On the Bayes Risk in Information-Hiding Protocols*

Konstantinos Chatzikokolakis Catuscia Palamidessi
INRIA and LIX, École Polytechnique
Palaiseau, France
{kostas,catuscia}@lix.polytechnique.fr

Prakash Panangaden
McGill University
Montreal, Quebec, Canada
prakash@cs.mcgill.ca

Abstract

Randomized protocols for hiding private information can be regarded as noisy channels in the information-theoretic sense, and the inference of the concealed information can be regarded as a hypothesis-testing problem. We consider the Bayesian approach to the problem, and investigate the probability of error associated to the MAP (Maximum Aposteriori Probability) inference rule. Our main result is a constructive characterization of a convex base of the probability of error, which allows us to compute its maximum value (over all possible input distributions), and to identify upper bounds for it in terms of simple functions. As a side result, we are able to improve the Hellman-Raviv and the Santhi-Vardy bounds expressed in terms of conditional entropy. We then discuss an application of our methodology to the Crowds protocol, and in particular we show how to compute the bounds on the probability that an adversary break anonymity.

1 Introduction

Information-hiding protocols try to hide the relation between certain facts, that we wish to maintain hidden, and the *observable* consequences of these facts. Example of such protocols are anonymity protocols like Crowds [23], Onion Routing [29], and Freenet [8]. Often these protocols use randomization to obfuscate

*This work has been partially supported by the INRIA DREI Équipe Associée PRINTEMPS. The work of Konstantinos Chatzikokolakis and Catuscia Palamidessi has been also supported by the INRIA ARC project ProNoBiS.

the link between the information that we wish to keep hidden and the observed events. Crowds, for instance, tries to conceal the identity of the originator of a message by forwarding the message randomly until it reaches its destination, so that if an attacker intercepts the message, it cannot be sure whether the sender is the originator or just a forwarder.

In most cases, protocols like the ones above can be regarded as information-theoretic channels, where the inputs are the facts to keep hidden and the outputs are the observables. In information theory channels are typically *noisy*, which means that for a given input we may obtain several different outputs, each with a certain probability. A channel is then characterized by what is called *transfer matrix*, whose elements are the conditional probabilities of obtaining a certain output given a certain input. In our case, the matrix represents the correlation between the facts and the observables. An adversary can try to infer the facts from the observables using the *Bayesian method* for *hypothesis testing*, which is based on the principle of assuming an *a priori* probability distribution on the hidden facts (*hypotheses*), and deriving from that (and from the matrix) the *a posteriori* distribution after a certain event has been observed. It is well known that the best strategy is to apply the MAP (Maximum Aposteriori Probability) criterion, which, as the name says, dictates that one should choose the hypothesis with the maximum *a posteriori* probability given the observation. "Best" means that this strategy induces the smallest probability of guessing the wrong hypothesis. The probability of error, in this case, is also called *Bayes risk*.

Intuitively, the Bayes risk is maximum when the rows of the channel's matrix are all the same; this case corresponds indeed to capacity 0, which means that the input and the output are independent, i.e. we do not learn anything about the inputs by observing the outputs. This is the ideal situation, from the point of view of information-hiding protocols. In practice, however, it is difficult to achieve such degree of privacy. We are then interested in maximizing the Bayes risk, so to characterize quantitatively the protection offered by the protocol. The interest in finding good bounds for the probability of error is motivated also by the fact that in some case the decision region can have a complicated geometry, or the decision function can be very sensitive to small variations in the input distribution, thus making it difficult to compute the probability of error. Some examples of such situations are illustrated in [26]. Good bounds based on "easy" functions (i.e. functions easy to compute, and not too sensitive to computational errors) are therefore very useful in such situations as they can be used as an approximation of the probability of error. It is particularly nice to have convex bounds since they bound any estimate based on linear interpolation.

The main purpose of this paper is to investigate the Bayes risk, in relation to the channel's matrix, and to produce tight bounds on it.

There are many bounds known in literature for the Bayes risk. One of these is the *equivocation bound*, due to Rényi [24], which states that the probability of error is bounded by the conditional entropy of the channel's input given the output. Later, Hellman and Raviv improved this bound by half [15]. Recently, Santhi and Vardy have proposed a new bound, that depends exponentially on the (opposite of the) conditional entropy, and which considerably improves the

Hellman-Raviv bound in the case of multi-hypothesis testing [26]. The latter is better, however, in the case of binary hypothesis testing.

The Bayes approach to hypothesis testing is often criticized because it assumes the knowledge of the *a priori* distribution, or at least of a good approximation of it, which is often an unjustified assumption. However, even if the adversary does not know the *a priori* distribution, the method is still valid asymptotically, under the condition that the matrix's rows are all pairwise distinguished. Under such condition indeed, as shown in [3], by repeating the experiment the contribution of the *a priori* probability becomes less and less relevant for the computation of the Bayesian risk, and it "washes out" in the limit. Furthermore, the Bayesian risk converges to 0. At the other extreme, when the rows are all equal, the Bayes risk does not converge to 0 and its limit is bound from below by a constant that depends on the input distribution. In the present paper we continue this investigation by considering what happens in the intermediate case when some of the rows (not necessarily all) are equal.

1.1 Contribution

The main contributions of this paper are the following:

1. We consider what we call "the corner points" of a piecewise linear function, and we propose criteria to compute the maximum of the function, and to identify concave functions that are upper bounds for the given piecewise linear function, based on the analysis of its corner points only.

2. We consider the hypothesis testing problem in relation to an information-theoretic channel. In this context, we show that the probability of error associated to the MAP rule is piecewise linear, and we give a constructive characterization of a set of corner points, which turns out to be finite. Together with the results of the previous paragraph, this leads to algorithms to compute the maximum Bayes risk over all the channel's input distributions, and to a method to improve functional upper bounds of the error probability. The improved functions are tight at at least one point.

3. By using the above results about concave functions and corner points, we give an alternative proof of the Hellman-Raviv and the Santhi-Vardy bounds on the Bayes risk in terms of conditional entropy. Our proof is intuitive and works exactly in the same way for both bounds, which were proven using different techniques in the corresponding papers.

4. Thanks to our characterization of the maximum Bayes risk, we are able to improve on the Hellman-Raviv and the Santhi-Vardy bounds. These two bounds are tight (i.e. coincide with the Bayes risk) on the corner points only for channels of capacity 0. Our improved bounds are tight at at least one corner point for *every* channel.

5. We consider the case of protocol re-execution, and we show that in the intermediate case in which at least two rows are equal the Bayes risk does

not converge to 0. Furthermore we give a precise lower bound for the limit of the Bayes risk.

6. We show how to apply the above results to randomized protocols for information hiding. In particular, we present an analysis of Crowds using two different network topologies, and derive the maximum Bayes risk for an adversary who tries to break anonymity, and improved bounds on this probability in terms of conditional entropy, for any input distribution.

1.2 Related work

Probabilistic notions of anonymity and information-hiding have been explored in [5, 14, 1, 2]. We discuss the relation with these works in detail in Section 5.

Several authors have considered the idea of using information theory to analyze anonymity. A recent line of work is due to [27, 12]. The main difference with our approach is that in these works the anonymity degree is expressed in terms of input entropy, rather than conditional entropy. More precisely, the emphasis is on the lack of information of the attacker about the distribution of the inputs, rather than on the capability of the protocol to prevent the attacker from determining this information from a statistical analysis of the observables which are visible to the attacker. Moreover, a uniform input distribution is assumed, while in this paper we abstract from the input distribution.

In [19, 20] the ability to have covert communication as a result of non-perfect anonymity is explored. These papers focus on the possibility of constructing covert channels by the users of the protocol, using the protocol mechanisms, and on measuring the amount of information that can be transferred through these channels. In [20] the authors also suggest that the channel's capacity can be used as an asymptotic measure of the worst-case information leakage. Note that in [20] the authors warn that in certain cases the notion of capacity might be too strong a measure to compare systems with, because the holes in the anonymity of a system might not behave like text book discrete memoryless channels.

Another information-theoretical approach is the one of [11]. The authors propose a probabilistic process calculus to describe protocols for ensuring anonymity, and use the *Kullback-Leibler distance* (aka *relative entropy*) to measure the degree of anonymity these protocols can guarantee. More precisely, the degree of anonymity is defined as the distance between the distributions on the observable traces produced by the original runs of the protocol, and those produced by the runs after permuting the identities of he users. Furthermore, they prove that the operators in the probabilistic process calculus are non-expansive with respect to the Kullback-Leibler distance.

A different approach, still using the Kullback-Leibler distance, is taken in [9]. In this paper, the authors define as information leakage the difference between the a priori accuracy of the guess of the attacker, and the a posteriori one, after the attacker has made his observation. The accuracy of the guess is defined as the Kullback-Leibler distance between the *belief* (which is a weight

attributed by the attacker to each input hypothesis) and the true distribution on the hypotheses.

In the field of information flow and non-interference there is a line of research which is related to ours. There have been various papers [18, 13, 6, 7, 16] in which the so-called *high information* and the *low information* are seen as the input and output respectively of a channel. The idea is that "high" information is meant to be kept secret and the "low" information is visible; the point is to prevent the high information from being deduced by observing the low information. From an abstract point of view, the setting is very similar; technically it does not matter what kind of information one is trying to conceal, what is relevant for the analysis is only the probabilistic relation between the input and the output information. We believe that our results are applicable also to the field of non-interference.

The connection between the adversary's goal of inferring a secret from the observables, and the field of hypothesis testing, has been explored in other papers in literature, see in particular [17, 21, 22, 3]. To our knowledge, however, [3] is the only work exploring the Bayes risk in connection to the channel associated to an information-hiding protocol. More precisely, [3] considers a framework in which anonymity protocols are interpreted as particular kinds of channels, and the degree of anonymity provided by the protocol as the converse of the channel's capacity (an idea already suggested in [20]). Then, [3] considers a scenario in which the adversary can enforce the re-execution of the protocol with the same input, and studies the Bayes risk on the statistics of the repeated experiment. The question is how the adversary can approximate the MAP rule when the *a priori* distribution is not known, and the main results of [3] on this topic is that the approximation is possible when the rows of the matrix are pairwise different, and impossible when they are all equal (case of capacity 0). Furthermore, in the first case the Bayes risk converges to 0, while in the second case it does not. In the present paper the main focus is on the Bayes risk as a function of the *a priori* distribution, and on the computation of its bounds. However we also continue the investigation of [3] on the protocol re-execution, and we give a lower bound to the limit of the Bayes risk in the intermediate case in which some of the rows (not necessarily all) coincide.

Part of the results of this paper were presented (without proofs) in [4].

1.3 Plan of the paper

Next section recalls some basic notions about information theory, hypothesis testing and the probability of error. Section 3 proposes some methods to identify bounds for a function that is generated by a set of corner points; these bounds are tight on at least one corner point. Section 4 presents the main result of our work, namely a constructive characterization of the corner points of Bayes risk. In Section 5 we discuss the relation with some probabilistic information-hiding notions in literature. Section 6 illustrates an application of our results to the anonymity protocol Crowds. In Section 7 we study the convergence of the Bayes risk in the case of protocol re-execution. Section 8 concludes.

2 Information theory, hypothesis testing and the probability of error

In this section we briefly review some basic notions in information theory and hypothesis testing that will be used throughout the paper. We refer to [10] for more details.

A *channel* is a tuple $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ where $\mathcal{A}, \mathcal{O}$ are the sets of input and output values respectively and $p(o|a)$ is the conditional probability of observing output $o \in \mathcal{O}$ when $a \in \mathcal{A}$ is the input. In this paper, we assume that both $\mathcal{A}$ and $\mathcal{O}$ are finite with cardinality n and m respectively. We will also sometimes use indices to represent their elements: $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ and $\mathcal{O} = \{o_1, o_2, \ldots, o_m\}$. The $p(o|a)$'s constitute the *transfer matrix* (which we will simply call *matrix*) of the channel. The usual convention is to arrange the a's by rows and the o's by columns.

In general, we consider the input of a channel as *hidden information*, and the output as *observable information*. The set of input values can also be regarded as a set of *mutually exclusive* (hidden) *facts* or *hypotheses*. A probability distribution $p(\cdot)$ over $\mathcal{A}$ is called *a priori probability*, and it induces a probability distribution over $\mathcal{O}$ called the *marginal probability* of $\mathcal{O}$. In fact,

$$p(o) = \sum_a p(a, o) = \sum_a p(o|a)\, p(a)$$

where $p(a, o)$ represents the joint probability of a and o, and we use its definition $p(a, o) = p(o|a)p(a)$.

When we observe an output o, the probability that the corresponding input has been a certain a is given by the conditional probability $p(a|o)$, also called *a posteriori probability* of a given o, which in general is different from $p(a)$. This difference can be interpreted as the fact that observing o gives us evidence that changes our degree of belief in the hypothesis a. The *a priori* and the *a posteriori* probabilities of a are related by Bayes' theorem:

$$p(a|o) = \frac{p(o|a)\, p(a)}{p(o)}$$

In hypothesis testing we try to infer the *true* hypothesis (i.e. the input fact that really took place) from the observed output. In general, it is not possible to determine the right hypothesis with certainty. We are interested in minimizing the *probability of error*, i.e. the probability of making the wrong guess. Formally, the probability of error is defined as follows. Given the *decision function* $f : \mathcal{O} \to \mathcal{A}$ adopted by the observer to infer the hypothesis, let $E_f : \mathcal{A} \to 2^{\mathcal{O}}$ be the function that gives the *error region* of f when $a \in \mathcal{A}$ has occurred, namely:

$$E_f(a) = \{o \in \mathcal{O} \mid f(o) \neq a\}$$

Let $\eta_f : \mathcal{A} \to [0, 1]$ be the function that associates to each $a \in \mathcal{A}$ the probability

that f gives the wrong input when $a \in \mathcal{A}$ has occurred, namely:

$$\eta_f(a) = \sum_{o \in E_f(a)} p(o|a)$$

The probability of error for f is then obtained as the sum of the probability of error for each possible input, averaged over the probability of the input:

$$P_f = \sum_a p(a)\, \eta_f(a)$$

In the Bayesian framework, the best possible decision function f_B, namely the decision function that minimizes the probability of error, is obtained by applying the MAP (*Maximum Aposteriori Probability*) criterion, that chooses an input a with a maximum $p(a|o)$. Formally:

$$f_B(o) = a \quad \Rightarrow \quad \forall a' \;\; p(a|o) \geq p(a'|o)$$

The probability of error associated with f_B, also called the *Bayes risk*, is then given by (we will use the notation P_e instead than P_{f_B} for simplicity)

$$P_e = 1 - \sum_o p(o) \max_a p(a|o) = 1 - \sum_o \max_a p(o|a)\, p(a)$$

Note that f_B, and the Bayes risk, depend on the inputs' *a priori* probability. The input distributions can be represented as the elements $\vec{x} = (x_1, x_2, \ldots, x_n)$ of the domain $D^{(n)}$ defined as

$$D^{(n)} = \{\vec{x} \in \mathbb{R}^n \mid \textstyle\sum_i x_i = 1 \text{ and } \forall i \; x_i \geq 0\}$$

(also called an $(n-1)$-simplex) where the correspondence is given by $x_i = p(a_i)$ for all i's. In the rest of the paper we will take the MAP rule as decision function and view the Bayes risk as a function $P_e : D^{(n)} \to [0, 1]$ defined by

$$P_e(\vec{x}) = 1 - \sum_i \max_j p(o_i|a_j) x_j \tag{1}$$

We will identify probability distributions and their vector representation freely throughout the paper.

There are some notable results in literature relating the Bayes risk to the information-theoretic notion of *conditional entropy*, also called *equivocation*. Let us first recall the concept of *random variable* and its *entropy*. A random variable A is determined by a set of values $\mathcal{A}$ and a probability distribution $p(\cdot)$ over $\mathcal{A}$. The entropy of A, $H(A)$, is given by

$$H(A) = -\sum_a p(a) \log p(a)$$

The entropy measures the uncertainty of a random variable. It takes its maximum value $\log n$ when A's distribution is uniform and its minimum value 0 when

A is constant. We usually consider the logarithm in base 2 and thus measure entropy in *bits*.

Now let A, O be random variables. The *conditional entropy* $H(A|O)$ is defined as

$$H(A|O) = -\sum_o p(o) \sum_a p(a|o) \log p(a|o)$$

The conditional entropy measures the amount of uncertainty of A when O is known. It can be shown that $0 \leq H(A|O) \leq H(A)$. It takes its maximum value $H(A)$ when O reveals no information about A, i.e. when A and O are independent, and its minimum value 0 when O completely determines the value of A.

Comparing $H(A)$ and $H(A|O)$ gives us the concept of *mutual information* $I(A;O)$, which is defined as

$$I(A;O) = H(A) - H(A|O)$$

Mutual information measures the amount of information that one random variable contains about another random variable. In other words, it measures the amount of uncertainty about A that we lose when observing O. It can be shown that it is symmetric ($I(A;O) = I(O;A)$) and that $0 \leq I(A;O) \leq H(A)$. The maximum mutual information between A and O over all possible input distributions $p(\cdot)$ is known as the channel's *capacity*:

$$C = \max_{p(\cdot)} I(A;O)$$

The capacity of a channel gives the maximum rate at which information can be transmitted using this channel without distortion.

Given a channel, let $\vec{x}$ be the *a priori* distribution on the inputs. Recall that $\vec{x}$ also determines a probability distribution on the outputs. Let A and O be the random variables associated to the inputs and outputs respectively. The Bayes risk is related to $H(A|O)$ by the Hellman-Raviv bound [15]:

$$P_e(\vec{x}) \leq \frac{1}{2} H(A|O) \tag{2}$$

and by the Santhi-Vardy bound [26]:

$$P_e(\vec{x}) \leq 1 - 2^{-H(A|O)} \tag{3}$$

We remark that, while the bound (2) is tighter than (3) in case of binary hypothesis testing, i.e. when $n = 2$, (3) gives a much better bound when n becomes larger. In particular the bound in (3) is always limited by 1, which is not the case for (2).

3 Convexly generated functions and their bounds

In this section we characterize a special class of functions on probability distributions, and we present various results regarding their bounds which lead to

methods to compute their maximum, to prove that a concave function is an upper bound, and to derive an upper bound from a concave function. The interest of this study is that the probability of error will turn out to be a function in this class.

We start by recalling some basic notions of convexity: let $\mathbb{R}$ be the set of real numbers. The elements $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$ constitute a set of *convex coefficients* iff $\forall i\ \lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. Given a vector space V, a *convex combination* of $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_k \in V$ is any vector of the form $\sum_i \lambda_i \vec{x}_i$ where the λ_i's are convex coefficients. A subset S of V is *convex* if and only if every convex combination of vectors in S is also in S. It is easy to see that for any n the domain $D^{(n)}$ of probability distributions of dimension n is convex. Given a subset S of V, the *convex hull* of S, which we will denote by $ch(S)$, is the smallest convex set containing S. Since the intersection of convex sets is convex, it is clear that $ch(S)$ always exists.

We now introduce (with a slight abuse of terminology) the concept of *convex base*: Intuitively, a convex base of a set S is a subset of S whose convex hull contains S.

Definition 3.1. *Given the vector sets S, U, we say that U is a convex base for S if and only if $U \subseteq S$ and $S \subseteq ch(U)$.*

In the following, given a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$, and a function f from n-dimensional vectors to reals, we will use the notation $(\vec{x}, f(\vec{x}))$ to denote the $(n+1)$-dimensional vector $(x_1, x_2, \ldots, x_n, f(\vec{x}))$. Similarly, given a vector set S in a n-dimensional space, we will use the notation $(S, f(S))$ to represent the set of vectors $\{(\vec{x}, f(\vec{x})) \mid \vec{x} \in S\}$ in a $(n+1)$-dimensional space. The notation $f(S)$ represents the image of S under f, i.e. $f(S) = \{f(\vec{x}) \mid \vec{x} \in S\}$.

We are now ready to introduce the class of functions that we mentioned at the beginning of this section:

Definition 3.2. *Given a vector set S, a convex base U of S, and a function $f : S \to \mathbb{R}$, we say that $(U, f(U))$ is a set of* corner points *of f if and only if $(U, f(U))$ is a convex base for $(S, f(S))$. We also say that f is* convexly generated *by $(U, f(U))$.*

Of particular interest are the functions that are convexly generated by a finite number of corner points. This is true for *piecewise linear functions* in which S can be decomposed into finitely many convex polytopes (n-dimensional polygons) and f is equal to a linear function on each of them. Such functions are convexly generated by the finite set of vertices of these polytopes.

We now give a criterion for computing the maximum of a convexly generated function.

Proposition 3.3. *Let U be a convex base of S and let $f : S \to \mathbb{R}$ be convexly generated by $(U, f(U))$. If $f(U)$ has a maximum element b, then b is the maximum value of f on S.*

Proof. Let b be the maximum of $f(U)$. Then for every $\vec{u} \in U$ we have that $f(\vec{u}) \leq b$. Consider now a vector $\vec{x} \in S$. Since f is convexly generated by

$(U, f(U))$, there exist $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_k$ in U such that $f(\vec{x})$ is obtained by convex combination from $f(\vec{u}_1)$, $f(\vec{u}_2)$, ..., $f(\vec{u}_k)$ via some convex coefficients λ_1, λ_2, ..., λ_k. Hence:

$$\begin{array}{rcll} f(\vec{x}) & = & \sum_i \lambda_i f(\vec{u}_i) & \\ & \leq & \sum_i \lambda_i b & \text{since } f(\vec{u}_i) \leq b \\ & = & b & \lambda_i\text{'s being convex coefficients} \end{array}$$

□

Note that if U is finite then $f(U)$ always has a maximum element.

Next, we propose a method for establishing functional upper bounds for f, when they are in the form of *concave* functions.

We recall that, given a vector set S, a function $g : S \to \mathbb{R}$ is *concave* if and only if for any $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_k \in S$ and any set of convex coefficients $\lambda_1, \lambda_2, \ldots, \lambda_k \in \mathbb{R}$ we have

$$\sum_i \lambda_i \, g(\vec{x}_i) \leq g(\sum_i \lambda_i \vec{x}_i)$$

Proposition 3.4. *Let U be a convex base of S, let $f : S \to \mathbb{R}$ be convexly generated by $(U, f(U))$, and let $g : S \to \mathbb{R}$ be concave. Assume that for all $\vec{u} \in U$ $f(\vec{u}) \leq g(\vec{u})$ holds. Then we have that g is an upper bound for f, i.e.*

$$\forall \vec{x} \in S \;\; f(\vec{x}) \leq g(\vec{x})$$

Proof. Let $\vec{x}$ be an element of S. Since f is convexly generated, there exist $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_k$ in U such that $(\vec{x}, f(\vec{x}))$ is obtained by convex combination from $(\vec{u}_1, f(\vec{u}_1))$, $(\vec{u}_2, f(\vec{u}_2))$, ..., $(\vec{u}_k, f(\vec{u}_k))$ via some convex coefficients λ_1, λ_2, ..., λ_k. Hence:

$$\begin{array}{rcll} f(\vec{x}) & = & \sum_i \lambda_i f(\vec{u}_i) & \\ & \leq & \sum_i \lambda_i g(\vec{u}_i) & \text{since } f(\vec{u}_i) \leq g(\vec{u}_i) \\ & \leq & g(\sum_i \lambda_i \vec{u}_i) & \text{by the concavity of } g \\ & = & g(\vec{x}) & \end{array}$$

□

We also give a method to obtain functional upper bounds, that are tight on at least one corner point, from concave functions.

Proposition 3.5. *Let U be a convex base of S, let $f : S \to \mathbb{R}$ be convexly generated by $(U, f(U))$, and let $g : S \to \mathbb{R}$ be concave and non-negative. Let $R = \{c \mid \exists \vec{u} \in U : f(\vec{u}) \geq c \, g(\vec{u})\}$. If R has an upper bound c_o, then the function $c_o \, g$ is a functional upper bound for f satisfying*

$$\forall \vec{x} \in S \;\; f(\vec{x}) \leq c_o \, g(\vec{x})$$

Furthermore, if $c_o \in R$ then f and $c_o \, g$ coincide at least at one point.

Proof. We first show that $f(\vec{u}) \leq c_o\, g(\vec{u})$ for all $\vec{u} \in U$. Suppose, by contradiction, that this is not the case. Then there exists $\vec{u} \in U$ such that $f(\vec{u}) > c_o\, g(\vec{u})$. If $g(\vec{u}) = 0$ then for all $c \in \mathbb{R} : f(\vec{u}) > c\, g(\vec{u}) = 0$ so the set R is not bounded, which is a contradiction. Considering the case $g(\vec{u}) > 0$ (g is assumed to be non-negative), let $c = \frac{f(\vec{u})}{g(\vec{u})}$. Then $c > c_o$ and again we have a contradiction since $c \in R$ and c_o is an upper bound of R . Hence by Proposition 3.4 we have that $c_o\, g$ is an upper bound for f.

Furthermore, if $c_o \in R$ then there exists $\vec{u} \in U$ such that $f(\vec{u}) \geq c_o\, g(\vec{u})$, so $f(\vec{u}) = c_o\, g(\vec{u})$ and the bound is tight as this point. □

Corollary 3.6. *If U is finite and $\forall \vec{u} \in U : g(\vec{u}) = 0 \Rightarrow f(\vec{u}) \leq 0$, then the maximum element of R always exists and is equal to*

$$\max_{\vec{u} \in U, g(\vec{u}) > 0} \frac{f(\vec{u})}{g(\vec{u})}$$

Finally, we develop a proof technique that will allow us to prove that a certain set is a set of corner points of a function f. Let S be a set of vectors. The *extreme points* of S, denoted by $extr(S)$, is the set of points of S that cannot be expressed as the convex combination of two distinct elements of S. A subset of $\mathbb{R}^n$ is called *compact* if it is closed and bounded. Our proof technique uses the Krein-Milman theorem which relates a compact convex set to its extreme points.

Theorem 3.7 (Krein-Milman)**.** *A compact and convex vector set is equal to the convex hull of its extreme points.*

We refer to [25] for the proof. Now since the extreme points of S are enough to generate S, to show that a given set $(U, f(U))$ is a set of corner points, it suffices to show that it includes all its extreme points.

Proposition 3.8. *Let S be a compact vector set, U be a convex base of S and $f : S \to \mathbb{R}$ be a continuous function. Let $T = S \setminus U$. If all elements of $(T, f(T))$ can be written as the convex combination of two distinct elements of $(S, f(S))$ then $(U, f(U))$ is a set of corner points of f.*

Proof. Let $S_f = (S, f(S))$ and $U_f = (U, f(U))$. Since S is compact and continuous maps preserve compactness then S_f is also compact, and since the convex hull of a compact set is compact then $ch(S_f)$ is also compact (note that we did not require S to be convex). Then $ch(S_f)$ satisfies the requirements of the Krein-Milman theorem, and since the extreme points of $ch(S_f)$ are clearly the same as those of S_f, we have

$$\begin{aligned} ch(extr(ch(S_f))) = ch(S_f) \Rightarrow \\ ch(extr(S_f)) = ch(S_f) \end{aligned} \tag{4}$$

Now all points in $S_f \setminus U_f$ can be written as convex combinations of other (distinct) points, so they are not extreme. Thus all extreme points are contained

in U_f, that is $extr(S_f) \subseteq U_f$, and since $ch(\cdot)$ is monotone with respect to set inclusion, we have

$$ch(extr(S_f)) \subseteq ch(U_f)$$

and by (4),

$$S_f \subseteq ch(S_f) \subseteq ch(U_f)$$

which means that U_f is a set of corner points of f. □

The advantage of the above proposition is that it only requires to express points outside U as convex combinations of any other points, not necessarily of points in U (as a direct application of the definition of corner points would require).

3.1 An alternative proof for the Hellman-Raviv and Santhi-Vardy bounds

Using Proposition 3.4 we can give an alternative, simpler proof for the bounds in (2) and (3). Let $f : D^{(n)} \to \mathbb{R}$ be the function $f(\vec{y}) = 1 - \max_j y_j$. We start by identifying a set of corner points of f, using Proposition 3.8 to prove that they are indeed corner points.

Proposition 3.9. *The function f defined above is convexly generated by $(U, f(U))$ with $U = U_1 \cup U_2 \cup \ldots \cup U_n$ where, for each k, U_k is the set of all vectors that have value $1/k$ in exactly k components, and 0 everywhere else.*

Proof. We have to show that for any point $\vec{x}$ in $D^{(n)} \setminus U$, $(\vec{x}, f(\vec{x}))$ can be written as a convex combination of two points in $(D^{(n)}, f(D^{(n)}))$. Let $w = \max_i x_i$. Since $\vec{x} \notin U$ then there is at least one element of $\vec{x}$ that is neither w nor 0, let x_i be that element. Let k the number of elements equal to w. We create two vectors $\vec{y}, \vec{z} \in D^{(n)}$ as follows

$$y_j = \begin{cases} x_i + \epsilon & \text{if } i = j \\ w - \frac{\epsilon}{k} & \text{if } x_j = w \\ x_j & \text{otherwise} \end{cases} \qquad z_j = \begin{cases} x_i - \epsilon & \text{if } i = j \\ w + \frac{\epsilon}{k} & \text{if } x_j = w \\ x_j & \text{otherwise} \end{cases}$$

where ϵ is a small positive number, such that $y_j, z_j \in [0, 1]$ for all j, and such that $w - \frac{\epsilon}{k}$, $w + \frac{\epsilon}{k}$ are "still" the maximum elements of $\vec{y}, \vec{z}$ respectively[1]. Clearly $\vec{x} = \frac{1}{2}\vec{y} + \frac{1}{2}\vec{z}$ and since $f(\vec{x}) = 1 - w$, $f(\vec{y}) = 1 - w + \frac{\epsilon}{k}$ and $f(\vec{y}) = 1 - w - \frac{\epsilon}{k}$ we have $f(\vec{x}) = \frac{1}{2}f(\vec{y}) + \frac{1}{2}f(\vec{z})$. Since f is continuous and $D^{(n)}$ is compact, the result follows from Proposition 3.8. □

[1] Taking $\epsilon = \min\{a, w - b\}/2$ is sufficient, where a is the minimum positive element of $\vec{x}$ and b is the maximum element smaller than w.

Consider now the functions $g, h : D^{(n)} \to \mathbb{R}$ defined as

$$g(\vec{y}) = \frac{1}{2}H(\vec{y}) \quad \text{and} \quad h(\vec{y}) = 1 - 2^{-H(\vec{y})}$$

where (with a slight abuse of notation) H represents the entropy of the distribution $\vec{y}$, i.e. $H(\vec{y}) = -\sum_j y_j \log y_j$. From the concavity of $H(\vec{y})$ ([10]) follows that both g, h are concave.

We now compare g, h with$f(\vec{y}) = 1 - \max_j y_j$ on the corner points on f. A corner point $\vec{u}_k \in U_k$ (defined in Proposition 3.9) has k elements equal to $1/k$ and the rest equal to 0. So $H(\vec{u}_k) = \log k$ and

$$\begin{aligned} f(\vec{u}_k) &= 1 - \frac{1}{k} \\ g(\vec{u}_k) &= \frac{1}{2}\log k \\ h(\vec{u}) &= 1 - 2^{-\log k} = 1 - \frac{1}{k} \end{aligned}$$

So $f(\vec{u}_1) = 0 = g(\vec{u}_1)$, $f(\vec{u}_2) = 1/2 = g(\vec{u}_2)$, and for $k > 2$, $f(\vec{u}_k) < g(\vec{u}_k)$. On the other hand, $f(\vec{u}_k) = h(\vec{u}_k)$, for all k.

Thus, both g and h are greater or equal than f on all its corner points, and since they are concave, from Proposition 3.4 we have

$$\forall \vec{y} \in D^{(n)} \;\; f(\vec{y}) \leq g(\vec{y}) \;\text{ and }\; f(\vec{y}) \leq h(\vec{y}) \tag{5}$$

The rest of the proof proceeds as in [15] and [26]: Let $\vec{x}$ represent an *a priori* distribution on $\mathcal{A}$ and let the above $\vec{y}$ denote the *a posteriori* probabilities on $\mathcal{A}$ with respect to a certain observable o, i.e. $y_j = p(a_j|o) = (p(o|a_j)/p(o))\, x_j$. Then $P_e(\vec{x}) = \sum_o p(o) f(\vec{y})$, so from (5) we obtain

$$P_e(\vec{x}) \leq \sum_o p(o)\frac{1}{2}H(\vec{y}) = \frac{1}{2}H(A|O) \tag{6}$$

and

$$P_e(\vec{x}) \leq \sum_o p(o)(1 - 2^{-H(\vec{y})}) \leq 1 - 2^{-H(A|O)} \tag{7}$$

where the last step in (7) is obtained by observing that $1 - 2^x$ is concave and applying Jensen's inequality. This concludes the alternative proof of (2) and (3).

We end this section with two remarks. First, we note that g coincides with f only on the points of U_1 and U_2, whereas h coincides with f on all U. This explains, intuitively, why (3) is a better bound than (2) for dimensions higher than 2.

Second, we observe that, although h is a good bound for f in the sense that they coincide in all corner points of f, $1 - 2^{-H(A|O)}$ is not necessarily a tight bound for $P_e(\vec{x})$. This is due to the averaging of h, f over the outputs to obtain

$\sum_o p(o)(1 - 2^{-H(\vec{y})})$ and $P_e(\vec{x})$ respectively, and also due to the application of the Jensen's inequality. In fact, we always loosen the bound unless the channel has capacity 0 (maximally noisy channel), as we will see in some examples later. In the general case of non-zero capacity, however, this means that if we want to obtain a better bound we need to follow a different strategy. In particular, we need to find directly the corner points of P_e instead than those of the f defined above. This is what we are going to do in the next section.

4 The corner points of the Bayes risk

In this section we present our main contribution, namely we show that P_e is convexly generated by $(U, P_e(U))$ for a finite U, and we give a constructive characterization of U, so that we can apply the results of the previous section to compute tight bounds on P_e.

The idea behind the construction of such U is the following: recall that the Bayes risk is given by $P_e(\vec{x}) = 1 - \sum_i \max_j p(o_i|a_j)x_j$. Intuitively, this function is linear as long as, for each i, the j which gives the maximum $p(o_i|a_j)x_j$ remains the same while we vary $\vec{x}$. When, for some i and k, the maximum becomes $p(o_i|a_k)x_k$, the function changes its inclination and then it becomes linear again. The exact point in which the inclination changes is a solution of the equation $p(o_i|a_j)x_j = p(o_i|a_k)x_k$. This equation actually represents a hyperplane (a space in $n-1$ dimensions, where n is the cardinality of $\mathcal{A}$) and the inclination of P_e changes in all its points for which $p(o_i|a_j)x_j$ is maximum, i.e. it satisfies the inequality $p(o_i|a_j)x_j \geq p(o_i|a_\ell)x_\ell$ for each ℓ. The intersection of $n-1$ hyperplanes of this kind, and of the one determined by the equation $\sum_j x_j = 1$, is a vertex $\vec{v}$ such that $(\vec{v}, P_e(\vec{v}))$ is a corner point of P_e.

Definition 4.1. *Given a channel $\mathcal{C} = (\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$, the family $\mathbb{S}(\mathcal{C})$ of systems generated by $\mathcal{C}$ is the set of all systems of inequalities of the following form:*

$$\begin{array}{rcll}
p(o_{i_1}|a_{j_1})x_{j_1} & = & p(o_{i_1}|a_{j_2})x_{j_2} & \\
p(o_{i_2}|a_{j_3})x_{j_3} & = & p(o_{i_2}|a_{j_4})x_{j_4} & \\
 & \vdots & & \\
p(o_{i_r}|a_{j_{2r-1}})x_{j_{2r-1}} & = & p(o_{i_r}|a_{j_{2r}})x_{j_{2r}} & \\
x_j & = & 0 \quad \text{for } j \notin \{j_1, j_2, \ldots, j_{2r}\} & \\
x_1 + x_2 + \ldots + x_n & = & 1 & \\
p(o_{i_h}|a_{j_{2h}})x_{j_{2h}} & \geq & p(o_{i_h}|a_\ell)x_\ell & \text{for } 1 \leq h \leq r \\
 & & & \text{and } 1 \leq \ell \leq n
\end{array}$$

such that all the coefficients $p(o_{i_h}|a_{j_{2h-1}})$, $p(o_{i_h}|a_{j_{2h}})$ are strictly positive ($1 \leq h \leq r$), and the equational part has exactly one solution. Here n is the cardinality of $\mathcal{A}$, and r ranges between 0 and $n-1$.

The variables of the above systems of inequalities are $x_1, \ldots, x_n$. Note that for $r = 0$ the system consists only of $n-1$ equations of the form $x_j = 0$, plus the equation $x_1 + x_2 + \ldots + x_n = 1$. A system is called *solvable* if it has solutions.

By definition, a system of the kind considered in the above definition has at most one solution.

The condition on the uniqueness of solution requires to (attempt to) solve more systems than they are actually solvable. Since the number of systems of equations of the form given in Definition 4.1 increases very fast with n, it is reasonable to raise the question of the effectiveness of our method. Fortunately, we will see that the uniqueness of solution can be characterized by a simpler condition (cf. Proposition 4.7), however still producing a huge number of systems. We will investigate the complexity of our method in Section 4.1.

We are now ready to state our main result:

Theorem 4.2. *Given a channel $\mathcal{C}$, the Bayes risk P_e associated with $\mathcal{C}$ is convexly generated by $(U, P_e(U))$, where U is the set of solutions to all solvable systems in $\mathbb{S}(\mathcal{C})$.*

Proof. We need to prove that, for every $\vec{u} \in D^{(n)}$, there exist $\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_t \in U$, and convex coefficients $\lambda_1, \lambda_2, \ldots, \lambda_t$ such that

$$\vec{u} = \sum_i \lambda_i \vec{u}_i \quad \text{and} \quad P_e(\vec{u}) = \sum_i \lambda_i P_e(\vec{u}_i)$$

Let us consider a particular $\vec{u} \in D^{(n)}$. In the following, for each i, we will use j_i to denote the index j for which $p(o_i|a_j)u_j$ is maximum. Hence, we can rewrite $P_e(\vec{u})$ as

$$P_e(\vec{u}) = 1 - \sum_i p(o_i|a_{j_i})u_{j_i} \tag{8}$$

We proceed by induction on n. All conditional probabilities $p(o_i|a_j)$ that appear in the proof are assumed to be strictly positive: we do not need to consider the ones which are zero, because we are interested in maximizing the terms of the form $p(o_i|a_j)x_j$.

Base case ($n = 2$) In this case U is the set of solutions of all the systems of the form

$$\{p(o_i|a_1)x_1 = p(o_i|a_2)x_2 \ , \ x_1 + x_2 = 1\}$$

or

$$\{x_j = 0 \ , \ x_1 + x_2 = 1\}$$

and $\vec{u} \in D^{(2)}$. Let c be the minimum $x \geq 0$ such that

$$p(o_i|a_1)(u_1 - x) = p(o_i|a_2)(u_2 + x) \quad \text{for some } i$$

or let c be u_1 if such x does not exist. Analogously, let d be the minimum $x \geq 0$ such that

$$p(o_i|a_2)(u_2 - x) = p(o_i|a_1)(u_1 + x) \quad \text{for some } i$$

or let d be u_2 if such x does not exist.

Note that $p(o_i|a_2)(u_2+c) \geq 0$, hence $u_1 - c \geq 0$ and consequently $u_2 + c \leq 1$. Analogously, $u_2 - d \geq 0$ and $u_1 + d \leq 1$. Let us define $\vec{v}$, $\vec{w}$ (the corner points of interest) as

$$\vec{v} = (u_1 - c, u_2 + c) \qquad \vec{w} = (u_1 + d, u_2 - d)$$

Consider the convex coefficients

$$\lambda = \frac{d}{c+d} \qquad \mu = \frac{c}{c+d}$$

A simple calculation shows that

$$\vec{u} = \lambda\vec{v} + \mu\vec{w}$$

It remains to prove that

$$P_e(\vec{u}) = \lambda P_e(\vec{v}) + \mu P_e(\vec{w}) \tag{9}$$

To this end, it is sufficient to show that P_e is defined in $\vec{v}$ and $\vec{w}$ by the same formula as (8), i.e. that $P_e(\vec{v})$, $P_e(\vec{w})$ and $P_e(\vec{u})$ are obtained as values, in $\vec{v}$, $\vec{w}$ and $\vec{u}$, respectively, of the same linear function. This amounts to show that the coefficients are the same, i.e. that for each i and k the inequality $p(o_i|a_{j_i})v_{j_i} \geq p(o_i|a_k)v_k$ holds, and similarly for $\vec{w}$.

Let i and k be given. If $j_i = 1$, and consequently $k = 2$, we have that $p(o_i|a_1)u_1 \geq p(o_i|a_2)u_2$ holds. Hence for some $x \geq 0$ the equality $p(o_i|a_1)(u_1 - x) = p(o_i|a_2)(u_2 + x)$ holds. Therefore:

$$\begin{array}{rcll} p(o_i|a_1)v_1 & = & p(o_i|a_1)(u_1 - c) & \text{by definition of } \vec{v} \\ & \geq & p(o_i|a_1)(u_1 - x) & \text{since } c \leq x \\ & = & p(o_i|a_2)(u_2 + x) & \text{by definition of } x \\ & \geq & p(o_i|a_2)(u_2 + c) & \text{since } c \leq x \\ & = & p(o_i|a_1)v_2 & \text{by definition of } \vec{v} \end{array}$$

If, on the other hand, $j_i = 2$, and consequently $k = 1$, we have:

$$\begin{array}{rcll} p(o_i|a_2)v_2 & = & p(o_i|a_2)(u_2 + c) & \text{by definition of } \vec{v} \\ & \geq & p(o_i|a_2)u_2 & \text{since } c \geq 0 \\ & \geq & p(o_i|a_1)u_1 & \text{since } j_i = 2 \\ & \geq & p(o_i|a_1)(u_1 - c) & \text{since } c \geq 0 \\ & = & p(o_i|a_1)v_1 & \text{by definition of } \vec{v} \end{array}$$

The proof that for each i and k the inequality $p(o_i|a_{j_i})w_{j_i} \geq p(o_i|a_k)w_k$ holds is analogous.

Hence we have proved that

$$P_e(\vec{v}) = 1 - \sum_i p(o_i|a_{j_i})v_{j_i} \quad \text{and} \quad P_e(\vec{w}) = 1 - \sum_i p(o_i|a_{j_i})w_{j_i}$$

and a simple calculation shows that (9) holds.

Inductive case Let $\vec{u} \in D^{(n)}$. Let c be the minimum $x \geq 0$ such that for some i and k

$$\begin{aligned} p(o_i|a_{j_i})(u_{j_i} - x) &= p(o_i|a_n)(u_n + x) \quad j_i = n-1 \\ &\text{or} \\ p(o_i|a_{j_i})(u_{j_i} - x) &= p(o_i|a_k)u_k \quad j_i = n-1 \text{ and } k \neq n \\ &\text{or} \\ p(o_i|a_{j_i})u_{j_i} &= p(o_i|a_n)(u_n + x) \quad j_i \neq n-1 \end{aligned}$$

or let c be u_{n-1} if such x does not exist. Analogously, let d be the minimum $x \geq 0$ such that for some i and k

$$\begin{aligned} p(o_i|a_{j_i})(u_{j_i} - x) &= p(o_i|a_{n-1})(u_{n-1} + x) \quad j_i = n \\ &\text{or} \\ p(o_i|a_{j_i})(u_{j_i} - x) &= p(o_i|a_k)u_k \quad j_i = n \text{ and } k \neq n-1 \\ &\text{or} \\ p(o_i|a_{j_i})u_{j_i} &= p(o_i|a_{n-1})(u_{n-1} + x) \quad j_i \neq n \end{aligned}$$

or let d be u_n if such x does not exist. Similarly to the base case, define $\vec{v}$, $\vec{w}$ as

$$\vec{v} = (u_1, u_2, \ldots, u_{n-2}, u_{n-1} - c, u_n + c)$$

and

$$\vec{w} = (u_1, u_2, \ldots, u_{n-2}, u_{n-1} + d, u_n - d)$$

and consider the same convex coefficients

$$\lambda = \frac{d}{c+d} \qquad \mu = \frac{c}{c+d}$$

Again, we have $\vec{u} = \lambda\vec{v} + \mu\vec{w}$.

By case analysis, and following the analogous proof given for $n = 2$, we can prove that for each i and k the inequalities $p(o_i|a_{j_i})v_{j_i} \geq p(o_i|a_k)v_k$ and $p(o_i|a_{j_i})w_{j_i} \geq p(o_i|a_k)w_k$ hold, hence, following the same lines as in the base case, we derive

$$P_e(\vec{u}) = \lambda P_e(\vec{v}) + \mu P_e(\vec{w})$$

We now prove that $\vec{v}$ and $\vec{w}$ can be obtained as convex combinations of corner points of P_e in the hyperplanes (instances of $D^{(n-1)}$) defined by the equations that give, respectively, the c and d above. More precisely, if $c = u_{n-1}$ the equation is $x_{n-1} = 0$. Otherwise, the equation is of the form

$$p(o_i|a_k)x_k = p(o_i|a_\ell)x_\ell$$

and analogously for d. We develop the proof for $\vec{w}$; the case of $\vec{v}$ is analogous.

If $d = u_n$, then the hyperplane is defined by the equation $x_n = 0$, and it consists of the set of vectors of the form $(x_1, x_2, \ldots, x_{n-1})$. The Bayes risk is

defined in this hyperplane exactly in the same way as P_e (since the contribution of x_n is null) and therefore the corner points are the same. By inductive hypothesis, those corner points are given by the solutions to the set of inequalities of the form given in Definition 4.1. To obtain the corner points in $D^{(n)}$ it is sufficient to add the equation $x_n = 0$.

Assume now that d is given by one of the other equations. Let us consider the first one, the cases of the other two are analogous. Let us consider, therefore, the hyperplane $\mathcal{H}$ (instance of $D^{(n-1)}$) defined by the equation

$$p(o_i|a_n)x_n = p(o_i|a_{n-1})x_{n-1} \tag{10}$$

It is convenient to perform a transformation of coordinates. Namely, represent the elements of $\mathcal{H}$ as vectors $\vec{y}$ with

$$y_j = \begin{cases} x_j & 1 \leq j \leq n-2 \\ x_{n-1} + x_n & j = n-1 \end{cases} \tag{11}$$

Consider the channel

$$\mathcal{C}' = \langle \mathcal{A}', \mathcal{O}, p'(\cdot|\cdot) \rangle$$

with $\mathcal{A}' = \{a_1, a_2, \ldots, a_{n-1}\}$, and

$$p'(o_k|a_j) = \begin{cases} p(o_k|a_j) & 1 \leq j \leq n-2 \\ \max\{p_1(k), p_2(k)\} & j = n-1 \end{cases}$$

where

$$p_1(k) = p(o_k|a_{n-1}) \frac{p(o_i|a_n)}{p(o_i|a_{n-1}) + p(o_i|a_n)}$$

($p(o_i|a_n)$ and $p(o_i|a_{n-1})$ are from (10)), and

$$p_2(k) = p(o_k|a_n) \frac{p(o_i|a_{n-1})}{p(o_i|a_{n-1}) + p(o_i|a_n)}$$

The Bayes risk in $\mathcal{H}$ is defined by

$$P_e(\vec{y}) = \sum_k \max_{1 \leq j \leq n-1} p'(o_k|a_j) y_j$$

and a simple calculation shows that $P_e(\vec{y}) = P_e(\vec{x})$ whenever $\vec{x}$ satisfies (10) and $\vec{y}$ and $\vec{x}$ are related by (11). Hence the corner points of $P_e(\vec{x})$ over $\mathcal{H}$ can be obtained from those of $P_e(\vec{y})$.

The systems in $\mathbb{S}(\mathcal{C})$ are obtained from those in $\mathbb{S}(\mathcal{C}')$ in the following way. For each system in $\mathbb{S}(\mathcal{C}')$, replace the equation $y_1 + y_2 + \ldots + y_{n-1} = 1$ by $x_1 + x_2 + \ldots + x_{n-1} + x_n = 1$, and replace, in each equation, every occurrence of y_j by x_j, for j from 1 to $n-2$. Furthermore, if y_{n-1} occurs in an equation E of the form $y_{n-1} = 0$, then replace E by the equations $x_{n-1} = 0$ and $x_n = 0$. Otherwise, it must be the case that for some k_1, k_2, $p'(o_{k_1}|a_{n-1})y_{n-1}$ and $p'(o_{k_2}|a_{n-1})y_{n-1}$

occur in two of the other equations. In that case, replace $p'(o_{k_1}|a_{n-1})y_{n-1}$ by $p(o_{k_1}|a_{n-1})x_{n-1}$ if $p_1(k_1) \geq p_2(k_1)$, and by $p(o_{k_1}|a_n)x_n$ otherwise. Analogously for $p'(o_{k_2}|a_{n-1})y_{n-1}$. Finally, add the equation $p(o_i|a_n)x_n = p(o_i|a_{n-1})x_{n-1}$. It is easy to see that the uniqueness of solution is preserved by this transformation. The conversions to apply on the inequality part are trivial. □

Note that $\mathbb{S}(\mathcal{C})$ is finite, hence the U in Theorem 4.2 is finite as well.

4.1 An alternative characterization of the corner points

In this section we give an alternative characterization of the corner points of the Bayes risk. The reason is that the new characterization considers only systems of equations that are guaranteed to have a unique solution (for the equational part). As a consequence, we need to solve much less systems than those of Definition 4.1. We characterize these systems in terms of graphs.

Definition 4.3. *A* labeled undirected multigraph *is a tuple $G = (V, L, E)$ where V is a set of* vertices*, L is a set of* labels *and $E \subseteq \{(\{v, u\}, l) \mid v, u \in V, l \in L\}$ is a set of* labeled edges *(note that multiple edges are allowed between the same vertices). A graph is* connected *iff there is a path between any two vertices. A* tree *is a connected graph without cycles. We say that a tree $T = (V_T, L_T, E_T)$ is a tree of G iff $V_T \subseteq V, L_T \subseteq L, E_T \subseteq E$.*

Definition 4.4. *Let $\mathcal{C} = (\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ be a channel. We define its associated graph $G(\mathcal{C}) = (V, L, E)$ as $V = \mathcal{A}$, $L = \mathcal{O}$ and $(\{a, a'\}, o) \in E$ iff $p(o|a)$, $p(o|a')$ are both positive.*

Definition 4.5. *Let $\mathcal{C} = (\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ be a channel, let $n = |\mathcal{A}|$ and let $T = (V_T, L_T, E_T)$ be a tree of $G(\mathcal{C})$. The system of inequalities generated by T is defined as*

$$\begin{aligned} p(o_i|a_j)x_j &= p(o_i|a_k)x_k \\ p(o_i|a_j)x_j &\geq p(o_i|a_l)x_l \qquad\qquad \forall\ 1 \leq l \leq n \end{aligned}$$

for all edges $(\{a_j, a_k\}, o_i) \in E_T$, plus the equalities

$$\begin{aligned} x_j &= 0 \qquad\qquad \forall a_j \notin V_T \\ x_1 + \ldots + x_n &= 1 \end{aligned}$$

Let $\mathbb{T}(\mathcal{C})$ be the set of systems generated by all trees of $G(\mathcal{C})$.

An advantage of this characterization is that it allows an alternative, simpler proof of Theorem 4.2. The two proofs differ substantially. Indeed, the new one is non-inductive and uses the proof technique of Proposition 3.8.

Theorem 4.6. *Given a channel $\mathcal{C}$, the Bayes risk P_e associated to $\mathcal{C}$ is convexly generated by $(U, P_e(U))$, where U is the set of solutions to all solvable systems in $\mathbb{T}(\mathcal{C})$.*

Proof. Let $J = \{1, \ldots, |\mathcal{A}|\}, I = \{1, \ldots, |\mathcal{O}|\}$. We define

$$\begin{aligned}
\mathrm{m}(\vec{x}, i) &= \max_{k \in J} p(o_i|a_k)x_k && \text{Maximum for column } i \\
\Psi(\vec{x}) &= \{i \in I \mid \mathrm{m}(\vec{x}, i) > 0\} && \text{Columns with non-zero maximum} \\
\Phi(\vec{x}, i) &= \{j \in J \mid p(o_i|a_j)x_j = \mathrm{m}(\vec{x}, i)\} && \text{Rows giving the maximum for col. } i
\end{aligned}$$

The probability of error can be written as

$$P_e(\vec{x}) = 1 - \sum_{i \in I} p(o_i|a_{j(\vec{x},i)})x_{j(\vec{x},i)} \quad \text{where } j(\vec{x}, i) = \min \Phi(\vec{x}, i) \tag{12}$$

We now fix a point $\vec{x} \notin U$ and we are going to show that there exist $\vec{y}, \vec{z} \in D^{(n)}$ different than $\vec{x}$ such that $(\vec{x}, P_e(\vec{x})) = t(\vec{y}, P_e(\vec{y})) + \bar{t}(\vec{z}, P_e(\vec{z}))$. Let $M(\vec{x})$ be the indexes of the non-zero elements of $\vec{x}$, that is $M(\vec{x}) = \{j \in J \mid x_j > 0\}$ (we will simply write M if $\vec{x}$ is clear from the context. The idea is that we will "slightly" modify some elements in M without affecting any of the sets $\Phi(\vec{x}, i)$. We first define a relation $\sim$ on the set M as

$$j \sim k \quad \text{iff} \quad \exists i \in \Psi(\vec{x}) : j, k \in \Phi(\vec{x}, i)$$

and take $\approx$ as the reflexive and transitive closure of $\sim$ ($\approx$ is an equivalence relation). Now assume that $\approx$ has only one equivalence class, equal to M. Then we can create a tree T as follows: we start from a single vertex a_j, $j \in M$. At each step, we find a vertex a_j in the current tree such that $j \sim k$ for some $k \in M$ where a_k is not yet in the tree (such a vertex always exist since M is an equivalence class of $\approx$). Then we add a vertex a_k and an edge $(\{a_j, a_k\}, o_i)$ where i is the one from the definition of $\sim$. Note that since $i \in \Psi(\vec{x})$ we have that $p(o_i|a_j), p(o_i|a_k)$ are positive so this edge also belongs to $G(\mathcal{C})$. Repeating this procedure creates a tree of $G(\mathcal{C})$ such that $\vec{x}$ is a solution to its corresponding system of inequalities, which is a contradiction since $\vec{x} \notin U$.

So we conclude that $\approx$ has at least two equivalence classes, say C, D. The idea is that we will add/subtract an ϵ from all elements of the class simultaneously, while preserving the relative ratio of the elements. We choose an $\epsilon > 0$ small enough such that $0 < x_j - \epsilon$ and $x_j + \epsilon < 1$ for all $j \in M$ and such that subtracting it from any element does not affect the relative order of the quantities $p(o_i|a_j)x_j$, that is

$$p(o_i|a_j)x_j > p(o_i|a_k)x_k \;\Rightarrow\; p(o_i|a_j)(x_j - \epsilon) > p(o_i|a_k)(x_k + \epsilon) \tag{13}$$

for all $i \in I, j, k \in M$.[2] Then we create two points $\vec{y}, \vec{z} \in D^{(n)}$ as follows:

$$y_j = \begin{cases} x_j - x_j\epsilon_1 & \text{if } j \in C \\ x_j + x_j\epsilon_2 & \text{if } j \in D \\ x_j & \text{otherwise} \end{cases} \qquad z_j = \begin{cases} x_j + x_j\epsilon_1 & \text{if } j \in C \\ x_j - x_j\epsilon_2 & \text{if } j \in D \\ x_j & \text{otherwise} \end{cases}$$

[2] Let $\delta_{i,j,k} = p(o_i|a_j)x_j - p(o_i|a_k)x_k$. It is sufficient to take

$$\epsilon < \min(\{\frac{\delta_{i,j,k}}{p(o_i|a_j) + p(o_i|a_j)} \mid \delta_{i,j,k} > 0\} \cup \{x_j \mid j \in M\})$$

where $\epsilon_1 = \epsilon / \sum_{j\in C} x_j$ and $\epsilon_2 = \epsilon / \sum_{j\in D} x_j$ (note that $x_j\epsilon_1, x_j\epsilon_2 \leq \epsilon$) . It is easy to see that $\vec{x} = \frac{1}{2}\vec{y} + \frac{1}{2}\vec{z}$, it remains to show that $P_e(\vec{x}) = \frac{1}{2}P_e(\vec{y}) + \frac{1}{2}P_e(\vec{z})$.

We notice that $M(\vec{x}) = M(\vec{y}) = M(\vec{z})$ and $\Psi(\vec{x}) = \Psi(\vec{y}) = \Psi(\vec{z})$ since $x_j > 0$ iff $y_j > 0, z_j > 0$. We now compare $\Phi(\vec{x}, i)$ and $\Phi(\vec{y}, i)$. If $i \notin \Psi(\vec{x})$ then $p(o_i|a_k) = 0,\ \forall k \in M$ so $\Phi(\vec{x}, i) = \Phi(\vec{y}, i) = J$. Assuming $i \in \Psi(\vec{x})$, we first show that $p(o_i|a_j)x_j > p(o_i|a_k)x_k$ implies $p(o_i|a_j)y_j > p(o_i|a_k)y_k$. This follows from (13) since

$$p(o_i|a_j)y_j \geq p(o_i|a_j)(x_j - \epsilon) > p(o_i|a_k)(x_k + \epsilon) \geq p(o_i|a_k)y_k$$

This means that $k \notin \Phi(\vec{x}, i) \Rightarrow k \notin \Phi(\vec{y}, i)$, in other words

$$\Phi(\vec{x}, i) \supseteq \Phi(\vec{y}, i) \tag{14}$$

Now we show that $k \in \Phi(\vec{x}, i) \Rightarrow k \in \Phi(\vec{y}, i)$. Assume $k \in \Phi(\vec{x}, i)$ and let $j \in \Phi(\vec{y}, i)$ (note that $\Phi(\vec{y}, i) \neq \emptyset$). By (14) we have $j \in \Phi(\vec{x}, i)$ which means that $p(o_i|a_k)x_k = p(o_i|a_j)x_j$. Moreover, since $i \in \Psi(\vec{x})$ we have that j, k belong to the same equivalence class of $\approx$. If $j, k \in C$ then

$$\begin{aligned} p(o_i|a_k)y_k &= p(o_i|a_k)(x_k - x_k\epsilon_1) \\ &= p(o_i|a_j)(x_j - x_j\epsilon_1) \qquad\qquad p(o_i|a_k)x_k = p(o_i|a_j)x_j \\ &= p(o_i|a_j)y_j \end{aligned}$$

which means that $k \in \Phi(\vec{y}, i)$. Similarly for $j, k \in D$. If $j, k \notin C \cup D$ then $x_k = y_k$, $x_j = y_j$ and the same result is immediate. So we have $\Phi(\vec{x}, i) = \Phi(\vec{y}, i),\ \forall i \in I$. And symmetrically we can show that $\Phi(\vec{x}, i) = \Phi(\vec{z}, i)$. This implies that $j(\vec{x}, i) = j(\vec{y}, i) = j(\vec{z}, i)$ (see (12)) so we finally have

$$\begin{aligned} \frac{1}{2}P_e(\vec{y}) + \frac{1}{2}P_e(\vec{z}) &= \frac{1}{2}\big(1 - \sum_{i\in I} p(o_i|a_{j(\vec{y},i)})y_{j(\vec{y},i)} + 1 - \sum_{i\in I} p(o_i|a_{j(\vec{z},i)})z_{j(\vec{z},i)}\big) \\ &= 1 - \sum_{i\in I} p(o_i|a_{j(\vec{x},i)})(\frac{1}{2}y_{j(\vec{x},i)} + \frac{1}{2}z_{j(\vec{x},i)}) \\ &= P_e(\vec{x}) \end{aligned}$$

Applying Proposition 3.8 completes the proof. □

We now show that both characterizations give the same systems of equations, that is $\mathbb{S}(\mathcal{C}) = \mathbb{T}(\mathcal{C})$.

Proposition 4.7. *Consider a system of inequalities of the form given in Definition 4.1. Then, the equational part has a unique solution if and only if the system is generated by a tree of $G(\mathcal{C})$.*

Proof. **if)** Assume that the system is generated by a tree of $G(\mathcal{C})$. Consider the variable corresponding to the root, say x_1. Express its children x_2, ..., x_k in terms of x_1. That is to say that, if the equation is $ax_1 = bx_2$, then we express x_2 as a/bx_1. At the next step, we express the children of x_2

in terms of x_2 an hence in terms of x_1, ... etc. Finally, we replace all $x_i's$ by their expressions in terms of x_1 in the equation $\sum_i x_i = 1$. This has exactly one solution.

only if) Assume by contradiction that the system is not generated by a tree. Then we we can divide the variables in at least two equivalence classes with respect to the equivalence relation $\approx$ defined in the proof of Theorem 4.6, and we can define the same $\vec{y}$ defined a few paragraphs later. This $\vec{y}$ is a different solution of the same system (also for the inequalities).

□

The advantage of Definition 4.5 is that it constructs directly solvable systems, in contrast to Definition 4.1 which would oblige us to solve all systems of the given form and keep only the solvable ones. We finally give the complexity of computing the corner points of P_e using the tree characterization, which involves counting the number of trees of $G(\mathcal{C})$.

Proposition 4.8. *Let $\mathcal{C} = (\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ be a channel and let $n = |\mathcal{A}|, m = |\mathcal{O}|$. Computing the set of corner points of P_e for $\mathcal{C}$ can be performed in $O(n(nm)^{n-1})$ time.*

Proof. To compute the set of corner points of P_e we need to solve all the systems of inequalities in $\mathbb{T}(\mathcal{C})$. Each of those is produced by a tree of $G(\mathcal{C})$. In the worst case, the matrix of the channel is non-zero everywhere, in which case $G(\mathcal{C})$ is the complete multigraph K_n^m of n vertices, each pair of which is connected by exactly m edges. Let K_n^1 be the complete graph of n vertices (without multiple edges). Cayley's formula ([**?**]) gives its number $\sigma(K_n^1)$ of spanning trees:

$$\sigma(K_n^1) = n^{n-2} \tag{15}$$

We now want to compute the total number $\tau(K_n^1)$ of trees of K_n^1. To create a tree of k vertices, we have $\binom{n}{k}$ ways to select k out of the n vertices of K_n^1 and $\sigma(K_k^1)$ ways to form a tree with them. Thus

$$\begin{aligned}
\tau(K_n^1) &= \sum_{k=1}^{n} \binom{n}{k} \sigma(K_k^1) \\
&= \sum_{k=1}^{n} \frac{n!}{k!(n-k)!}\, k^{k-2} && (15) \\
&= \sum_{k=1}^{n} \frac{1}{(n-k)!}\, (k+1) \cdot \ldots \cdot n \cdot k^{k-2} \\
&\leq \sum_{k=1}^{n} \frac{1}{(n-k)!}\, n^{n-k} \cdot n^{k-2} && k+i \leq n \\
&= n^{n-2} \sum_{l=0}^{n-1} \frac{1}{l!} && \text{set } l = n-k \\
&\leq e \cdot n^{n-2} && \text{since } \textstyle\sum_{l=0}^{\infty} \frac{1}{l!} = e
\end{aligned}$$

thus $\tau(K_n^1) \in O(n^{n-2})$. Each tree of K_n^m can be produced by a tree of K_n^1 by exchanging the edge between two vertices with any of the m available edges in K_n^m. Since a tree of K_n^m has at most $n-1$ edges, for each tree of K_n^1 we can produce at most m^{n-1} trees of K_n^m. Thus

$$\tau(K_n^m) \leq m^{n-1}\tau(K_n^1) \in O(m^{n-1}n^{n-2})$$

Finally, for each tree we have to solve the corresponding system of inequalities. Due to the form of this system, computing the solution can be done in $O(n)$ time by expressing all variables x_i in terms of the root of the tree, and then replace them in the equation $\sum_i x_i = 1$. On the other hand, for each solution we have to verify as many as $n(n-1)$ inequalities, so in total the solution can be found in $O(n^2)$ time. Thus, computing all corner points takes $O(n^2 m^{n-1} n^{n-2}) = O(n(nm)^{n-1})$ time. □

Note that, to improve a bound using Proposition 3.5, we need to compute the maximum ratio $f(\vec{u})/g(\vec{u})$ of all corner points $\vec{u}$. Thus, we need only to compute these points, not to store them. Still, as shown in the above proposition, the number of the systems we need to solve in the general case is huge. However, as we will see in Section 6.1, in certain cases of symmetric channel matrices the complexity can be severely reduced to even polynomial time.

4.2 Examples

Example 4.9 (Binary hypothesis testing)**.** *The case $n = 2$ is particularly simple: the systems generated by $\mathcal{C}$ are all those of the form*

$$\{p(o_i|a_1)x_1 = p(o_i|a_2)x_2 \ , \ x_1 + x_2 = 1\}$$

plus the two systems

$$\{x_1 = 0 \ , \ x_1 + x_2 = 1\}$$
$$\{x_2 = 0 \ , \ x_1 + x_2 = 1\}$$

These systems are always solvable, hence we have $m + 2$ corner points, where we recall that m is the cardinality of $\mathcal{O}$.

Let us illustrate this case with a concrete example: let $\mathcal{C}$ be the channel determined by the following matrix:

	o_1	o_2	o_3
a_1	1/2	1/3	1/6
a_2	1/6	1/2	1/3

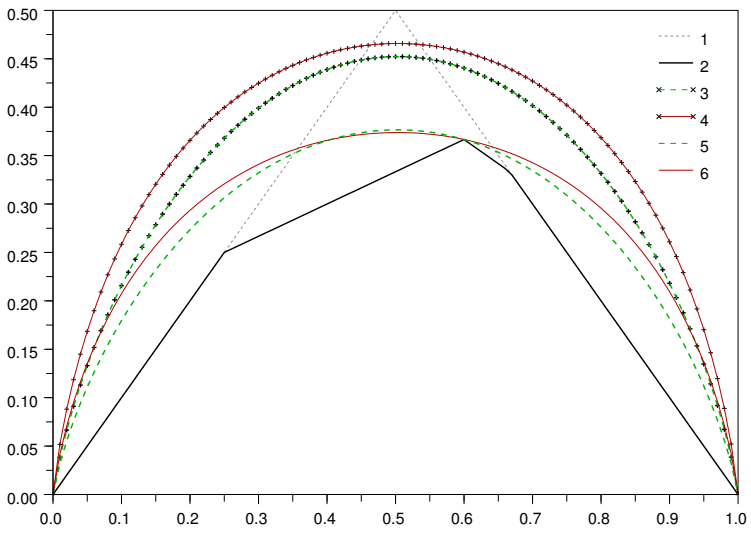

Figure 1: The graph of the Bayes risk for the channel in Example 4.9 and various bounds for it. Curve 1 represents the probability of error if we ignore the observables, i.e. the function $f(\vec{x}) = 1 - \max_j x_j$. Curve 2 represents the Bayes risk $P_e(\vec{x})$. Curve 3 represents the Hellman-Raviv bound $\frac{1}{2}H(A|O)$. Curve 4 represents the Santhi-Vardy bound $1-2^{-H(A|O)}$. Finally, Curves 5 and 6 represent the improvements on 3 and 4, respectively, that we get by applying the method induced by our Proposition 3.5.

The systems generated by $\mathcal{C}$ are:

$$\begin{aligned} \{x_1 = 0 \quad &, \quad x_1 + x_2 = 1\} \\ \{\tfrac{1}{2}x_1 = \tfrac{1}{6}x_2 \quad &, \quad x_1 + x_2 = 1\} \\ \{\tfrac{1}{3}x_1 = \tfrac{1}{2}x_2 \quad &, \quad x_1 + x_2 = 1\} \\ \{\tfrac{1}{6}x_1 = \tfrac{1}{3}x_2 \quad &, \quad x_1 + x_2 = 1\} \\ \{x_1 = 0 \quad &, \quad x_1 + x_2 = 1\} \end{aligned}$$

The solutions of these systems are: $(0,1)$, $(1/4, 3/4)$, $(3/5, 2/5)$, $(2/3, 1/3)$, *and* $(1,0)$, *respectively. The value of P_e on these points is* 0, 1/4, 3/10 *(maximum),* 1/3, *and* 0 *respectively, and P_e is piecewise linear between these points, i.e. it can be generated by convex combination of these points and its value on them. Its graph is illustrated in Figure 1, where x_1 is represented by x and x_2 by $1-x$.*

Example 4.10 (Ternary hypothesis testing)**.** *Let us consider now a channel $\mathcal{C}$ with three inputs. Assume the channel has the following matrix:*

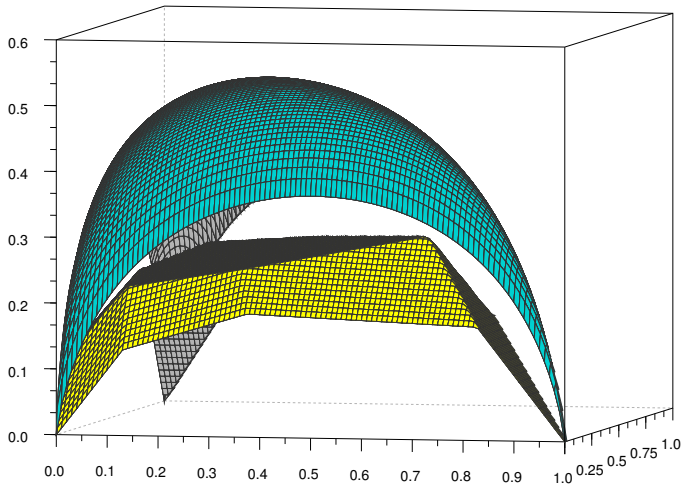

Figure 2: Ternary hypothesis testing. The lower curve represents the Bayes risk for the channel in Example 4.10, while the upper curve represents the Santhi-Vardy bound $1 - 2^{-H(A|O)}$.

	o_1	o_2	o_3
a_1	2/3	1/6	1/6
a_2	1/8	3/4	1/8
a_3	1/10	1/10	4/5

The following is an example of a solvable system generated by $\mathcal{C}$:

$$\begin{aligned} \tfrac{2}{3}x_1 &= \tfrac{1}{8}x_2 \\ \tfrac{1}{8}x_2 &= \tfrac{4}{5}x_3 \\ x_1 + x_2 + x_3 &= 1 \\ \tfrac{2}{3}x_1 &\geq \tfrac{1}{10}x_3 \\ \tfrac{1}{8}x_2 &\geq \tfrac{1}{6}x_1 \end{aligned}$$

Another example is

$$\begin{aligned} \tfrac{1}{6}x_1 &= \tfrac{3}{4}x_2 \\ x_3 &= 0 \\ x_1 + x_2 + x_3 &= 1 \end{aligned}$$

The graph of P_e is depicted in Figure 2, where x_3 is represented by $1-x_1-x_2$.

5 Maximum Bayes risk and relation with strong anonymity

In this section we discuss the Bayes risk in the extreme cases of maximum and minimum (i.e. 0) capacity, and, in the second case, we illustrate the relation with the notion of probabilistic strong anonymity existing in literature.

5.1 Maximum capacity

If the channel has no noise, which means that for each observable o there exists at most one a such that $p(o|a) \neq 0$, then the Bayes risk is 0 for every input distribution. In fact

$$\begin{aligned} P_e(\vec{x}) &= 1 - \textstyle\sum_o \max_j p(o|a_j)x_j \\ &= 1 - \textstyle\sum_j \sum_o p(o|a_j)x_j \\ &= 1 - \textstyle\sum_j x_j = 0 \end{aligned}$$

5.2 Capacity 0

The case in which the capacity of the channel is 0 is by definition obtained when $I(A;O) = 0$ for all possible input distributions of $\mathcal{A}$. From information theory we know that this is the case iff A and O are independent (cf. [10, p.27]). Hence we have the following characterization:

Proposition 5.1 ([10])**.** *The capacity of a channel $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ is 0 iff all the rows of the matrix are the same, i.e. $p(o|a) = p(o) = p(o|a')$ for all $o \in \mathcal{O}$ and $a, a' \in \mathcal{A}$.*

The condition $p(o|a) = p(o|a')$ *for all* o, a, a' has been called *strong probabilistic anonymity* in [1] and it is equivalent to the condition $p(a|o) = p(a)$ *for all* o, a. The latter was considered as a definition of anonymity in [5] and it is called *conditional anonymity* in [14].

Capacity 0 is the optimal case also with respect to the incapability of the adversary of inferring the hidden information. In fact, the Bayes risk achieves its highest possible value, for a given n (cardinality of $\mathcal{A}$), when the rows of the matrix are all the same and the distribution is uniform. To prove this, let $\vec{x} \in D^{(n)}$ and let x_k be the maximum component of $\vec{x}$. We have

$$\begin{aligned} P_e(\vec{x}) &= 1 - \textstyle\sum_o \max_j p(o|a_j)x_j \\ &\leq 1 - \textstyle\sum_o p(o|a_k)x_k \\ &= 1 - x_k \textstyle\sum_o p(o|a_k) \\ &= 1 - x_k \end{aligned}$$

Now, the minimum possible value for x_k is $1/n$, which happens in the case of uniform input distribution. We have therefore

$$P_e(\vec{x}) \leq 1 - \frac{1}{n} = \frac{n-1}{n}$$

namely, $n - 1/n$ is an upper bound of the probability of error. It remains to show that it is a maximum and that it is obtained when the rows are all the same ($p(o|a_j) = p(o|a)$ for all o and j, and some a) and the input distribution is uniform. This is indeed the case, as proven by the following:

$$\begin{aligned} P_e(\tfrac{1}{n}, \tfrac{1}{n}, \ldots, \tfrac{1}{n}) &= 1 - \textstyle\sum_o \max_j p(o|a_j)\tfrac{1}{n} \\ &= 1 - \textstyle\sum_o p(o|a)\tfrac{1}{n} \\ &= 1 - \tfrac{1}{n}\textstyle\sum_o p(o|a) \\ &= \tfrac{n-1}{n} \end{aligned}$$

An example of protocol with capacity 0 is the *dining cryptographers* in a connected graph [5], under the assumption that the payer is always one of the cryptographers, and that the coins are fair.

6 Application: Crowds

In this section we show how to apply the results of the previous sections to the analysis of a security protocol, in order to obtain improved bounds on the attacker's probability of error. This involves modeling the protocol, computing its channel matrix either analytically or using model-checking tools, and using it to compute the corner points of the probability of error. We illustrate our ideas on Crowds, a well-known anonymity protocol from the literature.

In this protocol, introduced by Reiter and Rubin in [23], a user (called the *initiator*) wants to send a message to a web server without revealing its identity. To achieve that, he routes the message through a crowd of users participating in the protocol. The routing is performed in the following way: in the beginning, the initiator randomly selects a user (called a *forwarder*), possibly himself, and forwards the request to him. A forwarder, upon receiving a message, performs a probabilistic choice. With probability p_f (a parameter of the protocol) he selects a new user and forwards once again the message. With probability $1 - p_f$ he sends the message directly to the server.

It is easy to see that the initiator is strongly anonymous with respect to the server, as all users have the same probability of being the forwarder who finally delivers the message. However, the more interesting case is when the attacker is one of the users of the protocol (called a *corrupted* user) which uses his information to find out the identity of the initiator. A corrupted user has more information than the server since he sees other users forwarding the message through him. The initiator, being the first in the path, has greater probability of forwarding the message to the attacker than any other user, so strong anonymity cannot hold. However, under certain conditions on the number of corrupted users, Crowds can be shown to satisfy a weaker notion of anonymity called *probable innocence*.

In our analysis, we consider two network topologies. In the first, used in the original presentation of Crowds, all users are assumed to be able to communicate with any other user, in other words the network graph is a clique. In this case,

the channel matrix is symmetric and easy to compute. Moreover, due to the symmetry of the matrix, the corner points of the probability of error are fewer in number and have a simple form.

However, having a clique network is not always feasible in practice, as it is the case for example in distributed systems. As the task of computing the matrix becomes much harder in a non-clique network, we employ model-checking tools to perform it automatically. The set of corner points, being finite, can also be computed automatically by solving the corresponding systems of inequalities.

6.1 Crowds in a clique network

We consider an instance of Crowds with m users, of which n are honest and $c = m - n$ are corrupted. To construct the matrix of the protocol, we start by identifying the set of anonymous facts, which depends on what the system is trying to hide. In protocols where one user performs an action of interest (like initiating a message in our example) and we want to protect his identity, the set $\mathcal{A}$ would be the set of the users of the protocol. Note that the corrupted users should not be included in this set, since we cannot expect the attacker's own actions to be hidden from him. So in our case we have $\mathcal{A} = \{u_1, \dots u_n\}$ where u_i means that user i is the initiator.

The set of observables should also be defined, based on the visible actions of the protocol and on the various assumptions made about the attacker. In Crowds we assume that the attacker does not have access to the entire network (such an attacker would be too powerful for this protocol) but only to the messages that pass through a corrupted user. Each time a user i forwards the message to a corrupted user we say that he is *detected* which corresponds to an observable action in the protocol. Along the lines of other studies of Crowds (e.g. [28]) we suppose that an attacker will not forward a message himself, since by doing so he would not gain more information. So at each execution of the protocol there is at most one detected user and we have $\mathcal{O} = \{d_1, \dots, d_n\}$ where d_j means that user j was detected.

Now we need to compute the probabilities $p(d_j|u_i)$ for all $1 \leq i, j \leq n$. We first observe some symmetries of the protocol. First, the probability of observing the initiator is the same, independently of who is the initiator. We denote this probability by α. Moreover, the probability of detecting a user other than the initiator is the same for all other users. We denote this probability by β. It can be shown ([23]) that

$$\alpha = c\,\frac{1 - \frac{n-1}{m}p_f}{m - np_f} \qquad\qquad \beta = \alpha - \frac{c}{m}$$

Note that there is also the possibility of not observing any user, if the message arrives to a server without passing through any corrupted user. To compute the matrix, we condition on the event that some user was observed, which is reasonable since otherwise anonymity is not an issue. Thus the conditional

	d_1	d_2	$\dots$	d_{20}
u_1	0.468	0.028	$\dots$	0.028
u_2	0.028	0.468	$\dots$	0.028
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
u_{20}	0.028	0.028	$\dots$	0.468

Figure 3: The channel matrix of Crowds for $n = 20, c = 5, p_f = 0.7$. The events u_i, d_j mean that user i is the initiator and user j was detected respectively.

probabilities of the matrix are:

$$p(d_j|u_i) = \begin{cases} \frac{\alpha}{s} & \text{if } i = j \\ \frac{\beta}{s} & \text{otherwise} \end{cases}$$

where $s = \alpha + (n-1)\beta$. The matrix for $n = 20, c = 5, p_f = 0.7$ is shown in Figure 3.

An advantage of the symmetry is that the corner points of the probability of error for such a matrix have a simple form.

Proposition 6.1. *Let $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ be a channel. Assume that all values of the matrix $p(\cdot|\cdot)$ are either α or β, with $\alpha, \beta > 0$, and that there is at most one α per column. Then all solutions to the systems of Definition 4.5 have at most two distinct non-zero elements, equal to x and $\frac{\alpha}{\beta}x$ for some $x \in (0, 1]$.*

Proof. Since all values of the matrix are either α or β, the equations of all the systems in Definition 4.5 are of the form $x_i = x_j$ or $\alpha \cdot x_i = \beta \cdot x_j$.[3] Assume that a solution of such a system has three distinct non-zero elements $x_1 > x_2 > x_3 > 0$. We consider the following two cases:

1. x_2, x_3 are related to each other by an equation. Since $x_2 > x_3$ this equation can only be $\alpha \cdot x_2 = \beta \cdot x_3$, where $p(o|a_2) = \alpha$ for some observable o. Since there is at most one α per column we have $p(o|a_1) = \beta$ and thus $p(o|a_1)x_1 = \beta\, x_1 > \beta\, x_3 = \alpha\, x_2 = p(o|a_2)x_2$ which violates the inequalities of Definition 4.5.

2. x_2, x_3 are not related to each other. Thus they must be related to x_1 by two equations (assuming $\alpha > \beta$) $\beta \cdot x_1 = \alpha \cdot x_2$ and $\beta \cdot x_1 = \alpha \cdot x_3$. This implies that $x_2 = x_3$ which is a contradiction.

Similarly for more than three non-zero elements. □

The above proposition allows us to efficiently compute the scaling factor of Proposition 3.5 to improve the Santhi-Vardy bound.

[3]Note that by construction of $G(\mathcal{C})$ the coefficients of all equations are non-zero, so in our case either α or β.

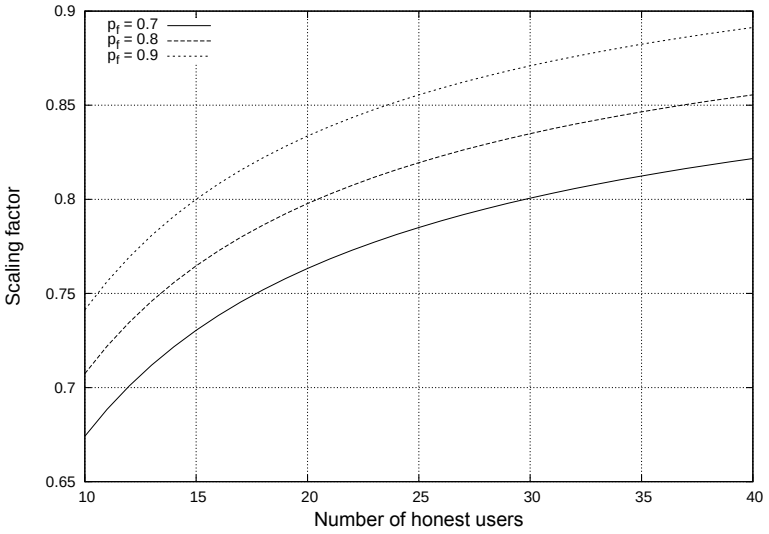

Figure 4: The improvement (represented by the scaling factor) with respect to the Santhi-Vardy bound for various instances of Crowds.

Proposition 6.2. *Let $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ be a channel with $n = |\mathcal{A}|$. Assume that all columns and all rows of the matrix $p(\cdot|\cdot)$ have exactly one element equal to $\alpha > 0$ and all others equal to $\beta > 0$. Then the scaling factor of Proposition 3.5 can be computed in $O(n^2)$ time.*

Proof. By Proposition 6.1, all corner points of P_e have two distinct non-zero elements x and $\frac{\alpha}{\beta}x$. If we fix the number k_1 of elements equal to x and the number k_2 of elements equal to $\frac{\alpha}{\beta}x$ then x can be uniquely computed in constant time. Due to the symmetry of the matrix, P_e as well as the Santhi-Vardy bound will have the same value for all corner points with the same k_1, k_2. So it is sufficient to compute the ratio in only one of them. Then by varying k_1, k_2, we can compute the best ratio without even computing all the corner points. Note that there are $O(n^2)$ possible values of k_1, k_2 and since we need to compute one point for each of them, the total computation can be performed in $O(n^2)$ time. □

We can now apply the algorithm described above to compute the scaling factor $c_o \leq 1$. Multiplying the Santhi-Vardy bound by c_o will give us an improved bound for the probability of error. The results are shown in Figure 4. We plot the obtained scaling factor while varying the number of honest users, for $c = 5$ and for various values of the parameter p_f. A lower scaling factor means a bigger improvement with respect to the Santhi-Vardy bound. We remind that we probability of error, in this case, gives the probability that the attacker "guesses" the wrong sender. The higher it is, the more secure is the protocol. It is worth noting that the scaling factor increases when the number of

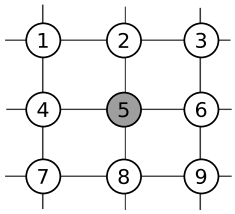

Figure 5: An instance of Crowds with nine users in a grid network. User 5 is the only corrupted one.

honest users increases or when the probability of forwarding increases. In other words, the improvement is better when the probability of error is smaller (and the system is less anonymous). When increasing the number of users (without increasing the number c of corrupted ones), the protocol offers more anonymity and the capacity increases. In this case the Santhi-Vardy bound becomes closer to the corner points of P_e and there is little room for improvement.

6.2 Crowds in a grid network

We now consider a grid-shaped network as shown in Figure 5. In this network there is a total of nine users, each of whom can only communicate with the four that are adjacent to him. We assume that the network "wraps" at the edges, so user 1 can communicate with both user 3 and user 7. Also, we assume that the only corrupted user is user 5.

In this example we have relaxed the assumption of a clique network, showing that a model-checking approach can be used to analyze more complicated network topologies (but of course is limited to specific instances). Moreover, the lack of homogeneity in this network creates a situation where the maximum probability of error is given by a non-uniform input distribution. This emphasizes the importance of abstracting from the input distribution: assuming a uniform one would be not justified in this example.

Similarly to the previous example, the set of anonymous events will be $\mathcal{A} = \{u_1, u_2, u_3, u_4, u_6, u_7, u_8, u_9\}$ where u_i means that user i is the initiator. For the observable events we notice that only the users 2, 4, 6 and 8 can communicate with the corrupted user. Thus we have $\mathcal{O} = \{d_2, d_4, d_6, d_8\}$ where d_j means that user j was detected.

To compute the channel's matrix, we have modeled Crowds in the language of the PRISM model-checker ([**?**]), which is essentially a formalism to describe Markov Decision Processes. PRISM can compute the probability of reaching a specific state starting from a given one. Thus, each conditional probability $p(d_j|u_i)$ is computed as the probability of reaching a state where the attacker has detected user j, starting from the state where i is the initiator. Similarly to the previous example, we compute all probabilities conditioned on the fact that some observation was made, which corresponds to normalizing the rows of the matrix.

	d_2	d_4	d_6	d_8
u_1	0.33	0.33	0.17	0.17
u_3	0.33	0.17	0.33	0.17
u_7	0.17	0.33	0.17	0.33
u_9	0.17	0.17	0.33	0.33
u_2	0.68	0.07	0.07	0.17
u_4	0.07	0.68	0.17	0.07
u_6	0.07	0.17	0.68	0.07
u_8	0.17	0.07	0.07	0.68

Figure 6: The channel matrix of the examined instance of Crowds. The symbols u_i, d_j mean that user i is the initiator and user j was detected respectively.

In Figure 6 the channel matrix is displayed for the examined Crowds instance, computed using probability of forwarding $p_f = 0.8$. We have split the users in two groups, the ones who cannot communicate directly with the corrupted user, and the ones who can. When a user of the first group, say user 1, is the initiator, there is a higher probability of detecting the users that are adjacent to him (users 2 and 4) than the other two (users 6 and 8) since the message needs two steps to arrive to the latters. So $p(d_2|u_1) = p(d_4|u_1) = 0.33$ are greater than $p(d_6|u_1) = p(d_8|u_1) = 0.17$. In the second group users have direct communication to the attacker, so when user 2 is the initiator, the probability $p(d_2|u_2)$ of detecting him is high. From the remaining three observables d_8 has higher probability since user 8 can be reached from user 2 in one step, while users 4 and 6 need two steps. Inside each group the rows are symmetric since the users behave similarly. However between the groups the rows are different which is caused by the different connectivity to the corrupted user 5.

We can now compute the probability of error for this instance of Crowds, which is displayed in the lower curve of Figure 7. Since we have eight users, to plot this function we have to map it to the three dimensions. We do this by considering the users 1, 3, 7, 9 to have the same probability x_1, the users 2, 8 to have the same probability x_2 and the users 4, 6 to have the same probability $1 - x_1 - x_2$. Then we plot P_e as a function of x_1, x_2 in the ranges $0 \leq x_1 \leq 1/4$, $0 \leq x_2 \leq 1/2$. Note that when $x_1 = x_2 = 0$ there are still two users (4, 6) among whom the probability is distributed, so P_e is not 0. The upper curve of Figure 7 shows the Santhi and Vardy's bound on the probability of error. Since all the rows of the matrix are different the bound is not tight, as illustrated.

We can obtain a better bound by applying Proposition 3.5. The set of corner points, characterized by Theorem 4.2, is finite and can be automatically constructed by solving the corresponding systems of inequalities. After finding the corner points, we compute the scaling factor $c_o = \max_u P_e(\vec{u})/h(\vec{u})$, where h is the original bound, and take $c_o \cdot h$ as the improved bound. In our example we found $c_o = 0.925$ which was given for the corner point $\vec{u} =$

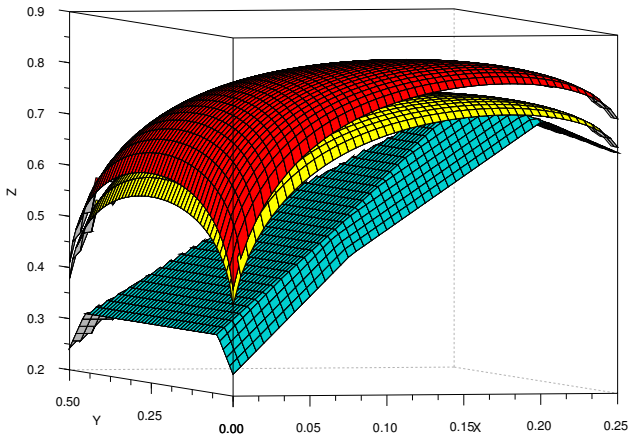

Figure 7: The lower curve is the probability of error in the examined instance of Crowds. The upper two are the Santhi and Vardy's bound and its improved version.

$(0.17, 0.17, 0.17, 0.17, 0.08,\ 0.08, 0.08, 0.08)$.

7 Protocol re-execution

In this section we consider the case in which a protocol is executed multiple times with the same input, either forced by the attacker himself or by some external factor. For instance, in Crowds users send messages along randomly selected routes. For various reasons this path might become unavailable, so the user will need to create a new one, thus re-executing the protocol. If the attacker is part of the path, he could also cause it to fail by not forwarding messages, thus obliging the sender to recreate it (unless measures are taken to prevent this, as it is done in Crowds).

From the point of view of hypothesis testing, the above scenario corresponds to repeating the experiment multiple times while the same hypothesis holds through the repetition. We assume that the the outcomes of the repeated experiments are independent. This corresponds to assuming that the protocol is memoryless, i.e. each time it is reactivated, it works according to the same probability distribution, independently from what happened in previous sessions.

The Bayesian approach to hypothesis testing requires the knowledge of the matrix of the protocol and of the *a priori* distribution of the hypotheses. The first assumption (knowledge of the matrix of the protocol) is usually granted in our setting, because the way the protocol works is public. The second assumption, on the contrary, is not obvious, since the attacker does not usually

know the distribution of the information that is supposed to be concealed by the protocol. However it was showed in [3] that, under certain conditions, the *a priori* distribution becomes less and less relevant with the repetition of the experiment, and it "washes out" at the limit. In this section, we recall briefly the results in [3] and we extend them by proving a lower bound on the limit of the Bayes risk.

Let $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$ be the channel of a protocol S. The experiment obtained by re-executing the protocol n times with the same event a as input will be denoted by S^n. The observables in S^n are sequences $\vec{o} = (o_1, \ldots, o_n)$ of observables of S and, since we consider the repetitions to be independent, the conditional probabilities for S^n will be given by[4]

$$p(\vec{o}|a) = \prod_{i=1}^{n} p(o_i|a) \tag{16}$$

Let $f_n : \mathcal{O}^n \to \mathcal{A}$ be the decision function adopted by the adversary to infer the anonymous action from the sequence of observable. Also let $E_{f_n} : \mathcal{A} \to 2^{\mathcal{O}^n}$ be the error region of f_n and let $\eta_{f_n} : \mathcal{A} \to [0, 1]$ be the function that associates to each $a \in \mathcal{A}$ the probability of inferring the wrong input event on the basis of f_n, namely $\eta_{f_n}(a) = \sum_{\vec{o} \in E_{f_n}(a)} p(\vec{o}|a)$. Then the probability of error of f_n will be the expected value of $\eta_{f_n}(a)$:

$$P_{f_n} = \sum_{a \in A} p(a)\eta_{f_n}(a)$$

The MAP rule and the notion of MAP decision function can be extended to the case of protocol re-execution in the obvious way. Namely a MAP decision function in the context of protocol repetition is a function f_n such that for each $\vec{o} \in \mathcal{O}^n$ and $a, a' \in \mathcal{A}$

$$f_n(\vec{o}) = a \Rightarrow p(\vec{o}|a)p(a) \geq p(\vec{o}|a')p(a')$$

Also in the case of protocol repetition the MAP rule gives the best possible result, namely if f_n is a MAP decision function then $P_{f_n} \leq P_{h_n}$ for any other decision function h_n.

The following definition establishes a condition on the matrix under which the knowledge of the input distribution becomes irrelevant for hypothesis testing.

Definition 7.1 ([3])**.** *Given a protocol with channel $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$, we say that the protocol is determinate iff all rows of the matrix p are pairwise different, i.e. the probability distributions $p(\cdot|a)$, $p(\cdot|a')$ are different for each pair a, a' with $a \neq a'$.*

[4] With a slight abuse of notations we denote by p the probability matrix of both S and S^n. It will be clear from the context to which we refer to.

Next proposition shows that if a protocol is determinate, then it can be approximated by a decision function which compares only the elements along the column corresponding to the observed event, without considering the input probabilities.

Proposition 7.2 ([3]). *Given a determinate protocol $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$, for any distribution on $\mathcal{A}$, any MAP decision functions f_n and any decision function $g_n : \mathcal{O}^n \to \mathcal{A}$ such that*

$$g_n(\vec{o}) = a \;\Rightarrow\; p(\vec{o}|a) \geq p(\vec{o}|a') \quad \forall \vec{o} \in \mathcal{O}^n \forall a, a' \in \mathcal{A}$$

we have that g_n approximates f_n. Namely, for any $\epsilon > 0$, there exists n such that the probability of the set $\{\vec{o} \in \mathcal{O}^n \mid f_n(\vec{o}) \neq g_n(\vec{o})\}$ is smaller than ϵ.

The conditional probability $p(o|a)$ (resp. $p(\vec{o}|a)$) is called *likelihood* of a given o (resp. $\vec{o}$). The criterion for the definition of g_n used in Proposition 7.2 is to choose the a which maximizes the likelihood of o (resp. $\vec{o}$), and it is known in literature as the *Maximum Likelihood criterion* (ML). This rule is quite popular in statistic, its advantage over the Bayesian approach being that it does not require any knowledge of the *a priori* probability on A.

When the protocol is determinate, the probability of error associated to the ML rule converges to 0, as shown by the following proposition. The same holds, of course, for the MAP rule, because of Proposition 7.2.

Proposition 7.3 ([3]). *Given a determinate protocol $(\mathcal{A}, \mathcal{O}, p(\cdot|\cdot))$, for any distribution $p_{\mathcal{A}}$ on $\mathcal{A}$ and for any $\epsilon > 0$, there exists n such that the property*

$$g_n(\vec{o}) = a \;\Rightarrow\; p(\vec{o}|a) \geq p(\vec{o}|a') \quad \forall a' \in \mathcal{A}$$

determines a unique decision function g_n on a set of probability greater than $1 - \epsilon$, and the probability of error P_{g_n} is smaller than ϵ.

One extreme case of determinate matrix is when the capacity is maximum. In this case the probability of error of the MAP and ML rules is always 0, independently from n. The proof is analogous to the one of Section 5.1.

Consider now the case in which determinacy does not hold, i.e. when there are at least two identical rows in the matrix, say a_1 and a_2. In such case, for the sequences $\vec{o} \in \mathcal{O}^n$ such that $p(\vec{o}|a_1)$ (or equivalently $p(\vec{o}|a_2)$) is maximum, the value of a ML function g_n is not uniquely determined, because we could choose either a_1 or a_2. Hence we have more than one ML decision function.

More generally, if there are k identical rows corresponding to $a_1, a_2, \ldots, a_k$, the ML criterion gives k different possibilities every time we get an observable $\vec{o} \in \mathcal{O}^n$ for which $p(\vec{o}|a_1)$ is maximum. Intuitively this is a situation which may induce an error which is difficult to get rid of, even by repeating the protocol many times.

The situation is different if we know the *a priori* distribution and we use a MAP function f_n. In this case we have to maximize $p(a)p(\vec{o}|a)$ and even in case of identical rows, the *a priori* knowledge can help to make a sensible guess about the most likely a.

Both in the case of the ML and of the MAP functions, however, we can show that the probability of error is bound from below by an expression that depends on the probabilities of $a_1, a_2, \ldots, a_k$ only. In fact, we can show that this is the case for *any* decision function, whatever criterion they use to select the hypothesis.

Proposition 7.4. *If the matrix has identical rows corresponding to $a_1, a_2, \ldots, a_k$ then for any n and any decision function h_n we have that*

$$P_{h_n} \geq (k-1)\min_{1\leq i\leq k}\{p(a_i)\}$$

Proof. Assume that $p(a_\ell) = \min_{1\leq i\leq k}\{p(a_i)\}$. We have:

$$\begin{aligned}
P_{h_n} &= \sum_{a\in\mathcal{A}} p(a)\eta_{f_n}(a) \\
&\geq \sum_{1\leq i\leq k} p(a_i)\eta_{f_n}(a_i) \\
&\geq \sum_{1\leq i\leq k} p(a_\ell)\eta_{f_n}(a_i) && (p(a_\ell) = \min_{1\leq i\leq k}\{p(a_i)\}) \\
&= \sum_{1\leq i\leq k} p(a_\ell) \sum_{h_n(\vec{o})\neq a_i} p(\vec{o}|a_i) \\
&= \sum_{1\leq i\leq k} p(a_\ell) \sum_{h_n(\vec{o})\neq a_i} p(\vec{o}|a_\ell) && (p(\vec{o}|a_i) = p(\vec{o}|a_\ell)) \\
&= p(a_\ell) \sum_{1\leq i\leq k} \sum_{h_n(\vec{o})\neq a_i} p(\vec{o}|a_\ell) \\
&= p(a_\ell) \sum_{1\leq i\leq k} (1 - \sum_{h_n(\vec{o})=a_i} p(\vec{o}|a_\ell)\ ) \\
&\geq (k-1)p(a_\ell) && (\textstyle\sum_{1\leq i\leq k}\sum_{h_n(\vec{o})=a_i} p(\vec{o}|a_\ell) \leq 1)
\end{aligned}$$

□

Note that the expression $(k-1)p(a_\ell)$ does not depend on n. Assuming that the a_i's have positive probability, from the above proposition we derive that the probability of error is always greater than a constant strictly greater than 0. Hence the probability of error does not converge to 0.

Corollary 7.5. *If there exist $a_1, a_2, \ldots, a_k$ with positive probability, $k \geq 2$, and whose corresponding rows in the matrix are identical, then for any n and any decision function h_n the probability of error is bound from below by a positive constant.*

Remark 7.6. *In Proposition 7.4 we are allowed to consider any subset of identical rows. In general it is not necessarily the case that a larger subset gives a better bound. In fact, as the subset increases, k increases too, but the minimal*

$p(a_i)$ may decrease. To find the best bound in general one has to consider all the possible subsets of identical rows.

Capacity 0 is the extreme case of identical rows: it corresponds, in fact, to the situation in which all the rows of the matrix are identical. This is, of course, the optimal case with respect to information-hiding. All the rows are the same, consequently the observations are of no use for the attacker to infer the input event, i.e. to define the "right" $g_n(\vec{o})$, since all $p(\vec{o}|a)$ are maximum.

The probability of error of any decision function is bound from below by $(|\mathcal{A}| - 1) \min_i p(a_i)$. Note that by Remark 7.6 we may get better bounds by considering subsets of the rows instead than all of them.

8 Conclusion and future work

In this paper we have investigated the hypothesis testing problem from the point of view of an adversary playing against an information-hiding protocol, seen as a channel in the information-theoretic sense. We have considered the Bayesian approach to hypothesis testing, and specifically the Maximum Aposteriori Probability (MAP) rule. We have shown that the function P_e expressing the probability of error for the MAP rule is piecewise linear, and we have given a constructive characterization of a special set of points which allows computing the maximum P_e over all probability distributions on the channel's inputs. This set of points is determined uniquely by the matrix associated to the channel. As a byproduct of this study, we have also improved both the Hellman-Raviv and the Santhi-Vardy bounds.

A common objection to the Bayesian approach to hypothesis testing is that it requires the knowledge of the input distribution (*a priori* probability). This is a valid criticism in our setting as well, since in general the adversary does not have *a priori* knowledge of the hidden information. Under certain conditions depending on the protocol's matrix, however, the adversary may be able to infer the input distribution with arbitrary precision by repeatedly observing the outcome of consecutive sessions. Our plans for future work include the investigation of the conditions under which such inference is possible, and the study of the corresponding probability of error as a function of the matrix.

References

[1] Mohit Bhargava and Catuscia Palamidessi. Probabilistic anonymity. In Martín Abadi and Luca de Alfaro, editors, *Proceedings of CONCUR*, volume 3653 of *Lecture Notes in Computer Science*, pages 171–185. Springer, 2005.

[2] Konstantinos Chatzikokolakis and Catuscia Palamidessi. Probable innocence revisited. *Theoretical Computer Science*, 367(1-2):123–138, 2006.

[3] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. Anonymity protocols as noisy channels. *Information and Computation*, 2007. To appear.

[4] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Prakash Panangaden. Probability of error in information-hiding protocols. In *Proceedings of the 20th IEEE Computer Security Foundations Symposium (CSF20)*, pages 341–354. IEEE Computer Society, 2007.

[5] David Chaum. The dining cryptographers problem: Unconditional sender and recipient untraceability. *Journal of Cryptology*, 1:65–75, 1988.

[6] David Clark, Sebastian Hunt, and Pasquale Malacaria. Quantitative analysis of the leakage of confidential data. In *Proc. of QAPL 2001*, volume 59 (3) of *Electr. Notes Theor. Comput. Sci*, pages 238–251. Elsevier Science B.V., 2001.

[7] David Clark, Sebastian Hunt, and Pasquale Malacaria. Quantified interference for a while language. In *Proc. of QAPL 2004*, volume 112 of *Electr. Notes Theor. Comput. Sci*, pages 149–166. Elsevier Science B.V., 2005.

[8] Ian Clarke, Oskar Sandberg, Brandon Wiley, and Theodore W. Hong. Freenet: A distributed anonymous information storage and retrieval system. In *Designing Privacy Enhancing Technologies, International Workshop on Design Issues in Anonymity and Unobservability*, volume 2009 of *Lecture Notes in Computer Science*, pages 44–66. Springer, 2000.

[9] Michael R. Clarkson, Andrew C. Myers, and Fred B. Schneider. Belief in information flow. *Journal of Computer Security*. To appear. Available as Cornell Computer Science Department Technical Report TR 2007-207.

[10] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[11] Yuxin Deng, Jun Pang, and Peng Wu. Measuring anonymity with relative entropy. In *Proceedings of the 4th International Workshop on Formal Aspects in Security and Trust (FAST)*, Lecture Notes in Computer Science. Springer, 2006. To appear.

[12] Claudia Díaz, Stefaan Seys, Joris Claessens, and Bart Preneel. Towards measuring anonymity. In Roger Dingledine and Paul F. Syverson, editors, *Proceedings of the workshop on Privacy Enhancing Technologies (PET) 2002*, volume 2482 of *Lecture Notes in Computer Science*, pages 54–68. Springer, 2002.

[13] J. W. Gray, III. Toward a mathematical foundation for information flow security. In *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy (SSP '91)*, pages 21–35, Washington - Brussels - Tokyo, May 1991. IEEE.

[14] Joseph Y. Halpern and Kevin R. O'Neill. Anonymity and information hiding in multiagent systems. *Journal of Computer Security*, 13(3):483–512, 2005.

[15] M.E. Hellman and J. Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Trans. on Information Theory*, IT–16:368–372, 1970.

[16] Gavin Lowe. Quantifying information flow. In *Proc. of CSFW 2002*, pages 18–31. IEEE Computer Society Press, 2002.

[17] Ueli M. Maurer. Authentication theory and hypothesis testing. *IEEE Transactions on Information Theory*, 46(4):1350–1356, 2000.

[18] John McLean. Security models and information flow. In *IEEE Symposium on Security and Privacy*, pages 180–189, 1990.

[19] Ira S. Moskowitz, Richard E. Newman, Daniel P. Crepeau, and Allen R. Miller. Covert channels and anonymizing networks. In Sushil Jajodia, Pierangela Samarati, and Paul F. Syverson, editors, *WPES*, pages 79–88. ACM, 2003.

[20] Ira S. Moskowitz, Richard E. Newman, and Paul F. Syverson. Quasi-anonymous channels. In *IASTED CNIS*, pages 126–131, 2003.

[21] Alessandra Di Pierro, Chris Hankin, and Herbert Wiklicky. Approximate non-interference. *Journal of Computer Security*, 12(1):37–82, 2004.

[22] Alessandra Di Pierro, Chris Hankin, and Herbert Wiklicky. Measuring the confinement of probabilistic systems. *Theoretical Computer Science*, 340(1):3–56, 2005.

[23] Michael K. Reiter and Aviel D. Rubin. Crowds: anonymity for Web transactions. *ACM Transactions on Information and System Security*, 1(1):66–92, 1998.

[24] Alfred Rényi. On the amount of missing information and the Neyman-Pearson lemma. In *Festschriftf for J. Neyman*, pages 281–288. Wiley, New York, 1966.

[25] H. L. Royden. *Real Analysis*. Macmillan Publishing Company, New York, third edition, 1988.

[26] Nandakishore Santhi and Alexander Vardy. On an improvement over Rényi's equivocation bound, 2006. Presented at the 44-th Annual Allerton Conference on Communication, Control, and Computing, September 2006. Available at http://arxiv.org/abs/cs/0608087.

[27] Andrei Serjantov and George Danezis. Towards an information theoretic metric for anonymity. In Roger Dingledine and Paul F. Syverson, editors, *Proceedings of the workshop on Privacy Enhancing Technologies (PET) 2002*, volume 2482 of *Lecture Notes in Computer Science*, pages 41–53. Springer, 2002.

[28] V. Shmatikov. Probabilistic model checking of an anonymity system. *Journal of Computer Security*, 12(3/4):355–377, 2004.

[29] P.F. Syverson, D.M. Goldschlag, and M.G. Reed. Anonymous connections and onion routing. In *IEEE Symposium on Security and Privacy*, pages 44–54, Oakland, California, 1997.