T.E. Feenstra

Conditional Prediction without a Coarsening at Random condition

Master's thesis, September 7, 2012

Thesis advisor: Prof. Dr. P. D. Grünwald

Mathematical Institute, Leiden University

Abstract

Suppose a Decision Maker wants to make a prediction about the value of a random variable. He knows the distribution of the random variable, and he is also told that the outcome is contained in some given subset of the domain of the random variable. The Decision Maker is then asked to give his best guess of the true value of the random variable. The knee-jerk reflex of a probabilist is to use conditioning, if the probability of all outcomes is known. However, this reflex may well be incorrect if the specific outcome of the random variable is contained in more than one of the subsets that may be revealed.
This situation has been analysed in the literature in the case of a single (random) selection procedure. When the selection procedure satisfies a condition called *Coarsening at Random*, standard conditioning does the trick. However, in many cases this condition cannot be satisfied. We analyse the situation in which the selection procedure is unknown. We use a minimax approach of the Decision Maker against Nature, which can choose from a set of selection procedures. The loss of the Decision Maker is modeled by the logarithmic loss.
We give a minimax theorem applicable in our setting. This enables us to give a characterisation of the best prediction in all cases. Surprisingly, our results show that for certain cases this characterisation is a kind of reverse Coarsening at Random conditioning.

Acknowledgement

I would like to thank my supervisor, Prof. Dr. Peter Grünwald, for his continuous support throughout this research project.

Contents

Chapter 1

Introduction

There are many situations in which a statistician receives incomplete data and still has to reach conclusions about this data. One possibility of incomplete data is coarse data. This means that one does not observe the real outcome of a random event, but only a subset of all possible outcomes. An example frequently occurs in questionnaires. People have to state if their date of birth lies between 1950 and 1960 or between 1960 and 1970 et cetera. Their exact year of birth is unknown, but at least we now know for sure in which decade they are born. We introduce another instance of coarse data with the following example.

Example 1. [**Fair die**] Suppose I throw a fair die. I get to see the result of the die, but you do not. Now I tell you that the result lies in the set {1, 2, 3, 4}. This is an example of coarse data. The outcome is unknown to you, but it is not missing. At least you now know that the outcome is certainly not 5 or 6. You know that I used a fair die and would not lie to you. Now you are asked to predict the probabilities of each outcome. Probably, you would predict that the probability of each of the remaining possible results is 1/4. This is the knee-jerk reaction of someone who studied probability theory, since this is standard *conditioning*. But is this always correct?
Suppose that there is one alternative set I could tell you, namely the set {3, 4, 5, 6}. I can now use a selection procedure to decide which set I tell you, once I know the result of the die. If the outcome is 1, 2, 5 or 6 there is nothing to choose. If the outcome is 3 or 4, I randomly select set {1, 2, 3, 4} with probability 1/2 and also set {3, 4, 5, 6} with probability 1/2. If I throw the die 6000 times, I expect to see the outcome 3 a thousand times. Therefore I expect to report the set {1, 2, 3, 4} five hundred times after I see the outcome 3. By symmetry it is clear that I expect to report the set {1, 2, 3, 4} 3000 times in total. So the probability that the outcome is 3, when you know that the outcome lies in {1, 2, 3, 4} is actually 1/6 with this selection procedure. We see that the prediction in the first paragraph was not correct, in the sense that the probabilities do not correspond to the long-run relative frequencies. We conclude that the knee-jerk reaction is not always correct. ■

In Example 1 we have seen that standard conditioning is not always correct. The question when standard conditioning in situations in which the revealed subset is not unique, is correct, is answered by Heitjan and Rubin [1991]. They discovered a necessary and sufficient condition of the selection procedure, called CAR (Coarsening at Random). The selection procedure describes which subset is revealed given a particular value of the random variable.

Loosely speaking, this condition states that the probabilities to observe a subset y giving the outcomes $x \in y$ and $x' \in y$ respectively must be equal. The CAR condition is a property of the selection procedure. In many situations, however, this condition cannot hold, as proved by Grünwald and Halpern [2003]. It depends on the arrangement of possible revealed subsets if there exists a selection procedure satisfying CAR.

What is the best prediction conditional on the revealed subset, if CAR does not hold? Neither the paper by Grünwald and Halpern nor other literature on Coarsening at Random answers this question. The goal of this thesis is to identify the best prediction without making assumptions about the selection procedure, even in cases where CAR cannot hold.

We continue with an example generating much debate among both the general public and professional probabilists: the Monty Hall puzzle, discussed in Ask Marilyn, a weekly column in "Parade Magazine" [vos Savant, 1990, Gill, 2011]. In this example the CAR condition holds only for two special selection procedures.

Example 2. [The Monty Hall puzzle] Suppose you are on a game show and you may choose one of three doors. Behind one of the doors a car can be found, but the other two only hide a goat. Initially the car is equally likely to be behind each of the doors. After you have picked one of the doors, the host Monty Hall, who knows the location of the prizes, will open one of the doors revealing a goat. Now you are asked if you would like to switch to the other door, which is still closed. Is it a good idea to switch?
At this moment we will not answer this question, but we show that the Monty Hall problem is an example of the problems considered in this thesis. The true value of the random variable is the door concealing the car. When the host does open a door, different from the one you picked, revealing a goat, this is equivalent to revealing a subset. The subset he reveals is the set of the two still closed doors. For example, if he opens door 2, he reveals that the true value, the car, is in the subset of door 1 and door 3. Note that if you have by chance picked the correct door, there are two possible doors Monty Hall can open, so also two subsets he can reveal. This implies that Monty has a choice in revealing a subset. How does Monty's selection procedure influence your prediction of the true location of the car?
The CAR condition is only satisfied if Monty Hall uses a non-random strategy [Grünwald and Halpern, 2003]. That means he will either open always door 1 or open always door 2. So for any other selection procedure standard conditioning will not be best. ■

In order to reach the goal of this thesis we need to define what we mean by the best prediction. We consider the prediction as a game between the Decision Maker and Nature, which can choose a selection procedure from a given set of procedures. In general terms, we use a minimax approach, which is a standard approach in game theory [Von Neumann et al., 1944]. The best prediction minimises the worst possible loss among all possible predictions. We call this problem the conditional prediction problem. We prove a generalised form of the minimax theorem in Theorem 2 in Chapter 2. The loss function we use is the logarithmic loss. This loss function is chosen, because it is appropriate in so many different scientific fields: (quantum) information theory, coding theory and even actuarial science [Robertson et al., 2005]. Moreover, using the minimax theorem, we find that the minimax of the logarithmic loss is equal to the maximum conditional entropy. This is a well-studied problem. Other loss functions, such as a 0/1 loss function, or distance measures, such as the total variation

distance, are not considered. In chapter 6 we discuss areas of future research with these loss functions and we conjecture what the best prediction would be under different loss functions.

We do not make any assumption concerning the selection procedure, but we do have to make three other assumptions in order to get significant results.

(a) We make the assumption that the underlying probability space is finite.
(b) We assume that the subsets which can be revealed, are known by the Decision Maker beforehand.
(c) The probability of a particular outcome of the random variable is known.

Assumption (c) is the most restrictive assumption, especially concerning applications in statistics, where (c) is almost always violated. However, in game theory this assumption is regularly made. Note that the Monty Hall problem and Example 1 above satisfy all these assumptions.

Under assumptions (a), (b) and (c), we can find the best prediction for the random variable conditional on the revealed subset. A complete characterisation of the best prediction can be given. This is done in Theorem 4 and Theorem 5 in Chapter 4. The first result shows that the best prediction satisfies a property that has a striking similarity with the CAR condition, but switches the role of outcome and set in this condition. The strenght of the results come from the fact that the characterisation is valid in every set-up satisfying the three assumptions, and for any distribution of the random variable. As an illustration, we show this characterisation in the setting of the fair die, introduced in Example 1.

Example 3. [Fair die, continued] The best prediction conditional on the revealed subset is found with the help of Theorem 4. The best prediction given that you observe the set {1, 2, 3, 4} is: predict the outcome 1 and 2 each with probability 1/3; and 3 and 4 both with probability 1/6. Given that you observe the set {3, 4, 5, 6} the best prediction is: 3 and 4 both with probability 1/6; and 5 and 6 with probability 1/3.
These probabilities correspond with the selection procedure mentioned in Example 1. However, the best prediction is independent of the selection procedure. Why then, is this the best solution? Here is the intuitive idea. If I want, I can choose a very extreme selection procedure: if I see a 3 or 4, I always reveal the set {1, 2, 3, 4}. The other extreme selection procedure of course never reveals this set, if the result is 3 or 4. The best prediction given above hedges against both possibilities.

■

Caveats on the use of the word *conditioning*

Since this thesis is concerned with making the best prediction conditional on a set of outcomes, we want to highlight the use of the word *conditioning*. We use the word *standard* conditioning for using the conditional information in the standard way. This means that the a-priori probability of an outcome is divided by the a-priori probabilities of all outcomes in the set. In Example 1 standard conditioning assigns probability 1/4 to all possible results in the set $\{1, 2, 3, 4\}$.
Above we stated that standard conditioning is not always correct. But standard conditioning is correct in a larger space, namely the space of all combinations of outcome and set. But

there you do not have the probabilities, since the selection procedure is unknown. So in some sense standard conditioning is correct, but it does not help us.
We use the word conditional prediction for the prediction of the Decision Maker conditional on the revealed set of outcomes. For each set of outcomes, it is a probability distribution. But it is not necessarily the standard conditional probability distribution. Finally, we call the probability assigned by a conditional prediction to one outcome conditional on a set, the conditional predictive probability.

Overview

Chapter 2 casts the problem of obtaining the best prediction in mathematical terms. This is done by introducing our mathematical definition of the best prediction. We use the common minimax approach. The logarithmic loss is our choice of loss function. The first result is a generalised minimax theorem, applicable in this specific setting (Theorem 2). The minimax theorem is needed to recast the problem in the form of maximising the conditional entropy. Maximising the conditional entropy is used extensively in the remainder of the thesis, because it is much easier to handle than minimising the worst-case loss.
Chapter 3 begins with several examples of our problem, one of which is the Monty Hall problem. It discusses the best conditional prediction in each case. Most examples are then generalised in the form of a lemma. These lemmata give (part of) the solution for small classes of problems. If the possible subsets form a partition, then standard conditioning is the best you can do. Another intuitive result is the following: If Nature can choose between revealing a set or subset of this set, then it chooses the larger set, hence giving the Decision Maker a harder time predicting. Also we find the general solution when the whole domain of the random variable can be revealed, which is the case when there is missing data. In this case Nature only gives you missing data. In Example 1 this would amount to telling you that the outcome is in the set $\{1, \ldots, 6\}$. Other examples and lemmata are less obvious and are surprising results in themselves.
In Chapter 4 a broad characterisation of the best prediction, valid in a wide range of situations, is presented (Theorem 4). Surprisingly, this characterisation is similar to a reverse CAR condition. The best prediction gives each specific outcome the same conditional probability, independent of the revealed subset. This is then generalised to give a characterisation of the best prediction in the general case (Theorem 5). For this general case an innovative trick is used. We extend the structure of the problem by adding an outcome to each possible subset and giving each new outcome a small probability. The characterisation in Theorem 4 yields a solution. Upon taking the probability of the added outcomes to zero, in the limit the best prediction of the original problem is obtained.
In the final chapter of this thesis a summary of all results is given. In that chapter the implications of the results and possible paths for further research are discussed. We provide some detail to the importance of the choice of loss function. Moreover, the best conditional prediction using other possible loss functions is hypothesised.

Chapter 2

Modelling of the conditional prediction problem

In this chapter we discuss the mathematical formulation of the prediction problem. In the first section a straightforward formalisation is given. In Section 2.2 we state and prove useful properties of the object function: finiteness, continuity and convexity. These properties are then used in the third section of this chapter to recast the problem in terms of the conditional entropy. This is done using a special form of the minimax theorem. In the final part of the chapter an interpretation of our model of the conditional prediction problem is given.

2.1 Conditional Predictions and Best Conditional Predictions

2.1.1 No-Lie and Marginal Consistency Axioms

First, we again state the problem. We formalise the following question: What is the best prediction of a random variable, conditional on a revealed subset containing the true value? The prediction takes the form of a probability distribution. We assume that the probability space is finite and the probability distribution of the random variable is known to the Decision Maker. Moreover, all possible subsets, i.e. the subsets that may possibly be revealed, are known. In summary, the whole structure of the problem is known to the Decision Maker, except the process selecting which subset is revealed. The prediction is actually made before the subset is given. For each possible subset, a probability distribution of the random variable is predicted.

We formalise the problem formulated above. Consider a random variable X which takes its values in a finite set $\mathcal{X}$. X is not observed, instead the observation will be a subset of $\mathcal{X}$ denoted by Y. The distribution of X is known. Denote by p_x the probability of $X = x$. We assume that $p_x > 0$ for every x, otherwise one could just remove the element x. The collection of possible subsets which may be observed is known and denoted by $\mathcal{S}$, so $\mathcal{S}$ is a subset of the power set of $\mathcal{X}$. The procedure how a subset is chosen given X, which can be probabilistic, is unknown. We require only that the *'no lie'-axiom* holds: $X \in Y$ with probability 1. This is called the 'no lie'-axiom, since the axiom prohibits that the revealed subset does not contain the true value (in this case, the Decision Maker would be told a lie). We call

a probability distribution on $\mathcal{X} \times \mathcal{S}$ with the marginal distribution equal to the distribution of X, as given by $\{p_x : x \in \mathcal{X}\}$, *marginal consistent*. In other words $P(\{x\} \times \mathcal{S}) = p_x$. Denote by $\mathcal{P}$ all probability distributions on $\mathcal{X} \times \mathcal{S}$ which are marginal consistent and satisfy the 'no lie'-axiom. Note that for every x there is a set in $\mathcal{S}$ containing x, otherwise the 'no lie'-axiom could not hold.

2.1.2 Conditional and Best Conditional Predictions

Given the subset Y the problem is now to find the best prediction of X. To state the problem in a mathematically precise way we need the following definition:

Definition 1. *A* conditional prediction *of X is given by a function Q: $\mathcal{X} \times \mathcal{S} \to [0, 1]$ that satisfies the following two conditions:*

1. *For all $y \in \mathcal{S}$, $\sum_{x \in y} Q(X = x|Y = y) = 1$.*
2. *For all $y \in \mathcal{S}$, for all $x \notin y$, $Q(X = x|Y = y) = 0$.*

We adopt here the notation $Q(X = x|Y = y)$ for Q evaluated in x and y. To avoid confusion, we use this notation instead of $Q(x|y)$ whenever possible. Let $\mathcal{Q}$ be the set of all conditional predictions.

The first condition just states that for all $y \in \mathcal{S}$ $Q(X = \cdot|Y = y)$ is a probability distribution. The second condition is again the 'no lie'-axiom, but now it is a statement about the rationality of the Decision Maker, who makes the prediction. It is not rational to predict that the revealed subset does not contain the true value. This prediction could not hold under any possible probability distribution $P \in \mathcal{P}$. We only adopt this condition for simplicity. It is not strictly necessary: if we would drop this condition, the best condition prediction satisfies the second condition (see Section 3.2).

Finally we can state our problem in mathematical form. We use for the loss function the expectation of the logarithmic loss [Cover and Thomas, 1991]. We explain the choice of this loss function in Section 2.4. After a subset $y \in \mathcal{S}$ is revealed, the logarithmic loss of the conditional prediction is given by $\sum_x (-\log Q(X = x|Y = y))$. Note that the expected loss is given by

$$\mathbb{E}_P \left[-\log Q(X|Y)\right] := \sum_{x,y} P(X = x, Y = y) \left(-\log Q(X = x|Y = y)\right). \tag{2.1}$$

The logarithmic loss gives a large penalty if the Decision Maker is confident that a pair (x, y) has a low probability, but P assigns a large probability to that pair. Decision Maker suffers even an infinitely large loss if he assigns zero conditional predictive probability to a pair (x, y), which has positive probability under P.
We need some useful convention when we encounter in this some terms of the type $0 \log 0$. Given that $\lim_{a \to 0} a \log a = \lim_{b \to \infty} 1/b \log(1/b) = \lim_{b \to \infty} -1/b \log(b) = 0$, we use $0 \log 0 = 0$.

We are now ready to define which conditional predictions we call best.

Definition 2. *A conditional prediction Q is* best *if:*

$$\max_{P\in\mathcal{P}} \mathbb{E}_P\left[-\log Q(X|Y)\right] = \min_{Q'\in\mathcal{Q}} \max_{P\in\mathcal{P}} \mathbb{E}_P\left[-\log Q'(X|Y)\right]. \tag{2.2}$$

The loss consists of the cross entropy between the conditional prediction Q and a probability distribution P. In words, the above formulation of the problem means that a conditional prediction is the best if it minimises the maximum loss. As we will later show in Example 10 in Chapter 3, in general there can be several conditional predictions $Q \in \mathcal{Q}$ for which the above expression holds. When we later state that a conditional prediction is 'the' solution, it is just one of the minimising conditional predictions. But we do know that there always exists at least one best conditional prediction. This will be shown later in Lemma 1. In the following example the definitions, which are given above, are used in the fair die example (see Example 1 and Example 3).

Example 4. [Fair die, continued] We describe the fair die example using the definitions in Section 2.1 and 2.2. The random variable X is the (hidden) outcome of the die, which takes values in $\mathcal{X} = \{1, \dots 6\}$. The possible subsets I can tell you are in $\mathcal{S} = \{\{1,2,3,4\},\{3,4,5,6\}\}$. The procedure I use to choose a subset has to satisfy two conditions. The 'no lie'-axiom means that if the outcome is 1 or 2, I must tell you the subset $\{1,2,3,4\}$ and if the outcome is 5 or 6 the subset $\{3,4,5,6\}$. We defined the selection procedure as a probability distribution on $\mathcal{X}\times\mathcal{S}$. To be marginal consistent a probability distribution must assign total probability 1/6 to each outcome, i.e. $P(X = x|Y = \{1,2,3,4\}) + P(X = x|Y = \{3,4,5,6\}) = 1/6$.
In Example 3 we claimed that the best conditional prediction is the following. The best prediction given that you observe the set {1, 2, 3, 4} is: predict the outcome 1 and 2 each with probability 1/3; and 3 and 4 both with probability 1/6. Given that you observe the set {3, 4, 5, 6} the best prediction is: 3 and 4 both with probability 1/6; and 5 and 6 with probability 1/3. Firstly, this is a probability distribution for each observed set. Secondly, the best prediction does not assign positive probability to an impossible outcome. For example, given the set {1, 2, 3, 4}, we assign no positive probability to the outcomes 5 and 6. We conclude that both conditions of Definition 1 are satisfied and this is a conditional prediction. ■

2.1.3 Game Theoretic Interpretation

In this subsection the game theoretic interpretation of the above definition of best conditional prediction is given. In Section 2.4 we motivate the choice of the logarithmic loss function. Here we highlight the use of the minimax approach.
One can view the conditional prediction problem as a game between the Decision Maker, who chooses the Q, and Nature, which chooses P in such a way as to make the loss as large as possible. Note that in reality there often is no real opponent, trying to maximise the Decision Maker's loss. In the examples we provided up till this point one can see Nature as an opponent. In the fair die example (Example 1) or the Monty Hall problem (Example 2) there is a human being, who may be trying to maximise your loss. However, if you try to find patterns in data not generated by humans, there is no malevolent Nature. We still use the minimax approach, which is based on a worst-case scenario. The idea is that you really do not know

the selection procedure and you are not willing to make assumptions about this procedure.

Note that we do not require that Q is the result of conditioning on a probability distribution in $\mathcal{P}$. So the prediction the Decision Maker makes does not have to correspond with a possible selection procedure. An example is given in Example 5. On the other hand, we have that a probability distribution P in $\mathcal{P}$ gives rise to one or more Q in $\mathcal{Q}$ by setting $Q(X = x|Y = y) := P(X = x|Y = y)$ when $P(Y = y) > 0$ and any probability distribution over all x in y when $P(Y = y) = 0$. We call such a Q *generated* by P.

Example 5. [Prediction Q not generated by P] We show in this example that there are conditional predictions in $\mathcal{Q}$ which cannot be generated by a P in $\mathcal{P}$ with the method described above.
Let $\mathcal{X} = \{a, b, c\}$ and $\mathcal{S} = \{\{a, b\}, \{c\}, \{a, b, c\}\}$ and $p_x = 1/3$ for every x. We claim that Q given by $Q(X = a|Y = \{a, b\}) = 1/4$, $Q(X = a|Y = \{a, b, c\}) = 1/4$ and $Q(X = b|Y = \{a, b, c\}) = 1/2$ will be such a conditional prediction. Note that this indeed fully characterizes Q. For example, it must hold that $Q(X = c|Y = \{c\}) = 1$, by the requirements on Q.
Suppose by means of contradiction that P in $\mathcal{P}$ is such that $Q(X = x|Y = y) = P(X = x|Y = y)$ for all y with $P(y) > 0$. We will show that this P cannot have the right marginal distribution on $\mathcal{X}$, which implies that $P \notin \mathcal{P}$. It is impossible that $P(Y = \{a, b\}) = 0$, as we would then have that $P(X = a, Y = \{a, b, c\}) = 1/3 = P(X = b, Y = \{a, b, c\})$, hence $Q(X = a|Y = \{a, b, c\}) = P(X = a|Y = \{a, b, c\}) = P(X = b|Y = \{a, b, c\}) = Q(X = b|Y = \{a, b, c\})$. The same reasoning shows that $P(Y = \{a, b, c\}) = 0$ gives rise to a contradiction. Now we use the fact that $Q(X = a|Y = \{a, b\}) < Q(X = b|Y = \{a, b\})$ and $Q(X = a|Y = \{a, b, c\}) < Q(X = b|Y = \{a, b, c\})$ to conclude $P(X = a|Y = \{a, b\}) < P(X = b|Y = \{a, b\})$ and $P(X = a|Y = \{a, b, c\}) < P(X = b|Y = \{a, b, c\})$. This is impossible using that $p_a = p_b$ and the only sets containing a or b are $\{a, b\}$ and $\{a, b, c\}$. We conclude that this Q cannot be obtained by a P in $\mathcal{P}$. ■

2.2 Finiteness, convexity and continuity

In this section we prove several properties of conditional predictions. Firstly, we show that there is a conditional prediction with finite expected loss. Secondly, the problem has a concave-convex optimisation structure. Thirdly, we show that the loss we use is semi-continuous. These lemmata are needed in Section 2.3, where we will prove a minimax theorem.

The first lemma shows that there is always a conditional prediction which even in the worst case has a finite expected loss.

Lemma 1. *There exists a $Q \in \mathcal{Q}$ such that $\max_{P\in\mathcal{P}} \mathbb{E}_P [-\log Q(X|Y)] < \infty$.*

Proof. Denote for every $x \in \mathcal{X}$ by n_x the number of sets in $\mathcal{S}$ which contain x. By our requirements above we know that $n_x > 0$. Now define Q in the following way: for every $x \in \mathcal{X}$ and $y \in \mathcal{S}$ let $Q(X = x|Y = y) = p_x/n_x$ if $x \in y$ and 0 otherwise. Note that this Q satisfies the 'no lie'-axiom and $\sum_{y\in\mathcal{S}} Q(X = x|Y = y) = p_x$ for every x. We conclude that $Q \in \mathcal{Q}$. Moreover we see that $Q(X = x|Y = y) = 0$ only if $x \notin y$ and no $P \in \mathcal{P}$ can assign nonzero probability to such events by the 'no lie'-axiom. Hence $\mathbb{E}_P [-\log Q(X|Y)] \leq \max_{x\in\mathcal{X}} -\log p_x/n_x \leq \max_{x\in\mathcal{X}} \log n_x$ for every $P \in \mathcal{P}$. So for this Q we have that $\max_{P\in\mathcal{P}} \mathbb{E}_P [-\log Q(X|Y)] \leq \max_{x\in\mathcal{X}} \log n_x < \infty$. □

The next lemma shows that $\mathcal{P}$ is a convex set.

Lemma 2. *$\mathcal{P}$ is a convex set.*

Proof. Suppose P^1 and $P^2 \in \mathcal{P}$. Let $t \in [0,1]$. Denote $P = tP^1 + (1-t)P^2$. Clearly P is a probability distribution on $\mathcal{X} \times \mathcal{S}$. Moreover we have for all $x \in \mathcal{X}$:

$$P(X = x) = tP^1(X = x) + (1-t)P^2(X = x) = tp_x + (1-t)p_x = p_x. \tag{2.3}$$

Also P satisfies the 'no lie'-axiom, because P^1 and P^2 do. So we find that $P \in \mathcal{P}$. This proves that $\mathcal{P}$ is a convex set. □

For completeness we also give the nearly identical proof that $\mathcal{Q}$ is a convex set.

Lemma 3. *$\mathcal{Q}$ is a convex set.*

Proof. Suppose Q^1 and $Q^2 \in \mathcal{Q}$. Let $t \in [0,1]$. Define $Q = tQ^1 + (1-t)Q^2$. We have for all $x \in y$:

$$\sum_{x \in y} Q(X = x|Y = y) = t \sum_{x \in y} Q^1(X = x|Y = y) + (1-t) \sum_{x \in y} Q^2(X = x|Y = y) = t + (1-t) = 1. \tag{2.4}$$

Also Q satisfies the 'no lie'-axiom, because Q^1 and Q^2 do. So we find that $Q \in \mathcal{Q}$. This proves that $\mathcal{Q}$ is a convex set. □

Let us now investigate the function $f : \mathcal{P} \times \mathcal{Q} \to [0, \infty]$ given by $f(P, Q) = \mathbb{E}_P[-\log Q(X|Y)]$ which is part of equation (2.2). In the following two lemmata we show that this function f has the required properties,to apply a minimax theorem.

Lemma 4. *The function $f(P, Q)$ is concave in P for every Q and convex in Q for every P, but neither strictly concave nor strictly convex.*

Proof. The function is linear in P, as it is just an expectation, which directly shows that it is concave but not strictly concave. Also convexity is clear as $-\log$ is convex. To show that it is not strictly convex we give a counterexample. Let $\mathcal{X} = \{1,2,3\}$ and $\mathcal{S} = \{\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$ and $p_x = 1/3$ for every x. We define P, Q^1 and Q^2 using a table where each column is an element and each row a set in $\mathcal{S}$, where we give the joint probability for the set and element for P and the conditional probability for Q^1 and Q^2. For example, the upper left entry in the second table means that $Q^1(X = 1|Y = 1,2) = 1$. Note that the columns of the table for P sum to the marginal distribution of X. Moreover the rows of the table for Q sum to 1.

P	1	2	3
$\{1,2\}$	0	0	0
$\{1,3\}$	0	0	0
$\{2,3\}$	0	0	0
$\{1,2,3\}$	1/3	1/3	1/3

Q^1	1	2	3
$\{1,2\}$	1	0	0
$\{1,3\}$	0	0	1
$\{2,3\}$	0	1	0
$\{1,2,3\}$	1/3	1/3	1/3

Q^2	1	2	3
$\{1,2\}$	0	1	0
$\{1,3\}$	1	0	0
$\{2,3\}$	0	0	1
$\{1,2,3\}$	1/3	1/3	1/3

.

We constructed P such that it assigns all probability to the set $\{1,2,3\}$. Note that Q^1 and Q^2 assign equal probabilities on this set, hence the same holds for all convex combinations of the two. Therefore $f(P,Q)$ has the same value for each Q which is a convex combination of Q^1 and Q^2. We conclude that the function is not strictly convex. □

Next we investigate the continuity of $f(P,Q)$. The definition of convergence we use is the convergence in Euclidean space, where a distribution function is regarded as a vector. The length of this vector is the number of elements in $\mathcal{X}$ multiplied by the number of elements in $\mathcal{S}$. Because all spaces we consider are finite, this convergence is equivalent to convergence in distribution or almost surely convergence. Note that $\mathcal{P}$ and $\mathcal{Q}$ are compact, since they are closed and bounded in Euclidean space. Boundedness is clear from the fact that all elements are between zero and one. Closedness follows from the fact that all conditions on the elements are equalities.

Lemma 5. *The function $f(P,Q)$ is lower semi-continuous in P, but not continuous. It is continuous in Q. The discontinuities can only occur if $P(X=x,Y=y)=0$ for some $x \in y$.*

Proof. First fix Q and let $P_n \to P$. Now suppose that P is such that $P(X=x,Y=y)>0$ for all $x \in y$. We show that in this case no discontinuities occur. From the above assumption we know that there is an N such that for all $n > N$, we have that $P_n(X=x,Y=y)>0$ for all $x \in y$. With this condition we have that $\mathbb{E}_{P_n}[-\log Q(X|Y)] \to \mathbb{E}_P[-\log Q(X|Y)]$. So discontinuities can only occur if $P(X=x,Y=y)=0$ for some $x \in y$.

So suppose that $P(X=x,Y=y)=0$ for some $x \in y$. If $Q(X=x|Y=y)>0$, there are no problems: $P_n(X=x,Y=y)(-\log Q(X=x|Y=y)) \to 0 = P(X=x|Y=y)(-\log Q(X=x|Y=y))$. Let now $Q(X=x|Y=y)=0$. In this case we have that for all n $\mathbb{E}_{P_n}[-\log Q(X|Y)] \geq 0 = \mathbb{E}_P[-\log Q(X|Y)]$. We conclude that $f(P,Q)$ is lower semi-continuous in P.

Fix now P and let $Q_n \to Q$. As above there are no problems if $P(X=x|Y=y)>0$ for all $x \in y$. Suppose that $P(X=x,Y=y)=0$ for some $x \in y$. If $Q(X=x|Y=y)>0$, there are no problems: $P(X=x,Y=y)(-\log Q_n(X=x|Y=y)) \to 0 = P(X=x|Y=y)(-\log Q(X=x|Y=y))$. Let now $Q(X=x|Y=y)=0$. Then the term $P(X=x|Y=y)(-log(Q_n(X=x|Y=y))$ is zero for all n and also $P(X=x|Y=y)(-log(Q(X=x|Y=y))$ is zero. So $f(P,Q)$ is continuous in Q.

With a counterexample we show that $f(P,Q)$ does not have to be continuous in P. Let $\mathcal{X}=\{1,2\}$ and $\mathcal{S}=\{\{1\},\{1,2\}\}$ and the marginal distribution of X is $(1/2,1/2)$. Let P_n be given by $P_n(X=1,Y=\{1,2\})=1/(2n)$ for $n=1,2,\ldots$. Let the conditional prediction be given by $Q(X=1|Y=\{1,2\})=0$. Then we have that $\mathbb{E}_{P_n}[-\log Q(X|Y)]=\infty$ for all n. However $P_n \to P$ in distribution (and hence almost surely, as $\mathcal{X}$ is finite) for P given by $P(X=1,Y=\{1,2\})=0$ and $\mathbb{E}_P[-\log Q(X|Y)]=0<\infty$. We conclude that $f(P,Q)$ is not continuous. □

2.3 Characterising the Solution in terms of Conditional Entropy

The problem as stated in equation (2.2) has two variables P and Q. One must find the best conditional prediction Q, so one must investigate all conditional predictions. Moreover, for each conditional prediction all probability distributions P must be considered. This makes the problem difficult to solve. In the remainder of this chapter we simplify equation (2.2), by using a minimax theorem. With this minimax theorem the problem is restated as maximising conditional entropy. To this end we use the lemmata of Section 2.2.

First we prove that there always exists a best solution.

Theorem 1. *There exists a condition prediction $Q \in \mathcal{Q}$ that is a best conditional prediction having a finite expected loss.*

Proof. The function $\max_{P\in\mathcal{P}} \mathbb{E}_P\left[-\log Q'(X|Y)\right]$ is bounded from below, since the function cannot be negative. It is moreover semi-continuous by Lemma 5. We conclude that it obtains it minimum, i.e. there exists a $Q \in \mathcal{Q}$ that is a best conditional prediction. By Lemma 1 this Q has a finite expected loss. □

This enables us to switch the order of minimisation and maximisation:

Theorem 2. *The minimax theorem holds, i.e.*

$$\min_{Q\in\mathcal{Q}} \max_{P\in\mathcal{P}} \mathbb{E}_P\left[-\log Q(X|Y)\right] = \max_{P\in\mathcal{P}} \min_{Q\in\mathcal{Q}} \mathbb{E}_P\left[-\log Q(X|Y)\right]. \tag{2.5}$$

Proof. First we give an overview of the proof. We cannot directly use a standard minimax theorem, since the function can attain the value ∞. Standard minimax theorems either can not deal with that or require stronger continuity assumptions (see [Fan, 1953] and [Grünwald and Dawid, 2004]). Therefore we have to use a trick.
One inequality always holds, that is

$$\min_{Q\in\mathcal{Q}} \max_{P\in\mathcal{P}} \mathbb{E}_P\left[-\log Q(X|Y)\right] \geq \max_{P\in\mathcal{P}} \min_{Q\in\mathcal{Q}} \mathbb{E}_P\left[-\log Q(X|Y)\right]. \tag{2.6}$$

On the left hand side the maximising player reacts on the minimising player and on the right hand side the other way around. Therefore it is clear that the left hand has a higher value than the right hand side.
We will prove the theorem by showing that for all $\epsilon \geq 0$, we have

$$\min_{Q\in\mathcal{Q}} \max_{P\in\mathcal{P}} \mathbb{E}_P\left[-\log Q(X|Y)\right] \leq \max_{P\in\mathcal{P}} \min_{Q\in\mathcal{Q}} \mathbb{E}_P\left[-\log Q(X|Y)\right] - \epsilon. \tag{2.7}$$

We do this by considering not our set of conditional predictions $\mathcal{Q}$, but an appropriate $\mathcal{Q}' \subset \mathcal{Q}$, in which no conditional prediction can be 0. Using this $\mathcal{Q}'$ enables us to use an ordinary minimax theorem. We finalise the proof by arguing that the maximin loss using $\mathcal{Q}'$ is not much different from the maximin loss using $\mathcal{Q}'$. This finishes the overview of the proof.

First we prove the following claim: for all $\epsilon > 0$, we can find a $\delta > 0$ such that for all $P \in \mathcal{P}$

$$\min_{Q\in\mathcal{Q}} \mathbb{E}_P\left[-\log Q(X|Y)\right] \geq \min_{Q'\in\mathcal{Q}'} \mathbb{E}_P\left[-\log Q'(X|Y)\right] - \epsilon, \tag{2.8}$$

where $\mathcal{Q}'$ is defined below and $Q \in \mathcal{Q}' \subset \mathcal{Q}$ satisfies $Q(x|y) \geq \delta$, if $x \in y$. In words this means that a conditional prediction in $\mathcal{Q}'$ cannot be on the boundary of $\mathcal{Q}$.

Let $\epsilon > 0$ be given. We define for a conditional prediction $Q \in \mathcal{Q}$ a conditional prediction $Q' \in \mathcal{Q}'$ in the following way. Let $Q'(x|y) = \delta$, if $Q(x|y) < \delta$ and $x \in y$. For each $y \in \mathcal{S}$, find an $x \in y$ such that $Q(x|y)$ is maximal and define $Q'(x|y) = Q(x|y) - \sum_{x' \in y}(\delta - Q(x'|y))^+$. All other values remain unchanged. In words the Q' we constructed increased the value of all predictions that are too small to the cut-off value δ and takes the required mass from the largest prediction. Note that for δ small enough this Q' lies in $\mathcal{Q}'$.

Now we find that $-\log Q'(x|y)$ is only greater than $-\log Q(x|y)$ for the $x \in y$ that now have a smaller predictive probability. Let n be the number of elements in $\mathcal{X}$. We have that

$$\begin{aligned} -\log Q'(x|y) - -\log Q(x|y) &= -\log\left(1 - \frac{\sum_{x' \in y}(\delta - Q(x'|y))^+}{Q(x|y)}\right) && (2.9)\\ &< -\log\left(1 - \frac{\delta n}{Q(x|y)}\right) && (2.10)\\ &\leq -\log\left(1 - \delta n^2\right), && (2.11) \end{aligned}$$

since $Q(x|y) \geq 1/n$, because $Q(x|y)$ is maximal at x. Now the difference in (2.9) is smaller than ϵ for $\delta = (1 - e^{-\epsilon})/n^2$. This proves also that $\mathbb{E}_P[-\log Q(X|Y)] \geq \mathbb{E}_P[-\log Q'(X|Y)] - \epsilon$, since $P(x|y) \leq 1$. Since this holds for all $Q \in \mathcal{Q}$, we have proved the claim.

This claim enables us to prove the minimax theorem:

$$\begin{aligned} \min_{Q \in \mathcal{Q}} \max_{P \in \mathcal{P}} \mathbb{E}_P[-\log Q(X|Y)] &= \min_{Q' \in \mathcal{Q}'} \max_{P \in \mathcal{P}} \mathbb{E}_P\left[-\log Q'(X|Y)\right] && (2.12)\\ &= \max_{P \in \mathcal{P}} \min_{Q' \in \mathcal{Q}'} \mathbb{E}_P\left[-\log Q'(X|Y)\right] && (2.13)\\ &\leq \max_{P \in \mathcal{P}} \min_{Q \in \mathcal{Q}} \mathbb{E}_P[-\log Q(X|Y)] - \epsilon. && (2.14) \end{aligned}$$

Equation (2.12) holds, because the minimiser $Q \in \mathcal{Q}$ cannot assign a value 0 to any (x, y) with $\sup_{P \in \mathcal{P}} P(x, y) > 0$. If that would be the case, then the P-expected loss is infinitely large, contradicting Theorem 1. So for δ small enough Equation (2.12) holds.

Equation (2.13) is an application of Fan's minimax theorem [Fan, 1953]. Indeed, as stated before, $\mathcal{P}$ and $\mathcal{Q}'$ are compact and convex, as proved in Lemmata 2 and 3. Moreover on $\mathcal{Q}'$ our function $\mathbb{E}_P[-\log Q(X|Y)]$ is real-valued. This function is concave on $\mathcal{P}$ and lower semi-continuous and convex on $\mathcal{Q}'$, see Lemmata 4 and 5.

Equation (2.14) is the result of our claim. Since this holds for all $\epsilon > 0$, we conclude that the minimax theorem holds. □

Now that we have the minimax theorem, we can simplify equation (2.2) and restate it as maximising conditional entropy [Cover and Thomas, 1991]. Conditional entropy can be interpreted as the average number of bits needed to code the outcome given that the subset is known in the optimal way using lossless coding. To be precise, the number of bits does not need to be integer, so it is a generalised notion of bits. Concretely, by Theorem 2 we have that for each $P \in \mathcal{P}$ the minimum is attained with a Q generated by P:

$$\max_{P \in \mathcal{P}} \min_{Q \in \mathcal{Q}} \mathbb{E}_P[-\log Q(X|Y)] = \max_{P \in \mathcal{P}} \mathbb{E}_P[-\log P(X|Y)] = \max_{P \in \mathcal{P}} H(X|Y). \quad (2.15)$$

Here $H(X|Y)$ is the conditional entropy when $(X, Y) \sim P(x, y)$. And we use constantly

$$H(X|Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{S}} P(X = x, Y = y) \log \frac{P(Y = y)}{P(X = x, Y = y)} \tag{2.16}$$

We need some useful convention when we encounter in this some terms of the type $0 \log \frac{0}{0}$. Given that $\lim_{a \to 0} a \log a = \lim_{b \to \infty} 1/b \log(1/b) = \lim_{b \to \infty} -1/b \log(b) = 0$, we use $0 \log \frac{0}{0} = 0 \log 0 - 0 \log 0 = 0$.

This characterisation of the solution in terms of conditional entropy is used extensively in the next chapters. Instead of solving the minimax question, we can solely focus on finding the maximiser of the conditional entropy P. This is the optimal selection procedure for Nature. Then the best prediction for the Decision Maker is a conditional prediction generated by this P. In the next section we discuss why it is convenient to use the conditional entropy.

2.4 Discussion of the Model

In this section the model proposed in this chapter is discussed. In Subsection 2.1.3 we discussed the game theoretic interpretation and focused on the minimax approach. Here we discuss the choice of the logarithmic loss function, the minimax theorem of Section 2.3 and the equivalence with the *rate-distortion problem*.

Why do we choose to use the logarithmic loss function? The main reason that we choose the logarithmic loss function is a practical one. In Section 2.3 it is shown that the minimax of the conditional logarithmic loss is equivalent to maximising the conditional entropy. Entropy maximisation has a long tradition in information theory, cf. Jaynes [1957]. Entropy can be viewed as the expected number of bits needed to code the outcome of a random variable, when you use the best code. So using the logarithmic loss function enables us to use entropy maximisation. For the maximisation it is convenient that entropy and conditional entropy is concave [Cover and Thomas, 1991].

We want to give another interpretation of a conditional prediction, following Chapter 6 of Cover and Thomas [1991]. A good way to view a conditional prediction Q is the following. Suppose the setting is known and the set containing the outcome of the random variable is revealed. Then a third person independent from Nature shows up. This person, a bookmaker, offers you a bet on each outcome of the random variable. To decide how you spend your money, you just look at your conditional prediction.
This interpretation is different from a more frequentist view. The Decision Maker is not guarantied that the conditional prediction is the result of the *true* selection procedure. As in the example of the fair die (Example 3), the best conditional prediction assigns probability 1/6 to the outcome 4, after observing the set {1, 2, 3, 4}. However, the true selection procedure could be that this set is never shown, when the outcome is 4. The true frequency of outcome 4, after observing the set {1, 2, 3, 4}, is then 0, which differs from the prediction 1/6. So the best prediction does not predict the true frequency.

As stated above the reason for choosing the logarithmic loss function is the resulting conditional entropy maximisation. To get this result the minimax Theorem 2 is essential. The

first occurrence of a minimax theorem that equaled maximising conditional entropy and the minimax conditional log loss is found in Harremoës and Topsøe [2002]. We tried to use one of the general minimax theorems described in Grünwald and Dawid [2004], but they were not applicable in this setting.

Finally, we note that, interestingly, our problem is equivalent to the rate-distortion problem, a central problem in information theory. See chapter 10 of Cover and Thomas [1991] for a discussion of the rate-distortion problem and the associated definitions. The conditional prediction problem is equivalent to the rate-distortion problem with the following distortion measure: $d(x, y) = 0$ if $x \in y$ and $d(x, y) = 1$ if $x \notin y$, and maximum distortion $D = 0$. This models our 'no lie'-axiom.

The mutual information between two random variables X and Y, $I(X;Y)$, is given by $\sum_{x\in\mathcal{X},y\in\mathcal{S}} P(x, y) \log \frac{P(x,y)}{P(x)P(y)}$. Theorem 2.4.1 of Cover & Thomas gives that $I(X;Y) = H(X) - H(Y|X)$. $H(X)$ is a constant so we find that

$$\arg\max_{P\in\mathcal{P}} H(X|Y) = \arg\min_{P\in\mathcal{P}} H(X) - H(X|Y) = \arg\min_{P\in\mathcal{P}} I(X;Y). \tag{2.17}$$

And this is precisely the definition of the rate-distortion problem. Unfortunately, the known solutions in finding the rate distortion function do not work well in our situation. Interestingly, this thesis can therefore also be read as finding the rate distortion function for a specific distortion measure.

Chapter 3

Examples and their Generalisations

In this chapter we give several examples of solutions to the conditional prediction problem. We focus on the set $\mathcal{S}$, which contains all subsets that can be revealed. Interesting features of the solutions appear to depend on characteristics of $\mathcal{S}$. In the second section we generalise the features of the solutions we see in the examples. Section 3.1 can be read as an informal introduction to Section 3.2.

3.1 Examples

In this section three cases of interesting sets $\mathcal{S}$ are investigated by means of an example:

1. The case that $\mathcal{S}$ is a partition of $\mathcal{X}$. This is done in Example 6. We see that when $\mathcal{S}$ is a partition, then standard conditioning gives the best prediction.

2. The case that $\mathcal{S}$ contains subsets consisting of one element is treated in Example 7. An intuitive result is found: Nature does not choose subsets consisting of one element.

3. The case in which $\mathcal{S}$ is not a partition of $\mathcal{X}$ nor does it contain subsets consisting of one element. We investigate the simplest example of such a set $\mathcal{S}$ in Example 8. The result is surprising. The conditional predictive probability of an outcome is lower than the predictive probability for another outcome, although the unconditional probability is higher.

We use two techniques to solve the problem of finding the best conditional prediction. Sometimes we use the first formulation of our problem as given in equation (2.2), solving the minimax of the logarithmic loss. Other times we use equation (2.15) to solve our problem, maximising the conditional entropy. At the end of the section we return to the Monty Hall problem mentioned in the introduction. We restate the problem in this chapter formally and give our solution.

Example 6. [$\mathcal{S}$ **is partition**] In this example we investigate the solution if $\mathcal{S}$ is a partition of $\mathcal{X}$.
For example, let $\mathcal{X} = \{1, 2, 3, 4\}$ and $\mathcal{S} = \{\{1, 2\}, \{3, 4\}\}$ and let the marginal distribution of X be given by p_1, p_2, p_3 and p_4. The requirements on $\mathcal{P}$ imply that $\mathcal{P}$ contains only one distribution P. By the 'no lie' axiom it must hold that $P(X = 1, Y = \{3, 4\}) = 0$. The marginal distribution of P must be equal to the distribution of X, hence $P(X = 1, Y = \{1, 2\}) = p_1$.

For the other elements of $\mathcal{X}$ the same reasoning holds. So we see that we have a unique distribution P in $\mathcal{P}$. Using the notation of the first chapter, we represent this P with the table

P	1	2	3	4
$\{1,2\}$	p_1	p_2	0	0
$\{3,4\}$	0	0	p_3	p_4

The best conditional prediction Q must satisfy equation (2.2). We must solve

$$\arg\min_{Q\in\mathcal{Q}} \mathbb{E}_P\left[-\log Q(X|Y)\right]. \tag{3.1}$$

Q is characterized by two variables $a := Q(X = 1|Y = \{1,2\})$ and $b := Q(X = 3|Y = \{3,4\})$. With this notation we must solve

$$\arg\min_{0\leq a\leq 1, 0\leq b\leq 1} -p_1\log a - p_2\log(1-a) - p_3\log b - p_4\log(1-b). \tag{3.2}$$

Solving this for a and b yields $a = p_1/(p_1+p_2)$ and $b = p_3/(p_3+p_4)$. This is standard conditioning. The best conditional prediction Q is given by

Q	1	2	3	4
$\{1,2\}$	$p_1/(p_1+p_2)$	$p_2/(p_1+p_2)$	0	0
$\{3,4\}$	0	0	$p_3/(p_3+p_4)$	$p_4/(p_3+p_4)$

■

In Example 6 we studied the solution if $\mathcal{S}$ is a partition of $\mathcal{X}$. The best conditional prediction is just normal conditioning. This holds in general and is proved in Lemma 6. In the next example we investigate the case that $\mathcal{S}$ contains subsets with one element.

Example 7. [$\mathcal{S}$ **contains singletons**] Let $\mathcal{X} = \{1,2,3\}$ and $\mathcal{S} = \{\{1\},\{1,2\},\{3\}\}$ and let the marginal distribution of X be given by p_1, p_2 and p_3. So we have two subsets in $\mathcal{S}$ with one element: $\{1\}$ and $\{3\}$. Note that if the true value is 1, also the subset {1,2} can be revealed, whereas there is no choice if the true value of X is 3.
First we determine what distributions are in $\mathcal{P}$. By our two constraints on $\mathcal{P}$, the 'no lie'-axiom and the marginal consistency, we have that $P(X = 2, Y = \{1,2\}) = 1$ and $P(X = 3, Y = \{3\}) = 1$. We denote $a := P(X = 1, Y = \{1,2\})$. It must hold that $P(X = 1, Y = \{1\}) + P(X = 1, Y = \{1,2\}) = p_1$, so $P(X = 1, Y = \{1\}) = 1 - a$. Now we want to solve equation (2.15) using a:

$$\begin{aligned}
&\arg\max_{P\in\mathcal{P}} \mathbb{E}_P\left[-\log P(X|Y)\right] = \\
&\arg\max_{0\leq a\leq p_1} (1-a)\log 1 + a\log\frac{a+p_2}{a} + p_2\log\frac{a+p_2}{p_2} + p_3\log 1 = \\
&\arg\max_{0\leq a\leq p_1} a\log\frac{a+p_2}{a} + p_2\log(a+p_2).
\end{aligned}$$

The function is strictly increasing in a. The maximum is therefore obtained by $a = p_1$. So the conditional entropy maximising P is given by

P	1	2	3
$\{1\}$	0	0	0
$\{1,2\}$	p_1	p_2	0
$\{3\}$	0	0	p_3.

Now we calculate the best conditional prediction Q. By the two requirements on Q we have that $Q(X = 1|Y = \{1\}) = 1$ and $Q(X = 3|Y = \{3\}) = 1$. Furthermore the distribution given $Y = \{1, 2\}$ is determined by P: $Q(X = 1|Y = \{1, 2\}) = p_1/(p_1 + p_2)$ and $Q(X = 2|Y = \{1, 2\}) = p_2/(p_1 + p_2)$. The best conditional prediction Q is given by

Q	1	2	3
$\{1\}$	1	0	0
$\{1, 2\}$	$p_1/(p_1 + p_2)$	$p_2/(p_1 + p_2)$	0
$\{3\}$	0	0	1.

■

In Example 7 we see that it is possible to assign probability 0 to the combination $X = 1, Y = \{1\}$ for the maximum conditional entropy distribution P. This means that when the Decision Maker gives her best prediction, in the worst case if $X = 1$ the subset $\{1\}$ is never revealed. This is not the case for the combination $X = 3, Y = \{3\}$, since there is no other subset in $\mathcal{S}$ containing 3. In general it holds that subsets containing only one element are always assigned probability 0 by the maximiser of the conditional entropy, if there is another subset in $\mathcal{S}$ containing that element. This is shown in Lemma 7. In the next example we consider an example, in which $\mathcal{S}$ is not a partition of $\mathcal{X}$ and does not contain subsets with one element.

Example 8. [$\mathcal{S}$ not a partition and without singletons] Let $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{S} = \{\{1, 2\}, \{2, 3\}\}$ and the marginal distribution of X is $(p_1, p_2, p_3) = (1/6, 1/3, 1/2)$. First we look at the possible distributions in $\mathcal{P}$. By the two requirements on a distribution P in $\mathcal{P}$, the 'no lie'-axiom and the marginal consistency, we have that $P(X = 1, Y = \{1, 2\}) = 1$ and also $P(X = 3, Y = \{2, 3\}) = 1$. So we can identify a distribution P in $\mathcal{P}$ with just one number a. Denote $a := P(X = 2, Y = \{1, 2\})$, consequently $P(X = 2, Y = \{2, 3\}) = 1 - a$.
To find the distribution that maximises the conditional entropy we use again equation (2.15):

$$\arg\max_{P \in \mathcal{P}} \mathbb{E}_P \left[-\log P(X|Y) \right] =$$
$$\arg\max_{0 \leq a \leq 1/3} (1/6) \log \frac{1/6 + a}{1/6} + a \log \frac{1/6 + a}{a} + (1/3 - a) \log \frac{5/6 - a}{1/3 - a} + (1/2) \log \frac{5/6 - a}{1/2}.$$

This can be solved by taking the first derivative with respect to a, setting it equal to 0 and solving for a. We find that $a = 1/12$. So the conditional entropy maximising P is given by:

P	1	2	3
$\{1, 2\}$	1/6	1/12	0
$\{2, 3\}$	0	1/4	1/2.

We see that the best conditional prediction Q is given by:

Q	1	2	3
$\{1, 2\}$	2/3	1/3	0
$\{2, 3\}$	0	1/3	2/3.

We make two observations, which, taken together, are rather surprising. First, interestingly, we get that $Q(X = 2|Y = \{1, 2\}) = 1/3 = Q(X = 2|Y = \{2, 3\})$. The best conditional

prediction for the outcome 2 is independent of the revealed subset. Second, we note that the marginal probability of $X = 2$ is larger than the marginal probability of $X = 1$. Despite this fact, in the best conditional prediction, given you see the subset $\{1, 2\}$, you think it is more likely that the true value is 1 than that it is 2. Of course, once you realise that the outcome 2 occurs in both sets that can be revealed, it may not be so surprising any more. ■

In Example 8 we see that the best conditional prediction of the outcome that lies in two subsets is equal for both subsets. We show this in the general case (under some conditions) in Theorem 4 in Chapter 4. In the following example we show how the Monty Hall problem (see also Example 2) can be formally restated in the context of our problem.

Example 9. [Monty Hall] The set $\mathcal{X}$ is $\{1, 2, 3\}$, where the three numbers are associated with the three doors. The random variable X gives the door with the prize. The standing assumption is that the prize is equally likely to be behind any of the three doors. This means that the marginal distribution of X is $(1/3, 1/3, 1/3)$. The host opening a door is equivalent with telling us the subset consisting of the other two doors. Indeed, by the rules of the game, it is impossible that the prize is behind the opened door. So it must be behind one of the other two doors.
Without loss of generality we can assume that door 1 is the first pick of the candidate. This gives that the host will open the door numbered 2 or 3. The host opening door 2 is equivalent with telling us the subset $\{1, 3\}$ and in the same way opening door 3 is equivalent with telling us the subset $\{1, 2\}$. So $\mathcal{S} = \{\{1, 2\}, \{1, 3\}\}$. This concludes our restating of the Monty Hall problem and enables us to give a solution.
The solution we are after is the maximiser of the conditional entropy, i.e. we will solve equation (2.15). Note that the host is forced to open door 3 if the prize is behind door 2. This means that $P(X = 2, Y = \{1, 2\}) = P(X = 2) = 1/3$. The same argument gives that $P(X = 3, Y = \{1, 3\}) = P(X = 2) = 1/3$. Denote $a = P(X = 1, Y = \{1, 2\})$ and observe that a lies between 0 and $1/3$ as $a + P(X = 1, Y = \{1, 3\}) = P(X = 1) = 1/3$. It is clear that $P(\{1, 2\}) = 1/3 + a$ and $P(\{1, 3\}) = 2/3 - a$. The conditional entropy is hence only dependent on the variable a and the maximising problem is

$$\arg\max_{P \in \mathcal{P}} \mathbb{E}_P \left[-\log P(X|Y) \right] =$$
$$\arg\max_{0 \leq a \leq 1/3} (1/3) \log \frac{1/3 + a}{1/3} + a \log \frac{1/3 + a}{a} + (1/3 - a) \log \frac{2/3 - a}{1/3 - a} + (1/3) \log \frac{2/3 - a}{1/3}.$$

This can be solved by taking the first derivative with respect to a, setting equal to 0 and solving for a. We find that $a = 1/6$. This means that if you predict that $P(X = 1, Y = \{1, 2\}) = 1/6$ and also $P(X = 1, Y = \{1, 3\}) = 1/6$ the worst case is that the host indeed opens doors in this way. The $P \in \mathcal{P}$ that maximises the conditional entropy is given by:

P	1	2	3
$\{1, 2\}$	1/6	1/3	0
$\{1, 3\}$	1/6	0	1/3.

The best conditional prediction is then given by $Q(1|\{1, 2\}) = 1/3$ and $Q(1|\{1, 3\}) = 1/3$:

P	1	2	3
$\{1, 2\}$	1/3	2/3	0
$\{1, 3\}$	1/3	0	2/3.

This solution to the Monty Hall problem has some similarities to the solution proposed by Gill [2011]. His modelling of the Monty Hall problem uses also a game theoretic approach. He uses the model of a zero-sum game of the game host against the contestant. If the contestant finds the prize he receives a payoff of one, otherwise a payoff of zero. The game consists of four steps. We explain each step and directly give the optimal strategy in that step and compare that strategy with our solution.
In the first step the game host can select a distribution of the prize behind the three doors. In Gill's solution this is the uniform distribution (1/3, 1/3, 1/3). In contrast to Gill, we assumed from the start that this is the true unconditional distribution. In the second step the contestant can choose a door. The best he can do is choosing uniformly. In contrast to Gill, we did not model this step either, but remodeled the problem so that the contestant can pick door one without loss of generality. In the third step the game host has to choose which door to open. The optimal strategy dictates that, if there is a choice, the game host should open a door uniformly at random. This is exactly the strategy we found. In the final step the contestant can choose to switch. In the solution of Gill he does best to switch. Again this is implied by our solution, which gives a conditional prediction of (1/3, 2/3) with 1/3 the probability of the prize behind the original door. ■

We have seen via examples that interesting features of the solution depend on the structure of $\mathcal{S}$. In the next section we prove that these dependencies hold in general.

3.2 General Features of the Best Prediction depend on $\mathcal{S}$

In this section we prove that the assertions made in the section Examples hold in general. We focus on $\mathcal{S}$, the collection of subsets. We prove that certain features of the best prediction depend on the structure of $\mathcal{S}$. The structure of this section is as follows. First we present three lemmata:

1. Lemma 6 gives the solution when $\mathcal{S}$ forms a partition of $\mathcal{X}$. This generalises Example 6.

2. Lemma 7 gives the probability the conditional entropy maximiser assigns to subsets consisting of one element. This is a generalisation of Example 7

3. Lemma 8 gives the solution when $\mathcal{S}$ contains $\mathcal{X}$.

Then we prove Theorem 3 generalising the above lemmata. It demonstrates a feature of the conditional entropy maximiser when $\mathcal{S}$ contains sets A and B, where $A \subset B$. This theorem makes the lemmata of point 2 and 3 obsolete. We choose to give the lemmata anyway to highlight these special cases. Moreover the proofs of these lemmata make the proof of the theorem easier to grasp.

When we state our problem as generally as possible, it boils down to the question what the best prediction is of a random variable given that we only know a subset in which the variable lies. This problem has been studied centuries ago in the case of $\mathcal{S}$ forming a partition of $\mathcal{X}$. The solution is simply to use the standard conditional probabilities.

Lemma 6. *If $\mathcal{S}$ forms a partition of $\mathcal{X}$, then the solution to (2.2) is given by $Q(x|y) = p_x / \sum_{x' \in y} p_{x'}$ if $x \in y$ and 0 otherwise.*

Proof. Suppose that $\mathcal{S}$ is a partition of $\mathcal{X}$. We will proof that the above Q is the only one in $\mathcal{Q}$ generated by a $P \in \mathcal{P}$. Then it is clear that it is the solution to equation (2.15).

Let $P \in \mathcal{P}$. The 'no lie'-axiom gives that $P(x, y) = 0$ if $x \notin y$. Because $\mathcal{S}$ forms a partition, we have that for every x there is only one $y \in \mathcal{S}$ such that $x \in y$. The conditional distribution on $\mathcal{X}$ has to be equal to X and combining this yields that $P(x, y) = p_x$ if $x \in y$. This uniquely defines P and we find that $\mathcal{P} = \{P\}$. The conditional prediction Q is then $Q(x|y) = p_x / \sum_{x' \in y} p_{x'}$ if $x \in y$. □

An important point we want to stress is the following. Except in the case that $\mathcal{S}$ forms a partition, the best conditional prediction is in general not found by standard conditioning. However, standard conditioning is at least better than forgetting the received information altogether. It is not rational to predict an outcome outside the revealed set of possible outcomes.

Proposition 1. *Standard conditioning yields a lower or equal loss than forgetting the received information.*

Proof. Let Q be the conditional prediction obtained by standard conditioning, i.e. $\forall x \in y : Q(X = x|Y = y) = p_x / \sum_{x' \in y} p_{x'}$. Not using the received information yields a prediction Q' with $Q'(X = x|Y = y) = p_x$. Note that this Q' does not satisfy the definition of a conditional prediction. We have for all $P \in \mathcal{P}$ that $\mathbb{E}_P [-\log Q(X|Y)] \leq \mathbb{E}_P [-\log Q'(X|Y)]$. So standard conditioning yields a lower or equal loss than forgetting the received information. □

The following lemma proves that our observation in Example 7 holds in general. The conditional entropy maximiser assigns probability 0 to any subset consisting of only one outcome.

Lemma 7. *A subset consisting of only one element will be assigned probability 0 by the conditional entropy maximiser, if there is another subset containing this element.*

Proof. Without loss of generality assume that the element is 1. Let us look at the terms in equation (2.15) containing $a = P(1, \{1\})$. Suppose the maximum conditional entropy distribution has $a > 0$. The only term with a is $a \log \frac{a}{a} = 0$. So any probability assigned to a will not contribute to the conditional entropy. If there is another subset S containing 1, the entropy will increase if $P(1, \{1\})$ is set to 0 and $P(1, S)$ to the original value with a added.

Hence a subset consisting of only one element will be assigned probability 0, if there is another subset containing this element. □

Another case is the case where $\mathcal{S}$ contains the whole set $\mathcal{X}$. For example this is the case in a statistical setting with missing data, and missing data does not rule out any outcome. Then the conditional entropy maximiser assigns all probability to this set $\mathcal{X}$.

Lemma 8. *If $\mathcal{X} \in \mathcal{S}$, then the conditional entropy maximiser P has $P(\mathcal{X}) = 1$. Or equivalently, for all $x \in \mathcal{X} : P(x, \mathcal{X}) = p_x$.*

Proof. Note that the distribution characterized by $x \in \mathcal{X} : P(x, \mathcal{X}) = p_x$ is an element of $\mathcal{P}$. For this P we have that $H(X|Y) = H(X)$. So we find that $\max_{P^* \in \mathcal{P}} H(X|Y) \geq H(X)$.

On the other hand, per definition of this P, we have that $\max_{P^* \in \mathcal{P}} \mathbb{E}_{P^*} [-\log P(x|y)] = H(X)$. Now we find $\max_{P^* \in \mathcal{P}} H(X|Y) = \min_{Q \in \mathcal{P}} \max_{P^* \in \mathcal{P}} \mathbb{E}_{P^*} [-\log Q(x|y)] \leq H(X)$. Combining these two observations we find that P is the unique maximum conditional entropy distribution. □

The above lemma enables us to show by example that the best conditional solution is not unique.

Example 10. [Not unique] Let $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{S} = \{\{1, 2\}, \{1, 2, 3\}\}$ and the marginal distribution of X is $(p_1, p_2, p_3) = (1/3, 1/3, 1/3)$. By Lemma 8 the conditional entropy maximiser P is given by:

P	1	2	3
$\{1, 2\}$	0	0	0
$\{1, 2, 3\}$	1/3	1/3	1/3.

The best conditional prediction is then given by $Q(1|\{1, 2, 3\}) = Q(2|\{1, 2, 3\}) = 1/3$. But there is some freedom to $Q(1|\{1, 2\})$. The most intuitive choice is $Q(1|\{1, 2\}) = Q(2|\{1, 2\} = 1/2$, since 1 and 2 are interchangeable. If the Decision Maker uses this conditional prediction, Nature will play the above defined maximiser P. If the Decision Maker, however, predicts $Q(1|\{1, 2\}) = 0$, Nature does choose an other selection procedure, since it can cause an infinitely large loss. A small deviation from $Q(1|\{1, 2\}) = Q(2|\{1, 2\} = 1/2$ does not cause Nature to choose an other selection procedure. So for example $Q(1|\{1, 2\}) = 0.51$ is still a best conditional prediction. Verification of this result can easily be done by using mathematical software to find $\max_{P\in\mathcal{P}} \mathbb{E}_P [-\log Q(X|Y)]$. So in this example the best conditional solution is not unique. ■

The preceding lemmata are superseded by the following general theorem.

Theorem 3. *Suppose there are sets A, B in $\mathcal{S}$ with $A \subset B$. Then all maximum conditional entropy distributions will assign probability 0 to the set A.*

Proof. In the proof we will show that if all probability initially assigned to set A is assigned to set B, then the conditional entropy will not drop.

Let $P \in \mathcal{P}$ with $P(x, A) > 0$ for some $x \in A$. Define P^* in the following way: take P^* equal to P, but set $P^*(x, A) = 0$ and $P^*(x, B) = P(x, B) + P(x, A)$. Note that $P^* \in \mathcal{P}$. Now we need to prove that

$$\sum_{x\in\mathcal{X},y\in\mathcal{S}} P(x, y) \log \frac{P(y)}{P(x, y)} < \sum_{x\in\mathcal{X},y\in\mathcal{S}} P^*(x, y) \log \frac{P^*(y)}{P^*(x, y)}. \tag{3.3}$$

Using the definition of P^* we only need that

$$\begin{aligned} &\sum_{x\,\in A} P(x, A) \log \frac{P(A)}{P(x, A)} + \sum_{x\,\in B} P(x, B) \log \frac{P(B)}{P(x, B)} \\ &\quad < \sum_{x\in B} (P(x, A) + P(x, B)) \log \frac{P(A) + P(B)}{P(x, A) + P(x, B)}. \end{aligned} \tag{3.4}$$

The terms with x in B but not in A are clearly higher on the right-hand side. So we concentrate on $x \in A$.

Please note that

$$\frac{P(A) + P(B)}{P(x, A) + P(x, B)} = \frac{P(x, A)}{P(x, A) + P(x, B)} \cdot \frac{P(A)}{P(x, A)} + \frac{P(x, B)}{P(x, A) + P(x, B)} \cdot \frac{P(B)}{P(x, B)}. \tag{3.5}$$

This tells us that $(P(A)+P(B))/(P(x,A)+P(x,B))$ is the weighted average of $P(A)/P(x,A)$ and $P(B)/P(x,B)$. The concavity of the logarithm gives then for every $x \in A$

$$P(x,A)\log\frac{P(A)}{P(x,A)} + P(x,B)\log\frac{P(B)}{P(x,B)} \leq (P(x,A)+P(x,B))\log\frac{P(A)+P(B)}{P(x,A)+P(x,B)}. \tag{3.6}$$

We conclude that indeed equation (3.4) and hence equation (3.3) hold. The maximum conditional entropy distribution cannot assign any positive probability to the set A. □

One can interpret this theorem in the following way. The maximum conditional entropy is such that the uncertainty is highest for the Decision Maker, cf. the game theoretic interprestation in Section 2.1.3.

In this chapter we have seen several lemmata giving features of the maximum conditional entropy distribution already shown in examples. These features depend on the collection of sets $\mathcal{S}$. In the next chapter a complex characterisation of the best conditional prediction is given. That does not mean that our lemmata of this chapter can be discarded. The method in the next chapter only gives a recipe to find the best conditional prediction, but it does not highlight the features of the solution as we have done in this chapter.

Chapter 4

General Characterisation of a Solution to the Conditional Prediction Problem

In this chapter we give a general recipe to find a best conditional prediction. We need three steps to find this method. First we give a general characterisation of a solution to the prediction problem. The class of solutions for which the characterisation holds, consists of those solutions where the conditional entropy maximiser does not assign zero probability to any outcome. This characterisation is given in the first section. In the second section we conclude that there is not always such a conditional entropy maximiser. But, more importantly, we find a sufficient condition under which there exists such a maximiser. In the third section we describe the recipe to find the best conditional prediction. The trick is to extend the original structure in such a way that the sufficient condition holds. In the extended structure there is an appropriate conditional entropy maximiser, so that we can use the characterisation of the first section. Finally, the extended structure is again reduced to the original structure and we find a best conditional prediction for the original prediction problem.

4.1 Characterisation in the Non-Degenerate Case

In this first section we give a complete characterisation of the solution. However, it is only valid if the conditional entropy maximiser does not assign zero probability to an element and a set containing this element.This may be viewed as the non-degenerate case for which the solution does not lie on the boundary of all possible selection procedures. The characterisation shows some similarities with the Coarsening at Random condition, mentioned in Chapter 1. The next theorem gives the characterisation.

Theorem 4. *Suppose that the solution of equation (2.15) has* $\forall y\ \forall x \in y : P(x,y) > 0$. *Then the solution satisfies* $P(x|y) = P(x|y')$, *if* $x \in y$ *and* $x \in y'$.

Proof. Overview of the proof: We will assume an interior solution, which means $\forall y\ \forall x \in y : P(x,y) > 0$. The first order conditions, which sets the first derivative of the objective function with respect to a variable $P(x,y)$ equal to zero, then give the result.

Let us look again at equation (2.15) in the following form:

$$\arg\max_{P\in\mathcal{P}} \sum_{x\in\mathcal{X},y\in\mathcal{S}} P(x,y)\log\frac{P(y)}{P(x,y)}. \tag{4.1}$$

We use the constraints giving by the fact that $P \in \mathcal{P}$. $P(x,y) = 0$ for any pair (x,y) with $x \notin y$. So this is no longer a variable in our maximisation problem. Moreover we have that for all $x \sum_{y\in\mathcal{S}} P(x,y) = p_x$. When there are at least two sets containing x, we can eliminate one variable by writing $P(x,y') = p_x - \sum_{y\in\mathcal{S},y\neq y'} P(x,y)$. The last step before we can use the first order condition is to rewrite $P(y)$. $P(y)$ is equal to $\sum_{x\in y} P(x,y)$, keeping in mind that $P(x,y)$ is zero when x is not in y.

Now fix one $x \in \mathcal{X}$, which is contained in more than two sets. Mark y' as the special set, for which we eliminate $P(x,y')$ as above. Let y be any other set containing x. We are going to calculate the first derivative of the objective function with respect to $P(x,y)$. Let us extract the part of the objective function depending on $P(x,y)$.

$$\sum_{x'\in y} P(x',y)\log\frac{P(y)}{P(x',y)} + \sum_{x'\in y'} P(x',y')\log\frac{P(y')}{P(x',y')} \tag{4.2}$$

This holds, because $P(y)$ depends on $P(x,y)$. Also $P(y')$ depends on $P(x,y')$, which depends on $P(x,y)$.

Now we give some much-needed derivatives, where $x' \neq x$.

$$\begin{aligned}
\frac{\partial P(x,y)\log\frac{P(y)}{P(x,y)}}{\partial P(x,y)} &= \frac{P(x,y)}{P(y)} - 1 + \log\frac{P(y)}{P(x,y)} \\
\frac{\partial P(x',y)\log\frac{P(y)}{P(x',y)}}{\partial P(x,y)} &= \frac{P(x',y)}{P(y)} \\
\frac{\partial P(x,y')\log\frac{P(y')}{P(x,y')}}{\partial P(x,y)} &= -\frac{P(x,y')}{P(y')} + 1 - \log\frac{P(y')}{P(x,y')} \\
\frac{\partial P(x',y')\log\frac{P(y')}{P(x',y')}}{\partial P(x,y)} &= -\frac{P(x',y')}{P(y')}
\end{aligned}$$

Finally we are able to take the derivative of the objective function with respect to $P(x,y)$.

$$\begin{aligned}
\frac{\partial \sum_{x'\in\mathcal{X},y*\in\mathcal{S}} P(x',y^*)\log\frac{P(y^*)}{P(x',y^*)}}{\partial P(x,y)} &= \sum_{x'\in y}\frac{P(x',y)}{P(y)} - 1 + \log\frac{P(y)}{P(x,y)} \\
&\quad - \sum_{x'\in y'}\frac{P(x',y')}{P(y')} + 1 - \log\frac{P(y')}{P(x,y')} \\
&= \log\frac{P(y)}{P(x,y)} - \log\frac{P(y')}{P(x,y')}.
\end{aligned}$$

Suppose that the solution has $\forall y\ \forall x \in y : P(x,y) > 0$. This means it is an interior solution and, by the first order condition, this implies that the above is equal to zero. This gives us directly the result that $P(x|y) = \frac{P(x,y)}{P(y)} = \frac{P(x,y')}{P(y')} = P(x|y')$. □

This is an important result as it gives a full characterisation of some solutions to our problem. Remarkable is the fact that it actually assigns a single probability to each element x, independent of the set one observes, excluding the case where x is not in the observed set, when the assigned probability is zero. This is quite similar to the Coarsening at Random (CAR) condition mentioned in the introduction. The selection procedure is coarsening at random if the probabilities $P(y|x)$ and $P(y|x')$ are equal for $x, x' \in y$. As seen in Theorem 4 the conditional entropy maximising selection procedure must satisfy a *reverse* CAR condition: $P(x|y)$ and $P(x|y')$ are equal for $x \in y$ and $x \in y'$. We provide more comparison with CAR in Chapter 5.

Another appealing feature of this type of solution is that it is a so-called *equaliser* strategy. For a detailed discussion of equaliser strategies see Ferguson [1967]. An equaliser strategy P has the property that $\mathbb{E}_Q[-\log P(x|y)]$ is equal for all $Q \in \mathcal{P}$. This means that the expected loss is independent of the procedure (or the way Nature plays this game).

It is clear that to make the result interesting, we need to find out in what situations the precondition holds. When are we in the non-degenerate case? We now investigate that question.

4.2 Existence of Degenerate Case

In this section we prove an important lemma, which gives a condition that assures the maximum conditional entropy distribution assigns all feasible combinations a positive probability mass. The lemma tells us that if a set is assigned positive probability mass, then all combinations of this set and an element in the set must have positive probability mass. In the second part of the chapter we show by example that there are situations in which the conditional maximiser assigns zero probability.

The next lemma seems innocent, but is really important.

Lemma 9. *Let $S \in \mathcal{S}$ be a set containing more than one element and let $x, x^* \in S$. If the maximum conditional entropy distribution has $P(x, S) > 0$, then also $P(x^*, S) > 0$.*

Proof. We give a proof by contradiction. Suppose that we have a maximum conditional entropy distribution P with $P(x, S) > 0$, but $P(x^*, S) = 0$. Let us look at the problem in the original form using equation (1). There exists a $R \in \mathcal{P}$ with $R(x^*, S) > 0$ (An example is the distribution used in the proof of Lemma 2.2). For this P and R we have that $\mathbb{E}_R[-\log P(x'|y)] = \infty$, as $-\log P(x^*|S) = \infty$ and this has positive probability under R. Invoking Lemma 1 yields the contradiction: P cannot be the solution. □

Lemma 9 has a major implication in the case that each set in $\mathcal{S}$ contains a unique element, i.e. for each set $S \in \mathcal{S}$ there exists an $x_s \in \mathcal{S}$ such that x_s is not contained in any other set. In that case the lemma assures us that the maximum conditional entropy distribution assigns all feasible combinations a positive probability mass.

The strength of Theorem 4 is discerned when the theorem is combined with Lemma 9. Suppose $\mathcal{S}$ is such that in every set $y \in \mathcal{S}$ there are one or more elements which are not in

any other set, i.e. $y \nsubseteq \bigcup_{y' \neq y \in \mathcal{S}} y'$. By Lemma 9 we have for all y, $x \in y$ $P(x, y) > 0$ for P the maximum conditional entropy distribution. So the condition in Theorem 4 is met, and the maximum conditional entropy distribution has to satisfy $P(x|y) = P(x|y')$, if $x \in y$ and $x \in y'$. In the next section we use this principle to provide a way of obtaining a best solution for the prediction problem.

To show that the precondition of Lemma 9 is sometimes really needed for a non-degenerate solution, we construct a nontrivial example where the result of Theorem 4, which is $P(x|y) = P(x|y')$ if $x \in y$ and $x \in y'$, cannot hold. Theorem 4 then implies that one or more subsets will be assigned probability 0, i.e. the solution is not in the interior but on the boundary:

Example 11. [No reverse CAR condition] Let $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{S} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$ and $p_1 = 1/2$, $p_2 = 1/3$, $p_3 = 1/6$. The implied solution given by Theorem 4 requires $P(x|S) = 1/2$ for each $x \in S$. This is easily seen by noting that $P(1|\{1, 2\}) = P(1|\{1, 3\}) = 1-P(3|\{1, 3\}) = 1-P(3|\{2, 3\}) = P(2|\{2, 3\}) = P(2|\{1, 2\})$, where we used $P(x|y) = P(x|y')$ in the first, middle and last equation.

Now this implies that $P(1, \{1, 2\}) = P(2, \{1, 2\}), 1/2 - P(1, \{1, 2\}) = P(3, \{1, 3\})$ and $1/3 - P(2, \{1, 2\}) = 1/6 - P(3, \{1, 3\})$. We use the first and third equation to find that $1/3 - P(1, \{1, 2\}) = 1/6 - P(3, \{1, 3\})$ and this yields that $P(3, \{1, 3\}) = P(1, \{1, 2\}) - 1/6$. We find that $P(1, \{1, 2\}) = 1/3, P(2, \{1, 2\}) = 1/3, P(3, \{1, 3\}) = 1/6$. This is however no interior solution as $P(2, \{2, 3\}) = 0$.

Thus the solution has to assign probability 0 to one of the sets (See Lemma 9). There are three candidates which, after removing a subset, can actually be calculated using Theorem 4:

P^1	1	2	3
$\{1, 2\}$	1/3	1/3	0
$\{1, 3\}$	1/6	0	1/6
$\{2, 3\}$	0	0	0

P^2	1	2	3
$\{1, 2\}$	1/2	1/4	0
$\{1, 3\}$	0	0	0
$\{2, 3\}$	0	1/12	1/6

P^3	1	2	3
$\{1, 2\}$	0	0	0
$\{1, 3\}$	1/2	0	1/10
$\{2, 3\}$	0	1/3	1/15

.

One can calculate the conditional entropy of the three given distributions and find that P^1 has the maximum conditional entropy. In this case it is actually quite intuitive why P^1 should have the maximum conditional entropy. It seems best for Nature to split the probability of the element x with the highest p_x, which is in this case 1, in order to keep the Decision Maker in the dark.

Let us calculate what the best conditional prediction is generated by this P^1. We find directly that $Q(1|\{1, 2\}) = Q(2|\{1, 2\}) = Q(1|\{1, 3\}) = Q(3|\{1, 3\}) = 1/2$. For Q to be the best conditional prediction one needs that $Q(2|\{2, 3\}) \geq 1/2$ and also $Q(3|\{2, 3\}) \geq 1/2$. Hence there is only one best conditional prediction, given by $Q(2|\{2, 3\}) = Q(3|\{2, 3\}) = 1/2$.

■

In this section we have seen by example that the characterisation in Section 4.1 is not valid in all cases. However, Lemma 9 shows in which case the characterisation in Section 4.1 does hold. The recipe we describe in Section 4.3 modifies the problem so that we can use Lemma 9.

4.3 Recipe to find the Best Conditional Prediction

The idea of finding the best solution is to extend the structure of $\mathcal{X}$ and $\mathcal{S}$ to sets we call $\mathcal{X}'$ and $\mathcal{S}'$, defined below. For each set y in $\mathcal{S}$ we create a unique new element z_y. We define $\mathcal{X}' = \mathcal{X} \cup \{z_y : y \in \mathcal{S}\}$ and $\mathcal{S}' = \{y' : y' = y \cup \{z_y\} \text{ for all } y \in \mathcal{S}\}$. Thus, for the extended structure the new element is added to the original set, i.e. $y' := y \cup \{z_y\}$ is the new set. All these new elements are given a marginal probability mass of ϵ. Let $\#\mathcal{S}$ be the number of sets in $\mathcal{S}$. The probability of all the original elements are proportionally reduced, i.e. $p'_x = p_x\,(1 - \epsilon \cdot \#\mathcal{S})$. For each ϵ we find a conditional entropy maximiser via Theorem 4. Note that the precondition of Theorem 4 holds, because of Lemma 9 in the previous section. Each set contains a unique element, hence we are in the non-degenerate case. These conditional predictions go to a best prediction in the original problem when ϵ goes to 0.

We prove this idea rigorously and for this we need more notation. Let $\mathcal{P}_\epsilon$ be our set of probability distributions satisfying the 'no lie'-axiom and that are marginal consistent with the new probability distribution given by p'_x for an original element x and ϵ for a new element. Let P^ϵ be the conditional entropy maximiser, i.e. $\arg\max_{P \in \mathcal{P}_\epsilon} \mathbb{E}_P\left[-\log P(X|Y)\right]$. We now define $Q^\epsilon(x \mid y) := P^\epsilon(x, y')/P^\epsilon(y)$ for all original elements x and original sets y. Note that Q^ϵ is a conditional prediction on the original structure, so $Q^\epsilon \in \mathcal{Q}$. We would like to take the limit of these conditional predictions as $\epsilon \to 0$, however the limit does not always exist, since there could be several best conditional predictions (see Example 5). Therefore the following trick is needed. Note that (P^ϵ, Q^ϵ) can be viewed as a vector in $\mathbb{R}^m$, as explained in the paragraph before Lemma 5. This vector takes values in a bounded, closed subset, hence there exists a sequence $\epsilon_1, \epsilon_2, ...$ such that $\epsilon_n \to 0$ and $(P^{\epsilon_n}, Q^{\epsilon_n}) \to (P^*, Q^*)$. Here P^* and Q^* are some accumulation points. Their existence is an application of Bolzano-Weierstrass. With all these definitions we can now prove that such Q^* are a solution to the prediction problem in the original problem.

Theorem 5. *Every conditional prediction Q^*, defined as above, is a best conditional prediction in the original problem.*

Proof. First note that $P^* \in \mathcal{P}$ and $Q^* \in \mathcal{Q}$. The structure of the proof is as follows: first we make two observations on the conditional entropy of the probability distributions P^{ϵ_n}; second we use these two observations to find that

$$\max_{P \in \mathcal{P}} \mathbb{E}_P\left[-\log Q^*(X|Y)\right] \leq \min_{Q \in \mathcal{Q}} \max_{P \in \mathcal{P}} \mathbb{E}_P\left[-\log Q(X|Y)\right]. \tag{4.3}$$

From this we conclude that Q^* is a best conditional prediction.

The first observation is that the conditional entropy is continuous, so

$$\max_{P \in \mathcal{P}_{\epsilon_n}} \mathbb{E}_P\left[-\log P(X|Y)\right] \to \mathbb{E}_{P^*}\left[-\log P^*(X|Y)\right] \text{ as } n \to \infty. \tag{4.4}$$

The second observation uses the indicator function $\mathbf{1}_{x \text{ original}}$, which is 1 if x is an original element and 0 if x is added, i.e. $x \in \mathcal{X}' \backslash \mathcal{X}$. We prove via a series of inequalities a lower bound of the maximum conditional entropy for each ϵ:

$$\max_{P \in \mathcal{P}_\epsilon} \mathbb{E}_P \left[-\log P(X|Y) \right] = \max_{P \in \mathcal{P}_\epsilon} \mathbb{E}_P \left[-\log P(X|Y) \mathbf{1}_{x \text{ original}} \right] + \sum_{y' \in \mathcal{S}} \epsilon \left(-\log \frac{\epsilon}{P(y')} \right) \tag{4.5}$$

$$\geq \max_{P \in \mathcal{P}_\epsilon} \mathbb{E}_P \left[-\log P(X|Y) \mathbf{1}_{x \text{ original}} \right] \tag{4.6}$$

$$\geq \max_{P \in \mathcal{P}_\epsilon} \mathbb{E}_P \left[-\log Q^\epsilon(X|Y) \mathbf{1}_{x \text{ original}} \right] \tag{4.7}$$

$$\to \max_{P \in \mathcal{P}} \mathbb{E}_P \left[-\log Q^*(X|Y) \right], \text{ as } \epsilon \to 0. \tag{4.8}$$

Equation (4.5) holds, since for any $x \in \mathcal{X}' \backslash \mathcal{X}$ we have $P(X = x, Y = y') = \epsilon$. Equation (4.7) holds, because $Q^\epsilon(x|y) > P^\epsilon(x|y)$ by definition. Equation (4.8) holds, since we have the indicator function in the expectation, which reduces it to the original problem.

Now we are ready to prove equation (4.4):

$$\begin{aligned}
\max_{P \in \mathcal{P}} \mathbb{E}_P \left[-\log Q^*(X|Y) \right] &\leq \lim_{n \to \infty} \max_{P \in \mathcal{P}_\epsilon} \mathbb{E}_P \left[-\log P(X|Y) \right] && (4.9) \\
&= \mathbb{E}_{P^*} \left[-\log P^*(X|Y) \right] && (4.10) \\
&\leq \max_{P \in \mathcal{P}} \mathbb{E}_P \left[-\log P(X|Y) \right] && (4.11) \\
&= \min_{Q \in \mathcal{Q}} \max_{P \in \mathcal{P}} \mathbb{E}_P \left[-\log Q(X|Y) \right]. && (4.12)
\end{aligned}$$

Equation (4.9) holds by the second observation and Equation (4.10) by the first observation. Because $P^* \in \mathcal{P}$, Equation (4.11) holds. The final Equation (4.12) holds by the minimax theorem (Theorem 2 in Chapter 2).

We conclude that Q^* is a best conditional prediction. □

In this chapter we showed a recipe to find the best conditional prediction in the general case. The first step was a characterisation of the best conditional prediction in the case that the solution is not on the boundary. This characterisation is reminiscent of the Coarsening at Random condition. This is discussed further in the next chapter. For each outcome $x \in \mathcal{X}$ there must be a single probability that is the conditional prediction for every possible subset. Moreover we found that this solution is an equaliser strategy. A problem is that this characterisation can be invalid if the solution is on the boundary. However, we discovered a method to use the characterisation of the first step in the general case. In the second step one must extend the problem with a unique new outcome per subset and give each new outcome a small probability. In the extended problem the characterisation can be used and gives us a best conditional prediction. Taking the small probabilities all to zero gives actually the best conditional prediction for the original conditional prediction problem.

Chapter 5

Conclusion - Discussion - Future Work

5.1 Conclusion

The goal of this thesis was to find the best prediction of a random variable conditional on a revealed subset containing the true value of the random variable. We first cast this problem in a mathematical framework. We chose to work with a minimax approach with the expected log loss function. The best conditional prediction minimises the maximum loss among all possible predictions. The conditional prediction problem can be viewed as a game between the Decision Maker, i.e. the statistician, and Nature. First the Decision Maker gives her conditional prediction, then Nature can choose a selection procedure to maximise the expected log loss.

In Chapter 2 we showed that the minimax theorem also holds in our case, where the loss can become infinitely large. Therefore we can restrict the problem to maximising the conditional entropy. If the conditional entropy maximiser P does assigns positive probability to each feasible combination of outcome and subset, then the best conditional prediction is just the conditional distribution of P. This is a major step to reduce the complexity of finding the best conditional prediction.

In Chapter 3 we discovered the best conditional prediction in certain special cases and we found properties of the conditional entropy maximiser. As an example of the former, the best conditional prediction is easily found if the set $\mathcal{S}$ is a partition: one only needs to use ordinary conditioning. As an example of the latter, the maximiser does not assign positive probability to a set $y \in \mathcal{S}$, if there is another set y' such that $y \subset y' \in \mathcal{S}$.

We discovered a recipe to find the best conditional prediction in the general case in Chapter 5. First we found a characterisation for the best conditional prediction if the conditional entropy maximiser does not lie on the boundary. The solution is such that the conditional probability of an outcome should not depend on the observed set. This characterisation reminds us of the Coarsening at Random condition, where the conditional probability of the set should not depend on the outcome.
The most important lemma, also features a property of the conditional entropy maximiser. The main implication of this lemma concerns the case that each set $y \in \mathcal{S}$ contains a unique element, i.e. an element that is not part of any other set in $\mathcal{S}$. In that case the conditional

entropy maximiser does assign positive probability to each feasible combination of outcome and subset.

A combination of the lemma and the characterisation gives a method to find the best conditional prediction in the general case. First one extends the structure of the original problem with a unique element for each set $y \in \mathcal{S}$ and assigns each a small probability. For this case the above characterisation gives the solution. Upon taking the small probability to zero, one obtains the best conditional prediction of the original problem.

5.2 Discussion and Future Work regarding the loss function

An area of possible further research is the question whether the results in this thesis depend crucially on the chosen loss function. Another possible loss function is the 0/1 function. In this case, one has to predict exactly one element for each set. The loss is 0, if indeed this element was the true value of the random variable. The loss is 1, if the prediction was wrong. The notation δ is used for the point prediction and it plays a role similar to the probabilistic prediction Q. The expected loss in this case is given by $\sum_{x,y} P(X = x, Y = y)(1 - \delta(X = x|Y = y))$. But now $\delta(X = x|Y = y) \in \{0, 1\}$. Maximising this for a fixed δ is equal to minimising $\sum_{x,y} P(X = x, Y = y)\delta(X = x|Y = y)$. Intuitively, it seems true that the minimising P assigns as much probability as possible to sets with $\delta(X = x|Y = y) = 0$. Therefore the best conditional prediction should try to set $\delta(X = x|Y = y) = 1$ for the elements that must be assigned the most probability. This leads to the following conjecture.

Conjecture 1. *We hypothesize that the best conditional prediction under the 0/1 loss function is based on the best conditional prediction under the log loss function in the following way: For each $y \in \mathcal{S}$ we predict the element that has the highest conditional predictive probability under the log loss function.*

This hypothesis is supported by the fact that it works on all examples considered in this thesis.

A second possibility is to look at expected variational distance. The expected loss in this case is given by $\sum_{x,y} P(X = x, Y = y) \mid P(X = x|Y = y) - Q(X = x|Y = y) \mid$. It is not directly clear what the maximising P is for a fixed conditional prediction. A small example shows that it is not equal to the best prediction using log loss found in this thesis. We repeat the setting of Example 8:

Example 12. [Expected variational distance] Let $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{S} = \{\{1, 2\}, \{2, 3\}\}$ and the marginal distribution of X is $(p_1, p_2, p_3) = (1/6, 1/3, 1/2)$. As shown in Example 8, the best conditional prediction is given by:

Q	1	2	3
$\{1, 2\}$	2/3	1/3	0
$\{2, 3\}$	0	1/3	2/3.

The maximising probability P under the expected variational distance is given by $P(X = 2, Y = \{1, 2\}) = 0$. This can be easily computed using mathematical software. With the maximising probability an expected loss of 1/3 is found.

Now another conditional prediction in this example is given by:

Q	1	2	3
$\{1,2\}$	5/6	1/6	0
$\{2,3\}$	0	1/2	1/2.

For this conditional prediction we find that the maximising probability P under the expected variational distance is given by $P(X = 2, Y = \{1,2\}) = 1/3$, yielding an expected loss of 5/18. This expected loss is lower than 1/3, hence the best conditional prediction under relative entropy, is not best using the expected variational distance. ■

In Section 2.4 our general model was discussed. There it is mentioned that one should view Q as a prediction and that it has no direct frequentist interpretation. Therefore a distance measure such as the expected variational distance is perhaps not such a good tool to evaluate your prediction. At least we know that it does not result in the same best conditional prediction. This shows that the results of this thesis do not hold directly in such a different setting.

5.3 Discussion and Future Work on the Comparison with CAR

In Section 4.1 we found a characterisation of the best conditional prediction in certain cases. The best prediction satisfies $P(x|y) = P(x|y')$, if $x \in y$ and $x \in y'$. As mentioned in Chapter 4 this reminds us of the CAR condition. The selection procedure is Coarsening at Random if the probabilities $P(y|x)$ and $P(y|x')$ are equal for $x, x' \in y$. Because of this similarity we call our characterisation reverse CAR. The CAR condition has been studied extensively [Heitjan and Rubin, 1991, Gill, van der Laan, and Robins, 1997, Gill and Grünwald, 2008].

The paper by Grünwald and Halpern [2003] contains several results we can use. They study CAR via a 0/1 matrix, which they call the CARacterizing matrix. The definition of this matrix M is as follows. Each row is identified with an outcome $x \in \mathcal{X}$. Each column is identified with a set $y \in \mathcal{S}$. The entries of the matrix are 1 if the corresponding outcome is part of the corresponding set and zero otherwise.

Example 13. [CARacterising matrix] The CARacterising matrix in the Monty Hall problem (see Example 2) is given by

M	$\{1,2\}$	$\{1,3\}$
1	1	1
2	1	0
3	0	1

.

■

Grünwald and Halpern use this matrix in two ways. First, a condition is given to check whether CAR may hold. Second, there are easy to check conditions on this matrix, which guarantee that CAR can *not* hold. The latter one we can put to use in our setting of conditional prediction. We are interested if the reverse CAR condition can hold. Note that the CAR and reverse CAR condition switch the role of outcome and subset. That means that we should look at the transpose of the CARacterising matrix. If it is impossible that CAR

can hold for this transposed matrix, that implies that it is impossible that the reverse CAR formula may hold in the original problem.

Further research could focus on the first way Grünwald and Halpern use this matrix. If CAR can hold for the transpose of the CARacterising matrix, does that imply that the inverse CAR condition holds in the original problem?

We state one more open question for future research: In Theorem 4 in Section 4.1 we show that if the best conditional prediction does not lie on the boundary, then it satisfies the reverse CAR condition. An important question then is: If there exists a $P \in \mathcal{P}$ satisfying the reverse CAR condition and also lying in the interior, does that imply that the best prediction also lies in the interior? If this question could be answered affirmative, then the strength of Theorem 4 is increased significantly.

Bibliography

T. Cover and J. Thomas. Elements of information theory. 1991. [7, 13, 14, 15]

K. Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39, 1953. [12, 13]

T. Ferguson. Mathematical statistics: a decision theoretic approach. 1967. [26]

R. Gill. The Monty Hall problem is not a probability puzzle (It's a challenge in mathematical modelling). *Statistica Neerlandica*, 65, 2011. [3, 20]

R. D. Gill and P. D. Grünwald. An Algorithmic And A Geometric Characterization Of Coarsening At Random. *Annals of Statistics*, 36, 2008. [32]

R.D. Gill, M.J. van der Laan, and J.M. Robins. *Coarsening at random: Characterizations, conjectures, counter-examples.* Springer, 1997. [32]

P. Grünwald and A. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Mathematical Statistics*, 32, 2004. [12, 15]

P. Grünwald and J. Halpern. Updating Probabilities. *Journal of Artificial Intelligence Research*, 19, 2003. [3, 32]

P. Harremoës and F. Topsøe. Unified approach to optimisation techniques in Shannon theory. 2002. [15]

D. Heitjan and D. Rubin. Ignorability and coarse data. *Annals of Statistics*, 19, 1991. [2, 32]

E. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106, 1957. [14]

J. Robertson, E. Tallman, and C. Whiteman. Forecasting Using Relative Entropy. *Journal of Money, Credit and Banking*, 37, 2005. [3]

J. Von Neumann, O. Morgenstern, A. Rubinstein, and H.W. Kuhn. *Theory of Games and Economic Behavior.* Princeton University Press, 1944. [3]

M. vos Savant. Ask Marilyn. *Parade Magazine*, 15, 1990. [3]