

# Ignoring Data in Court: An Idealized Decision-Theoretic Analysis

Peter D. Grünwald  
CWI, P.O. Box 94079  
1090 GB Amsterdam  
www.grunwald.nl

## Abstract

We give a decision-theoretic analysis of a central issue regarding statistical evidence in court: are there circumstances under which it is reasonable to ignore the part of the data that gave rise to suspicion in the first place? We heuristically show that under a minimax/robust Bayesian analysis, this part of the data should in fact be treated differently from any additional data one might have. In some situations, even completely ignoring this part of the data can be a minimax optimal strategy.

Lucia de B. is a Dutch nurse who worked in the Juliana children’s hospital in The Hague from 1999 to 2001. She happened to be on duty whenever a patient in her ward suddenly died or suddenly needed to be reanimated. The hospital’s management became suspicious and notified the police. The police then gathered data about Lucia’s shifts in the Red Cross hospital, a hospital where she had worked a few years previously. Thus, these data were only taken into account later, after the investigation against Lucia began. In 2004, the Court of Appeals found her guilty of 7 murders and 3 murder attempts. Statistics played a crucial role in the verdict; a statistician calculated that what happened “could not have been a coincidence”. Despite warnings by the statistician, the court did not need much further evidence to change “not a coincidence” into “murder.” The statistical analysis itself was flawed in several respects. The question is: can we do better?

When presenting the case in the UCL evidence seminar (March 20th 2007), I claimed that a purely Bayesian approach is problematic here. From a Bayesian point of view, we would like to determine posterior probabilities that Lucia is innocent or guilty. This requires a prior probability that Lucia is guilty. The problem is that there are a broad range of priors that may be deemed “reasonable,” and these may differ by several orders of magnitude. Therefore, I argued, one should either adopt a “robust Bayesian” approach, adopting a set of priors rather than a single one; or one should adopt a Neyman-Pearson (NP) style hypothesis test, but, to avoid selection bias, one should then ignore the first data set, and only use the second one. The latter proposal generated a lot of resistance. I have now studied both suggestions in more detail, focusing on the question whether it can be sensible to ignore, or at least treat differently, the first data set. Following a suggestion by C. Manski, I have taken a decision-theoretic approach. The result is the present note.

# 1 Overview

Before going to the conclusions, let's first list the arguments against using a NP-approach on the second data set only.

1. In court cases such as these, we often do not have any second data set. In that case, ignoring the first data set is simply not an option.
2. If one adopts a Bayesian approach, then the decision is based on a posterior distribution, which is obtained by conditioning on both data sets. Thus, both data sets are treated in the same manner. Even if one is not a Bayesian, by the complete class theorem every admissible decision rule should be a Bayes rule, and therefore, one should act equivalently to a Bayesian with a particular prior. Therefore, ignoring part of the data cannot be an admissible decision rule; instead both data sets should be treated in the same manner (a point suggested by Phil Dawid).
3. The NP approach gives a frequentist guarantee on the relative frequency of making false decisions, when repeatedly applying it in courts or elsewhere. If one performs a NP approach at a level of 1 in 10000, one will on average make a type I error (reject the null hypothesis while it is true) at most 1 in 10000 times. But such a guarantee may be irrelevant for several reasons. Most importantly (a point raised by Ton Derksen), it might happen that even in a population in which everybody is innocent but prosecutors are zealous, 1 in 10000 people land in jail.

Here, we study a stylized and strongly simplified model of the situation that allows us to focus on these issues. Our main conclusions are that, at least under our model assumptions:

- 1. Minimax Optimality vs. Admissibility** Among the (frequentist) minimax optimal strategies, there is one that ignores the first data set. However, there are other minimax optimal strategies that behave never worse but sometimes (if the worst-case does not apply) considerably better. Thus, while ignoring the first part of the data can be minimax optimal, it cannot be admissible, thus confirming Dawid's conclusion above. On the other hand, all minimax strategies treat the second data set differently from the first: the first will have less influence on the decision than the second. This may seem to contradict Dawid's reasoning (but not his conclusion) above. This paradox is resolved at the end of Section 3.2.
- 2. Robust Bayes vs. Bayes** The same holds in the more general *robust* Bayesian analysis: *the first and second data set should not be treated in the same way*. Roughly, assuming both data sets are of the same size, the first data set will have less influence on the decision than the second.
- 3. Significance Levels** The frequentist minimax optimal strategy is formally equivalent to a NP hypothesis test on the second part of the data, but with a significance level that changes with the sample size, and becomes dramatically high (1 in many millions) even for moderate sample sizes. Thus, we don't have to worry that in a completely innocent population, we might send 1 in 10000 people to jail. This addresses Derksen's point.
- 4. Reliability** Both the frequentist minimax optimal strategy and every robust Bayes optimal strategy have an attractive property which I call *reliability*, which essentially means that they do not give an overly favourable impression of their own performance.

In my own view, the best approach is a robust Bayesian approach, since, by using prior probabilities, it is more informative than a frequentist minimax approach, and it can still be used if no second data set is available. Still, (a) the relevant set of prior probabilities needed for a robust Bayesian approach may be exceedingly hard to determine, (b) the frequentist minimax approach is a special case of the robust Bayesian approach, so *if* a second data set is available, and *if* both frequentists and Bayesians have to be convinced of the validity of the decision, then the minimax approach may be the method of choice. Thus, I still think that there is a some sense in ignoring part of the data, but, in contrast to my former self, I now reject (to use appropriate terminology) the NP-approach.

I should add that if no second data set is available, then I maintain that a purely frequentist approach is simply out of the question, even if one just wants to answer the simple question ‘is it just a coincidence that Lucia is always present or not?’ The greatest mystery to me in this whole episode is still that many frequentists seem to think that this question can be given a meaningful answer, that does not involve prior probabilities, and that is based on just the first data set. This, to me, seems fundamentally impossible.

## 2 An Idealized Model

Here I consider a modified version of the actual situation: I assume the data consist of 0s and 1s and follow a Bernoulli distribution. This modification greatly simplifies the analysis, while not affecting the answer to the questions posed above.

Specifically, I assume that the data follow a Bernoulli distribution with parameter  $\theta_{\mathbf{G}} = 3/4$  if the suspect is guilty, and a distribution with parameter  $\theta_{\mathbf{NG}} = 1/2$  if the suspect is innocent. There are two sets of data:  $\mathbf{x}$ , representing the suspect’s data in the Juliana hospital, in which suspicion initially arose; and  $\mathbf{y}$ , representing the data in the Red Cross hospital, in which she had worked a few years earlier. Simplifying further, we assume that  $\mathbf{x} = (x_1, \dots, x_n)$  and  $\mathbf{y} = (y_1, \dots, y_m)$  are two sequences of fixed, given (but not necessarily equal) lengths  $n$  and  $m$ . Thus, we define, for  $\mathbf{x} \in \{0, 1\}^n$  and  $\mathbf{y} \in \{0, 1\}^m$ ,

$$\begin{aligned} P(\mathbf{x} \mid \mathbf{G}) &= \theta_{\mathbf{G}}^{\sum_{i=1}^n x_i} (1 - \theta_{\mathbf{G}})^{n - \sum_{i=1}^n x_i} = \left(\frac{3}{4}\right)^{\sum_{i=1}^n x_i} \left(\frac{1}{4}\right)^{n - \sum_{i=1}^n x_i} \\ P(\mathbf{y} \mid \mathbf{G}) &= \theta_{\mathbf{G}}^{\sum_{i=1}^m y_i} (1 - \theta_{\mathbf{G}})^{m - \sum_{i=1}^m y_i} = \left(\frac{3}{4}\right)^{\sum_{i=1}^m y_i} \left(\frac{1}{4}\right)^{m - \sum_{i=1}^m y_i} \\ P(\mathbf{x}, \mathbf{y} \mid \mathbf{G}) &= P(\mathbf{x} \mid \mathbf{G}) \cdot P(\mathbf{y} \mid \mathbf{G}), \end{aligned} \tag{1}$$

and similarly for  $P(\cdot \mid \mathbf{NG})$ , in particular

$$P(\mathbf{x}, \mathbf{y} \mid \mathbf{NG}) = \theta_{\mathbf{NG}}^{\sum_{i=1}^n x_i + \sum_{i=1}^m y_i} (1 - \theta_{\mathbf{NG}})^{n+m - \sum_{i=1}^n x_i - \sum_{i=1}^m y_i} = \left(\frac{1}{2}\right)^{n+m}. \tag{2}$$

A crucial aspect of the situation is that the case went to the court *because* an extreme event had happened to some nurse in some hospital. There was no other incriminating evidence against the suspect at all. We take this into account by introducing a random variable  $T$  such that  $T = 1$  indicates that the case goes to court, and  $T = 0$  indicates that the case does not go to court. For now, we assume that  $T$  is a deterministic function of  $\mathbf{x}$ , and  $T = 1$  if and only if the data is “sufficiently extreme.” Thus, we assume that there exists some parameter

$0 < \theta \leq 1$ , such that

$$P(T = 1 \mid \mathbf{x}, \mathbf{y}, \theta) = P(T = 1 \mid \mathbf{x}, \theta) = \begin{cases} 1 & \text{if } n^{-1} \sum_{i=1}^n x_i \geq \theta, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

More sophisticated approaches will be briefly discussed later. Note that, without loss of generality, we can assume that  $\theta \in \Theta$ , which we define as  $\Theta := \{0, 1/n, 2/n, \dots, 1\}$ . Summarizing and formalizing, we have a parameter set  $\mathcal{S} \times \Theta$ , where  $\mathcal{S} := \{\mathbf{G}, \mathbf{NG}\}$ ; we define a sample space  $\Omega = \mathcal{X} \times \mathcal{Y} \times \mathcal{T}$ , where  $\mathcal{T} = \{0, 1\}$ ;  $\mathcal{X} = \{0, 1\}^n$  and  $\mathcal{Y} = \{0, 1\}^m$ ; and we define, for each  $s \in \mathcal{S}, \theta \in \Theta$ , a distribution  $P(\cdot \mid s, \theta)$  on  $\Omega$ , with mass function  $P(\mathbf{x}, \mathbf{y}, t \mid s, \theta)$  given by (1)-(3).

To prepare for the Bayesian treatment, we define  $\Pi$  to be the set of *prior distributions*.  $\Pi$  is just the set of distributions on  $\mathcal{S} \times \Theta$ . Each  $\pi \in \Pi$  is identified with its mass function  $\pi(s, \theta)$ .

Each  $\pi \in \Pi$  uniquely induces a distribution in the joint space  $\mathcal{S} \times \Theta \times \Omega$ , given by, for all  $s, \theta, \mathbf{x}, \mathbf{y}, t$ :

$$P(s, \theta, \mathbf{x}, \mathbf{y}, t) = \pi(s, \theta)P(\mathbf{x}, \mathbf{y}, t \mid s, \theta).$$

We let  $\mathcal{P}$  be the set of all distributions thus obtained. Each  $P \in \mathcal{P}$  is uniquely determined by a prior  $\pi \in \Pi$ , and vice versa. An important subclass of prior distributions is given by the set  $\Pi_{\text{ind}}$  of priors  $\pi$  under which  $S$  and  $\theta$  are independent, i.e. for all  $s, \theta$ ,  $\pi(s, \theta) = \pi(s)\pi(\theta)$ . Although I have no idea whether treating  $S$  and  $\theta$  as independent is reasonable, at least in the case of Lucia de B., all Bayesian analyses performed until now implicitly made this assumption.

**Purely Bayesian Approach** In a purely Bayesian approach, we assume that, given all additional knowledge we have about the suspect, we can identify a unique prior  $\pi$  on  $\mathcal{S} \times \Theta$ , and therefore, a unique  $P \in \mathcal{P}$ .

Given the data  $\mathbf{x}, \mathbf{y}$  and the fact that the case has gone to court ( $T = 1$ ), we can then use Bayes theorem to calculate the posterior probability of ‘‘guilt’’,  $P(\mathbf{G} \mid \mathbf{x}, \mathbf{y}, T = 1)$ . If we assume that  $S$  and  $\theta$  are independent under the prior (i.e.  $\pi \in \Pi_{\text{ind}}$  as above), then we get that

$$P(\mathbf{G} \mid \mathbf{x}, \mathbf{y}, T = 1) = P(\mathbf{G} \mid \mathbf{x}, \mathbf{y}). \quad (4)$$

To see this, note that for  $\pi \in \Pi_{\text{ind}}$ ,

$$\begin{aligned} P(\mathbf{G} \mid \mathbf{x}, \mathbf{y}, T = 1) &= \\ \frac{\sum_{\theta} \pi(\theta) \pi(\mathbf{G}) P(\sum_{i=1}^n x_i \geq \theta, \mathbf{x}, \mathbf{y} \mid \mathbf{G})}{\sum_{\theta} \pi(\theta) \sum_{s \in \mathcal{S}} \pi(s) P(\sum_{i=1}^n x_i \geq \theta, \mathbf{x}, \mathbf{y} \mid s)} &= \frac{\sum_{\theta \leq \sum_{i=1}^n x_i} \pi(\theta) \pi(\mathbf{G}) P(\mathbf{x}, \mathbf{y} \mid \mathbf{G})}{\sum_{\theta \leq \sum_{i=1}^n x_i} \pi(\theta) \sum_{s \in \mathcal{S}} \pi(s) P(\mathbf{x}, \mathbf{y} \mid s)} = \\ &P(\mathbf{G} \mid \mathbf{x}, \mathbf{y}). \end{aligned} \quad (5)$$

(4) shows that the selection bias inherent in  $T$  can safely be ignored. This confirms the intuition that in a purely Bayesian approach, both samples should be treated on an equal footing *if  $S$  and  $\theta$  are independent under the prior*. If  $S$  and  $\theta$  may be dependent, then it is clear that (4) does not necessarily hold.

### 3 Decision-Theoretic Approaches

We now extend the analysis by assuming that the statistician/judge wants to make the decision which minimizes her expected loss, according to some loss function  $\text{LOSS}$ .

For concreteness, consider the following loss function  $\text{LOSS} : \{\mathbf{G}, \mathbf{NG}\} \times \{\mathbf{G}, \mathbf{NG}\} \rightarrow \mathbb{R}$ . In the table,  $\widehat{\mathbf{G}}$  stands for the *assertion* “suspect is guilty”;  $\widehat{\mathbf{NG}}$  stands for the *assertion* “suspect is not guilty.”

	$\widehat{\mathbf{G}}$	$\widehat{\mathbf{NG}}$
$\mathbf{G}$	0	$10^3$
$\mathbf{NG}$	$10^6$	0

Thus,  $\text{LOSS}(\mathbf{NG}; \widehat{\mathbf{G}}) = 10^3 \text{LOSS}(\mathbf{G}; \widehat{\mathbf{NG}})$ , so the loss incurred when we proclaim that the suspect is guilty, whereas in reality she is not, is a factor 1000 larger than the loss incurred when we proclaim “innocent,” whereas in reality she is guilty. One may certainly disagree about whether such a loss function is reasonable; but all conclusions below essentially continue to hold for every other loss function with  $\text{LOSS}(s, s) = 0$  and  $\text{LOSS}(\mathbf{NG}; \widehat{\mathbf{G}}) > \text{LOSS}(\mathbf{G}; \widehat{\mathbf{NG}}) > 0$ .

#### 3.1 Bayesian Approach

The extension of the Bayesian approach to the decision-theoretic setting is straightforward. Assume that we have identified a single  $P \in \mathcal{P}$  (equivalently, we have established a prior  $\pi \in \Pi$ ). Then we can determine the posterior distribution  $P(\mathbf{G} \mid \mathbf{x}, \mathbf{y})$ . We should then take the decision  $\hat{s}$  that minimizes expected loss according to this posterior distribution. Thus, we should choose the  $\hat{s} \in \{\mathbf{G}, \mathbf{NG}\}$  that minimizes

$$E_{S \sim P(\cdot \mid \mathbf{x}, \mathbf{y}, T=1)}[\text{LOSS}(S; \hat{s})] = E_{S \sim P(\cdot \mid \mathbf{x}, \mathbf{y})}[\text{LOSS}(S; \hat{s})]. \quad (6)$$

Another way to formulate this is as follows: for each pair of samples  $\mathbf{x}, \mathbf{y}$  that may be observed, we have to make a decision. Thus, our decision rule can be described as a function  $\delta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{S}$  that maps each  $\mathbf{x}, \mathbf{y}$ , to a corresponding decision  $\delta(\mathbf{x}, \mathbf{y})$ . We should then adopt the *Bayes optimal decision rule* (henceforth abbreviated to “Bayes rule”). This is the rule (function)  $\delta_0$  that achieves<sup>1</sup>

$$\min_{\delta \in \Delta(\mathcal{X} \times \mathcal{Y})} E_{S, \mathbf{x}, \mathbf{y} \sim P(\cdot \mid T=1)}[\text{LOSS}(S, \delta(\mathbf{x}, \mathbf{y}))], \quad (7)$$

where the minimum is over  $\Delta(\mathcal{X} \times \mathcal{Y})$ , the set of all decision rules, i.e. all functions  $\delta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{S}$ . It is immediate that the decision rule that, for each  $\mathbf{x}, \mathbf{y}$ , minimizes (6) is equivalent to the decision rule that achieves (7), so both approaches are equivalent.

#### 3.2 Minimax Approach

Here, we view the situation as a game between Statistician and “Nature”. Nature can freely choose the value of the threshold  $\theta$  and the value of  $S$ , i.e. whether or not the suspect is guilty. Statistician should adopt a decision rule that is best for the worst-case choice of Nature, but

<sup>1</sup>For simplicity we assume that there exists a unique Bayes decision rule  $\delta_0$ . If for some  $\mathbf{x}, \mathbf{y}$ , the expected loss of  $\mathbf{G}$  and  $\mathbf{NG}$  is identical, then  $\delta_0$  chooses one of the two uniformly at random.

only in situations that the case goes to court, i.e. *given that*  $T = 1$ . This can be any decision rule  $\delta^* : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{S}$  that achieves

$$\min_{\delta \in \Delta(\mathcal{X}, \mathcal{Y})} \max_{\theta \in \Theta, s \in \mathcal{S}} E_{\mathbf{x}, \mathbf{y} \sim P(\cdot | \theta, s, T=1)} [\text{LOSS}(s, \delta(\mathbf{x}, \mathbf{y}))] = \max_{Q \in \mathcal{Q}} \min_{\delta \in \Delta(\mathcal{X}, \mathcal{Y})} E_{S, \mathbf{x}, \mathbf{y} \sim Q} [\text{LOSS}(S, \delta(\mathbf{x}, \mathbf{y}))]. \quad (8)$$

The equality follows from the minimax theorem. The set  $\mathcal{Q}$  is defined as the set of all distributions  $Q$  on  $\mathcal{S} \times \Theta \times \Omega$  such that, for some prior  $\pi \in \Pi$ , for all  $\mathbf{x}, \mathbf{y}$ :

$$Q(s, \theta, \mathbf{x}, \mathbf{y}) = \pi(s, \theta) P_{s, \theta}(\mathbf{x}, \mathbf{y} | T = 1).$$

$\mathcal{Q}$  is essentially the set of mixtures over the conditional set of distributions  $P_{s, \theta}(\cdot | T = 1)$ . Note that  $\mathcal{Q}$  is *not* necessarily equal to  $\{P(\cdot | T = 1) | P \in \mathcal{P}\}$ . However, if we let  $P_{S, Y}$  and  $Q_{S, Y}$  stand for the marginal distribution of  $(S, Y)$  under  $P$  and  $Q$  respectively, and if we define  $\mathcal{P}_{S, Y} = \{P_{S, Y} | P \in \mathcal{P}\}$ , and  $\mathcal{Q}_{S, Y} = \{Q_{S, Y} | Q \in \mathcal{Q}\}$ , then we do have that  $\mathcal{P}_{S, Y} = \mathcal{Q}_{S, Y}$ .

The complete class theorem [Ferguson 1967, page 87] shows that if the minimax optimal decision rule  $\delta^*$  is unique, then there exists some  $Q^* \in \mathcal{Q}$ , corresponding to some prior  $\pi^*$ , such that  $\delta^*$  is the Bayes rule relative to  $Q^*$ , where ‘‘Bayes rule’’ is defined as in (7). Thus, a decision-maker who bases decisions on  $\delta^*$  *acts just as if* the data would follow distribution  $Q^*$ . In our context, it is more useful to use the lesser known Theorem 4.1 of Grünwald and Dawid [2004], which in our context shows that for *each* minimax decision rule  $\delta^*$ , there is some  $Q^* \in \mathcal{Q}$  such that  $\delta^*$  is Bayes for  $Q^*$ .

**Ignoring  $\mathbf{x}$**  Another decision rule of interest is the minimax optimal decision rule *among all decision rules that ignore  $\mathbf{x}$* . Thus, we now look at the restricted set  $\Delta(\mathcal{Y}) \subset \Delta(\mathcal{X} \times \mathcal{Y})$  of all decision rules  $\delta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{S}$  that satisfy, for all  $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ , all  $\mathbf{y} \in \mathcal{Y}$ ,  $\delta(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x}', \mathbf{y})$ . Such  $\delta$  can be written as functions from  $\mathcal{Y}$  to  $\mathcal{S}$  only, and with some abuse of notation, if it is clear that  $\delta \in \Delta(\mathcal{Y})$ , we will write  $\delta(\mathbf{y})$  rather than  $\delta(\mathbf{x}, \mathbf{y})$ . From now on, we call the decision problem in which the statistician is restricted to some  $\delta \in \Delta(\mathcal{Y})$  the  $\mathbf{y}$ -decision problem. We call the original decision problem the  $\mathbf{x}, \mathbf{y}$ -decision problem.

Let us try to determine the minimax optimal rule  $\tilde{\delta}$  in the  $\mathbf{y}$ -decision problem. This is the  $\tilde{\delta} : \mathcal{Y} \rightarrow \mathcal{S}$  that achieves

$$\min_{\delta \in \Delta(\mathcal{Y})} \max_{\theta, s} E_{\mathbf{x}, \mathbf{y} \sim P(\cdot | \theta, s, T=1)} [\text{LOSS}(s, \delta(\mathbf{y}))] = \max_{Q \in \mathcal{Q}} \min_{\delta \in \Delta(\mathcal{Y})} E_{S, \mathbf{y} \sim Q} [\text{LOSS}(S, \delta(\mathbf{y}))]. \quad (9)$$

Analogously to the unrestricted  $\mathbf{x}, \mathbf{y}$ -problem, there exists some  $\tilde{Q} \in \mathcal{Q}$ , corresponding to some prior  $\pi$ , such that  $\tilde{\delta}$  is the Bayes rule relative to  $\tilde{Q}$  in the  $\mathbf{y}$ -game.

**Reliability**  $\tilde{\delta}$  has an interesting property: suppose a decision-maker who uses  $\tilde{\delta}$  wants to estimate the quality of the decisions she makes. Since she acts just as if the data would follow  $\tilde{Q} \in \mathcal{Q}$ , she would tend to estimate that performance quality by

$$E_{S, \mathbf{x}, \mathbf{y} \sim \tilde{Q}} [\text{LOSS}(S, \tilde{\delta}(\mathbf{y}))] = E_{S, \mathbf{y} \sim \tilde{Q}_{S, Y}} [\text{LOSS}(S, \tilde{\delta}(\mathbf{y}))]$$

Now it turns out that this estimate of  $\tilde{\delta}$ 's performance (a) cannot be overly optimistic, *irrespective of whether the data actually follow  $\tilde{Q}$  or not*, and (b), in the worst case, it is not overly pessimistic either: it is easy to see that

$$E_{S, \mathbf{x}, \mathbf{y} \sim \tilde{Q}}[\text{LOSS}(S, \tilde{\delta}(\mathbf{x}, \mathbf{y}))] = \max_{Q \in \mathcal{Q}} E_{S, \mathbf{x}, \mathbf{y} \sim Q}[\text{LOSS}(S, \tilde{\delta}(\mathbf{x}, \mathbf{y}))] = \max_{Q \in \mathcal{Q}_{S, Y}} E_{S, \mathbf{y} \sim Q}[\text{LOSS}(S, \tilde{\delta}(\mathbf{y}))]. \quad (10)$$

Thus, in the terminology of [Grünwald 1998], basing decisions on  $\tilde{Q}$  is *safe*: although  $\tilde{Q}$  may not be the actual distribution, and may not even be the minimax optimal distribution, if we make the decisions that would be optimal if  $\tilde{Q}$  were true, our expected loss will be no larger than it would be if  $\tilde{Q}$  were indeed true. Basing our decision on  $\tilde{Q}$ , we will get no overly optimistic impression of how good we will predict.

Using the reliability property, we will now show that in some situations, ignoring  $\mathbf{x}$  can be a minimax optimal strategy.

**Ignoring  $\mathbf{x}$  is minimax optimal** Let, for  $\theta_0 \in \Theta = \{0, 1/n, 2/n, \dots, 1\}$ ,  $\mathcal{Q}_{\theta_0} = \{Q(\cdot \mid \theta = \theta_0) \mid Q \in \mathcal{Q}, Q(\theta = \theta_0) > 0\}$  be the set of distributions in  $\mathcal{Q}$  conditioned on  $\theta$  being equal to the value  $\theta_0$ .

Recall that all distributions in  $\mathcal{Q}$  and  $\mathcal{Q}_{\theta}$  are conditional on  $T = 1$ . For  $\theta \leq 1 - 1/n$ , we have, for all  $Q \in \mathcal{Q}_{\theta}$ , that  $0 < Q(\mathbf{x} = (1, \dots, 1)) < 1$ , so that, given  $T = 1$ ,  $\mathbf{x}$  can take on more than one value with positive probability. From this we can infer:

$$\begin{aligned} \max_{Q \in \mathcal{Q}_{\theta}} \min_{\delta \in \Delta(\mathcal{X} \times \mathcal{Y})} E_{S, \mathbf{x}, \mathbf{y} \sim Q}[\text{LOSS}(S, \delta(\mathbf{x}, \mathbf{y}))] < \\ \max_{Q \in \mathcal{Q}_{\theta}} \min_{\delta \in \Delta(\mathcal{Y})} E_{S, \mathbf{x}, \mathbf{y} \sim Q}[\text{LOSS}(S, \delta(\mathbf{y}))] = \\ \max_{Q \in \mathcal{Q}} \min_{\delta \in \Delta(\mathcal{Y})} E_{S, \mathbf{y} \sim Q}[\text{LOSS}(S, \delta(\mathbf{y}))]. \quad (11) \end{aligned}$$

To see where the strict inequality comes from, note first that  $\Delta(\mathcal{Y})$  is a subset of  $\Delta(\mathcal{X} \times \mathcal{Y})$ . This already shows nonstrict inequality. Since we assume  $\theta \leq 1 - 1/n$  and therefore,  $\mathbf{x}$  can take on more than one value,  $\mathbf{x}$  gives actual information about  $S$  and can be taken advantage of (we omit the details of the proof).

On the other hand, if we fill in  $\theta = 1$  in (11), then  $\mathbf{x}$  can only become  $(1, \dots, 1)$ , and then the inequality in (11) becomes an equality. Since the rightmost expression in (11) does not depend on  $\theta$ , it follows that, for the full  $\mathbf{x}, \mathbf{y}$ -problem, every maximin strategy for Nature in  $\mathcal{Q}$  must be a member of  $\mathcal{Q}_1$ . Moreover, if nature plays a maximin strategy in  $\mathcal{Q}_1$ , then, again because the first inequality in (11) then becomes an equality,  $\tilde{\delta}$ , the minimax optimal strategy for the  $\mathbf{y}$ -decision problem, is still a minimax optimal strategy for the full  $\mathbf{x}, \mathbf{y}$ -decision problem. Summarizing, we have shown that *even in the full  $\mathbf{x}, \mathbf{y}$ -decision problem, ignoring  $\mathbf{x}$  is a minimax optimal strategy*.

Nevertheless, there will be other minimax strategies that are equally good against maximin optimal  $Q^* \in \mathcal{Q}_1$ , and are strictly better against other  $Q \in \mathcal{Q} \setminus \mathcal{Q}_1$ . Therefore, ignoring  $\mathbf{x}$  is *not* an admissible strategy [Ferguson 1967, page 54]. However, we now give a heuristic argument that shows that such alternative minimax strategies *cannot* be viewed as Bayes rules relative to any prior in the set  $\Pi_{\text{ind}}$  of priors under which  $S$  and  $\theta$  are independent, and selection bias is ignored. Therefore, *all* minimax optimal strategies weigh the evidence given by  $\mathbf{x}$  differently from the evidence given by  $\mathbf{y}$ : it turns out that  $\mathbf{y}$  has a stronger influence on the decision than  $\mathbf{x}$ .

**There is no  $\pi \in \Pi_{\text{ind}}$  which makes Bayes minimax optimal in the  $\mathbf{x}, \mathbf{y}$ -problem**

We show this using a (trivial) extension of a result of Cover and Thomas [1991, page 312, Chapter 12], summarized in Lemma 1 below. From now on we assume that  $m = n$  and that  $\theta_{\mathbf{G}} > \theta_{\mathbf{NG}}$ . First, fix some constants  $c_1, c_2 > 0$  and let  $\text{LOSS}$  be defined as  $\text{LOSS}(\mathbf{G}, \hat{\mathbf{N}}\mathbf{G}) = c_1, \text{LOSS}(\mathbf{NG}, \hat{\mathbf{G}}) = c_2, \text{LOSS}(s, \hat{s}) = 0$  if  $\hat{s} = s$ . Fix some  $0 \leq c_0 < D(\theta_{\mathbf{G}} \parallel \theta_{\mathbf{NG}})$ , and let  $f : \mathbb{N} \rightarrow \mathbb{R}$  be a function satisfying  $f(n) = c_0 n + o(n)$ . Let, for each  $n$ ,  $\pi_n(\mathbf{G}) = e^{-f(n)}$  and let  $P_n \in \mathcal{P}_{S,Y}$  be the distribution in  $\mathcal{P}_{S,Y}$  based on prior  $\pi_n$ . Let  $\delta_n \in \Delta(\mathcal{Y})$  be the Bayes decision rule relative to  $P_n$  in the  $\mathbf{y}$ -decision problem, i.e.  $\delta_n$  is arrived at by determining the posterior given  $\pi_n$  and  $\mathbf{y}$  (we return to the  $\mathbf{x}, \mathbf{y}$ -problem later). Clearly there exists some  $\theta_n$  such that

$$\delta_n(\mathbf{y}) = \begin{cases} \mathbf{G} & \text{if } n^{-1} \sum_{i=1}^n y_i > \theta_n \\ \mathbf{NG} & \text{if } n^{-1} \sum_{i=1}^n y_i < \theta_n \\ \mathbf{G}/\mathbf{NG} & \text{each with probability } 1/2, \text{ if } n^{-1} \sum_{i=1}^n y_i = \theta_n. \end{cases} \quad (12)$$

Note that

$$E_{S,\mathbf{y} \sim P_n}[\text{LOSS}(S, \delta_n(\mathbf{y}))] \approx P\left(n^{-1} \sum_{i=1}^n y_i > \theta_n \mid \mathbf{NG}\right) + e^{-c_0 n - o(n)} P\left(n^{-1} \sum_{i=1}^n y_i < \theta_n \mid \mathbf{G}\right).$$

Let  $D(p \parallel q) = p \ln(p/q) + (1-p) \ln((1-p)/(1-q))$  be the Kullback-Leibler divergence between the Bernoulli distributions with parameters  $p$  and  $q$ . Based on [Cover and Thomas 1991], we find:

**Lemma 1** 1. We have  $\theta_n \rightarrow \theta^\circ \in [\theta_{\mathbf{NG}}, \theta_{\mathbf{G}}]$ , where  $D(\theta^\circ \parallel \theta_{\mathbf{NG}}) = D(\theta^\circ \parallel \theta_{\mathbf{G}}) + c_0 + o(1)$ .

2. We have  $E_{S,\mathbf{y} \sim P_n}[\text{LOSS}(S, \delta_n(\mathbf{y}))] = e^{-nD(\theta^\circ \parallel \theta_{\mathbf{NG}}) + o(n)}$ .

3. We have  $\max_{s \in \{\mathbf{G}, \mathbf{NG}\}} E_{\mathbf{y} \sim P(\cdot \mid s)}[\text{LOSS}(s, \delta_n(\mathbf{y}))] = e^{o(n)} \cdot \max_{s \in \{\mathbf{G}, \mathbf{NG}\}} e^{-nD(\theta^\circ \parallel \theta_s)}$ .

4. We have

$$\max_{s \in \{\mathbf{G}, \mathbf{NG}\}} \{P(\delta_n(\mathbf{y}) = \mathbf{NG} \mid \mathbf{G}), P(\delta_n(\mathbf{y}) = \mathbf{G} \mid \mathbf{NG})\} = e^{o(n)} \cdot \max_{s \in \{\mathbf{G}, \mathbf{NG}\}} e^{-nD(\theta^\circ \parallel \theta_s)}.$$

Again by Theorem 4.1 of [Grünwald and Dawid 2004], the minimax optimal decision rule  $\tilde{\delta}_n$  in the  $\mathbf{y}$ -decision problem for data of size  $n$  must be a Bayes rule relative to some prior  $\tilde{\pi}_n$ . Therefore, it must also be of the form (12) for some  $\tilde{\theta}_n$ . Since, for all  $\theta, \theta' \in [0, 1]$ ,  $D(\theta \parallel \theta')$  is a strictly increasing function of  $\theta$  if  $\theta \geq \theta'$ , and a strictly decreasing function of  $\theta$  if  $\theta \leq \theta'$ , and 0 iff  $\theta = \theta'$ , it follows from part 3 of the lemma that the minimax optimal strategy  $\tilde{\delta}_n$  in the  $\mathbf{y}$ -problem achieves loss

$$\max_{s \in \{\mathbf{G}, \mathbf{NG}\}} E_{\mathbf{y} \sim P(\cdot \mid s)}[\text{LOSS}(s, \tilde{\delta}_n(\mathbf{y}))] = e^{-n\tilde{D} + o(n)}$$

for some  $\tilde{D}$  satisfying

$$\tilde{D} = D(\tilde{\theta} \parallel \theta_{\mathbf{G}}) = D(\tilde{\theta} \parallel \theta_{\mathbf{NG}}),$$

for some  $\theta_{\mathbf{NG}} < \tilde{\theta} < \theta_{\mathbf{G}}$ .  $\tilde{\theta}$  is the unique  $\theta$  which has equal KL divergence to  $\theta_{\mathbf{G}}$  and  $\theta_{\mathbf{NG}}$ .  $\tilde{D}$  is called the *Chernoff information* between  $\theta_{\mathbf{G}}$  and  $\theta_{\mathbf{NG}}$ . In our example,  $\theta_{\mathbf{G}} = 3/4$ ,

$\theta_{\mathbf{NG}} = 1/2$ , and we get that  $\tilde{\theta} \approx 0.63$ , and  $\tilde{D} \approx 0.034$  (note that  $\tilde{\theta}$  is approximately in the middle between  $1/2$  and  $3/4$ , as it should).

By Part 1 of the lemma, it now follows that the minimax optimal prior  $\tilde{\pi}_n$  at sample size  $n$  in the  $\mathbf{y}$ -game satisfies

$$\tilde{\pi}_n(\mathbf{G}) = e^{o(n)}, \quad (13)$$

i.e. if it decreases at all in  $n$ , it must decrease sublinearly in the exponent (in particular, if, for each  $n$ , we use the uniform prior, we will achieve the minimax expected loss in the  $\mathbf{y}$ -problem, up to first order in the exponent).

Now consider the large  $n$  situation and the  $\mathbf{x}, \mathbf{y}$ -decision problem. Let  $\Pi_{\text{ind}}$  be defined as before, as the set of priors under which  $S$  and  $\theta$  are independent. We will show that for all large  $n$ , there exists no prior  $\pi' \in \Pi_{\text{ind}}$  such that the Bayes rule relative to  $\pi'$  is a minimax optimal strategy in the  $\mathbf{x}, \mathbf{y}$ -problem.

Suppose then, by means of contradiction, that for all large  $n$ , there does exist some prior  $\pi'_n \in \Pi_{\text{ind}}$  such that, in the  $\mathbf{x}, \mathbf{y}$ -decision problem, the Bayes rule  $\delta'_n$  based on  $\pi'_n$  would be minimax optimal. Let  $Q'_n \in \mathcal{Q}$  be the distribution in  $\mathcal{Q}$  based on prior  $\pi'_n$ . We can obtain the posterior  $Q'_n | \mathbf{x}, \mathbf{y}$  by first conditioning  $Q'_n$  on  $\mathbf{x}$ , and then conditioning  $Q'_n | \mathbf{x}$  on  $\mathbf{y}$ .

We already established that  $\tilde{\delta}_n$  is minimax optimal, not just for the  $\mathbf{y}$ -problem but also for the  $\mathbf{x}, \mathbf{y}$ -problem. Thus, if  $\delta'_n$  were minimax optimal as well, then it must behave like  $\tilde{\delta}_n$  if nature chooses a maximin distribution  $Q^* \in \mathcal{Q}_1$ , so that  $\mathbf{x}$  consists only of 1s. Thus, if  $\mathbf{x}$  consists only of 1s, then by (13) we must have  $Q'_n(\mathbf{G} | \mathbf{x}) = \tilde{\pi}_n(\mathbf{G}) = e^{o(n)}$ , or equivalently,

$$\frac{Q'_n(\mathbf{G} | \mathbf{x})}{Q'_n(\mathbf{NG} | \mathbf{x})} = e^{o(n)}.$$

In our example, with  $\theta_{\mathbf{NG}} = 1/2, \theta_{\mathbf{G}} = 3/4$ , this means that  $\pi'_n(\mathbf{G})$  must satisfy

$$\frac{\pi'_n(\mathbf{G})}{\pi'_n(\mathbf{NG})} = \left( \frac{\theta_{\mathbf{NG}}}{\theta_{\mathbf{G}}} \right)^n e^{o(n)} = \left( \frac{2}{3} \right)^n e^{o(n)}.$$

We will now see that, if  $n/4$  is an integer, and nature chooses a particular  $Q^* \in \mathcal{Q}_{3/4}$ , i.e. the case is taken to court already when the frequency of incidents is  $3/4$ , then the loss incurred by  $\delta'_n(\mathbf{x}, \mathbf{y})$  can become much larger than the loss incurred by  $\tilde{\delta}(\mathbf{y})$ . Therefore,  $\delta'_n$  is not minimax optimal in the  $\mathbf{x}, \mathbf{y}$ -problem. Namely, let  $Q^* \in \mathcal{Q}_{3/4}$  with  $Q^*(S = \mathbf{G}) = 1$ . By the law of large numbers, we have with  $Q^*$ -probability 1,

$$Q'_n(\mathbf{G} | \mathbf{x}) = \left( \frac{2}{3} \right)^n \left( \frac{3}{2} \right)^{3n/4} \left( \frac{1}{2} \right)^{n/4} e^{o(n)} = e^{-(n/4) \ln 3 + o(n)}.$$

By Part 3 of Lemma 1, we thus have that, for some  $c > 0$ , with  $Q^*$ -probability 1,  $\mathbf{x}$  is such that

$$E_{S, \mathbf{y} \sim Q^*} [\text{LOSS}(S, \delta'_n(\mathbf{x}, \mathbf{y}))] = e^{-nD(\theta^\circ | \theta_{\mathbf{G}}) + o(n)}$$

for some  $\theta^\circ \in [\theta_{\mathbf{NG}}, \theta_{\mathbf{G}}]$  satisfying  $D(\theta^\circ | \theta_{\mathbf{NG}}) = (1/4) \ln 3 + D(\theta^\circ | \theta_{\mathbf{G}})$ . Because  $D(\theta | \theta_{\mathbf{NG}})$  is a strictly increasing function of  $\theta$  if  $\theta \geq \theta_{\mathbf{NG}}$  and  $D(\theta | \theta_{\mathbf{G}})$  is strictly decreasing in  $\theta$  for  $\theta \leq \theta_{\mathbf{G}}$ , it follows that  $D(\theta^\circ | \theta_{\mathbf{G}}) < D(\tilde{\theta} | \theta_{\mathbf{G}}) = \tilde{D}$ , so that, for some  $c' > 0$ , with  $Q^*$ -probability 1,  $\mathbf{x}$  is such that

$$E_{S, \mathbf{y} \sim Q^*} [\text{LOSS}(S, \delta'_n(\mathbf{x}, \mathbf{y}))] = e^{-n\tilde{D} + nc' + o(n)},$$

which very strongly suggests that

$$E_{S, \mathbf{x}, \mathbf{y} \sim Q^*}[\text{LOSS}(S, \delta'_n(\mathbf{x}, \mathbf{y}))] = e^{-n\tilde{D} + nc' + o(n)},$$

as well, so that the strategy  $\delta'_n$  has exponentially larger loss than the strategy  $\tilde{\delta}_n$ , which is minimax optimal for the  $\mathbf{x}, \mathbf{y}$ -problem. Thus, assuming there exists a minimax optimal Bayes strategy based on  $\Pi_{\text{ind}}$  leads to a contradiction.

**How can this be?** By the complete class theorem, there must exist some prior  $\pi^*$  such that *some* minimax optimal  $\delta^*$  in the  $\mathbf{x}, \mathbf{y}$ -problem is Bayes relative to  $\pi^*$ . However, the analysis above shows that this must be a prior under which  $S$  and  $\theta$  are highly dependent; in fact, the prior  $\pi(\mathbf{G} \mid \theta)$  will decrease dramatically with increasing  $\theta$ . For  $\theta$  close to 1, it will be equal to  $(2/3)^n$ , which effectively amounts to ignoring  $\mathbf{x}$  altogether and using a uniform prior for  $\mathbf{y}$ . (Note that, in the actual Lucia de B. case, the first data set observed was in fact the most extreme that could have been observed; so my intuition that we may want to ignore it wasn't completely off!)

With such a prior, the posterior distribution  $P(\cdot \mid T = 1, \mathbf{x}, \mathbf{y})$  (which does not ignore selection bias) will be substantially different from  $P(\cdot \mid \mathbf{x}, \mathbf{y})$  (which ignores selection bias), and this resolves the paradox.

### 3.3 Robust Bayesian Approach

The worst-case scenario that we sketched above is quite pessimistic: it assumed that we had no prior knowledge at all about the suspect's guilt and the threshold  $\theta$  needed for going to court.

Such prior knowledge may be built-in by doing a minimax analysis, not relative to the full set of distributions  $\mathcal{P}$ , but rather to some subset  $\mathcal{P}' \subset \mathcal{P}$  that reflects our prior knowledge about the situation. That is, in contrast to (8), we look for the decision rule  $\delta^*$  achieving

$$\min_{\delta \in \Delta(\mathcal{X}, \mathcal{Y})} \max_{P \in \mathcal{P}'} E_{S, \mathbf{x}, \mathbf{y} \sim P(\cdot \mid T=1)}[\text{LOSS}(S, \delta(\mathbf{x}, \mathbf{y}))]. \quad (14)$$

This is the so-called *robust Bayesian* approach [Berger 1985; Grünwald and Dawid 2004], which combines Bayesian ideas with minimax analysis. Whereas the minimax analysis above, relative to the full set  $\mathcal{P}$ , also had a frequentist interpretation, the interpretation becomes inherently Bayesian if we take a  $\mathcal{P}'$  that is a proper subset of  $\mathcal{P}$  (note that each  $P \in \mathcal{P}'$  corresponds to a particular prior on  $\mathcal{S} \times \Theta$ ).

This substantially changes the analysis. We can still apply the minimax theorem and Theorem 4.1 of Grünwald and Dawid [2004] to show that there is some  $P^*$  such that  $\delta^*$  is the Bayes rule relative to  $P^*$ , but now that theorem only shows that this  $P^*$  must be an element of the convex closure of the set  $\mathcal{R} = \{P(\cdot \mid T = 1) \mid P \in \mathcal{P}'\}$ . (even if  $\mathcal{P}'$  is convex, the set  $\mathcal{R}$  may not be convex).

It is important to note that *even if  $\mathcal{P}'$  only contains distributions corresponding to priors under which  $S$  and  $\theta$  are independent,  $P^*$  will typically be a distribution under which there is substantial dependence between  $S$  and  $\theta$* . To see this, note that, (a), as we have already seen, under the ordinary, frequentist, minimax analysis, the maximin  $Q^*$  will have substantial dependence between  $S$  and  $\theta$ ; and (b), the frequentist minimax analysis is equivalent to the robust Bayes analysis relative to the set of priors  $\pi$  satisfying  $\pi(s) = 1$  for some  $s$  and  $\pi(\theta) = 1$  for some  $\theta$ . Under all these priors,  $\theta$  and  $S$  are independent.

## 4 Other Approaches

### 4.1 Naive Robust Bayes

What if we do not want to commit to any particular loss function? We can then use the following variation of the robust Bayes approach: let  $\mathcal{P}'$  be our a priori set of distributions. We define

$$\mathcal{P}'_S | \mathbf{x}, \mathbf{y} = \{P(S = \cdot | \mathbf{x}, \mathbf{y}) | P \in \mathcal{P}'\}$$

to be the set of all posterior distributions on  $\mathcal{S} = \{\mathbf{G}, \mathbf{NG}\}$  that correspond to some prior in  $\mathcal{P}'$ . If  $\mathcal{P}'_S | \mathbf{x}, \mathbf{y}$  contains any distribution with  $P(S = \mathbf{G})$  lower than some threshold, we may say that “there is reasonable doubt that the suspect is guilty,” and decide that she is not guilty. Note that in this variation of the robust Bayes idea, there is no harm in conditioning on both  $\mathbf{x}$  and  $\mathbf{y}$ .

This seems an appealing variation of the robust Bayes idea, but it is in fact quite problematic. Namely, suppose that we first update  $\mathcal{P}'$  to  $\mathcal{P}'_S | \mathbf{x}, \mathbf{y}$ , and then, later, we decide on a loss function  $\text{LOSS}$ , and take the minimax optimal decision with respect to  $\mathcal{P}'_S | \mathbf{x}, \mathbf{y}$  and  $\text{LOSS}$ . The resulting decision rule is *not* equivalent to the minimax optimal decision rule relative to  $\mathcal{P}'$  and  $\text{LOSS}$ , and may, in general, be *much* worse. This is the so-called “dilation phenomenon” that has been noted in the literature on “imprecise probabilities” [Seidenfeld and Wasserman 1993]. See [Grünwald and Halpern 2004] for examples and an extended discussion of the phenomenon.

### 4.2 Standard Bayes

**Reliability as a Minimum Requirement** Decisions made by judges in court should be acceptable not just to themselves, but to large groups of people. All these people, even when given the same information in court, may employ different prior distributions. Thus, we almost inevitably end up with a *set* of prior distributions  $\Pi$ , resulting in a set of joint distributions  $\mathcal{P}$ . If we use any type of procedure which picks a single distribution  $\tilde{P}$  from the set  $\mathcal{P}$ , and then proceeds *as if* that were the right distribution to use, then it seems that “reliability” as defined in (10) is a minimum requirement:  $\tilde{P}$  should not give an overly optimistic impression of how well we predict if we base our predictions on  $\tilde{P}$ ; the expected loss of using  $\tilde{P}$  under  $\tilde{P}$  should not be smaller than the expected loss of using  $\tilde{P}$  under any other distribution in the set  $\mathcal{P}$  of considered distributions. Thus, if we adopt such a  $\tilde{P}$ , then, even though it might be wrong, the world will still behave as favourable to the decision-makers as it would if it were true.

My problem with a standard Bayesian approach, where, for example, the judge is allowed to use his own subjective point prior, is that this procedure is evidently not “reliable” in the sense above.

In fact, even in a robust Bayesian approach, although ignoring the first part of the data will in general not be minimax optimal, it may in some cases be the worst-case strategy among all strategies that are both (a) reliable and (b) easy to compute. For this reason, I still think that ignoring  $\mathbf{x}$  may sometimes be the method of choice after all. For example, in a more realistic approach than the one described here, one would probably also want to allow distributions  $P$  for which  $P(T = 1)$  does not just depend on  $\mathbf{x}$ , but also on  $\mathbf{y}$ ,  $S$ , and possibly some external random variables. In that case, the minimax optimal decision rule  $\tilde{\delta}$  in the  $\mathbf{y}$ -decision problem, which ignores  $\mathbf{x}$ , will not be minimax optimal in the  $\mathbf{x}, \mathbf{y}$ -problem,

but it remains reliable and easy to compute within the  $\mathbf{x}, \mathbf{y}$ -problem. On the other hand, the true minimax optimal decision rule in the  $\mathbf{x}, \mathbf{y}$ -problem may be exceedingly hard to find.

### 4.3 Neyman-Pearson Approach on $\mathbf{y}$

We have seen that the frequentist minimax optimal strategy  $\tilde{\delta}$  that ignores  $\mathbf{x}$  must be a Bayes rule relative to the second data set  $\mathbf{y}$  and some prior  $\pi(\mathbf{G})$ . Since  $P(\mathbf{y} | \mathbf{G})$  and  $P(\mathbf{y} | \mathbf{NG})$  are simple hypotheses, this means that the frequentist minimax optimal strategy is formally equivalent to a Neyman-Pearson test (a likelihood ratio test) with a particular significance level  $\alpha_m$ , in which  $\mathbf{NG}$  is viewed as the null hypothesis. A frequentist would, justifiably, object to calling the two decisions “guilty” and “not guilty”; in this section, it should rather be “ $\theta_{\mathbf{NG}}$  is rejected” (which in itself does not imply guilt!) and “ $\theta_{\mathbf{NG}}$  is not rejected”.

However, it should be noted that the significance  $\alpha_m$  depends on the sample size  $m$ ; in this sense, the minimax procedure is really very different from an NP test. It is of some interest to estimate  $\alpha_m$  as a function of  $m$ . This can be done using item 4 of Lemma 1. Below Lemma 1 we showed, using item 3 of the lemma, that, the minimax optimal rule  $\tilde{\delta}$  in the  $\mathbf{y}$ -game achieves minimax optimal loss  $e^{-m\tilde{D}+o(m)}$ . Exactly the same argument, using item 4, shows that  $\tilde{\delta}$  achieves both type-I and type-II error probabilities of exponential small size  $e^{-m\tilde{D}+o(m)}$ . In our example,  $\theta_{\mathbf{G}} = 3/4, \theta_{\mathbf{NG}} = 1/2, \tilde{D} \approx 0.034$ .

Thus, the minimax procedure resembles a NP test with a significance level  $\alpha_m \approx e^{-0.034m}$ , that goes to 0 exponentially fast as a function of the sample size, and is thus considerably different from a “real” NP test.

## References

- Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis* (revised and expanded 2nd ed.). Springer Series in Statistics. New York: Springer-Verlag.
- Cover, T. and J. Thomas (1991). *Elements of Information Theory*. New York: Wiley-Interscience.
- Ferguson, T. (1967). *Mathematical Statistics – a decision-theoretic approach*. San Diego: Academic Press.
- Grünwald, P. D. (1998, October). *The Minimum Description Length Principle and Reasoning under Uncertainty*. Ph. D. thesis, University of Amsterdam, The Netherlands. Available as ILLC Dissertation Series 1998-03; see [www.grunwald.nl](http://www.grunwald.nl).
- Grünwald, P. D. and A. P. Dawid (2004). Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics* 32(4), 1367–1433.
- Grünwald, P. D. and J. Y. Halpern (2004, July). When ignorance is bliss. In *Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2004)*, Banff, Canada.
- Seidenfeld, T. and L. Wasserman (1993). Dilation for convex sets of probabilities. *The Annals of Statistics* 21, 1139–1154.