

The Independence of Linear Approximations in Symmetric Cryptanalysis

S. Murphy

Abstract—A theoretical framework for the use of multiple linear approximations in the linear cryptanalysis of block ciphers is given. The covariance of two mask counts is derived, and it is shown that under appropriate conditions the mask counts in linear cryptanalysis are stochastically independent, whether or not the masks are linearly independent. Some consequences of these observations are also considered.

Index Terms—block cipher, linear cryptanalysis, multiple approximations, symmetric cryptology.

I. INTRODUCTION

The technique of linear cryptanalysis [6] of block ciphers is based on “linear approximations”. For plaintext \mathbf{p} , ciphertext \mathbf{c} and key \mathbf{k} (considered as binary vectors), a linear approximation in its most basic form is an expression of the form

$$\mathbf{a}^T \begin{pmatrix} \mathbf{p} \\ \mathbf{c} \end{pmatrix} = (\mathbf{a}^\dagger)^T \mathbf{k} \text{ with probability } p,$$

where \mathbf{a} is the data (plaintext-ciphertext) mask and \mathbf{a}^\dagger is the key mask. If the probability $p \neq \frac{1}{2}$, then the linear approximation can be used to give an estimate of one bit of key information.

It is obviously natural to consider using more than one such linear approximation [3], [4], [10], though there are of course many other methods of generalising linear cryptanalysis [7], [5], [2], [11]. In particular, it has been noted for some time that using a larger (linearly) dependent set of masks can give a more powerful analysis than just using the smaller linearly independent basis of these masks [4]. Such an analysis for the DES [8] was reported at CRYPTO 2004 [1]. We give a theoretical framework for this observation and use this framework to give theoretical results about linear cryptanalysis. In particular, we show that the covariance of two mask counts is proportional to the bias of the sum of the two masks, so that counts for linearly dependent masks can be considered as stochastically independent in some fairly general circumstances likely to exist in cryptanalysis. This allows us to demonstrate that linear independence of masks and the stochastic independence of mask counts are entirely unrelated concepts, in contradiction to a seemingly widely held belief. Furthermore, we discuss some statistical issues related to estimating the key in linear cryptanalysis and show that the theoretical analysis given in [1] is incorrect.

S. Murphy is with the Information Security Group, Royal Holloway, University of London, Egham, Surrey TW20 0EX, U.K.

II. A SMALL EXAMPLE

We begin by discussing a small example in order to illustrate simply some of the issues discussed in Section I. Suppose we have two linearly independent data masks \mathbf{a}_{01}^T and \mathbf{a}_{10}^T (so $\mathbf{a}_{01} \neq \mathbf{a}_{10}$). For a given key, we can then define four random variables $\mathbf{W}_{jl}^{(i)}$ ($j, l = 0, 1$) by:

$$\begin{aligned} \mathbf{W}_{jl}^{(i)} &= 1 && \text{if } \begin{pmatrix} \mathbf{a}_{01}^T \\ \mathbf{a}_{10}^T \end{pmatrix} \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \begin{pmatrix} j \\ l \end{pmatrix}, \\ \mathbf{W}_{jl}^{(i)} &= 0 && \text{if } \begin{pmatrix} \mathbf{a}_{01}^T \\ \mathbf{a}_{10}^T \end{pmatrix} \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} \neq \begin{pmatrix} j \\ l \end{pmatrix}. \end{aligned}$$

The probabilities for these random variables are given by

$$P\left(\mathbf{W}_{jl}^{(i)} = 1\right) = \frac{1}{4}(1 + d_{jl}),$$

for some d_{jl} where $\sum_{jl} d_{jl} = 0$. Thus d_{jl} is a measure of the difference from uniform probability for the “data class” (j, l). If we let $\mathbf{W}^{(i)}$ denote the random variable given by the vector of all four $\mathbf{W}_{jl}^{(i)}$, then it is clear that $\mathbf{W}^{(i)}$ has a multinomial distribution given by

$$\mathbf{W}^{(i)} \sim \text{Mult}\left(1, \frac{1}{4}(\mathbf{1} + \mathbf{d})\right),$$

where $\mathbf{1} = (1, 1, 1, 1)^T$ and $\mathbf{d} = (d_{00}, d_{01}, d_{10}, d_{11})^T$ [9]. Suppose now that we have N plaintext-ciphertext pairs, then we can define

$$\mathbf{W} = \sum_{i=1}^N \mathbf{W}^{(i)} \sim \text{Mult}\left(N, 2^{-2}(\mathbf{1} + \mathbf{d})\right),$$

so \mathbf{W} is a random variable giving the counts for each data class (j, l).

We can now consider the linear approximations defined by the masks \mathbf{a}_{01} and \mathbf{a}_{10} . We can define the two “mask” random variables for the i^{th} plaintext-ciphertext pair by:

$$\begin{aligned} \mathbf{V}_{01}^{(i)} &= \mathbf{W}_{00}^{(i)} + \mathbf{W}_{01}^{(i)} = \begin{cases} 1 & \text{if } \mathbf{a}_{01}^T \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \begin{cases} 0 \\ 1 \end{cases} \\ 0 & \text{if } \mathbf{a}_{01}^T \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \begin{cases} 1 \\ 0 \end{cases} \end{cases} \\ \mathbf{V}_{10}^{(i)} &= \mathbf{W}_{00}^{(i)} + \mathbf{W}_{10}^{(i)} = \begin{cases} 1 & \text{if } \mathbf{a}_{10}^T \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \begin{cases} 0 \\ 1 \end{cases} \\ 0 & \text{if } \mathbf{a}_{10}^T \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \begin{cases} 1 \\ 0 \end{cases} \end{cases}. \end{aligned}$$

Thus $\mathbf{V}_{01}^{(i)}$ is 1 if mask \mathbf{a}_{01} takes the value 0 for the i^{th} plaintext-ciphertext pair and so on. Probabilities for these random variables are given by

$$\begin{aligned} P\left(\mathbf{V}_{01}^{(i)} = 1\right) &= \frac{1}{2} + \frac{1}{4}(d_{00} + d_{01}) \\ \text{and } P\left(\mathbf{V}_{10}^{(i)} = 1\right) &= \frac{1}{2} + \frac{1}{4}(d_{00} + d_{10}). \end{aligned}$$

If we define $e_{jl} = \frac{1}{2}(d_{00} + d_{jl})$ ($(j, l) \neq (0, 0)$), then these random variables are binomial random variables [9], with distributions given by

$$\begin{aligned} \mathbf{V}_{01}^{(i)} &\sim \text{Bin}\left(1, \frac{1}{2}(1 + e_{01})\right) \\ \text{and } \mathbf{V}_{10}^{(i)} &\sim \text{Bin}\left(1, \frac{1}{2}(1 + e_{10})\right). \end{aligned}$$

Thus e_{jl} is a measure of the difference from the uniform probability of the linear approximation defined by mask \mathbf{a}_{jl} , and is twice the *bias* [6] of the linear approximation, termed the *imbalance* in [1].

The expected values of $\mathbf{V}_{01}^{(i)}$ and $\mathbf{V}_{10}^{(i)}$ are $\frac{1}{2}(1 + e_{01})$ and $\frac{1}{2}(1 + e_{10})$ respectively, whilst the variances are $\frac{1}{4} - \frac{1}{4}e_{01}^2$ and $\frac{1}{4} - \frac{1}{4}e_{10}^2$ respectively. If second order terms are negligible, then the covariance of these two random variables is given by

$$\begin{aligned} \text{Cov}\left(\mathbf{V}_{01}^{(i)}, \mathbf{V}_{10}^{(i)}\right) &= E\left(\mathbf{V}_{01}^{(i)}\mathbf{V}_{10}^{(i)}\right) - E\left(\mathbf{V}_{01}^{(i)}\right)E\left(\mathbf{V}_{10}^{(i)}\right) \\ &= P\left(\mathbf{V}_{01}^{(i)} = \mathbf{V}_{10}^{(i)} = 1\right) - P\left(\mathbf{V}_{01}^{(i)} = 1\right)P_0\left(\mathbf{V}_{10}^{(i)} = 1\right) \\ &= \frac{1}{4}(1 + d_{00}) - \left(\frac{1}{2} + \frac{1}{4}(d_{00} + d_{01})\right)\left(\frac{1}{2} + \frac{1}{4}(d_{00} + d_{10})\right) \\ &= -\frac{1}{8}(d_{01} + d_{10}) = \frac{1}{8}(d_{00} + d_{11}) = \frac{1}{4}e_{11}. \end{aligned}$$

Thus the covariance of $\mathbf{V}_{01}^{(i)}$ and $\mathbf{V}_{10}^{(i)}$ is zero if and only if $e_{11} = 0$. We note that for this joint distribution of two random variables, zero covariance implies (pairwise) stochastic independence. Thus the ‘‘proximity’’ to stochastic independence depends only on the size of the imbalance e_{11} of the linear approximation. The fact that the mask values arose from linearly independent masks is clearly not relevant to the independence of the mask values.

Suppose now that we have N plaintext-ciphertext pairs. We can then define the random variables \mathbf{V}_{01} and \mathbf{V}_{10} by

$$\begin{aligned} \mathbf{V}_{01} &= \sum_{i=1}^N \mathbf{V}_{01}^{(i)} \sim \text{Bin}\left(N, \frac{1}{2}(1 + e_{01})\right) \\ \mathbf{V}_{10} &= \sum_{i=1}^N \mathbf{V}_{10}^{(i)} \sim \text{Bin}\left(N, \frac{1}{2}(1 + e_{10})\right), \end{aligned}$$

so \mathbf{V}_{01} and \mathbf{V}_{10} are the mask counts for \mathbf{a}_{01} and \mathbf{a}_{10} respectively. We can use central limit theorem ideas with the above covariance result to show that, for large N and negligible second order terms, the joint distribution of \mathbf{V}_{01} and \mathbf{V}_{10} is a bivariate normal distribution [9]. Thus

$$\frac{2}{\sqrt{N}} \begin{pmatrix} \mathbf{V}_{01} - \frac{N}{2}(1 + e_{01}) \\ \mathbf{V}_{10} - \frac{N}{2}(1 + e_{10}) \end{pmatrix}$$

has the bivariate normal distribution

$$\mathbf{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & e_{11} \\ e_{11} & 1 \end{pmatrix}\right).$$

For normal random variables, zero covariance (diagonal covariance matrix) implies stochastic independence. Thus if e_{11} is small, then it may be possible to regard the mask counts \mathbf{V}_{01} and \mathbf{V}_{10} as ‘‘approximately’’ stochastically independent. Conversely, if e_{11} are large, then the mask counts are not stochastically independent.

Suppose now we define a third mask $\mathbf{a}_{11} = \mathbf{a}_{01} + \mathbf{a}_{10}$. Clearly the three masks are not linearly independent. We can define a random variable

$$\mathbf{V}_{11}^{(i)} = \mathbf{W}_{00}^{(i)} + \mathbf{W}_{11}^{(i)} = \begin{cases} 1 & \text{if } \mathbf{a}_{11}^T \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \begin{cases} 0 \\ 1 \end{cases}, \end{cases}$$

which has a $\text{Bin}\left(1, \frac{1}{2}(1 + e_{11})\right)$ distribution. Thus the expected value of $\mathbf{V}_{11}^{(i)}$ is $\frac{1}{2}(1 + e_{11})$, whilst the variance is $\frac{1}{4} - \frac{1}{4}e_{11}^2$. Furthermore, the covariance of $\mathbf{V}_{11}^{(i)}$ and $\mathbf{V}_{01}^{(i)}$ is $\frac{1}{4}e_{10}$, whilst the covariance of $\mathbf{V}_{11}^{(i)}$ and $\mathbf{V}_{10}^{(i)}$ is $\frac{1}{4}e_{01}$. Thus if e_{10} is small, it may be possible to regard $\mathbf{V}_{11}^{(i)}$ and $\mathbf{V}_{01}^{(i)}$ as ‘‘almost’’ (pairwise) stochastically independent. Similarly if e_{11} is small, it may be possible to regard $\mathbf{V}_{11}^{(i)}$ and $\mathbf{V}_{10}^{(i)}$ as ‘‘almost’’ stochastically independent. However, the linear (algebraic) dependence of the three masks means that the three random variables $\mathbf{V}_{01}^{(i)}$, $\mathbf{V}_{10}^{(i)}$ and $\mathbf{V}_{11}^{(i)}$ cannot be mutually independent. Knowledge of two of the three random variables determines $\mathbf{W}^{(i)}$ and so determines the third. To illustrate this point, we give an example that we return to in Section VI. For simplicity and without loss of generality suppose that all the imbalances $e_{jl} = 0$, so in this case the three random variables $\mathbf{V}_{01}^{(i)}$, $\mathbf{V}_{10}^{(i)}$ and $\mathbf{V}_{11}^{(i)}$ are pairwise independent. However, it is clear that $\mathbf{V}_{01}^{(i)}$, $\mathbf{V}_{10}^{(i)}$ and $\mathbf{V}_{11}^{(i)}$ cannot all simultaneously be 0, so

$$\begin{aligned} 0 &= P\left(\mathbf{V}_{01}^{(i)} = \mathbf{V}_{10}^{(i)} = \mathbf{V}_{11}^{(i)} = 0\right) \\ &\neq P\left(\mathbf{V}_{01}^{(i)} = 0\right)P\left(\mathbf{V}_{10}^{(i)} = 0\right)P\left(\mathbf{V}_{11}^{(i)} = 0\right) = \frac{1}{8}. \end{aligned}$$

We now consider the mask count for the mask \mathbf{a}_{11} over N plaintext-ciphertext pairs. This is given by the random variable \mathbf{V}_{11} , where

$$\mathbf{V}_{11} = \sum_{i=1}^N \mathbf{V}_{11}^{(i)} \sim \text{Bin}\left(N, \frac{1}{2}(1 + e_{11})\right).$$

Again we can use central limit theorem ideas to show that, for large N and second order terms negligible, so the joint distribution of

$$\frac{2}{\sqrt{N}} \begin{pmatrix} \mathbf{V}_{01} - \frac{N}{2}(1 + e_{01}) \\ \mathbf{V}_{10} - \frac{N}{2}(1 + e_{10}) \\ \mathbf{V}_{11} - \frac{N}{2}(1 + e_{11}) \end{pmatrix}$$

is given by the multivariate normal distribution

$$\mathbf{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \begin{pmatrix} 1 & e_{11} & e_{10} \\ e_{11} & 1 & e_{01} \\ e_{10} & e_{01} & 1 \end{pmatrix}\right).$$

Again it can be seen that if the e_{jl} are small, then all three counts can be regarded as approximately mutually stochastically independent. However, we have seen that stochastic independence does not hold for counts for a single plaintext-ciphertext pair, so such stochastic independence is fundamentally a large sample property. This example shows that whilst the three masks (\mathbf{a}_{01} , \mathbf{a}_{10} , $\mathbf{a}_{11} = \mathbf{a}_{01} + \mathbf{a}_{10}$) are linearly (algebraically) dependent, the corresponding mask counts can be (asymptotically) stochastically independent.

III. INDEPENDENCE OF MASK COUNTS

We generalise and abstract the above example and consider m masks spanning an l -dimensional subspace in this section, in order to discuss the stochastic independence of mask counts. This discussion of multiple linear approximations is given in terms of the underlying data classes, building on a theoretical approach given in [7].

Suppose now that we have m plaintext-ciphertext masks $\mathbf{a}_1^T, \dots, \mathbf{a}_m^T$, and suppose that these mask vectors span a space of dimension $l \leq m$. Without loss of generality, we may suppose that the first l masks $\mathbf{a}_1^T, \dots, \mathbf{a}_l^T$ are linearly independent, so any mask is a linear combination of the first l . Thus any mask can be defined in terms of the full-rank matrix $A = (\mathbf{a}_1 \dots \mathbf{a}_l)^T$. For any mask \mathbf{a} , there exists a ‘‘mask selection’’ vector $\mathbf{r}_\mathbf{a}$ of length l such that the mask \mathbf{a}^T can be expressed as $\mathbf{r}_\mathbf{a}^T A$, so the first l such vectors $\mathbf{r}_{\mathbf{a}_1}, \dots, \mathbf{r}_{\mathbf{a}_l}$ are the l standard basis vectors. We can define an $m \times l$ matrix R to be the matrix with rows given by $\mathbf{r}_\mathbf{a}^T$, so $R^T = (I|Q^T)$ for some $(m-l) \times l$ matrix Q . Thus R is a mask selection matrix and the entire set of masks is then given by the rows of RA .

Suppose that we have a plaintext-ciphertext pair $(\mathbf{p}_i, \mathbf{c}_i)$. All mask values for this plaintext-ciphertext pair are determined by the matrix A . Accordingly, we define the data class $\mathbf{X}^{(i)}$ for the i^{th} plaintext-ciphertext pair by

$$\mathbf{X}^{(i)} = A \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix},$$

a vector of length l , so there are 2^l data classes. We can define 2^l random variables $\mathbf{W}_\mathbf{x}^{(i)}$ ($\mathbf{x} \in \mathbb{Z}_2^l$) based on the data class for the i^{th} plaintext-ciphertext pair by

$$\mathbf{W}_\mathbf{x}^{(i)} = \begin{cases} 1 & \mathbf{X}^{(i)} = \mathbf{x} \\ 0 & \mathbf{X}^{(i)} \neq \mathbf{x} \end{cases}.$$

Furthermore, suppose we express the probability that data class \mathbf{x} occurs by

$$P(\mathbf{W}_\mathbf{x}^{(i)} = 1) = 2^{-l} (1 + d_\mathbf{x}),$$

where $\sum_\mathbf{x} d_\mathbf{x} = 0$, then $d_\mathbf{x}$ gives the difference from uniform of the probability of occurrence of data class \mathbf{x} . We now let $\mathbf{W}^{(i)}$ denote the random variable given by the vector of length 2^l of all such $\mathbf{W}_\mathbf{x}^{(i)}$ and \mathbf{d} denote the vector given by the $d_\mathbf{x}$ with $\mathbf{1}$ the vector with every entry 1. It is clear that $\mathbf{W}^{(i)}$ has a multinomial distribution so

$$\mathbf{W}^{(i)} \sim \text{Mult} (1, 2^{-l}(\mathbf{1} + \mathbf{d})).$$

Suppose now that we have N plaintext-ciphertext pairs, then we can define

$$\mathbf{W} = \sum_{i=1}^N \mathbf{W}^{(i)} \sim \text{Mult} (N, 2^{-l}(\mathbf{1} + \mathbf{d})),$$

so \mathbf{W} is a random variable giving the counts for each data class \mathbf{x} .

Having considered the distribution for the underlying data classes, we now turn to the distribution for the masks. Any mask \mathbf{a} defines a hyperplane $H_{\mathbf{r}_\mathbf{a}} = \{\mathbf{u} \in \mathbb{Z}_2^l | \mathbf{r}_\mathbf{a}^T \mathbf{u} = 0\}$, that is a plane of dimension $l-1$ in \mathbb{Z}_2^l . We can define the vector $\mathbf{t}_{\mathbf{r}_\mathbf{a}}$ of length 2^l to be the indicator vector for the hyperplane $H_{\mathbf{r}_\mathbf{a}}$. Thus half the entries of $\mathbf{t}_{\mathbf{r}_\mathbf{a}}$ are 1 and half 0, so $\mathbf{t}_{\mathbf{r}_\mathbf{a}}^T \mathbf{1} = 2^{l-1}$ and we have

$$\mathbf{t}_{\mathbf{r}_\mathbf{a}}^T \mathbf{t}_{\mathbf{r}_{\mathbf{a}'}} = \left| H_{\mathbf{r}_\mathbf{a}} \cap H_{\mathbf{r}_{\mathbf{a}'}} \right| = \begin{cases} 2^{l-1} & \mathbf{r}_\mathbf{a} = \mathbf{r}_{\mathbf{a}'} \\ 2^{l-2} & \mathbf{r}_\mathbf{a} \neq \mathbf{r}_{\mathbf{a}'} \end{cases}.$$

We define T to be the $m \times 2^l$ matrix with rows $\mathbf{t}_{\mathbf{r}_\mathbf{a}}$, so T is a partial Walsh-Hadamard transform matrix. For mask \mathbf{a} , we can define the random variable

$$\begin{aligned} \mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)} &= \mathbf{t}_{\mathbf{r}_\mathbf{a}}^T \mathbf{W}^{(i)} = \mathbf{1} + \mathbf{a}^T \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} \\ &= \mathbf{1} + \mathbf{r}_\mathbf{a}^T A \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \mathbf{1} + \mathbf{r}_\mathbf{a}^T \mathbf{X}^{(i)}. \end{aligned}$$

Thus $\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)}$ is an individual mask count, which takes the value 1 if the mask value defined by \mathbf{a} for the i^{th} plaintext-ciphertext pair is 0 and vice-versa. We define $\mathbf{V}^{(i)}$ to be the vector of length m with components are $\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)}$, so

$$\mathbf{V}^{(i)} = \mathbf{1} + RA \begin{pmatrix} \mathbf{p}_i \\ \mathbf{c}_i \end{pmatrix} = \mathbf{1} + R\mathbf{X}^{(i)}.$$

If we define $e_{\mathbf{r}_\mathbf{a}} = 2^{-(l-1)} \mathbf{t}_{\mathbf{r}_\mathbf{a}}^T \mathbf{d}$, then the probability for $\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)}$ is given by

$$\begin{aligned} P(\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)} = 1) &= P(\mathbf{r}_\mathbf{a}^T \mathbf{X}^{(i)} = 0) = P(\mathbf{X}^{(i)} \in H_{\mathbf{r}_\mathbf{a}}) \\ &= \mathbf{t}_{\mathbf{r}_\mathbf{a}}^T (2^{-l}(\mathbf{1} + \mathbf{d})) \\ &= 2^{-l} \mathbf{t}_{\mathbf{r}_\mathbf{a}}^T \mathbf{1} + 2^{-l} \mathbf{t}_{\mathbf{r}_\mathbf{a}}^T \mathbf{d} = \frac{1}{2} (1 + e_{\mathbf{r}_\mathbf{a}}). \end{aligned}$$

Thus $e_{\mathbf{r}_\mathbf{a}}$ is the imbalance or twice the bias of the linear approximation and $\mathbf{e} = 2^{-(l-1)} \mathbf{d}$ is a vector of imbalances. The distribution of $\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)}$ is thus given by

$$\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)} \sim \text{Bin} (1, \frac{1}{2}(1 + e_{\mathbf{r}_\mathbf{a}})),$$

so the expected value of $\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)}$ is $\frac{1}{2}(1 + e_{\mathbf{r}_\mathbf{a}})$, whilst the variance of $\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)}$ is $\frac{1}{4} - \frac{1}{4}e_{\mathbf{r}_\mathbf{a}}^2$. If second order terms are negligible, then the covariance of two individual mask counts is given in the theorem below, with details of the proof given in Appendix I.

Theorem 1. *The covariance of two individual mask counts $\mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)}$ and $\mathbf{V}_{\mathbf{r}_{\mathbf{a}'}}^{(i)}$ is $\frac{1}{4}e_{\mathbf{r}_{\mathbf{a}+\mathbf{a}'}}$, a quarter of the imbalance (half the bias) of the sum of the two masks.*

If we have N plaintext-ciphertext pairs, then for mask \mathbf{a} we can define a random variable $\mathbf{V}_{\mathbf{r}_\mathbf{a}}$ by

$$\mathbf{V}_{\mathbf{r}_\mathbf{a}} = \sum_{i=1}^N \mathbf{V}_{\mathbf{r}_\mathbf{a}}^{(i)} = \mathbf{t}_{\mathbf{r}_\mathbf{a}}^T \mathbf{W} \sim \text{Bin} (N, \frac{1}{2} (1 + e_{\mathbf{r}_\mathbf{a}})),$$

which gives the count for mask \mathbf{a} ($\mathbf{a} \neq \mathbf{0}$), with the covariance of two such random variables given by $\text{Cov}(\mathbf{V}_{\mathbf{r}_\mathbf{a}}, \mathbf{V}_{\mathbf{r}_{\mathbf{a}'}}) = \frac{N}{4}e_{\mathbf{r}_{\mathbf{a}+\mathbf{a}'}}$ if second order terms are negligible. Again, we can use central limit ideas to derive the asymptotic distribution for large N . If we let T denote the $m \times 2^l$ matrix given by the m rows $\mathbf{t}_{\mathbf{r}_\mathbf{a}}$, then $\mathbf{V} = T\mathbf{W} = \sum_i \mathbf{V}^{(i)}$ is the m -dimensional random variable with entries $\mathbf{V}_{\mathbf{r}_\mathbf{a}}$ giving the joint counts for the m masks. If we let Δ denote the $m \times m$ matrix with zero diagonal and $(\mathbf{a}, \mathbf{a}')$ -entry ($\mathbf{a} \neq \mathbf{a}'$) given by $e_{\mathbf{r}_{\mathbf{a}+\mathbf{a}'}}$, then the asymptotic distribution of \mathbf{V} is an m -dimensional multivariate normal distribution given by

$$\frac{2}{\sqrt{N}} (\mathbf{V} - \frac{N}{2} (\mathbf{1} + \mathbf{e})) \sim \text{N} (\mathbf{0}; I + \Delta).$$

A common situation in linear cryptanalysis is that the imbalances $e_{\mathbf{r}_\mathbf{a}}$ are very small, so we may disregard the matrix Δ as being negligible when compared with I . In this situation, the joint distribution of the mask counts \mathbf{V} is given by

$$\frac{2}{\sqrt{N}} (\mathbf{V} - \frac{N}{2} (\mathbf{1} + \mathbf{e})) \sim \text{N} (\mathbf{0}; I).$$

The obvious striking feature of this distribution is that the covariance matrix of the mask counts \mathbf{V} is proportional to the identity matrix. Thus the components of \mathbf{V} , namely the individual mask counts, are independent. This result is summarised in the following theorem.

Theorem 2. *Consider a collection of m linear approximation masks that form a subspace of dimension l . If all $(2^l - 1)$ imbalances are small and the number of plaintext-ciphertext pairs is large, then the m different mask counts are (approximately) stochastically independent.*

We have also demonstrated that the only factors in determining the stochastic independence of mask counts are the number of plaintext-ciphertext pairs and the size of the imbalances. Thus we have the following corollary.

Corollary 1. *The stochastic independence of mask counts and the linear (algebraic) independence of the masks are entirely unrelated concepts.*

IV. GOODNESS-OF-FIT TECHNIQUES

We now consider how the methodology of the well-known χ^2 goodness-of-fit test for the multinomial distribution can be applied within this cryptographic framework, and our discussion closely follows the standard justification for this test [9]. This enables us firstly to give a more formal derivation of the above distributional result and secondly to consider some consequences for linear cryptanalysis. Some related discussion of the use of the χ^2 distribution in linear cryptanalysis has also been given in [10].

We begin this discussion by more clearly stating the assumptions underlying linear cryptanalysis. Each key has its own associated vector \mathbf{d} of differences from uniform probability and hence its own associated random variable \mathbf{W} of data class counts. However, in linear cryptanalysis (under certain assumptions) many keys have (approximately) the same associated vector \mathbf{d} (and hence \mathbf{W}). Thus we can partition the key space into key classes such that for all keys \mathbf{k} in key class \mathbf{z} have the same associated vector of differences \mathbf{d} and the same associated random variable \mathbf{W} of data class counts. We therefore have a number of possible different multinomial distributions, one for each key class. We can use statistical goodness-of-fit techniques to find which of these possible distributions is the best fit to the observed data of mask counts. The best candidate for the true key class, that is the key class containing the true key, is the key class corresponding to this distribution. Linear cryptanalysis is a technique for estimating the true key class by estimating which of the possible multinomial distributions gives the best fit to the data. As we use standard statistical techniques to estimate the true key class, we use standard statistical notation. Thus \mathbf{z}^* denotes the true key class, $\hat{\mathbf{z}}$ denotes the best (maximum likelihood) estimate of the key class and $\hat{\mathbf{z}}$ a generic key class under consideration. Similarly, \mathbf{d}^* , $\hat{\mathbf{d}}$ and \mathbf{d} respectively denote the corresponding true vector of differences \mathbf{d} , the best estimate of \mathbf{d} and a generic vector \mathbf{d} under consideration.

Suppose now that \mathbf{w} is a vector of data class counts, that is a realisation of the multinomial random variable \mathbf{W} . Suppose we wished to test whether the multinomial probabilities $2^{-l}(\mathbf{1} + \mathbf{d}^*)$ are equal to some specified proportions

$2^{-l}(\mathbf{1} + \mathbf{d})$. This is of course equivalent to testing whether $\hat{\mathbf{d}} = \mathbf{d}^*$. The χ^2 goodness-of-fit test for multinomial distributions is the well-known and standard technique. This test depends on the maximum likelihood estimate of \mathbf{d} , which is given by

$$\hat{\mathbf{d}} = (N2^{-l})^{-1} \mathbf{w} - \mathbf{1}.$$

The standard derivation of distribution of this test statistic depends on the distribution of the random variable $\hat{\mathbf{d}} - \mathbf{d}$. However, the mask count distributions also depend simply on this random variable, so we can use the standard derivation of the χ^2 goodness-of-fit test to derive the mask count distributions. The mask counts are given by the random variable $\mathbf{V} = T\mathbf{W}$. A vector \mathbf{v} of mask counts is a realisation of this random variable \mathbf{V} and can be expressed as $\mathbf{v} = T\mathbf{w}$, where \mathbf{w} is a realisation of counts for the underlying data classes. We recall that $\mathbf{e} = 2^{-(l-1)}T\mathbf{d}$ is a vector of mask imbalances, so $2\mathbf{e}$ a vector of mask biases. The estimate of \mathbf{e} is then given by $\hat{\mathbf{e}} = \frac{2}{N}\mathbf{v} - \mathbf{1}$ (see Appendix II), so the components of $\hat{\mathbf{e}}$ are proportional to the difference of the mask counts \mathbf{v} from the uniform value $\frac{N}{2}\mathbf{1}$. The random variable $\hat{\mathbf{e}} - \mathbf{e}$ is given by $2^{-(l-1)}T(\hat{\mathbf{d}} - \mathbf{d})$ and is a measure of the difference of the imbalance under consideration $\hat{\mathbf{e}}$ and the estimated imbalance \mathbf{e} . The full derivation of the distribution of this random variable $\hat{\mathbf{e}} - \mathbf{e}$ is given in Appendix II. We now use these results to discuss how to test whether $\mathbf{e} = \mathbf{e}^*$ in the case

The hypothesis that $\hat{\mathbf{d}} = \mathbf{d}^*$ implies that $\hat{\mathbf{e}} = \mathbf{e}^*$. For small imbalances, the random variable $\hat{\mathbf{e}} - \mathbf{e}$ is given under this hypothesis by

$$\sqrt{N} \left(I + \frac{1}{2}\Delta \right)^{-1} (\hat{\mathbf{e}} - \mathbf{e}) \sim \mathbf{Y}_m,$$

where $\mathbf{Y}_m \sim N(0, I)$ is an m -vector of independent standard normal ($N(0, 1)$) random variables and Δ is the off-diagonal imbalance covariance matrix of Section III. Thus the associated quadratic form

$$Q = N (\hat{\mathbf{e}} - \mathbf{e})^T (I + \Delta)^{-1} (\hat{\mathbf{e}} - \mathbf{e})$$

has a χ^2 distribution with m degrees of freedom. Under an alternative hypothesis that $\hat{\mathbf{e}} \neq \mathbf{e}^*$, the distribution of

$$\sqrt{N} (I + \Delta)^{-1} (\hat{\mathbf{e}} - \mathbf{e})$$

is given by the normal distribution

$$N \left(\sqrt{N} \left(I + \frac{1}{2}\Delta \right)^{-1} (\mathbf{e}^* - \mathbf{e}); I \right),$$

which gives the distribution of $\hat{\mathbf{e}} - \mathbf{e}$. Furthermore, the associated quadratic form Q has a non-central χ^2 distribution [9] with m degrees of freedom and non-centrality parameter

$$N (\mathbf{e}^* - \mathbf{e})^T (I + \frac{1}{2}\Delta)^{-1} (\mathbf{e}^* - \mathbf{e}).$$

Thus the hypothesis that $\hat{\mathbf{e}} = \mathbf{e}^*$ can be tested by comparing empirical values of the above quadratic form Q with a χ^2 distribution with m degrees of freedom. The power of such a test is given by m (degrees of freedom) and the non-centrality parameter. Such a test and the related distributions are given directly by the theory of large sample likelihood ratio tests [9].

We noted in Section III that in many cryptanalytic situations the matrix Δ is negligible. In such situations, the distribution of $\hat{\mathbf{e}} - \dot{\mathbf{e}}$ is given by

$$\sqrt{N}(\hat{\mathbf{e}} - \dot{\mathbf{e}}) \sim \mathbf{Y}_m.$$

This distribution also directly gives the stochastic independence of mask counts noted in Section III. Furthermore, in such situations the quadratic form Q is given by $Q = N|\hat{\mathbf{e}} - \dot{\mathbf{e}}|^2$ with the non-centrality parameter of the corresponding χ^2 distribution given by $N|\mathbf{e}^* - \dot{\mathbf{e}}|^2$.

We conclude this section with a remark about the number N of plaintext-ciphertext pairs. We clearly require that N be large enough so that above distributional results are applicable. Thus we require each underlying data class to occur reasonably often. However, the data classes are very roughly equiprobable (small d_x) and a comparison with the related χ^2 goodness-of-fit statistic would suggest that $N > 10 \times 2^l$.

V. KEY CLASS ESTIMATES

We recall that we have partitioned the key space into key classes, and the aim of the analysis is to estimate the true key class \mathbf{z}^* , that is the key class containing the true key. We now discuss how to use the distributional information of Section IV to estimate the true key class \mathbf{z}^* . Without loss of generality, we assume that the matrix Δ is negligible. It is in any case a simple matter to include it in the following discussion.

For each key class \mathbf{z} , there is a corresponding 2^l -vector \mathbf{e} of differences from uniform probability (2^{-l}) of the linear approximations for that key classes. With a slight abuse of notation, we let Z denote the set of all such key classes \mathbf{z} or equivalently the set of all corresponding vectors \mathbf{e} where appropriate. The best (maximum likelihood) estimate $\hat{\mathbf{z}}$ for the true key class is the key class corresponding to the best estimate of the vector \mathbf{e} within Z . The argument of the previous section demonstrates that, for small \mathbf{e} , this is given by

$$\min_{\hat{\mathbf{e}} \in Z} N|\hat{\mathbf{e}} - \dot{\mathbf{e}}|^2.$$

This quadratic minimisation for the best estimate of the key class can also be derived directly from the likelihood function for the mask counts under the assumption of the independence of mask counts [1].

Previous research has given measures for the effectiveness of a collection of linear approximations [3], [4], [10], [1]. These are usually related to what is termed the *capacity* in [1] of a collection of linear approximations. Under certain assumptions, the imbalance of the j^{th} mask can be given as $\pm c_j$, where the sign depends only on the key class. In this case, $e_j^* - \dot{e}_j = 0, \pm 2c_j$. The *capacity* [1] of such a collection of linear approximations is defined by $\sum_j c_j^2$. However, we have shown that the above quadratic form has a χ^2 distribution with m degrees of freedom and non-centrality parameter $\rho_{\dot{\mathbf{e}}} = N|\mathbf{e}^* - \dot{\mathbf{e}}|^2$, so the capacity is given by

$$\text{Capacity} = \sum_{j=1}^m c_j^2 = \frac{1}{4N} \max_{\dot{\mathbf{e}} \in Z} \rho_{\dot{\mathbf{e}}}.$$

Thus the capacity is related to the largest non-centrality parameter. In the case where there are only two key classes, the

capacity is proportional to the non-centrality parameter. The capacity gives a measure of the maximum difference between distributions for the true key class and any other class, or equivalently capacity is a measure of the difference between the distributions for the true key class and the worse key class.

Some research [3], [4], [10] has considered finding the key class by taking an appropriate linear combination of the mask counts. In the language of this paper, this is equivalent to defining a unit (without loss of generality) vector \mathbf{u} and considering the random variable

$$\sqrt{N}\mathbf{u}^T(\hat{\mathbf{e}} - \dot{\mathbf{e}}) \sim N\left(\sqrt{N}\mathbf{u}^T(\mathbf{e}^* - \dot{\mathbf{e}}); 1\right).$$

This random variable can be analysed directly [3], [4] or the related χ^2 random variable obtained by forming the obvious quadratic form [10]. The two approaches are essentially equivalent and we use the latter approach and consider the random variable

$$N(\mathbf{u}^T(\hat{\mathbf{e}} - \dot{\mathbf{e}}))^T(\mathbf{u}^T(\hat{\mathbf{e}} - \dot{\mathbf{e}})) = N|\mathbf{u}^T(\hat{\mathbf{e}} - \dot{\mathbf{e}})|^2,$$

which has a non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter $N|\mathbf{u}^T(\mathbf{e}^* - \dot{\mathbf{e}})|^2$. Under the hypothesis that $\dot{\mathbf{e}} \neq \mathbf{e}^*$, we need to choose \mathbf{u} to maximise this non-centrality parameter that is to maximise $\mathbf{u}^T(\mathbf{e}^* - \dot{\mathbf{e}})$. Clearly such a \mathbf{u} is a unit vector in the direction $\mathbf{e}^* - \dot{\mathbf{e}}$, when the non-centrality parameter for this non-central χ^2 distribution with 1 degree of freedom is given by $N|\mathbf{e}^* - \dot{\mathbf{e}}|^2 = \rho_{\dot{\mathbf{e}}}$, where $\rho_{\dot{\mathbf{e}}}$ is the non-centrality parameter for the χ^2 distribution with m degrees of freedom discussed above. If there are more than two key classes, this non-centrality parameter can take many values depending on $\dot{\mathbf{e}}$, but is maximised by $\max_{\dot{\mathbf{e}}} \rho_{\dot{\mathbf{e}}}$. In the case discussed above where the j^{th} imbalance is given by $\pm c_j$, the maximal non-centrality parameter for this χ_1^2 random variable is proportional to the capacity. The parameter given in [3], [4], [10] for the effectiveness of a collection of linear approximations is the sum of squared biases. In this case, this is $4\sum_j c_j^2$, which is four times the capacity and so proportional to the non-centrality parameter.

Capacity though does not always give the full picture. Capacity clearly cannot directly address the issue of “neighbouring” key classes to the true key class when there are more than two key classes. Furthermore, the sole use of capacity in determining the effectiveness of a collection of linear approximations would suggest that there is never a role for linear approximations which have zero imbalance. However, the sampling distribution for the quadratic form has two parameters. The non-centrality parameter is related to the capacity and the number of degrees of freedom m is given by the number of masks. This indicates that the number of masks (even for the same capacity) may also be important factor. When we test whether $\dot{\mathbf{e}} = \mathbf{e}^*$ using m masks, we are in fact testing whether $\hat{\mathbf{d}}$ lies in a particular coset (containing \mathbf{d}^*) of a subspace of dimension $(2^l - 1) - m$. Of course such a test may be sufficient to discriminate between key classes. However, it is clear that the more masks we use (even with zero imbalance), the more exact our test of $\hat{\mathbf{d}}$ becomes. Thus for a direct test of whether a particular distribution (from key

class \mathbf{z}) is the true distribution (from the true key class \mathbf{z}^*), we should use all $m = (2^l - 1)$ masks. This gives a test that $\mathbf{d} = \mathbf{d}^*$ rather than whether \mathbf{d} and \mathbf{d}^* have a similar property (membership of a coset). We note that this observation does not directly relate to finding the maximum likelihood estimate of the key class which is a comparative process and does not directly involve the sampling distribution. However, this observation may have some relevance if we have to estimate the probability that a key class is the true key class, for example to establish some type of key ranking threshold.

VI. TRANSFORM METHODS

The use of transforms which are defined as the expectation of a function of the random variable is widespread in probability and statistics. Common examples of such transforms include the generating function, moment-generating function or the characteristic function. For a binary vector random variable of length m , such as the mask count random variable $\mathbf{V}^{(i)}$ of the i^{th} plaintext-ciphertext pair, an appropriate such transform would be given by

$$\Psi_{\mathbf{V}^{(i)}}(\mathbf{s}) = E \left((-1)^{\mathbf{s}^T \mathbf{V}^{(i)}} \right),$$

for a vector $\mathbf{s} \in \mathbb{Z}_2^m$. We show in Appendix III that the values of this transform are essentially (up to the sign) the imbalances, as we have

$$\Psi_{\mathbf{V}^{(i)}}(\mathbf{s}) = (-1)^{\mathbf{s}^T \mathbf{1}} e_{(R^T \mathbf{s})},$$

where for completeness we define the “null” imbalance e_0 to be 1 and R is the $m \times l$ matrix used to define the mask set given in Section III. For the components of $\mathbf{V}^{(i)}$ to be mutually stochastically independent, we would require $\Psi_{\mathbf{V}^{(i)}}$ to factorise into its component transforms. This is clearly not the case unless all non-null imbalances are 0. However, this transform can also be expressed as

$$\Psi_{\mathbf{V}^{(i)}}(\mathbf{s}) = \sum_{\mathbf{v} \in \mathbb{Z}_2^m} (-1)^{\mathbf{s}^T \mathbf{v}} P \left(\mathbf{V}^{(i)} = \mathbf{v} \right),$$

so this transform corresponds to the Walsh-Hadamard transform of the probability vector for $\mathbf{V}^{(i)}$. We can invert this transform to obtain the probability vector for $\mathbf{V}^{(i)}$. The linear dependencies between the masks mean that $\mathbf{V}^{(i)} \in \text{Im}(R)$. In Appendix III, it is shown that for such “valid” $\mathbf{V}^{(i)}$, the probability vector $P(\mathbf{V}^{(i)} = \mathbf{v})$ is given by

$$P \left(\mathbf{V}^{(i)} = \mathbf{v} \right) = 2^{-l} \sum_{\mathbf{s}' \in \mathbb{Z}_2^l} (-1)^{\mathbf{s}'^T (S\mathbf{v})} (\delta(\mathbf{s}') e_{\mathbf{s}'})$$

for $\mathbf{v} \in \text{Im}(R)$, where S is the $l \times m$ matrix $(I|0)$ that gives the first l components of a vector and $\delta(\mathbf{s}') = (-1)^{\mathbf{1}^T \mathbf{s}'}$. Thus $\delta(\mathbf{s}') = 1$ if \mathbf{s}' has even weight and $\delta(\mathbf{s}') = -1$ if \mathbf{s}' has odd weight. For such valid $\mathbf{V}^{(i)}$, the probability vector is the Walsh-Hadamard transform of a vector of imbalances (up to sign). Obviously for “invalid” $\mathbf{V}^{(i)}$ we have $P(\mathbf{V}^{(i)} = \mathbf{v}) = 0$ for $\mathbf{v} \notin \text{Im}(R)$.

Biryukov *et al* in their CRYPTO 2004 paper [1] give results for experiments using multiple linear approximations for the analysis of the DES [8]. The experiments find the key class

using the form of the log-likelihood given above. In [1], the log-likelihood is derived under the assumption of the probabilistic independence of different masks counts, and a justification for this assumption is given in Section 3.4 of [1]. This justification is the basis of the entire theoretical analysis of multiple linear approximations given by [1], yet, as we now discuss, it is fallacious.

The justification given in [1] concentrates on the mask counts for an individual plaintext-ciphertext pair, and “shows” such mask counts are independent. Clearly, as the example of Section II shows, such a set of mask counts cannot be independent. The justification gives a version of the above expression that gives the imbalances in terms of the Walsh-Hadamard transform of the probability vector for the mask counts, and then inverts this transform to give a probability vector for the mask counts as a Walsh-Hadamard transform of the imbalances. However the Walsh-Hadamard inversion given in [1] is not correct. It ignores both the maximal null imbalance e_0 and the restrictions given by the linear dependencies of the masks. Thus the probabilities given for $\mathbf{V}^{(i)}$ are highly erroneous. Firstly, a constant 2^{-l} is omitted, and, secondly, positive probabilities are assigned to events that cannot occur. The sum given in the inverse transform is then approximated by a product (with the null imbalance re-included) to give in our terminology

$$\begin{aligned} P(\mathbf{V}^{(i)} = \mathbf{v}) &\approx 2^m \cdot 2^{-l} \prod_{j=1}^m \frac{1}{2} (1 + e_{\mathbf{f}_j} (-1)^{v_j}) \\ &= 2^{m-l} \prod_{j=1}^m P(\mathbf{V}_j^{(i)} = v_j), \end{aligned}$$

where \mathbf{f}_j is the j^{th} standard basis vector. This expression is incorrect as it assigns significant positive probability to events with probability zero (see the example of Section II), so this product form is just not correct. Furthermore, even when we have no dependent masks, so all events have positive probability, the expression is not correct. It is then stated that: “Apart from an irrelevant constant factor of 2^{m-1} , this is exactly what we need: it implies that even with dependent masks, we can still multiply probabilities as we did to obtain [an expression for the joint probability of mask counts over N plaintext-ciphertext pairs]”. This assertion is simply incorrect. Thus not only does this probability for an individual plaintext-ciphertext not have the above product form, but the method given for combining probabilities is also invalid.

The justification given in [1] for the stochastic independence of mask counts is fundamentally flawed for two reasons. Firstly, stochastic independence for mask counts is essentially a large-sample (central limit) property (see Section III), so any reasoning based on one plaintext-ciphertext pair, such as that of [1], ignores the fundamental issue. Secondly, the justification given in [1] asserts the independence of mask counts for moderately large imbalances, but such mask counts can be far from independent (see Section II).

VII. CONCLUSIONS

We have given a theoretical analysis of linear cryptanalysis based on standard statistical theory. This has enabled us to handle multiple linear approximations in a systematic manner and so have a consistent method of finding the true key class

based on well-known statistical techniques. This theoretical approach has also led to new results about multiple linear approximations and highlighted severe shortcomings in the theoretical approach in other research.

APPENDIX I COVARIANCE OF $\mathbf{V}_{\mathbf{r}_a}^{(i)}$ AND $\mathbf{V}_{\mathbf{r}_{a'}}^{(i)}$

In this Appendix, we calculate the covariance of the two mask random variables $\mathbf{V}_{\mathbf{r}_a}^{(i)}$ and $\mathbf{V}_{\mathbf{r}_{a'}}^{(i)}$ ($\mathbf{a} \neq \mathbf{a}'$) discussed in Section III. We use the same technique for this calculation as we did for the covariance calculation as in Section II, but with sums of terms replacing individual terms. Thus we have

$$\begin{aligned} Cov\left(\mathbf{V}_{\mathbf{r}_a}^{(i)}, \mathbf{V}_{\mathbf{r}_{a'}}^{(i)}\right) &= E\left(\mathbf{V}_{\mathbf{r}_a}^{(i)} \mathbf{V}_{\mathbf{r}_{a'}}^{(i)}\right) - E\left(\mathbf{V}_{\mathbf{r}_a}^{(i)}\right) E\left(\mathbf{V}_{\mathbf{r}_{a'}}^{(i)}\right) \\ &= P\left(\mathbf{V}_{\mathbf{r}_a}^{(i)} = \mathbf{V}_{\mathbf{r}_{a'}}^{(i)} = 1\right) - P\left(\mathbf{V}_{\mathbf{r}_a}^{(i)} = 1\right) P_0\left(\mathbf{V}_{\mathbf{r}_{a'}}^{(i)} = 1\right) \\ &= P\left(\mathbf{x} \in H_{\mathbf{r}_a} \cap H_{\mathbf{r}_{a'}}\right) - P\left(\mathbf{x} \in H_{\mathbf{r}_a}\right) P\left(\mathbf{x} \in H_{\mathbf{r}_{a'}}\right), \end{aligned}$$

where \mathbf{x} is the underlying data class. Now $H_{\mathbf{r}_a} \cap H_{\mathbf{r}_{a'}}$ is a plane of codimension 2 (dimension $l - 2$), so we let H^{jl} ($j, l = 0, 1$) denote the four cosets of this plane. Thus we have $H^{00} = H_{\mathbf{r}_a} \cap H_{\mathbf{r}_{a'}}$, and $H^{01} = H_{\mathbf{r}_a} \cap H_{\mathbf{r}_{a'}}^C$, and so on. We can express $P(\mathbf{x} \in H_a)$ in this notation by

$$\begin{aligned} P(\mathbf{x} \in H_{\mathbf{r}_a}) &= \sum_{\mathbf{x} \in H_{\mathbf{r}_a}} 2^{-l}(1 + d_{\mathbf{x}}) \\ &= \frac{1}{2} + 2^{-l} \sum_{\mathbf{x} \in H^{00}} d_{\mathbf{x}} + 2^{-l} \sum_{\mathbf{x} \in H^{01}} d_{\mathbf{x}}, \end{aligned}$$

with $P(\mathbf{x} \in H_{\mathbf{r}_{a'}})$ being given by

$$\begin{aligned} P(\mathbf{x} \in H_{\mathbf{r}_{a'}}) &= \sum_{\mathbf{x} \in H_{\mathbf{r}_{a'}}} 2^{-l}(1 + d_{\mathbf{x}}) \\ &= \frac{1}{2} + 2^{-l} \sum_{\mathbf{x} \in H^{00}} d_{\mathbf{x}} + 2^{-l} \sum_{\mathbf{x} \in H^{10}} d_{\mathbf{x}}. \end{aligned}$$

If second order terms are negligible, then the product of these probabilities, $P(\mathbf{x} \in H_{\mathbf{r}_{a'}})P(\mathbf{x} \in H_{\mathbf{r}_a})$, is given by

$$\frac{1}{4} + 2^{-l} \sum_{\mathbf{x} \in H^{00}} d_{\mathbf{x}} + 2^{-(l+1)} \sum_{\mathbf{x} \in H^{01}} d_{\mathbf{x}} + 2^{-(l+1)} \sum_{\mathbf{x} \in H^{10}} d_{\mathbf{x}}.$$

However, $P(\mathbf{x} \in H_{\mathbf{r}_a} \cap H_{\mathbf{r}_{a'}})$ is given by

$$P(\mathbf{x} \in H^{00}) = \sum_{\mathbf{x} \in H^{00}} 2^{-l}(1 + d_{\mathbf{x}}) = \frac{1}{4} + 2^{-l} \sum_{\mathbf{x} \in H^{00}} d_{\mathbf{x}},$$

so the covariance is given by

$$Cov\left(\mathbf{V}_{\mathbf{r}_a}^{(i)}, \mathbf{V}_{\mathbf{r}_{a'}}^{(i)}\right) = -2^{-(l+1)} \left(\sum_{\mathbf{x} \in H^{01}} + \sum_{\mathbf{x} \in H^{10}} \right) d_{\mathbf{x}}.$$

Now H^{01} and H^{10} are disjoint cosets, so this sum is over the set $H^{01} \cup H^{10}$. The complement of this set is the union of the complementary cosets, namely $H^{00} \cup H^{11}$, which is the hyperplane $H_{\mathbf{r}_a + \mathbf{r}_{a'}} = H_{\mathbf{r}_{a+a'}}$. As $\sum_{\mathbf{x}} d_{\mathbf{x}} = 0$, the sum over the set $H^{01} \cup H^{10}$ and the sum over the complementary set $H_{\mathbf{r}_{a+a'}}$ add to give 0, so the covariance is given by

$$Cov\left(\mathbf{V}_{\mathbf{r}_a}^{(i)}, \mathbf{V}_{\mathbf{r}_{a'}}^{(i)}\right) = 2^{-(l+1)} \sum_{\mathbf{x} \in H_{\mathbf{r}_{a+a'}}} d_{\mathbf{x}}.$$

However, this sum is just a quarter of the imbalance of the mask $\mathbf{a} + \mathbf{a}'$, so (if second order terms are negligible) the covariance is given by

$$Cov\left(\mathbf{V}_{\mathbf{r}_a}^{(i)}, \mathbf{V}_{\mathbf{r}_{a'}}^{(i)}\right) = \frac{1}{4} e_{\mathbf{r}_{a+a'}}.$$

APPENDIX II

χ^2 MULTINOMIAL GOODNESS-OF-FIT STATISTIC

In Section IV, we consider how to test statistically whether a specific value $\hat{\mathbf{e}}$ for the vector of mask imbalances is equal to the true value for the vector of mask imbalance \mathbf{e}^* , that is whether $\hat{\mathbf{e}} = \mathbf{e}^*$ by considering the best (maximum likelihood) estimate for the vector of mask imbalances $\hat{\mathbf{e}} = \frac{2}{N} \mathbf{v} - 1$. We now show how to derive the distribution of the random variable $\hat{\mathbf{e}} - \mathbf{e}$ by using the standard χ^2 goodness-of-fit test for multinomial distributions to test whether a vector of data class probabilities is the true vector of data class probabilities, that is whether $\hat{\mathbf{d}} = \mathbf{d}^*$. The derivation of the distribution of this goodness-of-fit test statistic depends on the distribution of the random variable $\hat{\mathbf{d}} - \mathbf{d}$, where $\hat{\mathbf{d}} = (N2^{-l})^{-1} \mathbf{w} - 1$ is the best (maximum likelihood) estimate of the vector of data class probabilities. This distribution depends on two symmetric $2^l \times 2^l$ matrices B and P which satisfy $PBP = B$ [9]. The matrix B is the information matrix and is given for a multinomial distribution by the diagonal matrix with diagonal entries $(2^{-l}(1 + d_{\mathbf{x}}^*))^{-1}$. As B is a diagonal matrix, we define the matrix $B^{\frac{1}{2}}$ to be the diagonal matrix with diagonal entries $(2^{-l}(1 + d_{\mathbf{x}}^*))^{-\frac{1}{2}}$, so $B = B^{\frac{1}{2}} \cdot B^{\frac{1}{2}}$. The matrix P arises from the restriction that $\sum d_{\mathbf{x}} = 0$ and is given by

$$P = (2^{-l}(1 + \mathbf{d}^*)) (2^{-l}(1 + \mathbf{d}^*))^T.$$

We now discuss how to use the χ^2 goodness of fit test for a multinomial distribution to analyse the distributions that arise from a linear cryptanalysis with multiple masks. Under the hypothesis that $\mathbf{d} = \mathbf{d}^*$, the random variable $\hat{\mathbf{d}} - \mathbf{d}$ is given by

$$\sqrt{N}2^{-l} (\hat{\mathbf{d}} - \mathbf{d}) = (B^{-1} - P) B^{\frac{1}{2}} \mathbf{Y}_{2^l}$$

where $\mathbf{Y}_{2^l} \sim N(0, I)$ is a vector of 2^l independent standard normal ($N(0, 1)$) random variables [9]. This distributional form is used to give the χ^2 goodness-of-fit statistic for a multinomial distribution, namely the quadratic form

$$N2^{-2l} (\hat{\mathbf{d}} - \mathbf{d})^T B (\hat{\mathbf{d}} - \mathbf{d})$$

or as it is more commonly expressed

$$\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}.$$

This quadratic form has a χ^2 distribution with $2^l - 1$ degrees of freedom under the hypothesis that $\mathbf{d} = \mathbf{d}^*$. Under the alternative hypothesis that $\mathbf{d} \neq \mathbf{d}^*$, $\sqrt{N}2^{-l} (\hat{\mathbf{d}} - \mathbf{d})$ is approximately a multivariate normal random variable with the same covariance structure as before, but with mean $\sqrt{N}2^{-l} (\mathbf{d}^* - \mathbf{d})$. Under this hypothesis, the test statistic is approximately a non-central χ^2 distribution with $2^l - 1$ degrees of freedom and non-centrality parameter $N2^{-2l} |\mathbf{d}^* - \mathbf{d}|^2$.

We now consider the mask counts, which are given by a random variable \mathbf{V} , where $\mathbf{V} = T\mathbf{W}$. A vector \mathbf{v} of mask counts is a realisation of this random variable \mathbf{V} and can be expressed as $\mathbf{v} = T\mathbf{w}$, where \mathbf{w} is a realisation of counts for the underlying data classes. We recall that $\mathbf{e} = 2^{-(l-1)} T\mathbf{d}$ is

a vector of mask imbalances, so $2\mathbf{e}$ a vector of mask biases. The estimate of \mathbf{e} is then given by

$$\begin{aligned}\hat{\mathbf{e}} &= 2^{-(l-1)}T\hat{\mathbf{d}} = 2^{-(l-1)}T((2^{-l}N)^{-1}\mathbf{w} - \mathbf{1}) \\ &= 2N^{-1}T\mathbf{w} - 2^{-(l-1)}T\mathbf{1} \\ &= \frac{2}{N}T\mathbf{w} - \mathbf{1} = \frac{2}{N}\mathbf{v} - \mathbf{1}\end{aligned}$$

so the components of $\hat{\mathbf{e}}$ are proportional to the difference of the mask counts \mathbf{v} from the uniform value $\frac{N}{2}\mathbf{1}$. The random variable $\hat{\mathbf{e}} - \mathbf{e}$ is given by $2^{-(l-1)}T(\hat{\mathbf{d}} - \mathbf{d})$ and is a measure of the difference of the imbalance under consideration \mathbf{e} and the estimated imbalance $\hat{\mathbf{e}}$. The hypothesis that $\hat{\mathbf{d}} = \mathbf{d}^*$ implies that $\mathbf{e} = \mathbf{e}^*$. Under this hypothesis, the random variable $\hat{\mathbf{e}} - \mathbf{e}$ is given by

$$\begin{aligned}\sqrt{N}(\hat{\mathbf{e}} - \mathbf{e}) &= \sqrt{N}2^{-(l-1)}T(\hat{\mathbf{d}} - \mathbf{d}) \\ &= 2T(B^{-1} - P)B^{\frac{1}{2}}\mathbf{Y}_{2^l}.\end{aligned}$$

This random variable is an m -dimensional multivariate normal random variable with zero mean and $m \times m$ covariance matrix $Cov(\sqrt{N}(\hat{\mathbf{e}} - \mathbf{e}))$ is given by

$$4\left(T(B^{-1} - P)B^{\frac{1}{2}}\right)Cov(\mathbf{Y}_{2^l})\left(T(B^{-1} - P)B^{\frac{1}{2}}\right)^T.$$

Thus this covariance matrix is given by

$$\begin{aligned}Cov(\sqrt{N}(\hat{\mathbf{e}} - \mathbf{e})) &= 4T(B^{-1} - P)B(B^{-1} - P)T^T \\ &= 4T(B^{-1} - P)T^T = 4(TB^{-1}T^T - TPPT^T).\end{aligned}$$

The calculation of an entry in this covariance matrix is given by the calculation for $Cov(\mathbf{V}_{\mathbf{r}_a}^{(i)}, \mathbf{V}_{\mathbf{a}'}^{(i)})$ given in Appendix I. The first component of this covariance matrix, $TB^{-1}T^T$, corresponds to $E(\mathbf{V}_{\mathbf{r}_a}^{(i)}\mathbf{V}_{\mathbf{a}'}^{(i)})$, whilst the second component $TPPT^T$, corresponds to $E(\mathbf{V}_{\mathbf{r}_a}^{(i)})E(\mathbf{V}_{\mathbf{a}'}^{(i)})$, so the covariance can be calculated directly from the χ^2 goodness-of-fit sampling distribution by partitioning the various sums in the same manner as described in Appendix I. Thus, ignoring second order terms, the off-diagonal $(\mathbf{a}, \mathbf{a}')$ -entry ($\mathbf{a} \neq \mathbf{a}'$) of $TB^{-1}T^T - TPPT^T$ is $\frac{1}{4}e_{\mathbf{a}+\mathbf{a}'}$, whilst the diagonal entry (\mathbf{a}, \mathbf{a}) is $\frac{1}{4}$, so the covariance matrix is given by $I + \Delta$. Thus the distribution of $\hat{\mathbf{e}} - \mathbf{e}$ under the hypothesis that $\mathbf{e} = \mathbf{e}^*$ is given by

$$\sqrt{N}(\hat{\mathbf{e}} - \mathbf{e}) \sim N(\mathbf{0}; I + \Delta).$$

If second order terms are negligible, then we may disregard Δ^2 , so $I + \Delta = (I + \frac{1}{2}\Delta)^2$. In this case, the distribution of $\hat{\mathbf{e}} - \mathbf{e}$ is given by

$$\sqrt{N}(I + \frac{1}{2}\Delta)^{-1}(\hat{\mathbf{e}} - \mathbf{e}) \sim \mathbf{Y}_m,$$

where $\mathbf{Y}_m \sim N(0, I)$ is an m -vector of independent standard normal ($N(0, 1)$) random variables. Under an alternative hypothesis that $\mathbf{e} \neq \mathbf{e}^*$, the distribution of

$$\sqrt{N}\left(I + \frac{1}{2}\Delta\right)^{-1}(\hat{\mathbf{e}} - \mathbf{e})$$

is given by the normal distribution

$$N\left(\sqrt{N}\left(I + \frac{1}{2}\Delta\right)^{-1}(\mathbf{e}^* - \mathbf{e}); I\right),$$

which gives the distribution of $\hat{\mathbf{e}} - \mathbf{e}$.

APPENDIX III TRANSFORM INVERSION

The transform of a binary vector random variable of length m such as the mask count random variable $\mathbf{V}^{(i)}$ was defined in Section VI as

$$\Psi_{\mathbf{V}^{(i)}}(\mathbf{s}) = E\left((-1)^{\mathbf{s}^T\mathbf{V}^{(i)}}\right),$$

for a vector \mathbf{s} of length m . For technical reasons, it is more convenient to consider $\tilde{\mathbf{V}}^{(i)} = \mathbf{1} + \mathbf{V}^{(i)}$, so we have (see Section III)

$$\tilde{\mathbf{V}}^{(i)} = \mathbf{1} + \mathbf{V}^{(i)} = R\mathbf{X}^{(i)},$$

where $\mathbf{X}^{(i)}$ is the random variable of length l corresponding to the data class for the i^{th} plaintext-ciphertext pair and R is the $m \times l$ matrix used to define the mask set given in Section III. Thus we have

$$\begin{aligned}\Psi_{\tilde{\mathbf{V}}^{(i)}}(\mathbf{s}) &= E\left((-1)^{\mathbf{s}^T R\mathbf{X}^{(i)}}\right) = E\left((-1)^{(R^T\mathbf{s})^T\mathbf{X}^{(i)}}\right) \\ &= (-1)^{\mathbf{s}^T\mathbf{1}}\Psi_{\mathbf{V}^{(i)}}(\mathbf{s}) = \Psi_{\mathbf{X}^{(i)}}(R^T\mathbf{s}),\end{aligned}$$

where $\Psi_{\mathbf{X}^{(i)}}$ denotes the transform for $\mathbf{X}^{(i)}$. However, $\Psi_{\mathbf{X}^{(i)}}$ can be expressed in terms of the imbalances of the masks, because, for $\mathbf{s}' \in \mathbb{Z}_2^l$, we have

$$\begin{aligned}\Psi_{\mathbf{X}^{(i)}}(\mathbf{s}') &= E\left((-1)^{(\mathbf{s}')^T\mathbf{X}^{(i)}}\right) \\ &= P(\mathbf{X}^{(i)} \in H_{\mathbf{s}'}) - P(\mathbf{X}^{(i)} \notin H_{\mathbf{s}'}) \\ &= \frac{1}{2}(1 + e_{\mathbf{s}'}) - \frac{1}{2}(1 - e_{\mathbf{s}'}) = e_{\mathbf{s}'},\end{aligned}$$

where we define $e_0 = 1$ (the ‘‘null’’ imbalance) for completeness. Thus the mask imbalances are just this transform evaluated at the mask selection vector. Hence the transform of $\tilde{\mathbf{V}}^{(i)}$ is given by

$$\Psi_{\tilde{\mathbf{V}}^{(i)}}(\mathbf{s}) = (-1)^{\mathbf{s}^T\mathbf{1}}\Psi_{\mathbf{V}^{(i)}}(\mathbf{s}) = e_{(R^T\mathbf{s})},$$

so the transform of $\mathbf{V}^{(i)}$ is given by

$$\Psi_{\mathbf{V}^{(i)}}(\mathbf{s}) = (-1)^{\mathbf{s}^T\mathbf{1}}e_{(R^T\mathbf{s})}$$

The transform $\Psi_{\tilde{\mathbf{V}}^{(i)}}$ corresponds to the Walsh-Hadamard transform of the probability vector for $\tilde{\mathbf{V}}^{(i)}$ since

$$\Psi_{\tilde{\mathbf{V}}^{(i)}}(\mathbf{s}) = \sum_{\mathbf{v} \in \mathbb{Z}_2^m} (-1)^{\mathbf{s}^T\mathbf{v}}P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v}).$$

The probability vector for $\tilde{\mathbf{V}}^{(i)}$ can therefore be given in terms of the imbalances by the standard inversion method for the Walsh-Hadamard transform. Thus we have

$$\begin{aligned}P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v}) &= 2^{-m} \sum_{\mathbf{s} \in \mathbb{Z}_2^m} (-1)^{\mathbf{s}^T\mathbf{v}}\Psi_{\tilde{\mathbf{V}}^{(i)}}(\mathbf{s}) \\ &= 2^{-m} \sum_{\mathbf{s} \in \mathbb{Z}_2^m} (-1)^{\mathbf{s}^T\mathbf{v}}e_{(R^T\mathbf{s})}.\end{aligned}$$

We recall from Section III that R is an $m \times l$ matrix given by $(I|Q^T)^T$. We can define an $m \times (m - l)$ matrix \bar{R} by $\bar{R} = (0|I)^T$, so the $m \times m$ matrix

$$(R|\bar{R}) = \left(\begin{array}{c|c} I & 0 \\ \hline Q & I \end{array}\right)$$

is invertible. We can define the $l \times m$ matrix $S = (I|0)$ and the $l \times (m - l)$ matrix $\bar{S} = (0|I)$, so S and \bar{S} give the first l and last $m - l$ components of a vector respectively. For $\mathbf{v} \in \mathbb{Z}_2^m$,

we define vectors \mathbf{v}' and \mathbf{v}'' of length l and $m-l$ respectively by

$$\mathbf{v}' = S\mathbf{v} \text{ and } \mathbf{v}'' = \bar{S}\mathbf{v} + Q S\mathbf{v},$$

so \mathbf{v} is given in terms of \mathbf{v}' and \mathbf{v}'' by

$$\begin{aligned} \mathbf{v} &= \begin{pmatrix} S\mathbf{v}' \\ \bar{S}\mathbf{v}'' \end{pmatrix} = \begin{pmatrix} I & 0 \\ Q & I \end{pmatrix} \begin{pmatrix} S\mathbf{v}' \\ \bar{S}\mathbf{v}'' + Q S\mathbf{v}' \end{pmatrix} \\ &= (R|\bar{R}) \begin{pmatrix} \mathbf{v}' \\ \mathbf{v}'' \end{pmatrix} = R\mathbf{v}' + \bar{R}\mathbf{v}'' . \end{aligned}$$

Thus $\mathbf{v} \in \text{Im}(R)$ if and only if $\bar{R}\mathbf{v}'' = 0$ or equivalently $\mathbf{v}'' = 0$ (as \bar{R} is injective). The inversion transform is given in terms of \mathbf{v}' and \mathbf{v}'' , so $P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v})$ is given by

$$2^{-m} \sum_{\mathbf{s} \in \mathbb{Z}_2^m} (-1)^{(R^T \mathbf{s})^T \mathbf{v}'} (e_{(R^T \mathbf{s})}) (-1)^{\mathbf{s}^T \bar{R}\mathbf{v}''} .$$

We evaluate this sum by partitioning \mathbb{Z}_2^m into cosets of $K = \text{Ker}(R^T)$, a space of codimension l (dimension $m-l$), so there are 2^l cosets of K . For $\mathbf{s}' \in \mathbb{Z}_2^l$, we define the coset $K_{\mathbf{s}'}$ by $K_{\mathbf{s}'} = \{\mathbf{s} \in \mathbb{Z}_2^m | R^T \mathbf{s} = \mathbf{s}'\}$. The inversion transform is then calculated by summing within a coset of K and then across cosets, so $P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v})$ is given by

$$2^{-m} \sum_{\mathbf{s}' \in \mathbb{Z}_2^l} \sum_{\mathbf{s} \in K_{\mathbf{s}'}} (-1)^{(R^T \mathbf{s})^T \mathbf{v}'} (e_{(R^T \mathbf{s})}) (-1)^{\mathbf{s}^T \bar{R}\mathbf{v}''} .$$

By definition, $R^T \mathbf{s} = \mathbf{s}'$ is constant for $\mathbf{s} \in K_{\mathbf{s}'}$, so we have

$$P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v}) = 2^{-m} \sum_{\mathbf{s}' \in \mathbb{Z}_2^l} (-1)^{\mathbf{s}'^T \mathbf{v}'} (e_{\mathbf{s}'}) \sum_{\mathbf{s} \in K_{\mathbf{s}'}} (-1)^{\mathbf{s}^T \bar{R}\mathbf{v}''} .$$

The coset $K_{\mathbf{s}'}$ is given by

$$K_{\mathbf{s}'} = \left\{ \begin{pmatrix} \mathbf{s}' + Q\mathbf{s}'' \\ \mathbf{s}'' \end{pmatrix} \mid \mathbf{s}'' \in \mathbb{Z}_2^{m-l} \right\} ,$$

whilst $\bar{R}\mathbf{v}'' = \begin{pmatrix} 0 \\ \mathbf{v}'' \end{pmatrix}$, so the sum over the coset $K_{\mathbf{s}'}$ is given by

$$\begin{aligned} \sum_{\mathbf{s} \in K_{\mathbf{s}'}} (-1)^{\mathbf{s}^T \bar{R}\mathbf{v}''} &= \sum_{\mathbf{s}'' \in \mathbb{Z}_2^{m-l}} (-1)^{(\mathbf{s}'')^T \mathbf{v}''} \\ &= \begin{cases} 2^{m-l} & \mathbf{v}'' = 0 \\ 0 & \mathbf{v}'' \neq 0 \end{cases} . \end{aligned}$$

The condition that $\mathbf{v}'' = 0$ is equivalent to $\mathbf{v} \in \text{Im}(R)$, so in this case with $\mathbf{v}' = S\mathbf{v}$, the inversion transform gives

$$P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v}) = 2^{-l} \sum_{\mathbf{s}' \in \mathbb{Z}_2^l} (-1)^{\mathbf{s}'^T (S\mathbf{v})} (e_{\mathbf{s}'}),$$

whereas $P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v}) = 0$ for $\mathbf{v} \notin \text{Im}(R)$. For $\mathbf{v} \in \text{Im}(R)$, the probability vector for $\mathbf{V}^{(i)}$ can now be given by

$$\begin{aligned} P(\mathbf{V}^{(i)} = \mathbf{v}) &= P(\tilde{\mathbf{V}}^{(i)} = \mathbf{v} + \mathbf{1}) \\ &= 2^{-l} \sum_{\mathbf{s}' \in \mathbb{Z}_2^l} (-1)^{\mathbf{s}'^T (S\mathbf{v})} (-1)^{\mathbf{s}'^T (S\mathbf{1})} (e_{\mathbf{s}'}). \end{aligned}$$

However, $S\mathbf{1}$ is a vector $\mathbf{1}$ of length l , so $\mathbf{s}'^T (S\mathbf{1})$ gives the parity of \mathbf{s}' . We therefore define $\delta(\mathbf{s}') = (-1)^{\mathbf{s}'^T (S\mathbf{1})}$, so

$\delta(\mathbf{s}') = 1$ if \mathbf{s}' has even weight and $\delta(\mathbf{s}') = -1$ if \mathbf{s}' has odd weight. Thus for $\mathbf{v} \in \text{Im}(R)$, we have

$$P(\mathbf{V}^{(i)} = \mathbf{v}) = 2^{-l} \sum_{\mathbf{s}' \in \mathbb{Z}_2^l} (-1)^{\mathbf{s}'^T (S\mathbf{v})} (\delta(\mathbf{s}') e_{\mathbf{s}'}),$$

with $P(\mathbf{V}^{(i)} = \mathbf{v}) = 0$ for $\mathbf{v} \notin \text{Im}(R)$.

ACKNOWLEDGEMENTS

We wish to thank Matt Robshaw and the anonymous IEEE referees for their comments.

REFERENCES

- [1] C. De Cannière, A. Biryukov and M. Quisquater. On Multiple Linear Approximations. In M. Franklin, editor, *Advances in Cryptology - CRYPTO 2004*, volume 3152 of *LNCS*, pages 1–22. Springer-Verlag, 2004.
- [2] T. Baignères, P. Junod, and S. Vaudenay. How Far Can We Go Beyond Linear Cryptanalysis? In Pil Joong Lee, editor, *Advances in Cryptology - ASIACRYPT 2004*, volume 3329 of *LNCS*, pages 432–451. Springer-Verlag, 2004.
- [3] B.S. Kaliski Jr. and M.J.B. Robshaw. Linear Cryptanalysis using Multiple Approximations. In Y. Desmedt, editor, *Fast Software Encryption 1994*, volume 839 of *LNCS*, pages 26–39. Springer-Verlag, 1994.
- [4] B.S. Kaliski Jr. and M.J.B. Robshaw. Linear Cryptanalysis using Multiple Approximations and FEAL. In B. Preneel, editor, *Fast Software Encryption 1994*, volume 1008 of *LNCS*, pages 249–264. Springer-Verlag, 1995.
- [5] P. Junod and S. Vaudenay. Optimal Key Ranking Procedures in a Statistical Cryptanalysis. In T. Johansson, editor, *Fast Software Encryption 2003*, volume 2887 of *LNCS*, pages 235–246. Springer-Verlag, 2003.
- [6] M. Matsui. Linear Cryptanalysis for DES Cipher. In T. Hellese, editor, *Advances in Cryptology Eurocrypt 1993*, volume 765 of *LNCS*, pages 386–397. Springer-Verlag, 1993.
- [7] S. Murphy, F. Piper, M. Walker, and P. Wild. Maximum Likelihood Estimation for Block Cipher Keys. Technical report, Royal Holloway (University of London), 1994. <http://www.isg.rhul.ac.uk/~sean>.
- [8] National Institute of Standards and Technology. Federal Information Processing Standards Publication (FIPS) 46: The Data Encryption Standard. January, 1977.
- [9] S.D. Silvey. *Statistical Inference*. Chapman and Hall, 1975.
- [10] S. Vaudenay. An Experiment on DES Statistical Cryptanalysis. In *Proceedings of the 3rd ACM Conference on Computer Security*, pages 386–397. ACM Press, 1996.
- [11] D. Wagner. Towards a Unifying View of Block Cipher Cryptanalysis. In B. Roy and W. Meier, editors, *Fast Software Encryption 2004*, volume 3017 of *LNCS*, pages 16–33. Springer-Verlag, 2004.

Biography. Sean Murphy is a Professor at Royal Holloway, University of London. He has a B.A. and a Ph.D. in Mathematics.