An $\mathcal{O}(n)$ Algorithm for Coefficient Grouping

Fukang Liu, Libo Wang

University of Hyogo, Hyogo, Japan
liufukangs@gmail.com

Abstract. In this note, we study a specific optimization problem arising in the recently proposed coefficient grouping technique, which is used for the algebraic degree evaluation. Specifically, we show that there exists an efficient algorithm running in time $\mathcal{O}(n)$ to solve this basic optimization problem relevant to upper bound the algebraic degree. Moreover, the main technique in this efficient algorithm can also be used to further improve the performance of the off-the-shelf solvers to solve other optimization problems in the coefficient grouping technique. We expect that some results in this note can inspire more studies on the coefficient grouping technique.

Keywords: coefficient grouping · set equivalence · optimization problem

1 Notation

The following notations will be used throughout this paper.

1. $|\mathcal{S}|$ denotes the size of the set $\mathcal{S}$.
2. $a\%b$ represents a mod b.
3. $a|b$ denotes that a divides b.
4. $[a, b]$ is a set of integers i satisfying $a \leq i \leq b$.
5. $H(a)$ is the hamming weight of $a \in [0, 2^n - 1]$.
6. The function $\mathcal{M}_n(x)$ $(x \geq 0)$ is defined as follows:

$$\mathcal{M}_n(x) = \begin{cases} 2^n - 1 \text{ if } 2^n - 1 | x, x \geq 2^n - 1, \\ x\%(2^n - 1) \text{ otherwise.} \end{cases}$$

By the definition of $\mathcal{M}_n(x)$, we have $\mathcal{M}_n(x_1 + x_2) = \mathcal{M}_n(\mathcal{M}_n(x_1) + \mathcal{M}_n(x_2))$, $\mathcal{M}_n(2^i) = 2^{i\%n}$ and $\mathcal{M}_n(2^i x) = \mathcal{M}_n(2^{i\%n}\mathcal{M}_n(x))$ for $i \geq 0$.

2 Motivation

We have recently developed a technique called coefficient grouping to upper bound the algebraic degree for ciphers defined over $\mathbb{F}_{2^n}$. The main idea of that technique is to convert the degree evaluation into some optimization problems. Among them, one basic optimization problem can be described as follows:

$$\text{maximize} \qquad H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_i)),$$

$$\text{subject to} \qquad \gamma_i \in \mathbb{N}, 0 \leq \gamma_i \leq N_i \text{ for } i \in [0, n-1],$$

where $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$.

A more general problem related to upper bounding the algebraic degree in the multivariate case with m variables is

$$\begin{aligned} \text{maximize} \qquad & H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{m,i})), \\ \text{subject to} \qquad & \gamma_{j,i} \in \mathbb{N},\ 0 \leq \gamma_{1,i} + \gamma_{2,i} + \cdots + \gamma_{m,i} \leq N_i \text{ for } i \in [0, n-1]. \end{aligned}$$

where $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$.

In [1], the above problems are first encoded as an MILP problem and then solved with an off-the-shelf solver Gurobi. Using a general-purpose blackbox solver is indeed very convenient but we may lose some insight into these special problem.

Regarding why we do not put this note in [1], we just cannot find a good place. First, we feel it not suitable to place this note at the Appendix of [1] as few people may read it and then its importance will be neglected. Placing it at the main content of [1] also looks inappropriate because it may destroy the simplicity and structure of [1]. The most important reason is that we can only find an efficient algorithm for the above optimization problems, while there are other different optimization problems related to computing more accurate upper bounds for the algebraic degree in [1], for which we cannot find an efficient ad-hoc algorithm.

One purpose of this note is thus to share our ideas of some special optimization problems and we expect that they can inspire more studies. The technique in this note is of independent interest.

3 The Studied Optimization Problems

Let us formally state the studied problems in this paper. We study 2 problems called Problem-U and Problem-M. Specifically, given a vector of integers $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$, Problem-U is defined as follows:

$$\begin{aligned} \text{maximize} \qquad & H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_i)), \\ \text{subject to} \qquad & 0 \leq \gamma_i \leq N_i \text{ for } i \in [0, n-1]. \end{aligned}$$

Problem-M with dimension m is defined as follows:

$$\begin{aligned} \text{maximize} \qquad & H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{m,i})), \\ \text{subject to} \qquad & 0 \leq \gamma_{1,i} + \gamma_{2,i} + \cdots + \gamma_{m,i} \leq N_i \text{ for } i \in [0, n-1]. \end{aligned}$$

Obviously, Problem-U is a special case of Problem-M with $m = 1$. For convenience, in the following, we will omit the constraints $\gamma_{j,i} \in \mathbb{N}$ and $\gamma_i \in \mathbb{N}$.

4 On Set Equivalence

Problem-U can be equivalently stated as finding an element e with the maximal hamming weight from the following set

$$\mathcal{S}_U = \{e | e = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_i), 0 \leq \gamma_i \leq N_i \text{ for } i \in [0, n-1]\}.$$

where $(N_{n-1}, N_{n-2}, \ldots, N_0)$ is a given vector.

Similarly, Problem-M with dimension m can be equivalently stated as finding a tuple $(e_1, e_2, \ldots, e_m)$ with $\sum_{i=1}^{m} H(e_i)$ maximal from the following set

$$\begin{aligned}\mathcal{S}_M &= \{(e_1, e_2, \ldots, e_m) | e_j = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{j,i}), \\ &\quad 0 \leq \sum_{j=1}^{m} \gamma_{j,i} \leq N_i \text{ for } i \in [0, n-1], j \in [1, m]\}.\end{aligned}$$

To solve Problem-U, our main idea is to find an equivalent set $\mathcal{S}'_U$ such that $\mathcal{S}_U = \mathcal{S}'_U$ and we can simply find the solution by directly studying $\mathcal{S}'_U$.

To solve Problem-M, similarly, we aim to find another set $\mathcal{S}'_M$ such that $\mathcal{S}_M = \mathcal{S}'_M$ and $\mathcal{S}'_M$ is much easier to study.

4.1 The Set Equivalence Theorem for Problem-U

To find the desired equivalent sets, we first build some theorems to ensure the correctness.

Lemma 1. *Let*

$$\begin{aligned}\mathcal{S}_1 &= \{e | e = a + 2b, 0 \leq a \leq c_1, 0 \leq b \leq c_2\}, \\ \mathcal{S}_2 &= \{e | e = a + 2b, 0 \leq a \leq c'_1, 0 \leq b \leq c'_2\},\end{aligned}$$

where $2c_2 + c_1 = 2c'_2 + c'_1$. Then, when one of the following conditions hold:

1. $c_1 > 0, c'_1 > 0$,
2. $c_1 = c'_1 = 0$,

we have $\mathcal{S}_1 = \mathcal{S}_2$.

Proof. When $c_1 > 0$, the set $\mathcal{S}_1$ indeed corresponds to the set of numbers $0, 1, 2, \ldots, 2c_2 + c_1$, i.e.

$$\mathcal{S}_1 = \{e | 0 \leq e \leq 2c_2 + c_1\}.$$

When $c'_1 > 0$, we also have

$$\mathcal{S}_2 = \{e | 0 \leq e \leq 2c'_2 + c'_1\}.$$

Since $2c_2 + c_1 = 2c'_2 + c'_1$, we have $\mathcal{S}_1 = \mathcal{S}_2$ when $c_1 > 0$ and $c'_1 > 0$.

When $c_1 = c'_1 = 0$, we have $c_2 = c'_2$ and hence $\mathcal{S}_1 = \mathcal{S}_2$. □

Theorem 1. *Let t be a given positive integer. Let $(N'_{n-1}, N'_{n-2}, \ldots, N'_0) \in \mathbb{N}^n$ and $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$ be two given vectors where $N'_i = N_i$ for $i \in \mathcal{I} = \{0, 1, \ldots, n-1\} \setminus \{j, (j+1)\%n\}$ and $N_j \geq t > 0$. Moreover, when $(N_j - t)\%2 = 1$,*

$$\begin{cases} N'_j = t + 1, \\ N'_{(j+1)\%n} = \dfrac{N_j - t - 1}{2} + N_{(j+1)\%n}. \end{cases} \tag{1}$$

When $(N_j - t)\%2 = 0$,

$$\begin{cases} N'_j = t, \\ N'_{(j+1)\%n} = \dfrac{N_j - t}{2} + N_{(j+1)\%n}. \end{cases} \tag{2}$$

Then, for

$$\begin{aligned}\mathcal{S}_1 &= \{e|e=\mathcal{M}_n(\sum_{i=0}^{n-1}2^i\gamma_i), 0\le\gamma_i\le N_i \text{ for } i\in[0,n-1]\},\\ \mathcal{S}_2 &= \{e|e=\mathcal{M}_n(\sum_{i=0}^{n-1}2^i\gamma_i), 0\le\gamma_i\le N'_i \text{ for } i\in[0,n-1]\},\end{aligned}$$

we have $\mathcal{S}_1=\mathcal{S}_2$.

Proof. Let

$$\begin{aligned}\mathcal{S}_3 &= \{e|e=a+2b, 0\le a\le N_j, 0\le b\le N_{(j+1)\%n}\},\\ \mathcal{S}_4 &= \{e|e=a+2b, 0\le a\le N'_j, 0\le b\le N'_{(j+1)\%n}\}.\end{aligned}$$

Then, $\mathcal{S}_1$ and $\mathcal{S}_2$ can be rewritten as

$$\begin{aligned}\mathcal{S}_1 &= \{e|e=\mathcal{M}_n(2^je_0+\sum_{i\in\mathcal{I}}2^i\gamma_i), 0\le\gamma_i\le N_i \text{ for } i\in\mathcal{I}, e_0\in\mathcal{S}_3\},\\ \mathcal{S}_2 &= \{e|e=\mathcal{M}_n(2^je_1+\sum_{i\in\mathcal{I}}2^i\gamma_i), 0\le\gamma_i\le N_i \text{ for } i\in\mathcal{I}, e_1\in\mathcal{S}_4\},\end{aligned}$$

respectively.

Since $N_j>0$ and $N'_j>0, 2N_{(j+1)\%n}+N_j=2N'_{(j+1)\%n}+N'_j$ when either Equation 6 or Equation 7 holds, according to Lemma 1, we have $\mathcal{S}_3=\mathcal{S}_4$. Hence, $\mathcal{S}_1=\mathcal{S}_2$. □

Algorithm 1 Finding equivalent sets for Problem-U

```
procedure REDUCE(N_{n-1}, N_{n-2}, ..., N_0)
    for i in range (n) do
        if N_i ≥ 1 and (N_i − 1)%2 = 1 then
            N_{(i+1)%n} = N_{(i+1)%n} + (N_i − 2)/2
            N_i = 2
        else if N_i ≥ 1 and (N_i − 1)%2 = 0 then
            N_{(i+1)%n} = N_{(i+1)%n} + (N_i − 1)/2
            N_i = 1
    return (N_{n-1}, N_{n-2}, ..., N_0)
```

Application of Theorem 1. By consecutively applying Theorem 1 with $t=1$, an equivalent set of $\mathcal{S}_U$ can be easily found with the reduction algorithm, as shown in Algorithm 1. Specifically, let $(N'_{n-1},N'_{n-2},\ldots,N'_0)\leftarrow\texttt{REDUCE}(N_{n-1},N_{n-2},\ldots,N_0)$. Then, we have

$$\begin{aligned}&\mathcal{S}_U=\{e|e=\mathcal{M}_n(\sum_{i=0}^{n-1}2^i\gamma_i), 0\le\gamma_i\le N_i \text{ for } i\in[0,n-1]\}\\ =\ &\mathcal{S}'_U=\{e|e=\mathcal{M}_n(\sum_{i=0}^{n-1}2^i\gamma_i), 0\le\gamma_i\le N'_i \text{ for } i\in[0,n-1]\}.\end{aligned}$$

Let us analyze $\mathcal{S}'_U$. By following Algorithm 1, we can observe that $N'_i\in\{0,1,2\}$ for $i\in[1,n-1]$. However, N'_0 may be still large because at the last step we may update N'_0 with $N'_0=N'_0+(N_{n-1}-1)/2, N'_{n-1}=1$ or $N'_0=N'_0+(N_{n-1}-2)/2, N'_{n-1}=2$.

Then, let us consider the case when the reduction algorithm is applied twice to the vector $(N_{n-1}, N_{n-2}, \ldots, N_0)$, i.e. we consider

$$\begin{aligned}(N'_{n-1}, N'_{n-2}, \ldots, N'_0) &\leftarrow \texttt{REDUCE}(N_{n-1}, N_{n-2}, \ldots, N_0),\\ (N''_{n-1}, N''_{n-2}, \ldots, N''_0) &\leftarrow \texttt{REDUCE}(N'_{n-1}, N'_{n-2}, \ldots, N'_0).\end{aligned}$$

In this way, we can find that there are at most two possible forms for $(N''_{n-1}, N''_{n-2}, \ldots, N''_0)$, as shown below:

Form 1: $N''_i \in \{0, 1, 2\}, \forall i \in [0, n-1]$.

Form 2: $N''_i > 0, \forall i \in [0, n-1]$.

Form 1 is easy to explain. The main problem is how to explain Form 2. As already stated above, we have $N'_i \in \{0, 1, 2\}, \forall i \in [1, n-1]$. If N'_0 is too large, then at the reduction phase, each N''_i with $1 \leq i \leq n-1$ will be updated to either 1 or 2 and N''_0 can be still very large, which explains Form 2. Note that if each element in the vector is either 1 or 2, the vector can be either Form 1 or Form 2.

Due to this property, we compute the equivalent set S'_U by always running Algorithm 1 twice, i.e. we consider

$$\begin{aligned}&(N'_{n-1}, N'_{n-2}, \ldots, N'_0) \leftarrow \texttt{REDUCE}(\texttt{REDUCE}(N_{n-1}, N_{n-2}, \ldots, N_0)),\\ &\mathcal{S}'_U = \{e | e = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_i), 0 \leq \gamma_i \leq N'_i \text{ for } i \in [0, n-1]\}.\end{aligned}$$

4.2 Extending the Set Equivalence Theorem for Problem-M

For Problem-M with dimension m, we need to consider a set $\mathcal{S}_M$ where each of its elements is a tuple of m natural numbers. Intuitively, finding such an equivalent set $\mathcal{S}'_M = \mathcal{S}_M$ becomes much harder when $m > 1$.

Lemma 2. *Let t be a given positive integer. Let*

$$\begin{aligned}\mathcal{S} &= \{e | e = a + 2b, 0 \leq a \leq f_1 - \sum_{i=1}^{t} c_i, 0 \leq b \leq f_2 - \sum_{i=1}^{t} d_i\},\\ \mathcal{S}' &= \{e | e = a + 2b, 0 \leq a \leq f'_1 - \sum_{i=1}^{t} c'_i, 0 \leq b \leq f'_2 - \sum_{i=1}^{t} d'_i\}\},\end{aligned}$$

where

$$\begin{cases} 2f_2 + f_1 = 2f'_2 + f'_1,\\ c_i + 2d_i = c'_i + 2d'_i, \forall i \in [1, t],\\ 0 \leq \sum_{i=1}^{t} c_i \leq f_1,\ 0 \leq \sum_{i=1}^{t} d_i \leq f_2,\\ 0 \leq \sum_{i=1}^{t} c'_i < f'_1,\ 0 \leq \sum_{i=1}^{t} d'_i \leq f'_2. \end{cases}$$

Then, $\mathcal{S} \subseteq \mathcal{S}'$.

Proof. Note that due to the specified conditions, we always have

$$2(f_2 - \sum_{i=1}^{t} d_i) + (f_1 - \sum_{i=1}^{t} c_i) = 2f_2 + f_1 - (\sum_{i=1}^{t} (c_i + 2d_i))$$

$$= \quad 2f_2' + f_1' - (\sum_{i=1}^{t}(c_i' + 2d_i')) = 2(f_2' - \sum_{i=1}^{t} d_i') + (f_1' - \sum_{i=1}^{t} c_i').$$

Due to the condition $0 \leq \sum_{i=1}^{t} c_i \leq f_1$, we cannot directly use Lemma 1. Hence, we consider two cases. First, when $0 \leq \sum_{i=1}^{t} c_i = f_1$, we have $2(f_2 - \sum_{i=1}^{t} d_i) = 2(f_2' - \sum_{i=1}^{t} d_i') + (f_1' - \sum_{i=1}^{t} c_i')$ and hence

$$\begin{aligned} \mathcal{S} &= \{e | e = 2b, 0 \leq b \leq f_2 - \sum_{i=1}^{t} d_i\}, \\ \mathcal{S}' &= \{e | 0 \leq e \leq 2(f_2 - \sum_{i=1}^{t} d_i)\}, \end{aligned}$$

which implies $\mathcal{S} \subseteq \mathcal{S}'$.

Second, when $0 \leq \sum_{i=1}^{t} c_i < f_1$, due to Lemma 1, we directly have $\mathcal{S}' = \mathcal{S}$. Therefore, $\mathcal{S} \subseteq \mathcal{S}'$ always holds. □

Lemma 3. *Let m be a given positive integer. Let*

$$\begin{aligned} \mathcal{S} &= \{(e_1, e_2, \ldots, e_m) | e_t = a_t + 2b_t, 0 \leq \sum_{t=1}^{m} a_t \leq f_1, 0 \leq \sum_{t=1}^{m} b_t \leq f_2, t \in [1, m]\}, \\ \mathcal{S}' &= \{(e_1, e_2, \ldots, e_m) | e_t = a_t + 2b_t, 0 \leq \sum_{t=1}^{m} a_t \leq f_1', 0 \leq \sum_{t=1}^{m} b_t \leq f_2', t \in [1, m]\}, \end{aligned}$$

where $2f_2 + f_1 = 2f_2' + f_1'$ and $f_1 \geq m$, $f_1' \geq m$. Then, $\mathcal{S} = \mathcal{S}'$.

Proof. We prove Lemma 3 by induction. Let $\mathcal{E}_i$ and $\mathcal{E}_i'$ be the sets of all possible values of $(e_1, e_2, \ldots, e_i)$ in $\mathcal{S}$ and $\mathcal{S}'$, respectively. The proof by induction is to first prove $\mathcal{E}_1 = \mathcal{E}_1'$ and then prove $\mathcal{E}_t = \mathcal{E}_t'$ under the condition $\mathcal{E}_{t-1} = \mathcal{E}_{t-1}'$ where $t \in [1, m]$. Note that $\mathcal{S} = \mathcal{E}_m$ and $\mathcal{S}' = \mathcal{E}_m'$ by definition.

First, we prove $\mathcal{E}_1 = \mathcal{E}_1'$. In this case, we have

$$\begin{aligned} \mathcal{E}_1 &= \{e_1 | e_1 = a_1 + 2b_1, 0 \leq a_1 \leq f_1, 0 \leq b_1 \leq f_2\}, \\ \mathcal{E}_1' &= \{e_1 | e_1 = a_1 + 2b_1, 0 \leq a_1 \leq f_1', 0 \leq b_1 \leq f_2'\}. \end{aligned}$$

Since $f_1 \geq m > 0$, $f_1' \geq m > 0$ and $2f_2 + f_1 = 2f_2' + f_1'$ always hold, according to Lemma 1, we have $\mathcal{E}_1 = \mathcal{E}_1'$.

Next, we prove $\mathcal{E}_t = \mathcal{E}_t'$ under the condition $\mathcal{E}_{t-1} = \mathcal{E}_{t-1}'$ where $t \in [1, m]$. Denote the set of all possible values of e_t in $\mathcal{E}_t$ with the same prefix $(e_1, e_2, \ldots, e_{t-1})$ by $\mathcal{S}_{(e_1, e_2, \ldots, e_{t-1})}$, i.e.

$$\mathcal{E}_t = \{(e_1, e_2, \ldots, e_t) | (e_1, e_2, \ldots, e_{t-1}) \in \mathcal{E}_{t-1}, e_t \in \mathcal{S}_{(e_1, e_2, \ldots, e_{t-1})}\}$$

For each $(e_1, e_2, \ldots, e_{t-1}) \in \mathcal{E}_{t-1}$, we associate a vector

$$(C_{t-1}, D_{t-1}) = (c_1, c_2, \ldots, c_{t-1}, d_1, d_2, \ldots, d_{t-1})$$

where

$$\begin{aligned} & e_i = c_i + 2d_i, \ \forall i \in [1, t-1], \\ & 0 \leq \sum_{i=1}^{t-1} c_i \leq f_1, \ 0 \leq \sum_{i=1}^{t-1} d_i \leq f_2, \end{aligned}$$

In this case, we can have

$$\mathcal{S}_{(e_1, e_2, \ldots, e_{t-1})} \supseteq \{e | e = 2a + b, 0 \leq a \leq f_1 - \sum_{i=1}^{t-1} c_i, 0 \leq b \leq f_2 - \sum_{i=1}^{t-1} d_i\}.$$

Case-1. If $\sum_{i=1}^{t-1} c_i = f_1$ and $\sum_{i=1}^{t-1} d_i = f_2$, we have

$$\mathcal{S}_{(e_1,e_2,\ldots,e_{t-1})} \supseteq \{0\}.$$

Case-2. If $\sum_{i=1}^{t-1} c_i = f_1$ and $\sum_{i=1}^{t-1} d_i < f_2$, since $f_1 \geq m > t-1$, there must exist an index h such that $c_h \geq 2$. In this case, we can make

$$\begin{aligned}
&(C''_{t-1}, D''_{t-1}) = (c''_1, c''_2, \ldots, c''_{t-1}, d''_1, d''_2, \ldots, d''_{t-1}),\\
&c_i'' = c_i, d''_i = d_i, \forall i \in [1, h-1] \cup [h+1, t-1],\\
&c_h'' = c_h - 2, d''_h = d_h + 1,
\end{aligned}$$

due to $\sum_{i=1}^{t-1} c''_i = (\sum_{i=1}^{t-1} c_i) - 2 < f_1$ and $\sum_{i=1}^{t-1} d''_i = 1 + \sum_{i=1}^{t-1} d_i \leq f_2$. Note that we still have

$$e_i = c''_i + 2d''_i, \ \forall i \in [1, t-1].$$

According to Lemma 2, we have

$$\begin{aligned}
\mathcal{S}_{(e_1,e_2,\ldots,e_{t-1})} &\supseteq \{e | e = 2a + b, 0 \leq a \leq f_1 - \sum_{i=1}^{t-1} c''_i, 0 \leq b \leq f_2 - \sum_{i=1}^{t-1} d''_i\}\\
&\supseteq \{e | e = 2a + b, 0 \leq a \leq f_1 - \sum_{i=1}^{t-1} c_i, 0 \leq b \leq f_2 - \sum_{i=1}^{t-1} d_i\}\\
&\supseteq \{0\}.
\end{aligned}$$

The above two cases imply that to construct $\mathcal{S}_{(e_1,e_2,\ldots,e_{t-1})}$ for each $(e_1, e_2, \ldots, e_{t-1}) \in \mathcal{E}_{t-1}$, it is sufficient to consider the associated vector

$$(C_{t-1}, D_{t-1}) = (c_1, c_2, \ldots, c_{t-1}, d_1, d_2, \ldots, d_{t-1})$$

where

$$\begin{cases}
e_i = c_i + 2d_i, \ \forall i \in [1, t-1],\\
0 \leq \sum_{i=1}^{t-1} c_i < f_1, \ 0 \leq \sum_{i=1}^{t-1} d_i \leq f_2.
\end{cases} \tag{3}$$

In other words, we can just ignore those associated vectors which satisfy $\sum_{i=1}^{t-1} c_i = f_1$.

In this way, we can interpret $\mathcal{S}_{(e_1,e_2,\ldots,e_{t-1})}$ from another perspective. Denote the set of all possible e_t in $\mathcal{S}$ under the associated vector (C_{t-1}, D_{t-1}) satisfying Equation 3 by $\mathcal{P}_{(C_{t-1},D_{t-1})}$, i.e.

$$\mathcal{P}_{(C_{t-1},D_{t-1})} = \{e | e = a + 2b, 0 \leq a \leq f_1 - \sum_{i=1}^{t-1} c_i, 0 \leq b \leq f_2 - \sum_{i=1}^{t-1} d_i\}.$$

Moreover, denote the set of all possible vectors (C_{t-1}, D_{t-1}) satisfying Equation 3 by $\mathcal{V}_{C,D}$. Then, we have

$$\mathcal{S}_{(e_1,e_2,\ldots,e_{t-1})} = \bigcup_{(C_{t-1},D_{t-1}) \in \mathcal{V}_{C,D}} \mathcal{P}_{(C_{t-1},D_{t-1})}.$$

Due to the symmetry between $\mathcal{S}$ and $\mathcal{S}'$, we can also interpret $\mathcal{E}'_t$ as

$$\mathcal{E}_t = \{(e'_1, e'_2, \ldots, e'_t) | (e'_1, e'_2, \ldots, e'_{t-1}) \in \mathcal{E}'_{t-1}, e'_t \in \mathcal{S}'_{(e'_1,e'_2,\ldots,e'_{t-1})}\},$$

where $\mathcal{S}'_{(e'_1,e'_2,\ldots,e'_{t-1})}$ denotes the set of all possible e'_t in $\mathcal{E}'_t$ under the same prefix $(e'_1, e'_2, \ldots, e'_{t-1})$. Moreover, we can associate each $(e'_1, e'_2, \ldots, e'_{t-1}) \in \mathcal{E}'_t$ with a vector

$$(C'_{t-1}, D'_{t-1}) = (c'_1, c'_2, \ldots, c'_{t-1}, d'_1, d'_2, \ldots, d'_{t-1})$$

where

$$\begin{cases} e'_i = c'_i + 2d'_i, \ \forall i \in [1, t-1], \\ 0 \leq \sum_{i=1}^{t-1} c'_i < f'_1, \ 0 \leq \sum_{i=1}^{t-1} d'_i \leq f'_2. \end{cases} \tag{4}$$

Denote the set of all possible vectors (C'_{t-1}, D'_{t-1}) satisfying Equation 4 by $\mathcal{V}'_{C,D}$. Then, we have

$$\mathcal{S}'_{(e'_1,e'_2,\ldots,e'_{t-1})} = \bigcup_{(C'_{t-1},D'_{t-1}) \in \mathcal{V'}_{C,D}} \mathcal{P}'_{(C'_{t-1},D'_{t-1})},$$

where $\mathcal{P}'_{(C'_{t-1},D'_{t-1})}$ is defined as follows:

$$\mathcal{P}'_{(C'_{t-1},D'_{t-1})} = \{e | e = a + 2b, 0 \leq a_t \leq f'_1 - \sum_{i=1}^{t-1} c'_i, 0 \leq b \leq f'_2 - \sum_{i=1}^{t-1} d'_i\}.$$

Note that our aim is to prove $\mathcal{E}'_t = \mathcal{E}_t$ under the condition $\mathcal{E}'_{t-1} = \mathcal{E}_{t-1}$. In other words, we need to prove that for each

$$(e'_1, e'_2, \ldots, e'_{t-1}) = (e_1, e_2, \ldots, e_{t-1}) \in \mathcal{E}_{t-1} = \mathcal{E}_{t-1},$$

there is

$$\begin{aligned} & \mathcal{S}'_{(e'_1,e'_2,\ldots,e'_{t-1})} = \mathcal{S}_{(e_1,e_2,\ldots,e_{t-1})} \\ \Leftrightarrow \quad & \bigcup_{(C_{t-1},D_{t-1}) \in \mathcal{V}_{C,D}} \mathcal{P}_{(C_{t-1},D_{t-1})} = \bigcup_{(C'_{t-1},D'_{t-1}) \in \mathcal{V'}_{C,D}} \mathcal{P}'_{(C'_{t-1},D'_{t-1})} \end{aligned}$$

This can be further reduced to proving that for any (C_{t-1}, D_{t-1}) and (C'_{t-1}, D'_{t-1}) satisfying

$$\begin{cases} e_i = c_i + 2d_i = c'_i + 2d'_i = e'_i, \ \forall i \in [1, t-1], \\ 0 \leq \sum_{i=1}^{t-1} c_i < f_1, \ 0 \leq \sum_{i=1}^{t-1} d_i \leq f_2, \\ 0 \leq \sum_{i=1}^{t-1} c'_i < f'_1, \ 0 \leq \sum_{i=1}^{t-1} d'_i \leq f'_2, \end{cases} \tag{5}$$

there is always $\mathcal{P}_{(C_{t-1},D_{t-1})} = \mathcal{P}'_{(C'_{t-1},D'_{t-1})}$ where

$$\begin{aligned} \mathcal{P}_{(C_{t-1},D_{t-1})} &= \{e | e = a + 2b, 0 \leq a \leq f_1 - \sum_{i=1}^{t-1} c_i, 0 \leq b \leq f_2 - \sum_{i=1}^{t-1} d_i\}, \\ \mathcal{P}'_{(C'_{t-1},D'_{t-1})} &= \{e | e = a + 2b, 0 \leq a_t \leq f'_1 - \sum_{i=1}^{t-1} c'_i, 0 \leq b \leq f'_2 - \sum_{i=1}^{t-1} d'_i\}. \end{aligned}$$

Due to Equation 5 and $2f_2 + f_1 = 2f'_2 + f'_1$, according to Lemma 1, $\mathcal{P}_{(C_{t-1},D_{t-1})} = \mathcal{P}'_{(C'_{t-1},D'_{t-1})}$ always holds. Hence, we complete the proof. □

Theorem 2. *Let* $(N'_{n-1}, N'_{n-2}, \ldots, N'_0) \in \mathbb{N}^n$ *and* $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$ *be two given vectors where* $N'_i = N_i$ *for* $i \in \mathcal{I} = \{0, 1, \ldots, n-1\} \setminus \{j, (j+1)\%n\}$ *and* $N_j \geq m > 0$. *Moreover, when* $(N_j - m)\%2 = 1$,

$$\begin{cases} N'_j = m+1, \\ N'_{(j+1)\%n} = \dfrac{N_j - m - 1}{2} + N_{(j+1)\%n}. \end{cases} \tag{6}$$

When $(N_j - m)\%2 = 0$,

$$\begin{cases} N'_j = m, \\ N'_{(j+1)\%n} = \dfrac{N_j - m}{2} + N_{(j+1)\%n}. \end{cases} \tag{7}$$

Then, for

$$\begin{aligned} \mathcal{S}_M &= \{(e_1, e_2, \ldots, e_m) | e_t = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{t,i}), 0 \leq \sum_{t=1}^{m} \gamma_{t,i} \leq N_i \text{ for } i \in [0, n-1], t \in [1, m]\}, \\ \mathcal{S}'_M &= \{(e_1, e_2, \ldots, e_m) | e_t = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{t,i}), 0 \leq \sum_{t=1}^{m} \gamma_{t,i} \leq N'_i \text{ for } i \in [0, n-1], t \in [1, m]\}, \end{aligned}$$

we have $\mathcal{S}_M = \mathcal{S}'_M$.

Proof. Let

$$\begin{aligned} \mathcal{S}_5 &= \{(k_1, k_2, \ldots, k_m) | k_t = a_t + 2b_t, 0 \leq \sum_{t=1}^{m} a_t \leq N_j, 0 \leq \sum_{t=1}^{m} b_t \leq N_{(j+1)\%n}, t \in [1, m]\}, \\ \mathcal{S}_6 &= \{(k_1, k_2, \ldots, k_m) | k_t = a_t + 2b_t, 0 \leq \sum_{t=1}^{m} a_t \leq N'_j, 0 \leq \sum_{t=1}^{m} b_t \leq N'_{(j+1)\%n}, t \in [1, m]\}, \end{aligned}$$

Due to the specified conditions on $(N_j, N_{(j+1)\%n})$, according to Lemma 3, we have $\mathcal{S}_5 = \mathcal{S}_6$.

Then, we can rewrite $\mathcal{S}_M$ and $\mathcal{S}'_M$ as follows:

$$\begin{aligned} \mathcal{S}_M &= \{(e_1, e_2, \ldots, e_m) | e_t = \mathcal{M}_n(2^j k_t + \sum_{i \in \mathcal{I}} 2^i \gamma_{t,i}), 0 \leq \sum_{t=1}^{m} \gamma_{t,i} \leq N_i \text{ for } i \in \mathcal{I}, (k_1, k_2, \ldots, k_m) \in \mathcal{S}_5\}, \\ \mathcal{S}'_M &= \{(e_1, e_2, \ldots, e_m) | e_t = \mathcal{M}_n(2^j k_t + \sum_{i \in \mathcal{I}} 2^i \gamma_{t,i}), 0 \leq \sum_{t=1}^{m} \gamma_{t,i} \leq N_i \text{ for } i \in \mathcal{I}, (k_1, k_2, \ldots, k_m) \in \mathcal{S}_6\}, \end{aligned}$$

Hence, we have $\mathcal{S}_M = \mathcal{S}'_M$. □

Application of Theorem 2. With similar analysis as for Algorithm 1, by running Algorithm 2 twice, i.e. we consider

$$(N'_{n-1}, N'_{n-2}, \ldots, N_0) \leftarrow \texttt{REDUCE-M}(m, \texttt{REDUCE-M}(m, N_{n-1}, N_{n-2}, \ldots, N_0)),$$

the output will be of the following two possible forms:

Form 1: $N'_i \in [0, m+1],\ \forall i \in [0, n-1]$.

Form 2: $N'_i \geq m,\ \forall i \in [0, n-1]$.

In other words, we find an equivalent set $\mathcal{S}'_M$ for $\mathcal{S}_M$ where

$$\mathcal{S}'_M = \{(e_1, e_2, \ldots, e_m) | e_t = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{t,i}), 0 \leq \sum_{t=1}^{m} \gamma_{t,i} \leq N'_i \text{ for } i \in [0, n-1], t \in [1, m]\}.$$

Algorithm 2 Finding equivalent sets for Problem-M with dimension m

procedure REDUCE-M($m, N_{n-1}, N_{n-2}, \ldots, N_0$)
  for i in range (n) **do**
    if $N_i \geq m$ and $(N_i - m)\%2 = 1$ **then**
      $N_{(i+1)\%n} = N_{(i+1)\%n} + (N_i - m - 1)/2$
      $N_i = m + 1$
    else if $N_i \geq m$ and $(N_i - m)\%2 = 0$ **then**
      $N_{(i+1)\%n} = N_{(i+1)\%n} + (N_i - m)/2$
      $N_i = m$
  return $(N_{n-1}, N_{n-2}, \ldots, N_0)$

5 Solving Optimal Problems By Processing Equivalent Sets

It is now clear that Problem-U is just a special case of Problem-M with $m = 1$. Hence, in this section, we only focus on how to solve Problem-M with dimension m.

According to the above explanation, after running Algorithm 2 twice, we can find a "reduced" vector $(N_{n-1}, N_{n-2}, \ldots, N_0)$ to equivalently describe the set $\mathcal{S}_M$ and there are two possible forms of $(N_{n-1}, N_{n-2}, \ldots, N_0)$:

Form 1: $N_i \in [0, m+1],\ \forall i \in [0, n-1]$.

Form 2: $N_i \geq m,\ \forall i \in [0, n-1]$.

5.1 Proceeding Form 2

For Form 2, according to Lemma 4 specified below, we can directly obtain that the solution to the Problem-M with dimension m is nm.

Lemma 4. *For a given vector $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$ where $N_i \geq m$ for $\forall i \in [0, n-1]$, there exists an element $(e_1, e_2, \ldots, e_m) \in \mathcal{S}_M$ where*

$$\mathcal{S}_M \quad = \quad \{(e_1, e_2, \ldots, e_m) | e_t = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{t,i}), 0 \leq \sum_{t=1}^{m} \gamma_{t,i} \leq N_i \text{ for } i \in [0, n-1], t \in [1, m]\},$$

such that

$$\sum_{i=1}^{m} H(e_i) = nm.$$

Hence, the solution to Problem-M with dimension m in this case is nm.

Proof. By assigning $\gamma_{t,i} = 1$ for $\forall i \in [0, n-1], t \in [1, m]$, we obtain an element

$$(e_1, e_2, \ldots, e_m) = (2^n - 1, 2^n - 1, \ldots, 2^n - 1) \in \mathcal{S}_M$$

due to

$$2^n - 1 \quad = \quad \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i).$$

Hence, $\sum_{i=1}^{m} H(e_i) = nm$. As the upper bound for Problem-M with dimension m is nm and we find an assignment to satisfy this upper bound in this case, the solution to this optimization problem is nm. □

5.2 Proceeding Form 1

Next, we describe how to solve Problem-M when the vector $(N_{n-1}, N_{n-2}, \ldots, N_0)$ is of Form 1.

Lemma 5. *For any $a, b \in [0, 2^n - 1]$, we have $H(\mathcal{M}_n(a+b)) \leq H(a) + H(b)$.*

This lemma is critical to finding the upper bound for the optimization problem. Although it looks obvious, the proof requires significant efforts and we put it at Appendix due to its length.

Theorem 3. *For any $m_1, m_2, \ldots, m_t \in [0, 2^n - 1]$, we have*

$$H(\mathcal{M}_n(m_1 + m_2 + \cdots + m_t)) \leq H(m_1) + H(m_2) + \cdots + H(m_t).$$

Proof. According to Lemma 5, we have

$$\begin{aligned}
& H(\mathcal{M}_n(m_1 + m_2 + \cdots + m_t)) \\
= \quad & H(\mathcal{M}_n(m_1 + \mathcal{M}_n(\sum_{i=2}^{t} m_i))) \\
\leq \quad & H(m_1) + H(\mathcal{M}_n(\sum_{i=2}^{t} m_i)) \\
\leq \quad & H(m_1) + H(m_2) + H(\mathcal{M}_n(\sum_{i=3}^{t} m_i)) \\
\cdots & \\
\leq \quad & H(m_1) + H(m_2) + \cdots + H(m_t).
\end{aligned}$$

□

Lemma 6. *For any natural number a, we have $H(a) \leq a$. If $a \geq 2$, we further have $H(a) \leq a - 1$.*

Proof. For any $i \geq 0$ and $2^i \leq a \leq 2^{i+1} - 1$, we have $H(a) \leq i + 1 \leq 2^i \leq a$. Therefore, $H(a) \leq a$ always holds. Moreover, for any $2^i \leq a \leq 2^{i+1} - 1$ where $i \geq 2$, we have $H(a) \leq i + 1 \leq 2^i - 1 \leq a - 1$. In addition, $H(2) \leq 2 - 1$ and $H(3) \leq 3 - 1$. Hence, $H(a) \leq a - 1$ for $a \geq 2$. □

Lemma 7. *Let $m_1, m_2, \ldots, m_t \in \mathbb{N}$ and t be a positive integer. If*

$$\sum_{i=1}^{t} m_i = k \leq t,$$

we have $\sum_{i=1}^{t} H(m_i) \leq k$.

If

$$\sum_{i=1}^{t} m_i = t + 1,$$

we have $\sum_{i=1}^{t} H(m_i) \leq t$.

Proof. According to Lemma 6, we always have

$$\sum_{i=1}^{t} H(m_i) \leq \sum_{i=1}^{t} m_i.$$

If $\sum_{i=1}^{t} m_i = k \leq t$ holds, we immediately obtain $\sum_{i=1} H(m_i) \leq \sum_{i=1}^{t} m_i = k$.

If $\sum_{i=1}^{t} m_i = t+1$ holds, there will exist an index i' such that $m_{i'} \geq 2$. Hence, according to Lemma 6, we have $\sum_{i=1} H(m_i) \leq (\sum_{i=1}^{t} m_i) - 1 = t$. □

Theorem 4. *Let $(N_{n-1}, N_{n-2}, \ldots, N_0)$ be a vector where $N_i \in [0, m+1]$ for $\forall i \in [0, n-1]$ and $m+1 \leq 2^n - 1$. Let $\mathcal{I}$ be a set of indices such that $i \in \mathcal{I}$ if $N_i = m+1$. Then, the solution to the following optimization problem*

$$\begin{aligned} \text{maximize} \quad & H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{m,i})), \\ \text{subject to} \quad & 0 \leq \gamma_{1,i} + \gamma_{2,i} + \cdots + \gamma_{m,i} \leq N_i \text{ for } i \in [0, n-1]. \end{aligned}$$

is

$$(\sum_{i=0}^{n-1} N_i) - |\mathcal{I}|.$$

Proof. According to Theorem 3, we have

$$\begin{aligned} & H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{m,i})) \\ \leq & \sum_{i=0}^{n-1} H(\mathcal{M}_n(2^i \gamma_{1,i})) + \sum_{i=0}^{n-1} H(\mathcal{M}_n(2^i \gamma_{2,i})) + \ldots + \sum_{i=0}^{n-1} H(\mathcal{M}_n(2^i \gamma_{m,i})) \\ = & \sum_{i=0}^{n-1} H(\mathcal{M}_n(\gamma_{1,i})) + \sum_{i=0}^{n-1} H(\mathcal{M}_n(\gamma_{2,i})) + \ldots + \sum_{i=0}^{n-1} H(\mathcal{M}_n(\gamma_{m,i})) \\ = & \sum_{i=0}^{n-1} \sum_{j=1}^{m} H(\mathcal{M}_n(\gamma_{j,i})). \end{aligned}$$

Since

$$0 \leq \sum_{j=1}^{m} \gamma_{j,i} \leq N_i \leq m+1 \leq 2^n - 1,$$

we can remove the modular operation and obtain

$$\sum_{i=0}^{n-1} \sum_{j=1}^{m} H(\mathcal{M}_n(\gamma_{j,i})) = \sum_{i=0}^{n-1} \sum_{j=1}^{m} H(\gamma_{j,i})$$

Let $\mathcal{I}'$ be another set such that $i \in \mathcal{I}'$ if $N_i \leq m$. Then, we further have

$$\sum_{i=0}^{n-1} \sum_{j=1}^{m} H(\gamma_{j,i}) = \sum_{i \in \mathcal{I}} \sum_{j=1}^{m} H(\gamma_{j,i}) + \sum_{i \in \mathcal{I}'} \sum_{j=1}^{m} H(\gamma_{j,i})$$

According to Lemma 7, we then have

$$\begin{aligned} & \sum_{i \in \mathcal{I}} \sum_{j=1}^{m} H(\gamma_{j,i}) + \sum_{i \in \mathcal{I}'} \sum_{j=1}^{m} H(\gamma_{j,i}) \\ \leq & |\mathcal{I}| \times m + \sum_{i \in \mathcal{I}'} N_i \end{aligned}$$

$$
\begin{aligned}
&= |\mathcal{I}| \times (m+1) - |\mathcal{I}| + \sum_{i \in \mathcal{I}'} N_i \\
&= (\sum_{i=0}^{n-1} N_i) - |\mathcal{I}|.
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
& H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{m,i})) \\
\leq \; & (\sum_{i=0}^{n-1} N_i) - |\mathcal{I}|,
\end{aligned}
$$

which implies the upper bound for the optimization problem is $(\sum_{i=0}^{n-1} N_i) - |\mathcal{I}|$.

For each $i \in \mathcal{I}$, we can assign $\gamma_{j,i} = 1$ for each $j \in [1, m]$. For each $i \in \mathcal{I}'$, we can assign $\gamma_{j,i} = 1$ for each $j \in [1, N_i]$ and $\gamma_{j,i} = 0$ for each $j \in [N_i + 1, m]$. In this way, we find an element $(e_1, e_2, \ldots, e_m)$ belonging to the following set

$$
\mathcal{S}_M = \{(e_1, e_2, \ldots, e_m) | e_j = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{j,i}), 0 \leq \sum_{j=1}^{m} \gamma_{j,i} \leq N_i \text{ for } i \in [0, n-1], j \in [1, m]\}
$$

such that $\sum_{i=1}^{m} H(e_i) = (\sum_{i=0}^{n-1} N_i) - |\mathcal{I}|$ because for the above assignment, we have

$$
\begin{aligned}
\sum_{i=1}^{m} H(e_i) &= H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{m,i})) \\
&= H(\sum_{i=0}^{n-1} 2^i \gamma_{1,i}) + H(\sum_{i=0}^{n-1} 2^i \gamma_{2,i}) + \cdots + H(\sum_{i=0}^{n-1} 2^i \gamma_{m,i}) \\
&= \sum_{i=0}^{n-1} \gamma_{1,i} + \sum_{i=0}^{n-1} \gamma_{2,i} + \cdots + \sum_{i=0}^{n-1} \gamma_{m,i} \\
&= (\sum_{i=0}^{n-1} N_i) - |\mathcal{I}|.
\end{aligned}
$$

Hence, the solution to the optimization problem is $(\sum_{i=0}^{n-1} N_i) - |\mathcal{I}|$. □

5.3 The $\mathcal{O}(n)$ Algorithm

Based on the above analysis, we can write a simple algorithm to solve the general optimization problem Problem-M with dimension m, as shown in Algorithm 3. Since `REDUCE-M` runs in time $\mathcal{O}(n)$, Algorithm 3 also runs in time $\mathcal{O}(n)$.

6 Other Applications

We show how the idea to construct equivalent sets can be used to improve the performance of the off-the-shelf solvers for other optimization problems in [1]. Specifically, to compute more accurate upper bounds for the algebraic degree, we need to solve the following optimization problem:

$$
\text{maximize} \qquad H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i r_1 \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i r_2 \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i r_m \gamma_{m,i})),
$$

Algorithm 3 Finding the solution to Problem-M with dimension m

procedure DEGREE$(m, N_{n-1}, N_{n-2}, \ldots, N_0)$
  $(N_{n-1}, N_{n-2}, \ldots, N_0) \leftarrow$ REDUCE-M$(m, N_{n-1}, N_{n-2}, \ldots, N_0)$
  $(N_{n-1}, N_{n-2}, \ldots, N_0) \leftarrow$ REDUCE-M$(m, N_{n-1}, N_{n-2}, \ldots, N_0)$
  $a = 0$
  $b = 0$
  $f = 0$
  for i in range (n) **do**
    $b = b + N_i$
    if $N_i = m + 1$ **then**
      $a = a + 1$
    if $N_i = 0$ **then**
      $f = 1$
  if $f = 1$ **then**
    return $b - a$ (Theorem 4)
  else
    return $n \times m$ (Lemma 4)

subject to $\quad 0 \leq \gamma_{1,i} + \gamma_{2,i} + \cdots + \gamma_{m,i} \leq N_i$ for $i \in [0, n-1]$.

where $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$ and $(r_1, r_2, \ldots, r_m) \in \mathbb{Z}_+^m$ are given vectors. Note that in [1], since the algebraic degree of the S-box is 2, there is $1 \leq H(r_i) \leq 2$ for $\forall i \in [1, m]$. We emphasize that we will consider generic $(r_1, r_2, \ldots, r_m) \in \mathbb{Z}_+^m$ in the following. In [1], after obtaining the vector $(N_{n-1}, N_{n-2}, \ldots, N_0)$, the problem is directly encoded to an MILP problem and solved with the off-the-shelf solvers.

We find that the performance can be significantly improved if we first apply the reduction algorithm REDUCE-M twice to the original vector $(N_{n-1}, N_{n-2}, \ldots, N_0)$ and then construct the MILP model for the new vector

$$(N'_{n-1}, N'_{n-2}, \ldots, N'_0) \leftarrow \texttt{REDUCE-M}(m, \texttt{REDUCE-M}(m, N_{n-1}, N_{n-2}, \ldots, N_0)).$$

To show its correctness, we should observe that the above optimization problem is equivalent to finding an element $(e_1, e_2, \ldots, e_m)$ with $\sum_{i=1}^{m} H(r_i e_i)$ maximal from the following set

$$\begin{aligned} \mathcal{S}_M \quad = \quad & \{(e_1, e_2, \ldots, e_m) | e_j = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{j,i}), \\ & 0 \leq \sum_{j=1}^{m} \gamma_{j,i} \leq N_i \text{ for } i \in [0, n-1], j \in [1, m]\}. \end{aligned}$$

With the set equivalence theorem, it has been proved that

$$\begin{aligned} \mathcal{S}_M = \mathcal{S}'_M \quad = \quad & \{(e_1, e_2, \ldots, e_m) | e_j = \mathcal{M}_n(\sum_{i=0}^{n-1} 2^i \gamma_{j,i}), \\ & 0 \leq \sum_{j=1}^{m} \gamma_{j,i} \leq N'_i \text{ for } i \in [0, n-1], j \in [1, m]\}. \end{aligned}$$

Hence, we can indeed consider the following equivalent optimization problem

$$\text{maximize} \qquad H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i r_1 \gamma_{1,i})) + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i r_2 \gamma_{2,i})) + \cdots + H(\mathcal{M}_n(\sum_{i=0}^{n-1} 2^i r_m \gamma_{m,i})),$$

subject to $0 \leq \gamma_{1,i} + \gamma_{2,i} + \cdots + \gamma_{m,i} \leq N'_i$ for $i \in [0, n-1]$.

If $N'_i \geq m$ for $\forall i \in [0, n-1]$, the solution is also directly $n \times m$. This is due to $(e_1, e_2, \ldots, e_m) = (2^n - 1, 2^n - 1, \ldots, 2^n - 1) \in \mathcal{S}'_M$ in this case, which results in $\sum_{i=1}^{m} H(r_i e_i) = n \times m$.

In practical cryptographic applications, m is very small. If $N'_i \in [0, m+1]$, i.e. the maximal value of N'_i is not larger than $m+1$, we find that the solver can solve such instances much faster than solving the problem with the original vector $(N_{n-1}, N_{n-2}, \ldots, N_0) \in \mathbb{N}^n$ where many N_i are very large. This further shows the benefits to study the basic optimization problem, i.e. how to construct equivalent sets.

Acknowledgements. We thank Clémence Bouvier and Willi Meier for discussing the preliminary version of this note.

A Proof

The proof of Lemma 5 is shown below.

Proof. Let $(a_{n-1}, a_{n-2}, \ldots, a) \in \mathbb{F}_2^n$ and $(b_{n-1}, b_{n-2}, \ldots, b_0) \in \mathbb{F}_2^n$ be the binary representations of a and b, respectively. Let $\mathcal{I}_0 = \{i_{0,1}, i_{0,2}, \ldots, i_{0,p_0}\}$ and $\mathcal{I}_1 = \{i_{1,1}, j_{1,2}, \ldots, i_{1,p_1}\}$ be the sets of indices such that $a_i = 1$ and $b_j = 1$ for $i \in \mathcal{I}_0$ and $j \in \mathcal{I}_1$. In other words, $H(a) = p_0$ and $H(b) = p_1$. Let

$$\mathcal{I}_2 = \mathcal{I}_0 \cap \mathcal{I}_1 = \{i_{2,1}, i_{2,2}, \ldots, i_{2,p_2}\}.$$

Then, we have

$$p_2 \leq min\{p_0, p_1\}.$$

In this way, we have

$$\begin{aligned}
\mathcal{M}_n(a+b) &= \mathcal{M}_n\Big(\sum_{i \in \mathcal{I}_0 \setminus \mathcal{I}_2} 2^i + \sum_{i \in \mathcal{I}_1 \setminus \mathcal{I}_2} 2^i + 2\sum_{i \in \mathcal{I}_2} 2^i\Big) = \mathcal{M}_n(\alpha_3 + \alpha_4), \\
\alpha_3 &= \sum_{i \in \mathcal{I}_0 \setminus \mathcal{I}_2} 2^i + \sum_{i \in \mathcal{I}_1 \setminus \mathcal{I}_2} 2^i, \\
\alpha_4 &= \sum_{i \in \mathcal{I}_2} 2^{(i+1)\%n}.
\end{aligned}$$

Hence, we have

$$\begin{aligned}
H(\alpha_3) &= p_0 + p_1 - 2p_2, \\
H(\alpha_4) &= p_2 \leq min\{p_0, p_1\}.
\end{aligned}$$

Repeating the same analysis, i.e. for $k \geq 1$, let

$$\begin{aligned}
\mathcal{I}_{3k} &= (\mathcal{I}_{3(k-1)} \cup \mathcal{I}_{3(k-1)+1}) \setminus \mathcal{I}_{3(k-1)+2} = \{i_{3k,1}, i_{3k,2}, \ldots, i_{3k,p_{3k}}\}, \\
\mathcal{I}_{3k+1} &= \{j | j = (i+1)\%n, i \in \mathcal{I}_{3(k-1)+2}\} = \{i_{3k+1,1}, i_{3k+1,2}, \ldots, i_{3k+1,p_{3k+1}}\}, \\
\mathcal{I}_{3k+2} &= \mathcal{I}_{3k} \cap \mathcal{I}_{3k+1} = \{i_{3k+2,1}, i_{3k+2,2}, \ldots, i_{3k+2,p_{3k+2}}\}.
\end{aligned}$$

Then, we have

$$\mathcal{M}_n(a+b) = \mathcal{M}_n(\alpha_{3k} + \alpha_{3k+1})$$

$$
\begin{aligned}
&= \mathcal{M}_n\Big(\sum_{i\in\mathcal{I}_{3k}\setminus\mathcal{I}_{3k+2}} 2^i + \sum_{i\in\mathcal{I}_{3k+1}\setminus\mathcal{I}_{3k+2}} 2^i + 2\sum_{i\in\mathcal{I}_{3k+2}} 2^i\Big)\\
&= \mathcal{M}_n(\alpha_{3(k+1)} + \alpha_{3(k+1)+1}),\\
\alpha_{3(k+1)} &= \sum_{i\in\mathcal{I}_{3k}\setminus\mathcal{I}_{3k+2}} 2^i + \sum_{i\in\mathcal{I}_{3k+1}\setminus\mathcal{I}_{3k+2}} 2^i,\\
\alpha_{3(k+1)+1} &= \sum_{i\in\mathcal{I}_{3k+2}} 2^{(i+1)\%n}.
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
p_{3(k+1)} &= p_{3k} + p_{3k+1} - 2p_{3k+2},\\
p_{3(k+1)+1} &= p_{3k+2},\\
p_{3(k+1)+2} &\le min\{p_{3k} + p_{3k+1} - 2p_{3k+2}, p_{3k+2}\} \le p_{3k+2},\\
p_{3(k+1)} + p_{3(k+1)+1} &\le p_{3k} + p_{3k+1} \le \ldots \le p_0 + p_1.
\end{aligned}
$$

Therefore, $p_{3(k+1)+2} \le p_{3k+2} \le \ldots \le p_2 \le min\{p_0, p_1\}$ must hold. Moreover, it is impossible to have a sequence $p_{3(s+\ell)+2} = \cdots = p_{3(s+1)+2} = p_{3s+2} > 0$ for $s \ge 0$ and $\ell \ge p_0 + p_1$. If there is, we have

$$
\begin{aligned}
p_{3(s+\ell)+2} &= p_{3(s+\ell-1)+2} \le min\{p_{3(s+\ell-1)} + p_{3(s+\ell-1)+1} - 2p_{3(s+\ell-1)+2}, p_{3(s+\ell-1)+2}\}\\
&\Rightarrow\ p_{3(s+\ell-1)} + p_{3(s+\ell-1)+1} \ge 3p_{3(s+\ell-1)+2} = 3p_{3(s+\ell-2)+2}\\
&\Rightarrow\ p_{3(s+\ell-2)} + p_{3(s+\ell-2)+1} - p_{3(s+\ell-2)+2} \ge 3p_{3(s+\ell-2)+2}\\
&\Rightarrow\ p_{3(s+\ell-2)} + p_{3(s+\ell-2)+1} \ge 4p_{3(s+\ell-2)+2} = 4p_{3(s+\ell-3)+2}\\
&\Rightarrow\ p_{3(s+\ell-3)} + p_{3(s+\ell-3)+1} - p_{3(s+\ell-3)+2} \ge 4p_{3(s+\ell-3)+2}\\
&\Rightarrow\ p_{3(s+\ell-3)} + p_{3(s+\ell-3)+1} \ge 5p_{3(s+\ell-3)+2} = 5p_{3(s+\ell-4)+2}\\
&\Rightarrow\ \cdots\\
&\Rightarrow\ p_{3(s+1)} + p_{3(s+1)+1} \ge (\ell+1)p_{3s+2} \ge \ell + 1 \ge p_0 + p_1 + 1
\end{aligned}
$$

However, we also have $p_{3(s+1)} + p_{3(s+1)+1} \le p_0 + p_1$, which causes a contradiction. Therefore, p_{3k+2} cannot always remain the same value and it must decrease at some k. Hence, there must exist $\hat{k}$ such that $p_{3\hat{k}+2} = 0$, i.e. $\mathcal{I}_{3\hat{k}} \cap \mathcal{I}_{3\hat{k}+1} = \emptyset$. In particular, in this case, we have

$$
\mathcal{M}_n(a+b) = \mathcal{M}_n(\alpha_3 + \alpha_4) = \cdots = \mathcal{M}_n(\alpha_{3\hat{k}} + \alpha_{3\hat{k}+1}) = \alpha_{3\hat{k}} + \alpha_{3\hat{k}+1}.
$$

As $H(\alpha_{3\hat{k}}) = p_{3\hat{k}}$, $H(\alpha_{3\hat{k}+1}) = p_{3\hat{k}+1}$, $\mathcal{I}_{3\hat{k}} \cap \mathcal{I}_{3\hat{k}+1} = \emptyset$ and $p_{3\hat{k}} + p_{3\hat{k}+1} \le p_0 + p_1$, we have $H(\mathcal{M}_n(a+b)) = p_{3\hat{k}} + p_{3\hat{k}+1} \le p_0 + p_1 = H(a) + H(b)$. □

References

[1] F. Liu, R. Anand, L. Wang, W. Meier, and T. Isobe. Coefficient Grouping: Breaking Chaghri and More. 2022. https://eprint.iacr.org/2022/991.