Computing Science Group

THE UNIFORM PRIOR AND ZERO INFORMATION: A TECHNICAL NOTE

Andrew D. Ker

CS-RR-10-06

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford OX1 3QD

Abstract

Reasoning about maximal performance of hypothesis tests is difficult when there is supposed to be no information about some of the parameters. A common technique is to place a uniform prior on unknowns. In a recent steganography application [1], this gave a nonsensical answer, and we give an exposition and resolution of the paradox here. This note is a companion to [1].

1 Introduction

Theoretical results about the capacity of stegosystems requires a bound on the performance of any detector. Such a bound is usually obtained by computing Kullback-Leibler (KL) divergence (also known as relative entropy) and showing that it equals, or tends to, zero under certain conditions. Using KL divergence to bound the performance of hypothesis tests, effectively giving minimum false positive and/or false negative error rates, is well-established: the classic literature for the application to steganography is [2], but the general idea goes back at least to the 1960s [3].

However, the KL divergence bound applies to tests between two *simple* hypotheses, when there are no unknown parameters in either case. This is not always the case in steganography, for example if the detector knows that some objects may contain a hidden payload, but not which: an example of a paradox occurring from the use of KL divergence in precisely this situation is given in [4]. Recently, we worked on a steganalysis problem where the detector does not know the exact distribution of cover objects, but instead has to estimate it from a cover oracle [1]; this creates a similar issue.

A common technique, for reasoning about unknown parameters, is to place a uniform prior on them. This was the first line of attack we tried, when working on the problem in [1]. However, we found that it gave a paradoxical answer, and instead turned to other techniques (the statistical concept of unbiasedness) to reason about a lack of information. In this note, a companion to [1], we give an exposition of that paradox, which could not be included there for space reasons. The calculations are fairly horrible and even here we do not include every detail.

The paradox is resolved once we understand that the uniform prior does not correctly model a complete lack of information about unknown parameters. The uniform prior is, itself, a form of information, and is something that the embedding of hidden data alters.

Manuscript version 4 June 2010.

2 The Coin Toss Equal Probability Problem

This is an abstraction of the steganalysis problem, designed to illustrate the difficulties of bounding detector performance when the detector does not know the distribution of covers. We reduce the problem to independent identically distributed (i.i.d.) binary sequences, so that the distribution is specified by the probability of each bit taking value 1 and the sequences can be reduced to their sufficient statistic, which counts the total number of 1s.

Scenario. Suppose that a detector has two independent sequences of random bits. One, of length m, has each bit taking value 1 with probability p; the other, of length n, may contain a hidden payload and the effect is supposed to be that bits take value 1 with probability $\gamma q + (1-\gamma)p$, where γ parameterises the size of the payload and $q \neq p$[1]. All bits are independent of all others.

The detector wishes to construct an asymptotically perfect hypothesis test, in the sense that type I and type II error rates tend to zero as $m, n \to \infty$, distinguishing the cases of $\gamma = 0$ (no payload in the second stream) and $\gamma > 0$ (some payload in the second stream). But the detector knows nothing about the true value of p: they can only deduce imperfect information about the value of p from the first bit sequence. There are two subcases we will consider:

(A) the detector does know the value of q, or

(B) they have no information about q.

These are analogous to a detector which does, or does not, know the embedding operation potentially in use.

It is crucial that p be unknown to the detector, otherwise they can simply ignore the first stream (their cover oracle which informs them about the distribution of covers) and test whether the second stream matches that probability. See [1] for more details.

We make two observations about this scenario:

1. If $m = 0$ then it is impossible for the detector to behave any better than random, for they have no information about covers and therefore cannot distinguish a stego stream.

2. As $m \to \infty$, the detector can deduce more and more accurate information about p from their cover stream. So we expect the situation to approach that of the classic Square Root Law of capacity, with an asymptotically perfect detector possible if $\gamma\sqrt{n} \to \infty$ and all detectors asymptotically random if $\gamma\sqrt{n} \to 0$.

[1] Thus the embedding operation mixes the cover sequence, i.i.d. bits which are 1 with probability p, with payload bits, i.i.d. bits which are 1 with a different probability q, in proportion $(1-\gamma)$:γ. This is akin to overwriting proportion γ least significant bits of a pseudorandom sequence of pixels in a digital image.

The scenario is examined in [1] where it is shown that, no matter whether we consider case (A) or (B), such a detector is possible if $p \neq q$ and

$$\frac{\gamma}{\sqrt{\frac{1}{m} + \frac{1}{n}}} \to \infty. \tag{1}$$

Conversely, if the ratio instead tends to zero then every detector tends asymptotically to random output, subject to a fairness condition on the detector which, it is argued, imposes the condition of zero knowledge about p. This result quantifies how the amount of hidden information may grow, depending on the size of the sequences available to the detector; it is a modified square root law [5–7] and it accords with our expectations listed as 1. and 2., above.

Before coming to the fairness condition of [1], we first tried to model the lack of knowledge of p by placing a uniform prior on it, also for q in case (B). However, this leads to a false result, as we shall now explore. Consider, then, the following hypothesis tests as models for the scenario. Writing $\mathrm{U}[-,-]$ and $\mathrm{Bi}(-,-)$ for the uniform and binomial distributions, and $X|_A$ for the conditional distribution of X given A, we have:

(A)

$$\begin{aligned} q &\ \text{a known constant} \\ P &\sim \mathrm{U}[0,1], \\ X|_{P=p} &\sim \mathrm{Bi}(m,p), \\ Y|_{P=p} &\sim \mathrm{Bi}\big(n, \gamma q + (1-\gamma)p\big), \end{aligned}$$

$$\begin{aligned} H_0 &: \ \gamma = 0, \\ H_1 &: \ 0 < \gamma \leq 1. \end{aligned}$$

(B)

$$\begin{aligned} P &\sim \mathrm{U}[0,1], \\ Q &\sim \mathrm{U}[0,1], \\ X|_{P=p} &\sim \mathrm{Bi}(m,p), \\ Y|_{P=p,\,Q=q} &\sim \mathrm{Bi}\big(n, \gamma q + (1-\gamma)p\big), \end{aligned}$$

$$\begin{aligned} H_0 &: \ \gamma = 0, \\ H_1 &: \ 0 < \gamma \leq 1. \end{aligned}$$

As usual, the null hypothesis is that no payload is in the second sequence, and the alternative is that some positive proportion of pixels have been replaced by payload. We will argue that these hypothesis tests *do not* properly model the scenario we are considering.

2.1 Computing KL Divergence

It is natural to calculate KL divergence between the observations in the cases of H_0 and H_1, and hope to show that it tends to zero given sufficiently slow growth of m and n with respect to γ.

First, note that, unconditionally, in case (A) we have the probability distribution

$$p(x,y;\gamma) = \int_0^1 \binom{m}{x}\binom{n}{y} p^x (1-p)^{m-x} \big(\gamma q + (1-\gamma)p\big)^y \big(1 - \gamma q - (1-\gamma)p\big)^{n-y} \, \mathrm{d}p. \tag{2}$$

In case (B) we have the alternative distribution

$$p'(x,y;\gamma) = \int_0^1 p(x,y;\gamma)\, \mathrm{d}q,$$

but let us consider only case (A) for the moment.

The KL divergence between the observations available to the detector, which is the pair (X, Y), under the cases H_0 and H_1, can be written as

$$D_{\mathrm{KL}}\big((X,Y)|H_0 \,\|\, (X,Y)|H_1\big) = -\mathrm{E}_{H_0}\left[\log\left(\frac{p(X,Y;\gamma)}{p(X,Y;0)}\right)\right]. \tag{3}$$

However, the integrals in (2) mean that (3) cannot be written in terms of elementary functions (or indeed, apparently, evaluated at all). Instead we look for its asymptotic properties as $\gamma \to 0$, and we can make use of the fact that (2) is extremely well-behaved analytically at least as long as $q \neq 0, 1$: it is a continuous function on a bounded interval and remains defined in a region of its boundary. So we may take power series and interchange derivatives and integrals with freedom.

Write

$$Q_1(x,y) = \frac{\frac{\partial p}{\partial \gamma}\big|_{\gamma=0}}{p(x,y;0)}, \quad Q_2(x,y) = \frac{\frac{\partial^2 p}{\partial \gamma^2}\big|_{\gamma=0}}{p(x,y;0)},$$

and take a Taylor expansion of $\log\big(p(x,y;\gamma)/p(x,y;0)\big)$ about $\gamma = 0$. We reach

$$\begin{aligned} D_{\mathrm{KL}}\big((X,Y)|H_0 \,\|\, (X,Y)|H_1\big) &= -\mathrm{E}_{H_0}\Big[\log\big(1 + \gamma Q_1(X,Y) + \tfrac{\gamma^2}{2} Q_2(X,Y) + O(\gamma^3)\big)\Big] \\ &= -\mathrm{E}_{H_0}\big[\gamma Q_1(X,Y) + \tfrac{\gamma^2}{2} Q_2(X,Y)\big] + \tfrac{\gamma^2}{2}\mathrm{E}_{H_0}\big[Q_1(X,Y)^2\big] \\ &\quad + O(\gamma^3). \end{aligned}$$

(This calculation is similar to the expansion in [3, §2.6].) The first term is zero because, taking the partial derivative outside the expectation, it is the derivative of a constant. Thus

$$D_{\mathrm{KL}}\big((X,Y)|H_0 \,\|\, (X,Y)|H_1\big) \sim \tfrac{\gamma^2}{2}\mathrm{E}\big[Q_1(X,Y)^2\big], \tag{4}$$

with the expectation taken over $P \sim \mathrm{U}[0,1]$, $X|_{P=p} \sim \mathrm{Bi}(m,p)$, $Y|_{P=p} \sim \mathrm{Bi}(n,p)$. The next step is to calculate this expectation.

Differentiating (2), inside the integral, with respect to γ, and using

$$\int_0^1 p^a(1-p)^b \,\mathrm{d}p = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)},$$

gives, after elementary but tedious calculations,

$$Q_1(x,y) = (m+n+1)\left(q\frac{y}{x+y} - (q-1)\frac{n-y}{m+n-x-y}\right) - n. \tag{5}$$

There are special cases for $x, y = 0$ and $x = m$, $y = n$, but they may be dealt with concisely if we conventionally say that a fraction $0/0$ indicates 0, in which case (5) holds for all x and y.

We have already noted that $\mathrm{E}[Q_1(X,Y)] = 0$ so

$$\begin{aligned} \mathrm{E}\big[Q_1(X,Y)^2\big] &= (m+n+1)^2\Big(q^2\mathrm{E}\Big[\tfrac{Y^2}{(X+Y)^2}\Big] + (q-1)^2\mathrm{E}\Big[\tfrac{(n-Y)^2}{(m+n-X-Y)^2}\Big] \\ &\quad -2q(q-1)\mathrm{E}\Big[\tfrac{Y(n-Y)}{(X+Y)(m+n-X-Y)}\Big]\Big) - n^2. \end{aligned} \tag{6}$$

We need to know about the expectations of quotients like $\big(Y/(X+Y)\big)^2$, when X and Y are independent binomials. This turns out to be rather ugly, involving manipulation of combinatorial coefficients and solving a recurrence relation, so we will not include the proofs of the following statements. The author would be delighted to hear of an alternative, elementary, proof.

Lemma. *Let* $P \sim \mathrm{U}[0,1]$*,* $X|_{P=p} \sim \mathrm{Bi}(m,p)$*,* $Y|_{P=p} \sim \mathrm{Bi}(n,p)$*. Then, with the convention that* $0/0 = 0$*,*

$$\mathrm{E}_{X,Y}\left[\frac{Y^2}{(X+Y)^2}\right] = \frac{mn}{(m+n)(m+n-1)}\left(\sum_{i=1}^{m+n-1}\frac{(1-p)^i-(1-p)^{m+n}}{m+n-i}\right) + \frac{n^2}{(m+n)^2}\big(1-(1-p)^{m+n}\big), \tag{7}$$

and so

$$\mathrm{E}_{P,X,Y}\left[\frac{Y^2}{(X+Y)^2}\right] = \frac{n}{(m+n)(m+n+1)}\left(\frac{m}{m+n-1}\big(H(m+n)-1\big)+n\right), \tag{8}$$

where $H(k)$ *is the* k*-th Harmonic number,* $H(k) = \sum_{i=1}^{k} 1/i$*.*

Swapping p *and* $1-p$ *swaps* X *for* $m-X$ *and* Y *for* $n-Y$*, so,* $\mathrm{E}_{X,Y}\left[\frac{(n-Y)^2}{(m+n-X+Y)^2}\right]$ *is (7) with* $1-p$ *for* p*, and hence also*

$$\mathrm{E}_{P,X,Y}\left[\frac{(n-Y)^2}{(m+n-X-Y)^2}\right] = \frac{n}{(m+n)(m+n+1)}\left(\frac{m}{m+n-1}\big(H(m+n)-1\big)+n\right). \tag{9}$$

$$\mathrm{E}_{X,Y}\left[\frac{Y(n-Y)}{(X+Y)(m+n-X-Y)}\right] = \frac{n(n-1)}{(m+n)(m+n-1)}\big(1-p^{m+n}-(1-p)^{m+n}\big),$$

and so

$$\mathrm{E}_{P,X,Y}\left[\frac{Y(n-Y)}{(X+Y)(m+n-X-Y)}\right] = \frac{n(n-1)}{(m+n)(m+n+1)}. \tag{10}$$

Finally, we have the classic result

$$H(n) \sim \ln n + \gamma, \tag{11}$$

where γ is Euler's constant.

Substituting (8)–(10) into (6), simplifying, using (11) and dropping dominated terms, and plugging into (4), we finally obtain

$$D_{\mathrm{KL}}\big((X,Y)|H_0 \,\|\, (X,Y)|H_1\big) \sim \frac{\gamma^2}{2}(2q^2-2q+1)\left(\frac{n^2}{n+m}+\frac{nm\log(m+n)}{m+n}\right). \tag{12}$$

Delighted as we are to complete such a difficult calculation, (12) appears to be nonsense. For a start, even when $m = 0$, (12) is positive, proportional to n. Yet if $m = 0$, the detector in the scenario of Sect. 2 has *no* information about p and cannot possibly tell whether $\gamma > 0$ or not. Even worse, the KL divergence does not necessarily tend to zero if (1) holds, so it appears to contradict the result in [1].

The problem is that the value of P is not a random variable in the sense we modelled it, with any sort of prior, for it is supposed to be a constant. We discuss in Sect. 3 how this is incompatible with the use of KL divergence as a measure of security.

2.2 Simple Special Cases

It does not require the calculations of the previous section to see why the uniform prior is not correctly reflecting a complete lack of knowledge. Consider the scenario of Sect. 2 with $m = 0$: we have already noted that the detector then has nothing with which to compare their possible stego stream, and so cannot possibly have any information as to whether it contains a hidden payload mixed in. But, for small n, we can now compute (2) directly, and hence derive the KL divergence (3) exactly.

Take the case $m = 0$, $n = 1$ and hypothesis test (A). Then (2) reduces to

$$\begin{aligned} p(0,0;\gamma) &= \int_0^1 1-\gamma q-(1-\gamma)p \, \mathrm{d}p &= \tfrac{1}{2}-\gamma(q-\tfrac{1}{2}), \\ p(0,1;\gamma) &= \int_0^1 \gamma q+(1-\gamma)p \, \mathrm{d}p &= \tfrac{1}{2}+\gamma(q-\tfrac{1}{2}). \end{aligned}$$

So

$$\begin{aligned} D_{\mathrm{KL}}\big((X,Y)|H_0 \,\|\, (X,Y)|H_1\big) &= -\mathrm{E}_{H_0}\left[\log\left(\frac{p(X,Y;\gamma)}{p(X,Y;0)}\right)\right] \\ &= -\frac{1}{2}\Big(\log\big(1-\gamma(2q-1)\big)+\log\big(1+\gamma(2q-1)\big)\Big) \\ &= -\frac{1}{2}\Big(\log\big(1-\gamma^2(2q-1)^2\big)\Big) \\ &> 0 \end{aligned}$$

as long as $q \neq \frac{1}{2}$ and $\gamma > 0$. As before, we know this to be paradoxical because there is supposed to be zero information about the value of p when $m = 0$, so the detector would have no way of knowing whether the potential stego stream has been tampered with.

And if the reader is concerned that we have only considered case (A), especially since repeating the calculations for case (B) gives zero KL divergence above, take instead the case $m = 0$, $n = 2$. For then

$$\begin{aligned} p(0,0;\gamma) &= \int_0^1 (1-\gamma q-(1-\gamma)p)^2 \, \mathrm{d}p &= \tfrac{1}{3}(1+\gamma(1-3q)+\gamma^2(1-3q+3q^2)), \\ p(0,1;\gamma) &= \ldots &= \tfrac{1}{3}(1+\gamma-2\gamma^2(1-3q+3q^2)), \\ p(0,2;\gamma) &= \ldots &= \tfrac{1}{3}(1+\gamma(3q-2)+\gamma^2(1-3q+3q^2)), \end{aligned}$$

so

$$\begin{aligned} p'(0,0;\gamma) &= \tfrac{1}{6}(2-\gamma+\gamma^2), \\ p'(0,1;\gamma) &= \tfrac{1}{3}(1+\gamma-\gamma^2), \\ p'(0,2;\gamma) &= \tfrac{1}{6}(2-\gamma+\gamma^2). \end{aligned}$$

So, in case (B),

$$\begin{aligned} D_{\mathrm{KL}}\big((X,Y)|H_0 \,\|\, (X,Y)|H_1\big) &= -\mathrm{E}_{H_0}\left[\log\left(\frac{p'(X,Y;\gamma)}{p'(X,Y;0)}\right)\right] \\ &= -\frac{1}{3}\Big(\log\big((1+\gamma-\gamma^2)(1-\tfrac{1}{2}\gamma+\tfrac{1}{2}\gamma^2)^2\big)\Big) \\ &> 0 \end{aligned}$$

for all $\gamma \in (0,1)$, giving the same paradox. (The case $\gamma = 1$ corresponds to completely replacing one binomial sequence, with probability parameter uniformly drawn from $[0,1]$, with another, thus does not change the distribution.)

3 Discussion

These results show that the models

$$\text{(A)} \qquad q \text{ known}, \quad P \sim \mathrm{U}[0,1], \quad X|_{P=p} \sim \mathrm{Bi}(m,p), \quad Y|_{P=p} \sim \mathrm{Bi}\big(n, \gamma q + (1-\gamma)p\big)$$

and

$$\text{(B)} \qquad P, Q \sim \mathrm{U}[0,1], \quad X|_{P=p} \sim \mathrm{Bi}(m,p), \quad Y|_{P=p,Q=q} \sim \mathrm{Bi}\big(n, \gamma q + (1-\gamma)p\big)$$

do not properly reflect the situation described in the scenario of Sect. 2, for even when $m = 0$ the KL divergence suggests that it is possible to determine whether $\gamma = 0$ or $\gamma = \gamma_1 > 0$.

Why is it not impossible? The answer lies in the use of the uniform prior for P, instead of it being an unknown constant. Were the entire experiment repeated many times, the models say that a *different* value of P would be used on each repetition, and the detector can estimate the value of $\gamma q + (1-\gamma)P$ each time, as Y/n. If $\gamma = 0$, these estimates should conform to the uniform distribution of P. But if $\gamma > 0$, the distribution of $\gamma q + (1-\gamma)P$ is no longer uniform: in subcase (A), when q is known, the mixture distribution is skewed towards q, and in subcase (B) when we assumed that Q was uniform, the mixture distribution is skewed towards $1/2$. So a repeated experiment will eventually distinguish the case $\gamma = 0$ from $\gamma > 0$.

So, instead of using the cover oracle to estimate the value of p, the detector can check for uniformity of the probability in the second stream. That is not possible if the true situation is that p is a constant, but unknown. The uniform distribution for P is itself a piece of information that can be checked.

KL divergence, as a measure of security, always suffers from this problem. KL divergence can be viewed as the expected value of the log likelihood ratio statistic, but that expectation is taken over all the unknowns, including (in this model) P. And if we fix p instead, there is nothing to reflect a lack of knowledge of p on the part of the detector. We conclude that we cannot apply KL divergence, in the usual way, to measure security when there are parameters unknown to the detector. Instead, we must use an alternative, such as the unbiasedness argument of [1].

This problem is extremely relevant to other situations in information-theoretic analyses of steganography, particularly the *batch steganography* problem [8]. There, we must impose a condition on the detector that it does not know exactly which covers contain payload. Perhaps another unbiasedness argument can be used, or another statistical concept called *invariance* [9, Ch. 6], in which the detector is required to be invariant to changes in parameters it cannot observe.

References

[1] A. Ker, "The square root law in stegosystems with imperfect information," to appear in *Proc. 12th Information Hiding Workshop*, 2010.

[2] C. Cachin, "An information-theoretic model for steganography," *Information and Computation*, vol. 192, no. 1, pp. 41–56, 2004.

[3] S. Kullback, *Information Theory and Statistics.* New York: Dover, 1968.

[4] A. Ker, "Steganographic strategies for a square distortion function," in *Security, Forensics, Steganography and Watermarking of Multimedia Contents X*, ser. Proc. SPIE, vol. 6819, 2008, pp. 0401–0413.

[5] A. Ker, T. Pevný, J. Kodovský, and J. Fridrich, "The square root law of steganographic capacity," in *Proc. 10th ACM Workshop on Multimedia and Security*, 2008, pp. 107–116.

[6] A. Ker, "The Square Root Law requires a linear key," in *Proc. 11th ACM Workshop on Multimedia and Security*, 2009, pp. 85–92.

[7] T. Filler, A. Ker, and J. Fridrich, "The square root law of steganographic capacity for Markov covers," in *Media Forensics and Security XI*, ser. Proc. SPIE, vol. 7254, 2009, pp. 0801–0811.

[8] A. Ker, "Batch steganography and pooled steganalysis," in *Proc. 8th Information Hiding Workshop*, ser. Springer LNCS, vol. 4437, 2006, pp. 265–281.

[9] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*, 3rd ed. Springer, 2005.