IEICE TRANS. INF. & SYST., VOL.E103–D, NO.5 MAY 2020

LETTER

A New Upper Bound for Finding Defective Samples in Group Testing*

Jin-Taek SEONG[†a)], ***Member***

SUMMARY The aim of this paper is to show an upper bound for finding defective samples in a group testing framework. To this end, we exploit minimization of Hamming weights in coding theory and define probability of error for our decoding scheme. We derive a new upper bound on the probability of error. We show that both upper and lower bounds coincide with each other at an optimal density ratio of a group matrix. We conclude that as defective rate increases, a group matrix should be sparser to find defective samples with only a small number of tests.

key words: *defective samples, group testing, probability of error, upper bound*

1. Introduction

Group testing was introduced by Dorfman [1]. Its use has been extended to various applications for half a century. A main example is Compressed Sensing [2] introduced in the field of information theory. Group testing gives us a chance to reconsider its broad applications in DNA screening [3], security networks [4], blood screening [5], and quality testing [6]. Recently, there has been a move to study the performance of group testing more precisely. Furthermore, for noiseless and noisy cases, nearly optimal performance has been presented [7]–[9].

Group testing began with a project to find all men with syphilis in the US Public Health Service during World War II. At that time, syphilis test was performed using blood samples from individual soldiers to diagnose syphilis infection. However, because a large number of soldiers needed the syphilis test, the cost of testing was enormous. To this end, a group test was first proposed by Dorfman [1]. The initial group testing was performed by the following method. First, blood samples from several soldiers were mixed to see if they responded to syphilis. When the result was positive, at least one soldier in the group was infected with syphilis. Conversely, if negative, it meant that none of blood samples used in the syphilis group was infected with syphilis. Such syphilis tests were possible because most soldiers were not infected with syphilis while only a few soldiers were infected with syphilis. The problem of group testing is mainly focused on two issues: 1) how to choose samples to be included in one group; and 2) which detection method should be used to find defective samples among a plurality of samples.

Most of the bounds presented in group testing problems have shown meaningful results [8]–[14]. For example, in [8], the authors proposed the combination basis pursuit and the combination orthogonal matching pursuit algorithms for the noise and noise group testing frameworks. In addition, they derived the lower and upper bounds of performance for the proposed algorithms. However, previous works lacked a study of how the relationship between the density of a group matrix and the defective rates of signals is affected. This is the motivation for this paper. Therefore, the aim of this paper is to clarify the relationship between the density of the group matrix and the defect rate of the signal. This is the part that has not been studied in other existing papers and contributes to this paper.

In this paper, we consider a group testing framework. We derive an upper bound for finding defective samples out of input samples. In order to define our decoding, we use minimization of Hamming weights known in coding theory. We also define probability of error for our decoding scheme. We obtain a new upper bound on the probability of error which well-matches a lower bound obtained from information-theoretic approach. We show that both upper and lower bounds coincide with each other at optimal density ratio of a group matrix. In addition, we conclude that as defective rate increases, a group matrix should be sparser to find defective samples with only a small number of tests.

2. Group Testing Framework

2.1 Problem Statement

In this section, group testing problem is defined in detail. First, an input signal $\mathbf{x} = (x_1, x_2, \cdots, x_N)$ is a binary vector of size N, where $\mathbf{x} \in \{0,1\}^N$. Each element of the input signal is represented by 0 or 1. Note that the input signal has at most K defective samples where its size is actually very small compared to the size of the input signal $\mathbf{x}$, that is, $K << N$. The group testing problem is to accurately find defective samples for the input signal $\mathbf{x}$. We now define signal $\mathbf{x}$. Let $\mathcal{L}_{k_1}$ be the set of signals with the number of k_1 ones in $\mathbf{x}$, then $\|\mathbf{x}\|_0 = k_1$ and $\binom{N}{k_1} = |\mathcal{L}_{k_1}|$, where $\|\cdot\|_0$ is the Hamming weight and $|\cdot|$ is the cardinality of the set. We can define the set $\mathcal{L}$ of input signals as $\mathcal{L} = \bigcup_{k_1=1}^{K} \mathcal{L}_{k_1}$. Its size

Manuscript received October 16, 2019.
Manuscript revised December 31, 2019.
Manuscript publicized February 17, 2020.

[†]The author is with the Department of Convergence Software, Mokpo National University, Republic of Korea.

*This paper was supported by the National Research Foundation of Korea (NRF) grant funded by the korean government (NRF-2017R1C1B5075823).

a) E-mail: jtseong@mokpo.ac.kr

DOI: 10.1587/transinf.2019EDL8187

is $|\mathcal{L}| = \sum_{k_1=1}^{K} \binom{N}{k_1}$. In this paper, an input signal $\mathbf{x}$ is chosen randomly and uniformly from the set $\mathcal{L}$.

Let $\mathbf{A}$ be a group matrix with M rows and N columns where each row of its matrix refers to a set of elements of signal $\mathbf{x}$. These elements are then subjected to one group testing. In other words, if the ith element x_i of the input signal is included in the jth group and the group testing is performed, the corresponding element of the group matrix is represented as $A_{ji} = 1$. Otherwise, if the ith element of the signal $\mathbf{x}$ is excluded in the jth group testing, the element of the group matrix is expressed as $A_{ji} = 0$. In this paper, we assume that each element of group matrix $\mathbf{A}$ has the following probability distribution with identically independent distribution (i.i.d.),

$$\Pr\left(A_{ji} = \alpha\right) = \begin{cases} 1-\gamma & \text{if } \alpha = 0, \\ \gamma & \text{if } \alpha = 1, \end{cases} \tag{1}$$

where γ is the density ratio of the group matrix. If the density ratio is large, it means that the probability that the element of the group matrix has 1 is high. That is, most elements of vector $\mathbf{x}$ is pooled in each test. Note that collecting many elements from signal $\mathbf{x}$ is undesired and costly. In order to perform efficient group testing, density of a group matrix may need to be small.

Next, we explain how results of group testing are related to a group matrix and an input signal. First, to help readers understand clearly and definitely, we can express the relation between them in formulas as follows,

$$\mathbf{y} = \mathbf{A} \odot \mathbf{x}, \tag{2}$$

where $\mathbf{y} \in \{0,1\}^M$ is a vector of the testing result with size M and symbol $\odot$ denotes *element-wise logical* operation. Equation (2) is explained by showing the following simple example. Let the input signal $\mathbf{x}$ be $\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$ and the group matrix $\mathbf{A}$ be $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$. Given $\mathbf{x}$ and $\mathbf{A}$, we can obtain the vector $\mathbf{y}$ as $\begin{bmatrix} 0 & 1 \end{bmatrix}^T$. The reason why the first element of $\mathbf{y}$ becomes 0 is as follows. Both the first row of $\mathbf{A}$ and $\mathbf{x}$ are performed by element-wise logical operation,

$$\begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = (0 \wedge 1) \vee (1 \wedge 0) \vee (1 \wedge 0) = 0, \tag{3}$$

where two symbols $\wedge$ and $\vee$ denote AND and OR logical operations, respectively. By applying the same manner, the second element of $\mathbf{y}$ becomes 1 from the fact that: $(1 \wedge 1) \vee (1 \wedge 0) \vee (0 \wedge 0) = 1$. The result for group testing with one or more defective samples participating as shown in the example above is positive, i.e., $y_2 = 1$. Conversely, if all elements participating in the group testing are negative, the corresponding result is negative.

The aim of group testing is to find an unknown signal $\mathbf{x}$ from a group matrix $\mathbf{A}$ and the corresponding result vector $\mathbf{y}$. So far in past works, the main research direction of group testing problems is how many tests M can be done to successfully find defective samples of the input signal $\mathbf{x}$. Next, we define the probability of error on successful decoding to derive an upper bound of the performance.

2.2 Definition for Probability of Error

In this section, we define probability of error for finding defective samples of $\mathbf{x}$ in a group testing framework for given parameters, i.e., N, K, and M. Before defining the probability of error, we classify the input signal $\mathbf{x}$ as a set of signals according to the number of ones in $\mathbf{x}$.

We assume that decoding in our framework is to find a feasible solution $\hat{\mathbf{z}}$ using minimization of Hamming weights as follows:

$$\hat{\mathbf{z}} = \arg\min \|\mathbf{z}\|_0 \qquad \text{subject to } \mathbf{A} \odot \mathbf{z} = \mathbf{y}, \tag{4}$$

where $\mathbf{z} \in \mathcal{L}$ is a feasible signal. Let k_2 be the number of ones in $\mathbf{z}$ as $k_2 = \|\mathbf{z}\|_0$, so that $k_2 \le k_1$. We define the error as occurring when a feasible solution $\hat{\mathbf{z}}$ found by (4) is not equal to the input signal $\mathbf{x}$ which is desired, $\mathbf{y} = \mathbf{A} \odot \hat{\mathbf{z}}$ but $\mathbf{x} \neq \hat{\mathbf{z}}$. Let $\mathcal{E}_0(\mathbf{x}) \triangleq \{\mathbf{A} : \mathbf{x} \neq \hat{\mathbf{z}}\}$ be the exact error event of this decoder as a function of the group matrix $\mathbf{A}$. This error event $\mathcal{E}_0$ is a subset of the following feasible error event $\mathcal{E}$ since a feasible signal $\mathbf{z}$ is a potential candidate of decoded signals. We define the feasible error event $\mathcal{E}$ as follows,

$$\mathcal{E}(\mathbf{x}, \mathbf{z}) \triangleq \{\mathbf{A} : \mathbf{x} \neq \mathbf{z}, \mathbf{y} = \mathbf{A} \odot \mathbf{z}\}. \tag{5}$$

Note that $\mathcal{E}_0(\mathbf{x}) \subseteq \mathcal{E}(\mathbf{x}, \mathbf{z})$. Let $\Pr(\mathcal{E}_0)$ and $\Pr(\mathcal{E})$ be the probability of error for both events $\mathcal{E}_0(\mathbf{x})$ and $\mathcal{E}(\mathbf{x}, \mathbf{z})$, respectively. The following inequality is then satisfied as $\Pr(\mathcal{E}_0) \le \Pr(\mathcal{E})$. The probability of error $\Pr(\mathcal{E}_0)$ is upper bounded by

$$\begin{aligned} \Pr(\mathcal{E}_0) &\le \Pr(\mathcal{E}) \\ &= \frac{1}{|\mathcal{L}|} \sum_{\mathbf{x} \in \mathcal{L}} \sum_{\mathbf{z} \in \mathcal{L}, \mathbf{z} \neq \mathbf{x}} \Pr\left(\mathbf{A} \odot \mathbf{x} = \mathbf{A} \odot \mathbf{z} \mid (\mathbf{x}, \mathbf{z})\right). \end{aligned} \tag{6}$$

It is noteworthy that Eq. (6) is almost intractable to evaluate since $|\mathcal{L}|$ is typically very large. This brute-force approach can be avoided with what will be described next.

3. Upper Bound on Performance

Let us recall the upper bound on probability of error we defined in (6). Now we aim to drive (6) more concisely. The basic idea can be thought of as follows. We first think of the same error pattern. The next step is to find the probability for that error pattern. Finally, we can obtain total probability by adding all individual probabilities with the same error pattern.

To find the same error pattern, consider that two probabilities are the same, i.e., $\Pr\left(\mathbf{A}_j \odot \mathbf{x} = \mathbf{A}_j \odot \mathbf{z}_1\right) = \Pr\left(\mathbf{A}_j \odot \mathbf{x} = \mathbf{A}_j \odot \mathbf{z}_2\right)$, such that $\mathbf{z}_1 \neq \mathbf{z}_2 \in \mathcal{L}$ and $\|\mathbf{z}_1\|_0 = \|\mathbf{z}_2\|_0 = k_2$ where $\mathbf{A}_j$ is the jth row of $\mathbf{A}$. In other words, two probabilities for $\mathbf{z}_1$ and $\mathbf{z}_2$ having the same Hamming

weights are the same (further detail will be shown later). Then we can collect individual probabilities with the same Hamming weights with respect to two vectors $\mathbf{x}$ and $\mathbf{z}$. We can count the number of vectors with the same probability. Now the conditional probability in (6) with given condition of k_1 and k_2 Hamming weights for $\mathbf{x}$ and $\mathbf{z}$, can be rewritten in an independent row as follows,

$$\Pr\left(\mathbf{A}\odot\mathbf{x}=\mathbf{A}\odot\mathbf{z}\right)=\prod_{j=1}^{M}\Pr\left(\mathbf{A}_j\odot\mathbf{x}=\mathbf{A}_j\odot\mathbf{z}\right). \tag{7}$$

We therefore take probability of the first row of (7) into account and look at the probability in more detail,

$$\Pr\left(\mathbf{A}_j\odot\mathbf{x}=\mathbf{A}_j\odot\mathbf{z}\right)=\Pr\left(\bigcup_{i=1}^{N}\left(A_{ji}\wedge x_i\right)=\bigcup_{i=1}^{N}\left(A_{ji}\wedge z_i\right)\right). \tag{8}$$

Since our group testing problem uses logical operation defined in Sect. 2, we can consider an exclusive (XOR) operation on left and right sides in the equality of $\mathbf{A}_j\odot\mathbf{x}=\mathbf{A}_j\odot\mathbf{z}$, i.e., $0=0$ and $1=1$. Thus, Eq. (8) can be rewritten as follows,

$$\begin{aligned}\Pr\left(\mathbf{A}_j\odot\mathbf{x}=\mathbf{A}_j\odot\mathbf{z}\right)&=\Pr\left((\mathbf{A}_j\odot\mathbf{x})\oplus(\mathbf{A}_j\odot\mathbf{z})=0\right)\\&=\Pr\left((\mathbf{A}_j\odot\mathbf{x})=0\right)\Pr\left((\mathbf{A}_j\odot\mathbf{z})=0\right)\\&\quad+\Pr\left((\mathbf{A}_j\odot\mathbf{x})=1\right)\Pr\left((\mathbf{A}_j\odot\mathbf{z})=1\right).\end{aligned} \tag{9}$$

where symbol $\oplus$ deontes XOR operation, the first equality is from the property of XOR operation, and the second equality is due to the independent of $\mathbf{A}$, $\mathbf{x}$, and $\mathbf{z}$.

Note that two vectors $\mathbf{x}$ and $\mathbf{z}$ have k_1 and k_2 Hamming weights, respectively. The probability P_{k_1} that the sum of k_1 logical OR is 0 for $\mathbf{x}$ can be obtained as follows,

$$P_{k_1}\triangleq\Pr\left((\mathbf{A}_j\odot\mathbf{x})=0\right)=\Pr\left(\bigcup_{i=1}^{k_1}A_{ji}=0\right). \tag{10}$$

In the same manner, we obtain the probability P_{k_2} that the sum of k_2 logical OR is 0 for $\mathbf{z}$

$$P_{k_2}\triangleq\Pr\left((\mathbf{A}_j\odot\mathbf{z})=0\right)=\Pr\left(\bigcup_{i=1}^{k_2}A_{ji}=0\right). \tag{11}$$

We finally obtain the conditional probability (7) as follows,

$$\begin{aligned}\Pr\left(\mathbf{A}\odot\mathbf{x}=\mathbf{A}\odot\mathbf{z}\right)&=\prod_{j=1}^{M}\Pr\left(\mathbf{A}_j\odot\mathbf{x}=\mathbf{A}_j\odot\mathbf{z}\right)\\&=\left(P_{k_1}P_{k_2}+(1-P_{k_1})(1-P_{k_2})\right)^{M},\end{aligned} \tag{12}$$

where the probability $P_{k_1(k_2)}$ with k_1 (k_2) Hamming weights is

$$P_{k_1(k_2)}=\left(1-\gamma\right)^{k_1(k_2)}. \tag{13}$$

In summary, Eq. (6) is expressed as follows,

$$\Pr\left(\mathcal{E}_0\right)\le\frac{1}{|\mathcal{L}|}\sum_{k_1=1}^{K}\sum_{k_2=1}^{k_1}\binom{N}{k_1}\binom{N}{k_2}\Pr(\mathbf{A}\odot\mathbf{x}=\mathbf{A}\odot\mathbf{z}). \tag{14}$$

For a special case with $k_1=k_2=K$, we know exactly K defective samples in advance,

$$\begin{aligned}\Pr(\mathcal{E}_0)&\le\binom{N}{K}\Pr(\mathbf{A}\odot\mathbf{x}=\mathbf{A}\odot\mathbf{z})\\&=\binom{N}{K}\left(P_K^2+(1-P_K)^2\right)^{M}\\&\le 2^{NH_b\left(\frac{K}{N}\right)+M\log_2 P},\end{aligned} \tag{15}$$

where $H_b(\cdot)$ denotes binary entropy and $P:=P_K^2+(1-P_K)^2$. Note that since the probability P is a convex function, its minimum value can be found by the first derivative at $P_K=0.5$. And then, P has a value from 0.5 to 1, i.e., $0.5\le P\le 1$. In order for the probability of error in the left side of (15) to vanish, the following condition as $N\to\infty$ holds:

$$M>\frac{NH_b\left(\frac{K}{N}\right)}{\log_2 P^{-1}}\ge NH_b\left(\frac{K}{N}\right). \tag{16}$$

The minimum number of tests M required is obtained when the probability P is 0.5. This result is exactly the same as the necessary condition in [15].

4. Numerical Results and Discussion

Figure 1 shows a plot of different density ratios γ of group matrices versus the number of tests M with length $N=1000$. This plot is drawn from the expression given in (14) such that for $K=50$, 70, and 100, the number of tests M is obtained based on the fact that the probability of error is less than 10^{-5}, i.e, $\Pr(\mathcal{E}_0)\le 10^{-5}$. One interesting point of this result is that there is an optimal density ratio of group matrices to obtain the minimum number of tests. In addition, our proposed upper bounds are well matched in comparison with lower bounds from the information-theoretic theory. One more fact from Fig. 1 is that as defective rate (K/N) increases, group matrix should be the sparser to successfully find defective samples with only a small number of tests. This is an important meaning. For example, if the defective rate is very low, we should design more denser group matrices. Otherwise, performance of group testing framework will fail. As shown in Fig. 1, the permissible range of density ratio depends on defective rate. In other words, when defective rate is small, a wide range of density ratio can be used. However, when defective rate is not, a narrow density ratio should be used. This feature should be considered when designing group testing frameworks.

Figure 2 compares lower bounds [15] and upper bounds (14) with $N=1000$ evaluated at probability of error of 10^{-5}. Marks for upper bounds shown in Fig. 2 are obtained by applying optimal density ratios. As shown in Fig. 2, our proposed upper bounds coincide with lower bounds obtained from the information theory.

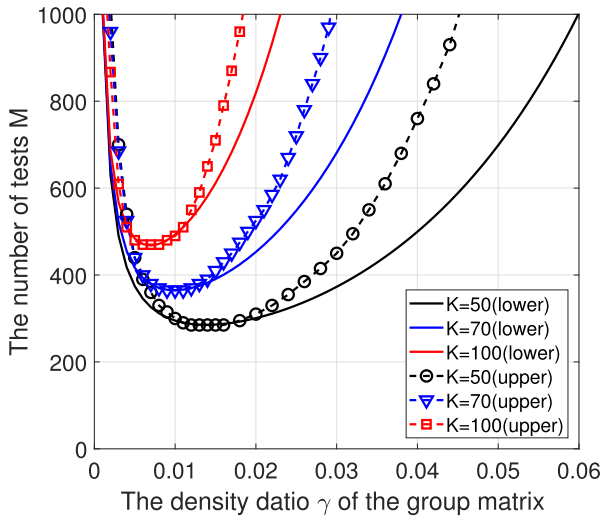

Fig. 1 Comparisons of upper (14) and lower bounds [15] (evaluated at 10^{-5}) for $N = 1000$. Solid lines indicate lower bounds and dashed lines indicate upper bounds.

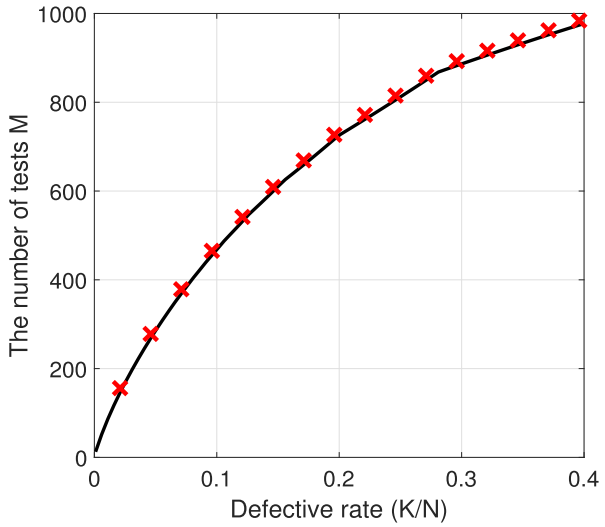

Fig. 2 Comparisons of upper (14) and lower bounds [15] (evaluated at 10^{-5}) with different defective rates for $N = 1000$.

5. Conclusion

In this paper, we considered a framework of group testing. We derived the upper bound for finding defective samples out of many samples. In order to define our decoding, we used minimization of Hamming weights known in coding theory. We defined the probability of error for our decoding scheme. We found that new upper bound on the probability of error well matched lower bounds obtained from the information-theoretic approach. We showed that upper and lower bounds coincided with each other at optimal density ratio of the group matrix. In addition, we concluded that as defective rate increased, the group matrix should be sparser to find defective samples with only a small number of tests. Our main results provide answer to an important question regarding how many tests are needed for finding defective samples in group testing frameworks.

References

[1] R. Dorfman, "The Detection of Defective Members of Large Populations," The Annals of Mathematical Statistics, vol.14, no.4, pp.436–440, Dec. 1943.
[2] D.L. Donoho, "Compressed Sensing," IEEE Trans. Inf. Theory, vol.52, no.4, pp.1289–1306, April 2006.
[3] D.-Z. Du and F.-K. Hwang, Pooling Designs and Nonadaptive Group Testing: Important Tools for DNA Sequencing, World Scientific, 2006.
[4] T. Laarhoven, "Efficient probabilistic group testing based on traitor tracing," 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, Oct. 2013.
[5] S.K. Bar-Lev, O. Boxma, I. Kleiner, D. Perry, and W. Stadje, "Recycled incomplete identification procedures for blood screening," European Journal of Operational Research, vol.259, no.1, pp.330–343, May 2017.
[6] S.K. Bar-Lev, A. Boneh, and D. Perry, "Incomplete identification models for group-testable items," Naval Research Logistics, vol.37, no.5, pp.647–659, Oct. 1990.
[7] V. Ganditota, E. Grigorescu, S. Jaggi, and S. Zhou, "Nearly Optimal Sparse Group Testing," IEEE Trans. Inf. Theory, vol.65, no.5, pp.2760–2773, May 2019.
[8] C.L. Chan, P.H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: near-optimal bounds with efficient algorithms," 49th Annual Allerton Conference on Communication, Control, and Computing, pp.1832–1839. Sept. 2011.
[9] J. Scarlett, "Noisy Adaptive Group Testing: Bounds and Algorithms," IEEE Trans. Inf. Theory, vol.65, no.6, pp.3646–3661, June 2019.
[10] G.K. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," IEEE Trans. Inf. Theory, vol.58, no.3, pp.1880–1901, March 2012.
[11] T. Wadayama, "Nonadaptive Group Testing Based on Sparse Pooling Graphs," IEEE Trans. Inf. Theory, vol.63, no.3, pp.1525–1534, March 2017.
[12] C. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri, "Non-Adaptive Group Testing: Explicit Bounds and Novel Algorithms," IEEE Trans. Inf. Theory, vol.60, no.5, pp.3019–3035, May 2014.
[13] M. Aldridge, L. Baldassini, and O. Johnson, "Group Testing Algorithms: Bounds and Simulations," IEEE Trans. Inf. Theory, vol.60, no.6, pp.3671–3687, June 2014.
[14] O. Johnson, M. Aldridge, and J. Scarlett, "Performance of group testing algorithms with near-constant tests per item," IEEE Trans. Inf. Theory, vol.65, no.2, pp.707–723, Feb. 2019.
[15] J.T. Seong, "Density of Pooling Matrices vs. Sparsity of Signals for Group Testing Problems," IEICE Trans. Inf. & Syst., vol.E102-D, no.5, pp.1081–1084, May 2019.