each case.

4 Conclusions

We studied $\mathrm{add}_k(f)$, a parameter describing the density of monomials of degree k in the Algebraic Normal Form of a Boolean function f, averaged over all functions which are affine equivalent to f. We obtained lower and upper bounds for $\mathrm{add}_k(f)$ for polynomials of any degree d (only the particular case $d = k$ having been dealt with in our previous work). A first consequence is that the $\deg(f) < k$ probabilistic test, introduced by us in previous work, is guaranteed to have high accuracy when the actual degree of f is not much higher than k. We also answered negatively the following natural question: does there exist a function f which has no monomials of a particular degree k (with $k < \deg(f)$) and, moreover, it still has no monomials of degree k after applying any affine invertible change of coordinates to f. Finally, we evaluated the bounds numerically in several typical situations of interest. For example, for functions in at least $n \geq 20$ variables, when $k \leq n - 10$ and $k < \deg(f)$ there are functions with a quite low value for $\mathrm{add}_k(f)$ (approximately $\frac{1}{2^{d-k}}$), but, somewhat surprisingly (seen that $\mathrm{add}_k(f)$ has mean 0.5) there are no functions where $\mathrm{add}_k(f)$ is higher than around 0.5005.

References

[1] Ming Duan, Mohan Yang, Xiaorui Sun, Bo Zhu, and Xuejia Lai. Distinguishing properties and applications of higher order derivatives of Boolean functions. *Information Sciences*, 271:224–235, 2014.

[2] Xuejia Lai. Higher order derivatives and differential cryptanalysis. In Richard E. Blahut, Daniel J. Costello, Jr., Ueli Maurer, and Thomas Mittelholzer, editors, *Communications and Cryptography*, volume 276 of *The Springer International Series in Engineering and Computer Science*, pages 227–233. Springer, 1994.

[3] Philippe Langevin. Classification of RM(4,7)/RM(2,7), January 2012. https://langevin.univ-tln.fr/project/rm742/rm742.html.

[4] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 1978.

[5] Ana Sălăgean and Percy Reyes-Paredes. Probabilistic estimation of the degree of Boolean function. The Selmer Center in Secure Communication, The 7th International Workshop on Boolean Functions and their Applications (BFA), September 2022.

[6] Ana Sălăgean and Percy Reyes-Paredes. Probabilistic estimation of the algebraic degree of Boolean functions. *Cryptography and Communications*, 15(6):1199–1215, 2023.