On linear cryptanalysis with many linear approximations

June 29, 2008

Abstract

In this paper we elaborate on the theoretical framework of [BCQ04] to quantify the information brought by several linear approximations of a block-cipher without putting any restriction on these approximations. This allows to estimate accurately how many plaintext-ciphertext pairs are needed in order to recover with good probability the vector $\tilde{\mathbf{K}}$ formed by the linear combinations of the key bits involved in the linear approximations. Moreover, we also show how decoding techniques can be used in this context in order to reduce significantly the time complexity of finding the most likely $\tilde{\mathbf{K}}$.

1 Introduction

Related work

Linear cryptanalysis is probably one of the most powerful tools available for attacking symmetric cryptosystems. It was invented by Matsui [Mat93, **?**] to break the DES cipher building upon ideas put forward in [TCG91, MM92]. It was quickly discovered that other ciphers can be attacked in this way, for instance FEAL [OA94], LOKI [TSM94], SAFER [MPWW95].

It is a known plaintext attack which takes advantage of probabilistic linear equations that involve bits of the plaintext $\mathbf{P}$, the ciphertext $\mathbf{C}$ and the key $\mathbf{K}$

$$\mathbf{Pr}(<\pi,\mathbf{P}>\oplus<\gamma,\mathbf{C}>\oplus<\kappa,\mathbf{K}>=b)=\frac{1}{2}+\epsilon. \tag{1}$$

ϵ is called the *bias* of the equation, π,γ and κ are linear masks and $<\pi,\mathbf{P}>$ denotes the following inner product between $\pi=(\pi_i)_{1\leq i\leq m}$ and $\mathbf{P}=(P_i)_{1\leq i\leq m}$, $<\pi,\mathbf{P}>\stackrel{\text{def}}{=}\bigoplus_{i=1}^{m}\pi_i P_i$. There might be several different linear approximations of this kind we have at our disposal and we let n be their number. We denote the corresponding key masks by $\kappa_i=(\kappa_i^j)_{1\leq j\leq k}$ and the corresponding biases by ϵ_i for $i\in\{1,\ldots,n\}$.

Such an attack can be divided in three parts:

- *Distillation phase:* it consists in extracting from the available plaintext-ciphertext pairs the relevant parts of the data. It basically consists in counting for each linear approximation how many times $<\pi,\mathbf{P}>\oplus<\gamma,\mathbf{C}>$ evaluates to zero.
- *Analysis phase:* It consists of extracting from the values taken by the counters some information on the key and testing whether some key guesses are correct or not by using the linear approximation(s) (1) as a distinguisher.
- *Search phase:* It typically consists in finding the rest of the key by exhaustive search.

In [Mat93] Matsui used only one approximation to distinguish wrong last round keys from the right one. One year later, he refined his attack by using a second approximation obtained by symmetry [**?**] and by also distinguishing with them the first round key. Later Vaudenay [Vau96] has presented a framework for statistical cryptanalysis where Matsui's attack is presented as a particular case. With Junod, he has also studied the optimal way of merging information from two (or more) approximations [JV03]. This kind of attack can use several approximations but the key masks must have disjoint supports. A second approach of using multiple equations is given by Kaliski and Robshaw [BSKR94]. They improved Matsui's first attack using several approximations which have the same key mask κ.

All these improvements have a common goal: reducing the amount of messages needed for the attack. Clearly, using several approximations should give more information than a single one. Biryukov and al. suggest in [BCQ04] a way of using multiple linear approximations without putting any restriction on them. We call this kind of attack *multilinear cryptanalysis*. Moreover, they bring in a simple probabilistic model which is quite convenient for studying linear cryptanalysis with multiple linear approximations. We will elaborate on the this theoretical framework here.

Our contribution

The purpose of this paper is to study how much multiple linear approximations may benefit linear cryptanalysis. Our purpose is twofold here:
- we wish to quantify accurately how much information is gained on the key from the knowledge of a certain amount of plaintext-ciphertext pairs and a certain number of probabilistic approximate linear equations of type (1).
- By using decoding techniques, we suggest a much faster way for recovering the linear combinations of the key bits $< \kappa, \mathbf{K} >$ than what has been proposed before.

How much information do we have on the key by using linear cryptanalysis statistics?

Several statistics have been proposed to study how many plaintext-ciphertext pairs we need to have in order to carry out succesfully a linear cryptanalysis. This includes for instance the probability of guessing incorrectly a linear combination of key bits by Matsui's Algorithm 1 [**?**], the ranking of the right subkey in the ordered list of candidates [**?**] or the expected size of the number of keys which are more likely than the right key [BCQ04]. Some of these statistics are either not relevant for linear cryptanalysis with several linear equations or are extremely difficult to calculate in the case of multilinear cryptanalysis (such as for instance the ranking statistics of [**?**]). This is not the case of the expected size of the number L of keys which are more likely than the right key considered in [BCQ04]. However, this kind of statistics also leads to pessimistic predictions concerning the number of plaintext-ciphertexts which are needed. This is related to the following probabilistic phenomenon which is detailled in Subsection 4.4: this expectation is in a rather wide range of amount of plaintext-ciphertext pairs exponential in the key size k, while for most plaintext-ciphertext pairs the most likely key is the right one. This comes from the fact that rare events (of exponentially small probability) yield values of L which are exponentially large in k. In other words, while for typical plaintext-ciphertext pairs L is equal to zero, for some rare occurrences of the plaintext-ciphertext pairs L is very large, and this accounts quite heavily in the expectation of L. This boils down to the fact that we take the expectation of a quantity which can vary between 0 and 2^k. It would be much better to take the expectation of a quantity which is much smaller.

A quantity which is of this type is the *entropy* $H(\mathbf{K}|\mathbf{Y})$ of the key $\mathbf{K}$ (or more generally

$H(\mathbf{K}'|\mathbf{Y})$, where $\mathbf{K}'$ is a certain subkey of K- for instance it can be the part of the key involved in a distinguisher attack) given the statistics $\mathbf{Y}$ we have derived from the plaintext-ciphertext pairs. This quantity is one of the most fundamental measure of the uncertainty on the key $\mathbf{K}$ given $\mathbf{Y}$ and displays many attractive features. First of all, it can be considered as the number of "truly" random bits left in $\mathbf{K}$ given $\mathbf{Y}$. Moreover, it quantifies precisely what happens if there is a subset T of possible $\mathbf{K}$'s such that:
(i) all the keys of T have roughly the same probability of being the right key given $\mathbf{Y}$,
(ii) T is of much smaller cardinality than the set of all possible keys $\mathbf{K}$,
(iii) and is such that the probability that the key is not in T given $\mathbf{Y}$ is small.

In such a case we wish to say that the number of random bits in the key given $\mathbf{Y}$ is of order $\log_2(|T|)$. The entropy functional captures with accuracy this behavior. If the aforementioned phenomenon occurs, then the size of T will be about $2^{H(\mathbf{K}|\mathbf{Y})}$. This quantity $H(\mathbf{K}|\mathbf{Y})$ can be really viewed in such a case as the expectation of the logarithm $\log_2(L)$ of of the list of candidates which are at least as likely as the right key. The logarithm of L varies much less than L and this why the typical size of $\log L$ coincides quite well with the expectation.

Despite the fact that it much more desirable to estimate the entropy than the expectation of L, it might seem that this quantity is much harder to calculate. Our main result is to give here a lower bound on this quantity (see Subsection**??**) which is quite sharp. The sharpness of the bound is illustrated by the results of Subsection 4.3. We apply this bound in three different scenarios: (i) the linear attack which recovers only the linear combination of the key bits, (ii) the usual linear distinguishing attack which recovers some linear combinations of the key bits of the first (and/or) last round, and (iii) the algorithm MK2 in [BCQ04]. We wish to emphasize the fact that the technique to derive the lower bound is quite general and applies in a very wide range of situations, and not only in the case where $\mathbf{Y}$ corresponds to counters of linear approximations (see Subsection 4.1). A second useful property of this lower bound on the entropy is that it gives an upper bound on the information we gain on the $\mathbf{K}$ when we know $\mathbf{Y}$ which is independent of the algorithm we use afterwards to extract this information.

A fast algorithm for recovering the linear combinations of the key bits $<\kappa, \mathbf{K}>$

Our second contribution is to suggest an algorithm for recovering the linear combinations $<\kappa_i, \mathbf{K}>$. It will be convenient to denote by $\tilde{\mathbf{K}} \stackrel{\text{def}}{=} (\tilde{K}_i)_{1\leq i\leq n}$ the vector of linear combinations of the key bits induced by the key masks, that is $\tilde{K}_i \stackrel{\text{def}}{=} \bigoplus_{j=1}^{k} \kappa_i^j K_j$. A particular quantity will play a fundamental role in this setting. It is the dimension (what we will denote by d) of the vector space generated by the κ_i's. It can be much smaller than the number n of different key masks.

We first show that finding the most likely value(s) for $\tilde{\mathbf{K}}$ reduces to the problem of decoding a linear code over the Gaussian channel. This problem has been studied in coding theory [Dum00, VF04], [PH98, chapter 7] and has lead to algorithms for fulfilling this task which are much more efficient than exhaustive search over the code space. This translates in our setting into algorithms which can search for the most likely value(s) for $\tilde{\mathbf{K}}$ without looking at the whole space of possible linear combinations. For instance, in the toy example we have considered, namely 8-round DES, we find the most likely values of 20 linearly independent combinations of the key bits by using only 2^{19} plaintext-ciphertext pairs and considering in the search phase only a few hundred possible linear combinations. This is comparable with the amount of plaintext-ciphertext pairs used in the best current linear cryptanalysis of the 8-round DES and improves significantly the time complexity of the attack.

This approach opens up the hope of (linearly) cryptanalyzing a cipher in a different way. Generally in linear cryptanalysis, the linear approximations for a cipher are mainly used as distinguishers for last round or first round key bits and yield only very limited information on the overall key bits. The approach which is suggested here is different in nature. It consists in trying to take advantage of the multiplicity of linear approximations and of the fact that there are rather efficient decoding algorithms to process this kind of information in order to find these key bits directly without peeling off the cipher by one or two rounds and using the linear approximations for distinguishing a right last or first round key guess from wrong ones. This is basically also what is considered in the first phase of the attack algorithm MK1 proposed in [BCQ04], but there is an important difference here: whereas the processing of the information provided by the distillation phase in the analysis phase requires there to consider all possible linear combinations of key bits given by the masks and to rank them according to the information available from the distillation phase, we proceed differently with our decoding algorithm. The time complexity of algorithm MK1 is namely at least of order $O(d2^d)$ for calculating the probabilities of each possible linear combination and for ordering them. Decoding algorithms here provide good candidates for the most likely values of $\tilde{\mathbf{K}}$ with complexity $2^{\alpha d}$ with an α which is significantly smaller than 1 which depends on the decoding algorithm and on several parameters (d, n, the noise, the probability of failure of not producing the best candidate(s) we tolerate).

2 The probabilistic model

We review in this section the probabilistic model suggested in [BCQ04]. We denote by Σ the set of N plaintext-ciphertext pairs. The information available after the distillation phase is modeled as follows.

Model 1 — *The attacker receives a vector* $\mathbf{Y} = (Y_i)_{1 \le i \le n}$ *such that:*

$$\forall\, i \in \{1, \ldots, n\}, \quad Y_i = (-1)^{\tilde{K}_i} + N_i \quad , \quad N_i \sim \mathcal{N}(0, \sigma_i^2), \tag{2}$$

where $\sigma_i^2 \stackrel{def}{=} \frac{1}{4N\epsilon_i^2}$.

We denote by $f(\mathbf{Y}|\tilde{\mathbf{K}})$ *the density function of the variable* $\mathbf{Y}$ *conditioned by the value taken by* $\tilde{\mathbf{K}}$ *and* $f_i(Y_i \mid \tilde{K}_i)$ *denotes the density of the variable* Y_i *conditioned by* $\tilde{K}_i$.
These conditional densities satisfy the independence relation

$$f(\mathbf{Y} \mid \tilde{\mathbf{K}}) = \prod_{i=1}^{n} f(Y_i \mid \tilde{K}_i) \tag{3}$$

$\mathbf{Y}$ is derived from Σ as follows. We first define for every i in $\{1, \ldots, n\}$ and every j in $\{1, \ldots, N\}$ the following quantity $D_i^j \stackrel{\text{def}}{=} < \pi_i, \mathbf{P}^j > \oplus < \gamma_i, \mathbf{C}^j > \oplus b_i$, where the plaintext-ciphertext pairs in Σ are indexed by $(\mathbf{P}_1, \mathbf{C}_1), \ldots, (\mathbf{P}_N, \mathbf{C}_N)$ and b_i is the constant appearing in the i-th linear approximation. Then for all i in $\{1, \ldots, n\}$ we set up the counters D_i with $D_i \stackrel{\text{def}}{=} \sum_{j=1}^{N} D_i^j$ from which we build the vector of counters $\mathbf{D} = (D_i)_{1 \le i \le n}$. D_i is a binomial random variable which is approximately distributed as a normal law $\mathcal{N}((1/2 - \epsilon_i(-1)^{\tilde{K}_i})N, (1/4 - \epsilon_i^2)N)$. This explains why the vector $\mathbf{Y} = (Y_i)_{1 \le i \le n}$ is defined as:

$$Y_i \stackrel{\text{def}}{=} \frac{N - 2D_i}{2N\epsilon_i} \tag{4}$$

and why Equation (2) holds. There is some debate about the independence relation (3). This point is discussed by Murphy in [Mur06] where he proves that even if some key masks are linearly dependent, the independence relation (3) holds asymptotically if for a fixed key the covariances $\mathrm{cov}(D^j_{i_1}, D^j_{i_2}) \stackrel{\text{def}}{=} \mathbf{Pr}(D^j_{i_1} = D^j_{i_2} = 1) - \mathbf{Pr}(D^j_{i_1} = 1)\mathbf{Pr}(D^j_{i_2} = 1)$ are negligible. We have checked whether this holds in our experimental study. We had 129 linear approximations with biases in the range $[1.45.10^{-4}, 5.96.10^{-4}]$ and we found empirical covariances in the range $[-2.10^{-7}, 2.10^{-7}]$ for 10^{12} samples. This corroborates the fact that the covariances are negligible and that the independence relation (3) approximately holds.

3 Performing the linear attack by decoding a linear code

The problem of finding the most likely $\tilde{\mathbf{K}}$'s for a given $\mathbf{Y}$ is exactly the problem of decoding a linear code of dimension d (the maximum number of independent masks) of length n (the number of linear approximations) over an additive white Gaussian noise channel with BPSK modulation [**?**] with different values of the noise for each bit: the i-bit is transmitted through a Gaussian channel with noise variance σ_i^2. Many methods have been suggested to perform this task faster than calculating all 2^r probabilities $\mathbf{Pr}(\tilde{\mathbf{K}}|\mathbf{Y})$ (which represents the probability that $\tilde{\mathbf{K}}$ corresponds to the right key given that $\mathbf{Y}$ has been received) and outputting the most likely ones. We present in what follows an algorithm for performing this task which was proposed in [Val00]. Contrarily to other decoding algorithms such as [Dum00, VF04] there is no proof that this algorithm performs faster than exhaustive search over the whole key space, but this algorithm has the advantage to be simple to explain and quite efficient in practice. For this purpose, we need to define a few quantities. The *log-likelihood* $\mathcal{L}_{\mathbf{Y}}(\tilde{\mathbf{K}})$ of a key $\tilde{\mathbf{K}}$ for a received vector $\mathbf{Y}$ is defined by $\mathcal{L}_{\mathbf{Y}}(\tilde{\mathbf{K}}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (-1)^{\tilde{K}_i} Y_i/\sigma_i^2$. This definition comes from the following lemma.

Lemma 1 *For a given* $\mathbf{Y}$*, the probabilities* $\mathbf{Pr}(\tilde{\mathbf{K}}|\mathbf{Y})$ *are ordered in the same way as the log-likelihoods* $\mathcal{L}_{\mathbf{Y}}(\tilde{\mathbf{K}})$*.*

This lemma implies that in order to find the most likely $\tilde{\mathbf{K}}$ we have to find the $\tilde{\mathbf{K}}$'s with the largest log-likelihoods. Many algorithms for performing this task operate on the *quantified version* $\widehat{\mathbf{Y}} = (\widehat{Y}_i)_{1 \le i \le n}$ of $\mathbf{Y}$ which is a binary vector of length n defined by

$$\widehat{Y}_i \stackrel{\text{def}}{=} 0 \text{ if } Y_i \ge 0 \text{ and 1 otherwise.} \tag{5}$$

$\widehat{Y}_i$ can be viewed as the most likely value for $\tilde{\mathbf{K}}_i$ given Y_i. It is therefore natural to expect that the most likely $\tilde{\mathbf{K}}$ is not too far away from $\widehat{\mathbf{Y}}$ with respect to the Hamming distance. However, it is by no means obvious to produce efficiently possible values for $\tilde{\mathbf{K}}$ which are the closest to $\widehat{Y}$, the problem being that only a fraction 2^{d-n} of binary words of length n can be taken by $\tilde{\mathbf{K}}$. Nevertheless, it is easy to set up $\tilde{\mathbf{K}}$'s which are at a specified Hamming distance on a subset of d positions when this subset forms what is called an *information set*. This is by definition a subset of d positions on which the $\tilde{\mathbf{K}}$'s take all possible values among $\{0,1\}^d$. It is also defined equivalently by a set of r independent columns of the *generating matrix* G with d rows and n columns which is such that $\tilde{\mathbf{K}} = \mathbf{K}G$. The columns of this matrix correspond to the key masks κ_i: such a set of r independent columns corresponds to a set of r independent key masks. $\tilde{\mathbf{K}}$ is completely determined by the values it takes on an information set. More specifically, if we denote for subset $P \subset \{1, \ldots, n\}$ by

$\tilde{\mathbf{K}}_P$ the vector $(\tilde{\mathbf{K}}_i)_{i\in P}$ and by $\bar{P}$ the complementary set of P, then for any information set I there exists an $d \times (n-d)$ matrix G_I such that $\tilde{\mathbf{K}}_{\bar{I}} = \tilde{\mathbf{K}}_I G_I$.

A rather good heuristic would then be to find an information set I, use the previous facts about an information set for generating all $\tilde{\mathbf{K}}$'s such that the Hamming distance between $\tilde{\mathbf{K}}_I$ and $\widehat{\mathbf{Y}}_I$ is at most d for some small d and hope that the most likely $\tilde{\mathbf{K}}$'s are generated in this way. This basically amounts to generate $1 + \binom{n}{1} + \cdots + \binom{n}{d}$ candidates for $\tilde{\mathbf{K}}$. However, even if $\tilde{\mathbf{K}}_I$ and $\widehat{\mathbf{Y}}$ are close to each other, many of these $\tilde{\mathbf{K}}$'s generated in this way are rather far away from $\mathbf{Y}$. It is possible to avoid many useless $\tilde{\mathbf{K}}$'s by proceeding slightly differently and asking for $\tilde{\mathbf{K}}$'s such that the Hamming distance between $\widehat{\mathbf{Y}}_J$ and $\tilde{\mathbf{K}}_J$ is at most d where J is a set of positions of size $d+h$ (where h is much smaller than d) which contains an information set. The problem of generating such $\tilde{\mathbf{K}}$'s is solved by bringing in parity-check considerations.

This time all 2^{d+h} possible binary values can not be taken by the $\tilde{\mathbf{K}}$'s on J. There exists for such a set J of positions a parity-check matrix H_J of size $h \times (d+h)$ which can be obtained from G by Gaussian elimination such that a vector $\mathbf{X} \in \{0,1\}^{d+h}$ is a possible $\tilde{\mathbf{K}}_J$ if and only if $H_J \mathbf{X}^T = 0$. To produce efficiently such $\tilde{\mathbf{K}}$'s we split J in two halves $J = J_1 \cup J_2$ and generate all couples of error patterns $(\mathbf{E}_1, \mathbf{E}_2) \in \{0,1\}^{d+h} \times \{0,1\}^{d+h}$ such that (i) the support of $\mathbf{E}_i$ is included in J_i (ii) the Hamming weights of $\mathbf{E}_1$ and $\mathbf{E}_2$ are at most $d/2$, (iii)$H_J(\widehat{\mathbf{Y}} \oplus \mathbf{E}_1 \oplus \mathbf{E}_2)^T = 0$. This can be achieved efficiently by storing the values of $H_J \mathbf{E}_1^T$ in a hash table. The algorithm has a time complexity of $O\left(2^{-h}\binom{d+h}{d/2}^2\right)$ and a memory complexity of $O\binom{d+h}{d/2}$. A good idea is to choose h such that $\binom{d+h}{d/2} \simeq 2^h$. The final algorithm deals then with $O\binom{d+h}{d/2}$ possible values for $\tilde{\mathbf{K}}$. This algorithm can be described by:

INPUTS:
- J: a set of positions of size $k+h$,
- $\mathbf{Y}$: the vector to decode.

OUTPUT:
res: the result of the decoding procedure.

MAXIMUM LIKELIHOOD DECODING ALGORITHM($\mathbf{Y}$, J)

```
1  Calculate H_J the parity-check matrix restricted to J
2  Set S ← H_J Ŷ_J^T , vmax ← −∞ , res ← 0
3  Split J in two halves J_1 and J_2
4  Generate all error patterns E_1 of maximal weight d/2 on J_1
5  Store them in a table at the address H_J E_1^T
6  for ALL error patterns E_2 of maximal weight d/2 on J_2
7      do Look at the address S ⊕ (H_J E_2^T)
8          for ALL error pattern E_1 stored here
9              do Complete Ŷ_J ⊕ E_1 ⊕ E_2 in a vector K̃
10                 v ← L_Y(K̃)
11                 if v > vmax
12                    then vmax ← v and res ← K̃
13 Return res
```

d is generally chosen to be quite small and with rather large probability we might miss the most likely $\tilde{\mathbf{K}}$. To avoid that, the previous algorithm is applied on many sets of size $d+h$. Obviously, we should look for sets which contain the least number of quantifications errors (i.e. be such that

for most i in J, $\widehat{Y}_i$ coincides with $\tilde{K}'_i$ where $\tilde{\mathbf{K}}'$ denotes now the most likely value for $\tilde{\mathbf{K}}$). The best set of this kind corresponds by Formula (15) to the set J of $d+h$ positions with the largest $\sum_{i\in J} |Y_i|/\sigma_i^2$. One might be tempted to look first for this set, then for the set with the second largest sum and so on. This is not a very good idea, since these sets are generally very close to each other, and if $\tilde{\mathbf{K}}'$ has too many errors on a given set J of positions, then this may also be the case with a neighboring set. It is better to find a way to produce sets J of size $d+h$ with a large $\sum_{i\in J} |Y_i|/\sigma_i^2$ but which are sufficiently different from each other. A nice way to do this is to use stochastic resonance as Valembois proposed in his PhD thesis [Val00]:

INPUTS:
- *NbRounds* the number of sets J we are going to check.
- h the size of J minus r.
OUTPUT: **res** the result of the decoding algorithm.

STOCHASTIC RESONANCE DECODING ALGORITHM(h , *NbRounds*)

$vmax \leftarrow -\infty$, **res** $\leftarrow 0$
for j from 1 to *NbRounds*
  do for i from 1 to n
    do $Y'_i \leftarrow Y_i + N_i$, where $N_i \sim \mathcal{N}(0, \sigma_i^2/4)$
  Choose the set J as the $d+h$ indices with the largest $|Y'_i/\sigma_i^2|$
  if J contains an information set
    then $\tilde{\mathbf{K}} \leftarrow$ Maximum Likelihood Decoding Algorithm($\mathbf{Y}', J$)
      $v \leftarrow \mathcal{L}_{\mathbf{Y}}(\tilde{\mathbf{K}})$
      if $v > vmax$
        then $vmax \leftarrow v$ and **res** $\leftarrow \tilde{\mathbf{K}}$
Return **res**

4 Bounds on the required amount of plaintext-ciphertext pairs

4.1 An information-theoretic lower bound

The purpose of this subsection is to derive a general lower bound on the amount of uncertainty $\mathcal{H}(\mathbf{K}|\mathbf{Y})$ we have on the key given the statistics $\mathbf{Y}$ derived from the plaintext-ciphertext pairs. We recall that the (binary) *entropy* $\mathcal{H}(X)$ of a random variable X is given by the expression:

$$\begin{aligned}\mathcal{H}(X) &\stackrel{\text{def}}{=} -\sum_x \mathbf{Pr}(X=x)\log_2 \mathbf{Pr}(X=x) \text{ (for discrete } X)\\ &\stackrel{\text{def}}{=} -\int f(x)\log_2(f(x))dx \text{ (for continuous } X \text{ of density } f)\end{aligned} \tag{6}$$

For a couple of random variables (X,Y) we denote by $\mathcal{H}(X|Y)$ the *conditional entropy of X given Y*. It is defined by $\mathcal{H}(X|Y) \stackrel{\text{def}}{=} \sum_y \mathbf{Pr}(Y=y)\mathcal{H}(X|Y=y)$, where $\mathcal{H}(X|Y=y) \stackrel{\text{def}}{=} -\sum_x \mathbf{Pr}(X=x|Y=y)\log_2 \mathbf{Pr}(X=x|Y=y)$ when X and Y are discrete variables and when $\mathbf{Y}$ is a continuous random variable taking its values over $\mathbb{R}^n$ it is given by $\mathcal{H}(X|\mathbf{Y}) = \int_{\mathbb{R}^n} \mathcal{H}(X|\mathbf{Y}=\mathbf{y})f(\mathbf{y})d\mathbf{y}$, where $f(\mathbf{y})$ is the density of the distribution of $\mathbf{Y}$ at the point $\mathbf{y}$. A related quantity is the *mutual*

information $\mathcal{I}(X;Y)$ *between* X and Y which is defined by

$$\mathcal{I}(X;Y) \stackrel{\text{def}}{=} \mathcal{H}(X) - \mathcal{H}(X|Y). \tag{7}$$

It is straightforward to check [CT91] that this quantity is symmetric and that

$$\mathcal{I}(X;Y) = \mathcal{I}(Y;X) = \mathcal{H}(Y) - \mathcal{H}(Y|X). \tag{8}$$

We will be interested in deriving a lower bound on $\mathcal{H}(\mathbf{K}'|\mathbf{Y})$ when $K' = (K'_1, \ldots, K'_n)$ is a subkey derived from $\mathbf{K}$ which satisfies:

(i) (conditional independence assumption)

$$f(\mathbf{Y} \mid \mathbf{K}') = \prod_{i=1}^{n} f(Y_i \mid K'_i), \tag{9}$$

where $f(\mathbf{Y}|\mathbf{K}')$ is the density function of the variable $\mathbf{Y}$ conditioned by the value taken by $\mathbf{K}'$ and $f_i(Y_i \mid K'_i)$ denotes the density of the variable Y_i conditioned by K_i.

(ii) $\mathbf{K}'$ may take $2^{k'}$ values and all are equally likely.

With these assumptions we have the following result

Lemma 2

$$I(\mathbf{K}';\mathbf{Y}) \leq \sum_{i=1}^{n} I(K_i;Y_i) \tag{10}$$

$$\mathcal{H}(\mathbf{K}'|\mathbf{Y}) \geq k' - \sum_{i=1}^{n} I(K_i;Y_i). \tag{11}$$

The proof is this lemma can be found in the appendix. It will be used in what follows in various scenarios for linear attacks, but it can obviously be used to cover many other cryptographic attacks. This lower bound is in general amazingly sharp as long as it is non-trivial, i.e when $k' \geq \sum_{i=1}^{n} I(K'_i;Y_i)$. We will prove this for one the application in what follows but this can also be done for the other cases. In what follows, when K is a discrete random variable and Y is a continuous one such that the conditional disributions of Y given K have density $f(Y|K)$ it will be convenient to use the following formula for the mutual information

$$I(K;Y) = \sum_{k} \mathbf{Pr}(K=k) \int f(y|k) \log \frac{f(y|k)}{\sum_k f(y|k)} dy. \tag{12}$$

4.2 Application to various scenarios

Attack 1 : It corresponds to the case where we do not use the linear equations as distinguishers but only want to recover the $< \kappa_i, \mathbf{K} >$'s. This corresponds in the case of a single equation to

Matsui's attack 1 and in the case of multiple equations to the attack MK1 in [BCQ04]. We have here

$$K'_i = < \kappa_i, \mathbf{K} >$$
$$Y_j = \frac{N - 2D_j}{2N\epsilon_j}.$$

$\mathbf{K}'$ and $\mathbf{Y}$ satisfy the required conditional independence assumption (see Equation 3) and a straightforward calculation using Formula (12) yields

$$I(K'_i; Y_i) = \mathbf{Cap}(\sigma_j^2)$$

where

$$\mathbf{Cap}(\sigma^2) \stackrel{\text{def}}{=} 1 - \frac{\sigma e^{-\frac{1}{2\sigma^2}}}{\sqrt{8\pi}} \int_{\infty}^{\infty} e^{-\frac{u^2\sigma^2}{8}} e^{\frac{u}{2}} \log_2\left(1 + e^{-u}\right) du.$$

and therefore by applying Lemma 2 we obtain

$$\mathcal{H}(\mathbf{K}'|\mathbf{Y}) \geq d - \sum_{j=1}^{n} \mathbf{Cap}(\sigma_j^2) \tag{13}$$

Attack 2: We consider a distinguisher attack where the approximate linear equations of the form (1) correspond to a first and last round reduced cipher. This means that they apply to pairs $(\mathbf{P}, \mathbf{C})$ for which $\mathbf{P}$ is the encrypted version of the plaintext with the first round key $\mathbf{K}_{\text{first}}$ and $\mathbf{C}$ being the inverse of the ciphertext corresponding to the last round key $\mathbf{K}_{\text{last}}$. The $< \pi_i, \mathbf{P} >$'s and the $< \gamma_i, \mathbf{C} >$'s might not depend on all the bits of $\mathbf{K}_{\text{first}}$ and $\mathbf{K}_{\text{last}}$. We denote by K' the vector formed by the bits of $\mathbf{K}_{\text{first}}$ and $\mathbf{K}_{\text{last}}$ on which the $< \pi_i, \mathbf{P} >$'s and the $< \gamma_i, \mathbf{C} >$'s depend. We assume that K' may take $2^{k'}$ values. We define $\mathbf{K}'$ by the vector $(K'_i)_{i=1}^n$ such that $K'_i = K'$ for all i's. We also assume that we make no assumption on the $< \kappa_i, \mathbf{K} >$'s (or consider all possible values for these quantities) and we just want to recover K' based on the values of the counters D_j^z for j in $\{1, \ldots, n\}$ and z ranging over all possible values for K'. These counters are defined similarly as in Section 2 with the difference being that we use the value $K' = z$ for deriving the relevant couples $(\mathbf{P}, \mathbf{C})$. The statistics $\mathbf{Y} = (Y_j)_{1 \leq j \leq n}$ we consider in this case is given by $Y_j \stackrel{\text{def}}{=} (Y_j^z)_z$ with $Y_j^z = \frac{|N - 2D_j^z|}{2N\epsilon_j}$. Again the conditional independence relation (3) is also satisfied in this case. With the help of Lemma 2, we can write $H(\mathbf{K}'|\mathbf{Y}) \geq k' - \sum_{i=1}^n I(K'; Y_i)$. We can again use Lemma 2 and obtain $I(K'; Y_i) \leq \sum_z I(K'; Y_i^z)$. The variable Y_j^z has density r_j if z corresponds to the right choice for K' and w_j otherwise, where $r_j(t) = \phi_j^1(t) + \phi_j^{-1}(t)$, $w_j(t) = 2\phi_j^0(t)$ for nonnegative t with $\varphi_j^{\alpha}(t) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(t-\alpha)^2}{2\sigma_j^2}}$ being the density of a normal variable of expectation α and variance σ_j^2. A straightforward application of Formula (12) gives

$$I(K'_j; Y_j) = \int_0^{+\infty} \frac{r_j(t)}{2^{k'}} \log\left(\frac{r_j(t)}{s_j(t)}\right) dt + \int_0^{\infty} (1 - 2^{-k'}) w_j(t) \log\left(\frac{w_j(t)}{s_j(t)}\right) dt, \tag{14}$$

with $s_j(t) \stackrel{\text{def}}{=} 2^{-k'} r_j(t) + (1 - 2^{-k'}) w_j(t)$. We denote this quantity by I_j and we finally obtain

$$H(\mathbf{K}'|\mathbf{Y}) \geq k' - 2^{k'} \sum_{j=1}^{n} I_j.$$

Attack 3: This corresponds to the attack MK2 in [BCQ04] which is a variation of the previously seen distinguisher attack. In this case, we wish to find simultaneously the K' defined in Attack 2 and also the vector $K'' = (< \kappa_i, \mathbf{K} >)_{1 \leq i \leq n}$. In this case, we let $K''' = (K', K'')$ and define $\mathbf{K'} \stackrel{\text{def}}{=} (K'_i)_{1 \leq i \leq n}$ by $K'_i = K'''$ for every i. We assume that $2^{k'''}$ is the number of all possible values for K''' and that $2^{k'}$ is the number of all possible values for K'. Here, we define the relevant statistics $\mathbf{Y} = (Y_i)_{1 \leq i \leq n}$ by $Y_i = (Y_i^z)_z$ where z ranges over all possible values for K' and where $Y_j^z = \frac{N - 2D_j^z}{2N\epsilon_j}$. We have again the desired independence relation (3) and as in the previous example we can use Lemma 2 twice to obtain

$$\mathcal{H}(\mathbf{K'}|\mathbf{Y}) \geq k''' - \sum_{j=1}^{n} I(K'''; Y_j) \geq k''' - 2^{k'} \sum_{j=1}^{n} I(K'''; Y_j^z)$$

A straightforward application of Formula (12) yields

$$I(K'''; Y_j^z) = \int_{-\infty}^{\infty} \frac{\varphi_{-1}(t)}{2^{k'}} \log\left(\frac{\varphi_{-1}(t)}{\psi_j(t)}\right) dt + \int_{-\infty}^{\infty} (1 - 2^{-k'}) \varphi_0(t) \log\left(\frac{\varphi_0(t)}{\psi_j(t)}\right) dt,$$

with $\varphi_j^{\alpha}(t)$ defined as in Attack 2 and $\psi_j(t) \stackrel{\text{def}}{=} (1 - 2^{-k'})\varphi_j^0(t) + 2^{-k'-1}[\varphi_j^{-1}(t) + \varphi_j^1(t)]$.

4.3 An upper bound

One might wonder whether or not the bounds given in the previous subsection are sharp or not. It is clear that these lower bounds become negative when the number of plaintext-ciphertext pairs is large enough and that they are worthless in this case (since mutual information is always nonnegative). However in all three cases it can be proved that as long the bound is non trivial it is quite sharp. We will prove this for the lower-bound (13). Similar techniques can be used for the other bounds but it would be too long to include them in this paper. To study how sharp (13) is we will consider the case when

$$\sum_{i=1}^{n} \mathbf{Cap}(\sigma_i^2) \approx d?$$

If the lower-bound is sharp, one might be tempted to say that the conditional entropy of $\mathbf{K'}$ given $\mathbf{Y}$ should be close to 0 which would mean that $\mathbf{K'}$ is determined from $\mathbf{Y}$ with probability close to 1. This is of course not always true, but it is the case *for most choices* of the coefficients κ_i^j. To give a precise meaning to this statement we will first consider what happens when the κ_i^j's are chosen *at random.*

Theorem 1 *Assume that the κ_i^j are chosen chosen uniformly at random and that $\sum_{i=1}^{n} \mathbf{Cap}(\sigma_i^2) \geq d + \delta n$ for some constant $\delta > 0$. Let P_{err} be the probability that the most likely value for $\mathbf{K'}$ given $\mathbf{Y}$ is not the right one. There exists a constant A such that*

$$P_{err} \leq \frac{A}{\delta^2 n} + 2^{-\delta n/2}.$$

The probability P_{err} is taken over $\mathbf{Y}$ but also over the choices of the κ_i^j's. It says nothing about a particular choice of the κ_i^j's. However it implies the aforementioned assertion about most choices of the κ_i's. Let us be more specific by bringing in $P_{\text{err}}(\mathcal{C})$ which is the probability that the most

likely key given $\mathbf{Y}$ is not the right one when the subspace of dimension d of the possible values for $\tilde{\mathbf{K}}$ is $\mathcal{C}$. A bound on P_{err} implies that for most choices of the κ_i's (and hence of $\mathcal{C}$) $P_{\text{err}}(\mathcal{C})$ is small by using the following lemma

Lemma 3 *Assume that $P_{err} \leq \epsilon$. Then for any $t > 0$:*

$$\mathbf{Pr}_{\mathcal{C}}\left(P_{err}(\mathcal{C}) \geq t\epsilon\right) \leq \frac{1}{t}$$

Proof. Let $P \stackrel{\text{def}}{=} \mathbf{Pr}_{\mathcal{C}}\left(P_{\text{err}}(\mathcal{C}) \geq t\epsilon\right)$. We observe that $P_{\text{err}} = \sum_{\mathcal{C}} P_{\text{err}}(\mathcal{C})\mathbf{Pr}(\mathcal{C})$ $\geq Pt\epsilon$. This implies that $P \leq \frac{1}{t}$. ∎

Remark: The notation $\mathbf{Pr}_{\mathcal{C}}$ means here that the probability is taken over the choices for $\mathcal{C}$. It actually denotes the proportion of choices for $\mathcal{C}$ which lead to the specified event inside the probability.

4.4 Entropy vs. expected number of $\tilde{\mathbf{K}}$'s which are more likely than the right one

The aim of this subsection is to explain that in a certain range of values of N (which is the number of plaintext-ciphertext pairs) the expected size $\mathcal{E}$ of the list of the $\tilde{\mathbf{K}}$'s which are more likely than the right one gives pessimistic estimates of the amount of plaintext-ciphertext pairs we need to mount an attack. We illustrate this with the following example. We assume that we have n linear approximations which are all with biases $\epsilon = 10^{-8}$ and that the dimension of the key masks is $d = n/2$. Let $\sigma^2 \stackrel{\text{def}}{=} \frac{1}{4N\epsilon^2}$. If these approximations behave like random approximations we expect from Theorem 1 that as soon $\mathbf{Cap}(\sigma^2)$ gets slightly larger than $\frac{1}{2}$ the probability that the most likely $\tilde{\mathbf{K}}$ is the right one approaches 1 as n goes to infinity. This is the case for σ^2 below 0.958, that is for $N \approx 2^{51}$.

On the other hand, it can be proved by classical calculations (see for instance [RU, exercise 1.21]) that for random linear key masks, $\mathcal{E}$ is at least of order $A\frac{(1+e^{-\frac{1}{2\sigma^2}})^n-1}{n2^{n/2}}$ for some constant A. This quantity is exponential large in n for σ^2 in the range $0.567 - 0.958$, that is for N in the range $2^{51} - 2^{52}$. This seems surprising because in this range the most likely $\tilde{\mathbf{K}}$ is with probability going to 1 the right one. It can be checked that this is due to the following phenomenon: most of the time the list of $\tilde{\mathbf{K}}$'s which are more likely than the right one is empty, but with exponentially small probability this list is of much larger size (and is exponentially large).

5 Experimental Results

We have benchmarked our algorithm on DES reduced to 8 rounds. We have used 129 linear approximations for the 8 rounds of DES which stem from the work of Loidreau and Tavernier [LT07]. The biases of all these approximations are in the range $1.45.10^{-4} - 5.96.10^{-4}$.

These equations can be split into two groups of 55 and 74 equations. The first group involves 9 key bits, the second 13 and there are 2 key bits which are involved in both groups. We have computed the amount of plaintext for which the bound in Theorem **??** becomes zero. It corresponds to $N \approx 2^{19.49}$ for the first group of equations and it corresponds to $N \approx 2^{19.84}$ for the second group.

The quality of the lower-bound of Theorem **??** can be checked by estimating empirically the entropy. Figure 1 displays the empirical conditional entropies of $\tilde{\mathbf{K}}$ given $\mathbf{Y}$ for both groups of equations. There is an excellent agreement between the lower bound and the empirical entropies up to when we approach the critical value of N for which the lower bound is equal to zero. This kind of lower bound is really suited to the case when the amount of plaintext-ciphertext pairs is some order of magnitude below this critical value. This is typically the case in attacks when we want to decrease the amount of plaintext-ciphertext pairs as much as possible at the expense of keeping a list of possible candidates for $\tilde{\mathbf{K}}$'s.

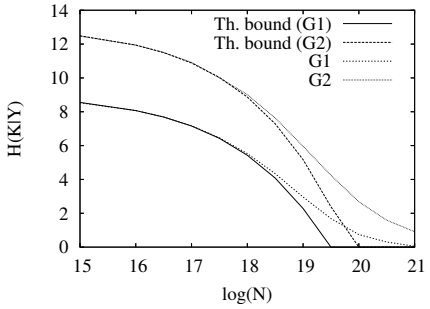

Figure 1:

We have implemented the linear attack described in Section 3 with these two groups of equations. Figure 2 displays the success rate as a function of $\log_2(N)$. We count here an attack as successful when the best $\tilde{\mathbf{K}}$ obtained by the decoding algorithm corresponds to the right key. We also show in the same graph the joint probability of success of the attack (i.e. when the two algorithms give the right answer together).

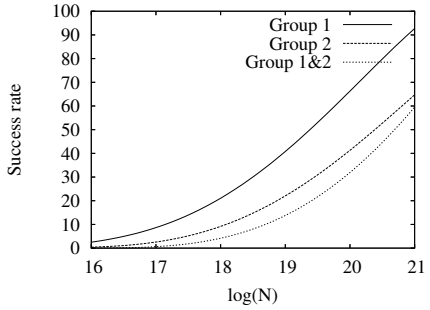

Figure 2:

There are many ways of improving the success rate of the algorithm. For instance, we may keep a list of the best candidates for $\tilde{\mathbf{K}}$ for both groups of equations and merge both lists in a single (final) list. For example, for 2^{19} plaintext-ciphertext pairs we have plotted the success rate of this procedure (where we say that the final list is a success iff it contains the right $\tilde{\mathbf{K}}$) against the size of the final list of candidates for $\tilde{\mathbf{K}}$ and obtain Figure 3 (where the size of the two intermediate lists has been fixed to 16).

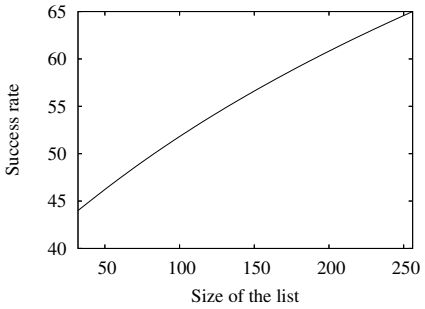

Figure 3:

References

[BCQ04] Alex Biryukov, Christophe De Cannière, and Michaël Quisquater. On Multiple Linear Approximations. In *CRYPTO 2004*, LNCS, pages 1–22. Springer–Verlag, 2004.

[BSKR94] Jr. Burton S. Kaliski and M. J. B. Robshaw. Linear Cryptanalysis Using Multiple Approximations. In *CRYPTO '94*, LNCS, pages 26–39. Springer–Verlag, 1994.

[CT91] T.M. Cover and J.A. Thomas. *Information theory*. Wiley series in communications. Wiley, 1991.

[Dum00] I. Dumer. Ellipsoid lists and maximum-likelihood decoding. *IEEE Trans. Info. Theory*, 46(2):649–656, March 2000.

[JV03] P. Junod and S. Vaudenay. Optimal key ranking procedures in a statistical cryptanalysis. In *FSE 2003*, LNCS, pages 235–246. Springer–Verlag, 2003.

[LT07] Pierre Loidreau and Cédric Tavernier. An algorithm for finding linear approximations and its application to 8-round of DES, 2007.

[Mat93] Mitsuru Matsui. Linear cryptanalysis method for DES cipher. In *EUROCRYPT '93*, LNCS, pages 386–397. Springer–Verlag, 1993.

[Mat94] Mitsuru Matsui. The First Experimental Cryptanalysis of the Data Encryption Standard. In *CRYPTO '94*, LNCS, pages 1–11. Springer–Verlag, 1994.

[MM92] Atsuhiro Yamagishi Mitsuru Matsui. A New Method for Known Plaintext Attack of FEAL Cipher. In *EUROCRYPT '92*, LNCS, page 81. Springer–Verlag, 1992.

[MPWW95] S. Murphy, F. Piper, M. Walker, and P. Wild. Likelihood estimation for block cipher keys, 1995.

[Mur06] Sean Murphy. The Independence of Linear Approximations in Symmetric Cryptology, 2006.

[OA94] Kazuo Ohta and Kazumaro Aoki. Linear Cryptanalysis of the Fast Data Encipherment Algorithm. *LNCS*, 839:12–16, 1994.

[PH98] V.S. Pless and W.C. Huffman, editors. *Handbook of coding theory*. North Holland, 1998.

[RU] T. Richardson and R. Urbanke. Modern coding theory. In preparation. see http://lthcwww.epfl.ch/papers/ics.ps.

[tB01] S. ten Brink. Convergence behaviour of iteratively decoded parallel concatenated code. *IEEE Trans. Commun.*, 49:1727–1737, Oct. 2001.

[TCG91] Anne Tardy-Corfdir and Henri Gilbert. A Known Plaintext Attack of FEAL-4 and FEAL-6. In *CRYPTO '91*, LNCS, pages 172–181. Springer–Verlag, 1991.

[TSM94] Toshio Tokita, Tohru Sorimachi, and Mitsuru Matsui. Linear Cryptanalysis of LOKI and s2DES. In *ASIACRYPT '94*, LNCS, pages 293–303. Springer–Verlag, 1994.

[Val00] Antoine Valembois. *Détection, Reconnaissance et Décodage des Codes Linéaires Binaires*. PhD thesis, Université de Limoges, 2000.

[Vau96] Serge Vaudenay. An Experiment on DES Statistical Cryptanalysis. In *ACM Conference on Computer and Communications Security*, pages 139–147, 1996.

[VF04] A. Valembois and M. Fossorier. Box and match techniques applied to soft-decision decoding. *IEEE Trans. Inform. Theory*, 50(5):796–810, May 2004.

A Proofs

A.1 Proof of lemma 1

First of all, note that for a given $\mathbf{Y}$, the probabilities $\mathbf{Pr}(\tilde{\mathbf{K}}|\mathbf{Y})$ are ordered in the same way as the $f(\mathbf{Y}|\tilde{\mathbf{K}})$'s since $\mathbf{Pr}(\tilde{\mathbf{K}}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\tilde{\mathbf{K}})\Pr(\tilde{\mathbf{K}})}{f(\mathbf{Y})} = \frac{f(\mathbf{Y}|\tilde{\mathbf{K}})}{2^r f(\mathbf{Y})}$. The logarithm is an increasing function therefore the $\ln f(\mathbf{Y}|\tilde{\mathbf{K}})$'s are also ordered in the same way. From the independence relation (3) we know that $\ln f(\mathbf{Y}|\tilde{\mathbf{K}}) = \sum_{i=1}^{n} \ln f_i(Y_i|\tilde{K}_i)$. All this implies that the probabilities $\mathbf{Pr}(\tilde{\mathbf{K}}|\mathbf{Y})$ are ordered in the same way as the sums $\sum_{i=1}^{n} \ln f_i(Y_i|\tilde{K}_i)$. We notice now that

$$\ln\left(\frac{f_i(Y_i|\tilde{K}_i=1)}{f_i(Y_i|\tilde{K}_i=0)}\right) = \ln\left(\frac{e^{-(Y_i+1)^2/(2\sigma_i^2)}/\sqrt{2\pi\sigma_i^2}}{e^{-(Y_i-1)^2/(2\sigma_i^2)}/\sqrt{2\pi\sigma_i^2}}\right) = -\frac{2Y_i}{\sigma_i^2} \tag{15}$$

We finish the proof by

$$\begin{aligned}\sum_{i=1}^{n}(-1)^{\tilde{K}_i}Y_i/\sigma_i^2 &= \tfrac{1}{2}\sum_{i:\tilde{K}_i=1}\ln\left(\tfrac{f_i(Y_i|\tilde{K}_i=1)}{f_i(Y_i|\tilde{K}_i=0)}\right) - \tfrac{1}{2}\sum_{i:\tilde{K}_i=0}\ln\left(\tfrac{f_i(Y_i|\tilde{K}_i=1)}{f_i(Y_i|\tilde{K}_i=0)}\right)\\ &= \sum_{i=1}^{n}\ln f_i(Y_i|\tilde{K}_i) - \tfrac{1}{2}\sum_{i=1}^{n}\ln f_i(Y_i|\tilde{K}_i=0) - \tfrac{1}{2}\sum_{i=1}^{n}\ln f_i(Y_i|\tilde{K}_i=1)\end{aligned}$$

This shows that the sums $\sum_{i=1}^{n}(-1)^{\tilde{K}_i}Y_i/\sigma_i^2$ are ordered in the same way as the sums $\sum_{i=1}^{n}\ln f_i(Y_i|\tilde{K}_i)$.

A.2 Proof of Lemma 2

Let us use Equation (8) and write in two different ways the mutual information between $\mathbf{K}'$ and $\mathbf{Y}$: $\mathcal{I}(\mathbf{K}';\mathbf{Y}) = \mathcal{H}(\mathbf{K}') - \mathcal{H}(\mathbf{K}'|\mathbf{Y}) = \mathcal{H}(\mathbf{Y}) - \mathcal{H}(\mathbf{Y}|\mathbf{K}')$. From this we deduce that

$$\begin{aligned}\mathcal{H}(\mathbf{K}'|\mathbf{Y}) &= \mathcal{H}(\mathbf{K}') - \mathcal{H}(\mathbf{Y}) + \mathcal{H}(\mathbf{Y}|\mathbf{K}') \\ &= k' - \mathcal{H}(Y_1,\ldots,Y_n) + \mathcal{H}(Y_1,\ldots,Y_n|\mathbf{K}'). \end{aligned} \tag{16}$$

Here Equation (16) is a consequence of the fact that the a priori distribution over $\mathbf{K}'$ is the uniform distribution and the entropy of a discrete random variable which is uniformly distributed is obviously nothing but the logarithm of the number of values it can take. Moreover (see [CT91, Theorem 2.6.6])

$$\mathcal{H}(Y_1,\ldots,Y_n) \leq \mathcal{H}(Y_1) + \cdots + \mathcal{H}(Y_n). \tag{17}$$

On the other hand, by the chain rule for entropy [CT91, Theorem 2.5.1]:

$$\mathcal{H}(Y_1,\ldots,Y_n|\mathbf{K}') = \mathcal{H}(Y_1|\mathbf{K}') + \mathcal{H}(Y_2|Y_1,\mathbf{K}') + \cdots + \mathcal{H}(Y_n|\mathbf{K}',Y_1,Y_2,\ldots,Y_{n-1}). \tag{18}$$

We notice now that $\mathcal{H}(Y_i|\mathbf{K}',Y_1\ldots Y_{i-1})$ can be written as

$$\textstyle\sum_{\mathbf{k}} \int_{\mathbb{R}^{i-1}} \mathcal{H}(Y_i|\mathbf{K}'=\mathbf{k},Y_1=y_1,\ldots,Y_{i-1}=y_{i-1}) f(y_1,\ldots,y_{i-1}|\mathbf{K}'=\mathbf{k})\mathbf{Pr}(\mathbf{K}'=\mathbf{k}) dy_1\ldots dy_{i-1}, \tag{19}$$

where the sum is taken over all 2^r possible values $\mathbf{k}$ of $\mathbf{K}$ and $f(y_1,\ldots,y_{i-1}|\mathbf{K}'=\mathbf{k})\mathbf{Pr}(\mathbf{K}'=\mathbf{k})$ is the density of the distribution of the vector $(Y_1,\ldots,Y_{i-1})$ given the value $\mathbf{k}$ of $\mathbf{K}'$ at the point $(y_1,\ldots,y_{i-1})$. From conditional independence assumption (9) we deduce that $\mathcal{H}(Y_i|\mathbf{K}'=\mathbf{k}, Y_1 = y_1,\ldots,Y_{i-1}=y_{i-1}) = \mathcal{H}(Y_i|K'_i)$. By summing in Expression (19) over $y_1,\ldots,y_{i-1}$ and all possible values of $K'_1,\ldots,K'_{i-1},K'_{i+1},\ldots,K'_n$ we obtain that

$$\mathcal{H}(Y_i|\mathbf{K}',Y_1,\ldots,Y_{i-1}) = \frac{1}{2}\mathcal{H}(Y_i|K'_i=0) + \frac{1}{2}\mathcal{H}(Y_i|K'_i=1) = \mathcal{H}(Y_i|K'_i=k_i) \tag{20}$$

Plugging in this last expression in Expression (18) we obtain that

$$\mathcal{H}(Y_1,\ldots,Y_n|K'_1,\ldots,K'_n) = \mathcal{H}(Y_1|K'_1) + \cdots + \mathcal{H}(Y_n|K'_n). \tag{21}$$

Using this last equation and Inequality (17) in (16) we finally deduce that

$$\begin{aligned}\mathcal{H}(\mathbf{K}'|\mathbf{Y}) &\geq r - \mathcal{H}(Y_1) - \cdots - \mathcal{H}(Y_n) + \mathcal{H}(Y_1|K'_1) + \cdots + \mathcal{H}(Y_n|K'_n) \\ &\geq k' - \sum_{i=1}^{n} \mathcal{H}(Y_i) - \mathcal{H}(Y_i|K'_i) \\ &\geq k' - \sum_{i=1}^{n} \mathcal{I}(K'_i;Y_i). \end{aligned} \tag{22}$$

A.3 Proof of Theorem 1

The proof of this theorem follows closely standard proofs of the direct part of Shannon's channel capacity theorem [CT91], however most of the proofs given for this theorem are asymptotic in nature and are not suited to our case. There are proofs which are not asymptotic, but they are

tailored for the case where all the σ_i's are equal and are rather involved. We prefer to follow a slightly different path here. The first argument we will use is an explicit form of the joint AEP (Asymptotic Equipartition Property) theorem.

For this purpose, we denote by $(\mathbf{X}, \mathbf{Y})$ a couple of random variables where $\mathbf{X} = (X_i)_{1 \leq i \leq n}$ is uniformly distributed over $\{0,1\}^n$ and $\mathbf{Y} = (Y_i)_{1 \leq i \leq n}$ is the output of the Gaussian channel described in Section 2 when $\mathbf{X}$ is sent through it. This means that

$$Y_i = (-1)^{X_i} + N_i, \tag{23}$$

where the N_i are independent centered normal variables of variance σ_i^2.

Let us first bring in the following definition.

Definition 1 *For $\epsilon > 0$, we define the set T_ϵ of ϵ-jointly typical sequences of $\{0,1\}^n \times \mathbb{R}^n$ by $T_\epsilon \stackrel{def}{=} \bigcup_{\mathbf{x} \in \{0,1\}^n} \{\mathbf{x}\} \times T_\epsilon(\mathbf{x})$ with*

$$T_\epsilon(\mathbf{x}) \stackrel{def}{=} \{\mathbf{y} \in \mathbb{R}^n : |-\log_2(f(\mathbf{y})) - \mathcal{H}(\mathbf{Y})| < n\epsilon \tag{24}$$

$$\left|-\log_2\left(f(\mathbf{y}|\mathbf{x})2^{-n}\right) - \mathcal{H}(\mathbf{X}, \mathbf{Y})\right| < n\epsilon\} \tag{25}$$

where $f(\mathbf{y})$ is the density distribution of $\mathbf{Y}$ and $f(\mathbf{y}|\mathbf{x})$ is the density distribution of $\mathbf{Y}$ given that $\mathbf{X}$ is equal to $\mathbf{x}$.

The entropies of $\mathbf{Y}$ and $(\mathbf{X}, \mathbf{Y})$ are given by the following expressions

Lemma 4

$$\mathcal{H}(\mathbf{Y}) = \sum_{i=1}^{n} \mathbf{Cap}(\sigma_i^2) + \frac{1}{2}\log_2(2\pi e \sigma_i^2)$$

$$\mathcal{H}(\mathbf{X}, \mathbf{Y}) = n + \sum_{i=1}^{n} \frac{1}{2}\log_2(2\pi e \sigma_i^2)$$

Proof. Notice that with our model the Y_i's are independent. Therefore $\mathcal{H}(\mathbf{Y}) = \sum_{i=1}^n \mathcal{H}(Y_i)$. Moreover, by the very definitions of entropy and mutual information: $\mathcal{H}(Y_i) = \mathcal{H}(Y_i|X_i) + \mathcal{I}(X_i; Y_i)$; X_i is uniformly distributed over $\{0,1\}$ and therefore by the definition of the capacity of a Gaussian channel and the fact that the capacity attains its maximum for a binary input which is uniformly distributed we have $\mathcal{I}(X_i; Y_i) = \mathbf{Cap}(\sigma_i^2)$. On the other hand $\mathcal{H}(Y_i|X_i)$ is obviously the same as $\mathcal{H}(N_i)$. The calculation of this entropy is standard (see [CT91]) and gives

$$\mathcal{H}(N_i) = \frac{1}{2}\log_2(2\pi e \sigma_i^2) \tag{26}$$

By putting all these facts together we obtain the expression for $\mathcal{H}(\mathbf{Y})$. Concerning the other entropy, with similar arguments we obtain

$$\begin{aligned}
\mathcal{H}(\mathbf{X}, \mathbf{Y}) &= \mathcal{H}(\mathbf{X}) + \mathcal{H}(\mathbf{Y}|\mathbf{X}) \\
&= n + \sum_{\mathbf{x} \in \{0,1\}^n} \frac{1}{2^n} \mathcal{H}(\mathbf{Y}|\mathbf{X} = \mathbf{x}) \\
&= n + \sum_{\mathbf{x} \in \{0,1\}^n} \frac{1}{2^n} \mathcal{H}(N_1, \ldots, N_n) \\
&= n + \sum_{i=1}^{n} \frac{1}{2}\log_2(2\pi e \sigma_i^2)
\end{aligned}$$

∎

"T_ϵ" stands for "typical set" since it is highly unlikely that $(\mathbf{X}, \mathbf{Y})$ does not belong to T_ϵ:

Lemma 5 *There exists a constant A such that*

$$\mathbf{Pr}\left((\mathbf{X}, \mathbf{Y}) \notin T_\epsilon\right) \leq \frac{A}{\epsilon^2 n}.$$

Before giving the proof of this lemma we will first give an interpretation of entropy which provides an explanation of why the probability of falling outside the typical set becomes smaller as n increases.

Lemma 6 *Let $U_i \stackrel{def}{=} -\log_2 f_i(Y_i)$ where f_i is the following probability density $f_i(y) \stackrel{def}{=} \frac{1}{2\sqrt{2\pi\sigma_i^2}}\left(e^{-\frac{(y-1)^2}{2\sigma_i^2}} + e^{-\frac{(y+1)^2}{2\sigma_i^2}}\right)$. We also denote by $V_i \stackrel{def}{=} -\log_2\left(\frac{g_i(Y_i-(-1)^{X_i})}{2}\right)$ where g_i is the density distribution of a centered Gaussian variable of variance σ_i^2.*

$$-\log_2(f(\mathbf{Y})) - \mathcal{H}(\mathbf{Y}) = \sum_{i=1}^n U_i - \mathbb{E}\left(\sum_{i=1}^n U_i\right)$$
$$-\log_2(f(\mathbf{Y}|\mathbf{X})2^n) - \mathcal{H}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^n V_i - \mathbb{E}\left(\sum_{i=1}^n V_i\right)$$

Proof. For the first equation we just have to notice that

$$-\log_2(f(\mathbf{Y})) = -\log_2\left(\Pi_{i=1}^n f_i(Y_i)\right) = -\sum_{i=1}^n \log_2(f_i(Y_i)) = \sum_{i=1}^n U_i$$

and that $\mathcal{H}(\mathbf{Y}) = \mathbb{E}(-\log_2 f(\mathbf{Y}))$, which follows directly from the definition of the entropy given in (6). The second equation can be obtained in a similar way. ∎

This implies that in order to estimate the probability that a point falls outside the typical set we have to estimate the probability that the deviation between a sum of n independent random variables and its expectation is at least of order ϵn. In our case, it can be proven that for fixed ϵ, this probability is exponentially small in n. However, we prefer to give a much weaker statement which is also easier to prove and which uses only Chebyschev's inequality, which we recall here

Lemma 7 *Consider a real random variable X of variance* $\mathrm{var}(X)$*. We have for any $t > 0$:*

$$\mathbf{Pr}\left(|X - \mathbb{E}(X)| \geq t\right) \leq \frac{\mathrm{var}(X)}{t^2}. \tag{27}$$

To use this inequality we have to estimate the variances of the U_i's and the V_i's. It can be checked that

Lemma 8 *There exists a constant A such that for any i we have*

$$\mathrm{var}(V_i) \leq A$$
$$\mathrm{var}(U_i) \leq A.$$

Proof. Let us prove the first statement. Recall that from (23): $N_i = Y_i - (-1)^{X_i}$.

$$\bar{V_i} \stackrel{\text{def}}{=} V_i - \mathbb{E}(V_i) = -\log_2\left(\frac{g_i(N_i)}{2}\right) - \mathbb{E}\left(\left(-\log_2\left(\frac{N_i}{2}\right)\right) = -\log_2\left(g_i(N_i)\right) - \frac{1}{2}\log_2(2e\pi\sigma_i^2).\right.$$

where the last equation follows from Expression (26). Hence:

$$\bar{V_i} = \log_2(e)\frac{N_i^2}{2\sigma_i^2} + \frac{1}{2}\log_2(2\pi\sigma_i^2) - \frac{1}{2}\log_2(2e\pi\sigma_i^2) = \frac{\log_2(e)}{2}\left(\frac{N_i^2}{\sigma_i^2} - 1\right),$$

and therefore

$$\begin{aligned}\text{var}(V_i) \stackrel{\text{def}}{=} \mathbb{E}\left[\bar{V_i}^2\right] &= \frac{\log_2(e)^2}{4}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma_i^2}}\left(\frac{u^2}{\sigma_i^2} - 1\right)^2 e^{-\frac{u^2}{2\sigma_i^2}}du \\ &= \frac{\log_2(e)^2}{4}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\left(v^2-1\right)^2 e^{-\frac{v^2}{2}}du\end{aligned}$$

where the last equation follows by the change of variable $v = \frac{u}{\sigma_i}$ in the integral. This shows that the variance of V_i is constant. For the second statement we will make use of the following inequalities. For nonnegative u we have

$$\frac{e^{-\frac{(u-1)^2}{2\sigma_i^2}}}{2\sqrt{2\pi\sigma_i^2}} \leq f_i(u) \leq \frac{e^{-\frac{(u-1)^2}{2\sigma_i^2}}}{\sqrt{2\pi\sigma_i^2}}. \tag{28}$$

Recall that $\mathbb{E}(U_i) = \mathcal{H}(Y_i) = \mathcal{H}(Y_i|X_i) + \mathcal{I}(X_i;Y_i) = \mathcal{H}(N_i) + \mathcal{I}(X_i;Y_i)$. Note that $0 \leq \inf(X_i;Y_i) \leq H(X_i) = 1$ by the properties of mutual information (see [CT91][chapter 2]). And since $\mathcal{H}(N_i) = \frac{1}{2}\log_2(2e\pi\sigma_i^2)$ we deduce that

$$\frac{1}{2}\log_2(2e\pi\sigma_i^2) \leq \mathbb{E}(U_i) \leq \frac{1}{2}\log_2(2e\pi\sigma_i^2) + 1. \tag{29}$$

To simplify the expressions below we let $u = Y_i$. Assume that U_i is greater that its expectation and that this expectation is nonnegative. This means that $-\log_2 f_i(u) \geq \mathbb{E}(U_i) \geq 0$. We notice that

$$\begin{aligned}\bar{U_i}^2 &= (U_i - \mathbb{E}(U_i))^2 \\ &= (-\log_2(f_i(u)) - \mathbb{E}(U_i))^2 \\ &\leq \left(\log_2(e)\frac{(u-1)^2}{2\sigma_i^2} + 1 + \frac{1}{2}\log_2(2\pi\sigma_i^2) - \frac{1}{2}\log_2(2e\pi\sigma_i^2)\right)^2 \\ &= \left(\log_2(e)\frac{(u-1)^2}{2\sigma_i^2} - \frac{1}{2}\log_2(e/2)\right)^2 \end{aligned} \tag{30}$$

by using inequations (28) and (29). Let us now write

$$\begin{aligned}\text{var}(U_i) = \mathbb{E}(\bar{U_i})^2 &= \int_{-\infty}^{\infty}\bar{U_i}^2 f_i(u)du \\ &= \int_{-\infty}^{0}\bar{U_i}^2 f_i(u)du + \int_{0}^{\mathbb{E}(U_i)}\bar{U_i}^2 f_i(u)du + \int_{\mathbb{E}(U_i)}^{\infty}\bar{U_i}^2 f_i(u)du\end{aligned}$$

From the previous upper-bound on ${\bar{U}_i}^2$ we deduce that

$$\begin{aligned}
\int_{\mathbb{E}(U_i)}^{\infty} {\bar{U}_i}^2 f_i(u)du &\leq \int_{\mathbb{E}(U_i)}^{\infty} \left(\log_2(e)\frac{(u-1)^2}{2\sigma_i^2} - \frac{1}{2}\log_2(e/2)\right)^2 f_i(u)du \\
&\leq \int_{\mathbb{E}(U_i)}^{\infty} \left(\log_2(e)\frac{(u-1)^2}{2\sigma_i^2} - \frac{1}{2}\log_2(e/2)\right)^2 \frac{e^{-\frac{(u-1)^2}{2\sigma_i^2}}}{\sqrt{2\pi\sigma_i^2}}du \qquad (31)\\
&= \int_{\frac{\mathbb{E}(U_i)-1}{\sigma_i}}^{\infty} \left(\log_2(e)\frac{v^2}{2} - \frac{1}{2}\log_2(e/2)\right)^2 \frac{e^{-\frac{v^2}{2}}}{\sqrt{2\pi}}dv \qquad (32)\\
&\leq \int_{-\infty}^{\infty} \left(\log_2(e)\frac{v^2}{2} - \frac{1}{2}\log_2(e/2)\right)^2 \frac{e^{-\frac{v^2}{2}}}{\sqrt{2\pi}}dv,
\end{aligned}$$

where Inequality (31) is a consequence of (28) and Equality (32) follows from the change of variable $v = \frac{u-1}{\sigma_i}$. The two other integrals in (30) can be treated similarly where instead of using (28) we use for negative values of u: $\frac{e^{-\frac{(u+1)^2}{2\sigma_i^2}}}{2\sqrt{2\pi\sigma_i^2}} \leq f_i(u) \leq \frac{e^{-\frac{(u+1)^2}{2\sigma_i^2}}}{\sqrt{2\pi\sigma_i^2}}$. This yields a constant upper-bound for all variances $\mathrm{var}(U_i)$. ∎

We are ready now to prove Lemma 5:

Proof. We start the proof by writing

$$\begin{aligned}
\mathbf{Pr}((\mathbf{X},\mathbf{Y})\notin T_\epsilon) &= \mathbf{Pr}\big(\{|-\log_2(f(\mathbf{Y}))-\mathcal{H}(\mathbf{Y})|\geq n\epsilon\}\cup\big\{\big|-\log_2\big(f(\mathbf{Y}|\mathbf{X})2^{-n}\big)-\mathcal{H}(\mathbf{X},\mathbf{Y})\big|\geq n\epsilon\big\}\\
&\leq \mathbf{Pr}(|-\log_2(f(\mathbf{Y}))-\mathcal{H}(\mathbf{Y})|\geq n\epsilon)+\mathbf{Pr}\big(\big|-\log_2\big(f(\mathbf{Y}|\mathbf{X})2^{-n}\big)-\mathcal{H}(\mathbf{X},\mathbf{Y})\big|\geq n\epsilon\big)\\
&= \mathbf{Pr}(|U-\mathbb{E}(U)|\geq n\epsilon)+\mathbf{Pr}(|V-\mathbb{E}(V)|\geq n\epsilon)
\end{aligned}$$

with $U \stackrel{\text{def}}{=} \sum_{i=1}^n U_i$ and $V \stackrel{\text{def}}{=} \sum_{i=1}^n V_i$. We use now Chebyschev's inequality (Lemma 27) together with the upper-bounds $\mathrm{var}(U) = \sum_{i=1}^n \mathrm{var}(U_i) \leq nA$ and $\mathrm{var}(V) = \sum_{i=1}^n \mathrm{var}(V_i) \leq nA$ to obtain $\mathbf{Pr}\left((\mathbf{X},\mathbf{Y}) \notin T_\epsilon\right) \leq \frac{2A}{n\epsilon^2}$. ∎

Moreover, not only is it unlikely that $(\mathbf{X},\mathbf{Y})$ does not fall in T_ϵ, but the Euclidean volume (which we denote by "Vol") of this set is not too large:

Lemma 9

$$\sum_{\mathbf{x}\in\mathbb{R}^n} \mathrm{Vol}(T_\epsilon(\mathbf{x})) \leq 2^{\mathcal{H}(\mathbf{X},\mathbf{Y})+\epsilon n}$$

Proof. Let us notice that

$$1 = \sum_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2^n}\int_{\mathbb{R}^n} f(\mathbf{y}|\mathbf{x})d\mathbf{y} \geq \sum_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{2^n}\int_{T_\epsilon(\mathbf{x})} f(\mathbf{y}|\mathbf{x})d\mathbf{y} \geq \sum_{\mathbf{x}\in\mathbb{R}^n} \mathrm{Vol}(T_\epsilon(\mathbf{x}))2^{-\mathcal{H}(\mathbf{X},\mathbf{Y})-\epsilon n}$$

where the last inequality follows from (25) ∎

We will use this result to show that

Proposition 1 *If $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ is a couple of independent random variables, where $\tilde{\mathbf{X}}$ is uniformly distributed and $\tilde{\mathbf{Y}}$ has the same distribution as $\mathbf{Y}$, then $\mathbf{Pr}\left((\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in T_\epsilon\right) \leq 2^{-C+2n\epsilon}$ with $C \stackrel{def}{=} \sum_{i=1}^{n} \mathbf{Cap}(\sigma_i^2)$.*

Proof. We evaluate $\mathbf{Pr}\left((\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in T_\epsilon\right)$ as follows

$$\mathbf{Pr}\left((\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in T_\epsilon\right) = \sum_{\mathbf{x} \in \{0,1\}^n} \frac{1}{2^n} \int_{T_{\mathbf{x}}(\epsilon)} f(\mathbf{y}) \leq \sum_{\mathbf{x} \in \{0,1\}^n} \frac{1}{2^n} \mathrm{Vol}(T_{\mathbf{x}}(\epsilon)) 2^{-\mathcal{H}(\mathbf{Y})+\epsilon n}$$

The last inequality follows from (24) in the definition of the typical set. We use now Lemma 9 to obtain

$$\mathbf{Pr}\left((\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in T_\epsilon\right) \leq \frac{1}{2^n} 2^{\mathcal{H}(\mathbf{X},\mathbf{Y})+\epsilon n} 2^{-\mathcal{H}(\mathbf{Y})+\epsilon n} \leq 2^{-n+\mathcal{H}(\mathbf{X};\mathbf{Y})-\mathcal{H}(\mathbf{Y})+2\epsilon n}$$

By using the expressions for $\mathcal{H}(\mathbf{X}, \mathbf{Y})$ and $\mathcal{H}(\mathbf{Y})$ given in Lemma 4 we deduce $-n + \mathcal{H}(\mathbf{X}, \mathbf{Y}) - \mathcal{H}(\mathbf{Y}) = -\sum_{i=1}^{n} \mathbf{Cap}(\sigma_i^2)$. This finishes the proof. ∎

These results can be used to analyze the following typical set decoder, which takes as inputs a vector $\mathbf{y}$ in $\mathbb{R}^n$ which is the output of the Gaussian channel described in Section 2 and a real parameter ϵ, and outputs either "Failure" or a possible key $\tilde{\mathbf{K}} \in \{0,1\}^n$.

TYPICAL SET DECODER($\mathbf{y}, \epsilon$)

```
1  counter ← 0
2  for all possible values k of K̃
3      do if y ∈ T_k(ε)
4            then counter ← counter + 1
5                 result ← k
6  if counter = 1
7    then return result
8    else  return failure
```

This algorithm is therefore successful if and only if $\mathbf{y}$ is in the typical set of the right key and if there is no other value $\mathbf{k}$ for $\tilde{\mathbf{K}}$ for which $\mathbf{y}$ belongs to the typical set associated to $\mathbf{k}$. Let us now finish the proof of Theorem 1.

Proof. Let $\mathbf{k}$ be right value of $\tilde{\mathbf{K}}$ and let $\mathcal{C}$ be the set of possible values of $\tilde{\mathbf{K}}$. The probability P_{err} that the typical decoder fails is clearly upper-bounded by

$$P_{\text{err}} \leq \mathbf{Pr}_{\mathbf{y},\mathcal{C}}(\overline{T_{\mathbf{k}}(\epsilon)}) + \sum_{\mathbf{k}' \in \mathcal{C}, \mathbf{k}' \neq \mathbf{k}} \mathbf{Pr}_{\mathbf{y},\mathcal{C}}\left(T_{\mathbf{k}'}(\epsilon)\right) \tag{33}$$

where $\overline{T_{\mathbf{k}}(\epsilon)}$ denotes the complementary set of $T_{\mathbf{k}}(\epsilon)$. On the one hand

$$\mathbf{Pr}_{\mathbf{y},\mathcal{C}}(\overline{T_{\mathbf{k}}(\epsilon)}) = \mathbf{Pr}((\mathbf{X}, \mathbf{Y}) \notin T_\epsilon) \leq \frac{A}{\epsilon^2 n}.$$

by Lemma 5, and on the other hand for $\mathbf{k}' \neq \mathbf{k}$:

$$\sum_{\mathbf{k}' \in \mathcal{C}, \mathbf{k}' \neq \mathbf{k}} \mathbf{Pr}_{\mathbf{y},\mathcal{C}}\left(T_{\mathbf{k}'}(\epsilon)\right) \leq \sum_{\mathbf{k}' \in \mathcal{C}} \mathbf{Pr}_{\mathbf{y},\mathcal{C}}\left(T_{\mathbf{k}'}(\epsilon)\right) = 2^r \mathbf{Pr}\left((\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in T_\epsilon\right) \leq 2^{r-\sum_{i=1}^{n} \mathbf{Cap}(\sigma_i^2)+2\epsilon n},$$

by Proposition 1. By plugging in these two upper bounds in the union bound (33) we obtain $P_{\text{err}} \leq \frac{A}{\epsilon^2 n} + 2^{r-\sum_{i=1}^{n} \mathbf{Cap}(\sigma_i^2)+2\epsilon n} \leq \frac{A}{\epsilon^2 n} + 2^{-\delta n+2\epsilon n}$. We finish the proof by choosing $\epsilon = \frac{\delta}{4}$. ∎