Discovering the truth by conducting experiments

MSc Thesis (*Afstudeerscriptie*)

written by

Wouter Michiel Koolen
(born July 2nd, 1982 in Groningen, The Netherlands)

under the supervision of **Dr. Peter Grünwald**, and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:
7th December, 2006

Members of the Thesis Committee:
Prof. Dr. Johan van Benthem
Lector Dr. Peter van Emde Boas
Dr. Nikos Vlassis
Dr. Peter Grünwald

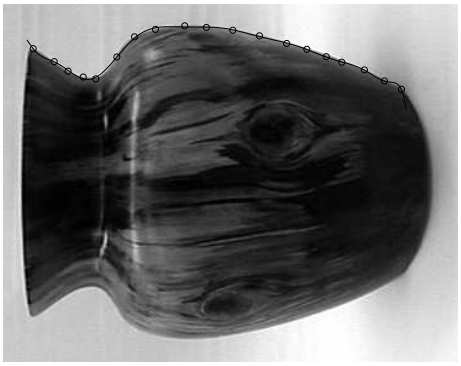
Prolusion

Paul Vitányi's 2003 Kolmogorov complexity lecture included a computer exercise in which a polynomial relation had to be learnt from samples.[1] The following data were provided: a sequence of pairs of numbers $(h_1, d_1), (h_2, d_2), \ldots, (h_n, d_n)$, supposedly noisy measurements of a classical urn, h_i being the height from the floor and d_i being the diameter of the urn at the height h_i. The goal was to infer a polynomial that represented the relation between height and diameter. For a given degree, this can easily be done using linear algebra. The crux of the exercise was finding the best degree. An example is shown in Figure 1.

To me, learning from given data is only part of a more general concept of learning, and I started to wonder whether the techniques that I learnt during my studies could be adapted to an interactive setting, allowing the learner to perform experiments. For example, when learning polynomials, the learner could be allowed to choose a point, and she would then receive the value of the polynomial at that point.

For this thesis, I started working on the interactive polynomial learning problem, but it turned out to be much too hard. I then devised the balance scale problem (see Chapter 4), a toy problem that conserves the important features of the polynomial learning problem: it is interactive, probabilistic, model-based, but finite. I had by then developed a slight aversion to subjective Bayesian methods, for my initial work on the polynomial learning problem suggested that they are not robust. It seemed that a subjective Bayesian learner can be tricked into assigning high posterior probability to a certain proposition while

Figure 1 Urn example



[1] The exercise by Troy Lee is still available online at `http://www.lri.fr/~lee/final.html`.

this proposition is false, and additionally, great confidence in this proposition leads to great confidence in the usefulness of experiments that in fact do not help to determine that this proposition is false.

With this in mind, I decided to perform a worst-case analysis of the balance scale problem, and of similar problems in general. This problem naturally decomposed into the truth-finding problem, where we want to find the true model from given data, and the experiment-design problem, where experiments have to be selected, whose outcomes subsequently serve as the data for truth finding.

I have yet to solve the balance scale problem completely. But I have already learned and discovered much more than I could initially imagine.

I hope that this thesis will provide inspiration to others.

Wouter Koolen-Wijkstra

Amsterdam
23$^{\text{rd}}$ November, 2006

Contents

List of Figures

List of Tables

List of Protocols

Chapter 1

Introduction

Science progresses by the performing of experiments to evaluate hypotheses.

1.1 Problem statement

This thesis is motivated by two important and interesting questions:

Question 1. How can data be used to learn about reality?

Question 2. How can experiments be used to accelerate learning?

These questions are important, because their answers should provide a solid foundation for rational, e.g. scientific, learning. The answers will most likely also provide new insights into natural, e.g. human, learning. These questions are interesting, because they have not yet been answered satisfactorily, in spite of their importance. Existing approaches require prior knowledge of a specific kind, focus on the learning of predictors, and are mostly analysed in non-sequential settings.

In this thesis, we formalise the first question in the framework of decision theory as the Nature versus Learner *truth-finding* game. This game focuses on finding the true model instead of on prediction. We perform a worst-case analysis of this game, requiring no prior knowledge. We provide the solution to the truth-finding game in the form of optimal strategies for both players, and interpret the value of the game as a measure of certainty.

We formalise the second question as the Nature versus Experimenter *experimentation game*: a straightforward generalisation of the truth-finding game that includes sequential experimentation. We also solve this game, and compare our solution to that of Bayesian experiment design.

Overview This introduction is structured as follows. We describe the components of the learning setting in §1.2, and introduce our two running examples. In §2.6.1 we explain how information can be quantified. This allows us to measure how much has been learned. Then in §1.3, we put forward the truth-finding problem: its interpretation as a game, and the first detailed example. In §1.4 we extend the truth-finding game with experiments, and give the second detailed

example. We summarise the contribution of this thesis in §1.5. We conclude with a list of related work, and words of gratitude.

1.2 Basic terminology

Experiment An experiment is a two-stage act. First, one influences the state of the world in a controlled way, for example by injecting a mouse with a certain dose of elixir. The action that is undertaken in this first stage is called the *input* of the experiment. Second, one observes the value of a predetermined quantity, for example the lifespan of said mouse measured in days. The value obtained in this second stage is called the *outcome* of the experiment. Note that in *controlled experiments*, for example clinical trials, one separately observes the outcome that occurs when no influence is exerted. In this thesis, we do not assume that there is such a special *null input*, but always compare the outcomes of several experiments. A series of experiments yields *data*, the concatenation of successive input/outcome pairs.

Hypothesis A hypothesis provides an explanation for a phenomenon; it relates that what one can *influence* to that what one can *observe*. A *deterministic* hypothesis predicts a single outcome for each possible input. A *probabilistic* hypothesis assigns a likelihood to each outcome for each possible input. Deterministic hypotheses are mathematically represented as functions; probabilistic hypotheses are represented as (conditional) probability distributions. The latter, more general type, will be used in this thesis, as we are interested in modelling phenomena that involve chance. Hypotheses of both types abound in science. We can regard science as the prime application of both learning from data and experimental learning.

In practice, hypotheses are used for different purposes:

- to describe regularities in past experiments,
- to predict the outcome of future experiments, and
- to explain, i.e. formally specify, a data-generating process.

In this thesis, we focus on the third interpretation.

Evaluation Experiments provide the empirical basis for the evaluation of hypotheses. A deterministic hypothesis can (theoretically) be disqualified on the basis of a single contradictory outcome. It is generally impossible to reject probabilistic hypotheses, but they lose credibility when they predict poorly, i.e., they assign low probability to subsequently observed outcomes of experiments. In the absence of prior knowledge of the generating process, there is no absolute quantitative scale to judge prediction quality, hence we can only evaluate prediction quality in comparison to the prediction quality of other hypotheses.

Model A collection of similar probabilistic hypotheses is called a *model*. A model has no direct relation to the phenomena that its member hypotheses explain; it is a cognitive device, created by the learner to structure the learning problem at hand. Models as such are not hypotheses in our strict sense; they

provide no means to combine the specific predictions made by the member hypotheses into a single prediction. There are ways to achieve such prediction though, for instance by using a *Bayesian universal model*. This is a model endowed with weights for each member hypothesis. A Bayesian universal model can be translated into a hypothesis, by taking the weighted average of the member hypotheses.

A collection of models is called *competing* if no pair of models shares a common hypothesis. As an example, consider the following scenarios, which are summarised in Table 1.1.

Scenario 1 (Biased coin)**.** A coiner has just minted a prototype coin showing the new queen's face. The new queen is rather gourmand, so he wonders whether the coin is loaded. To find out, he can perform experiments by flipping the coin, and observing the side that turns up, either heads or tails. Hypotheses are of the form: *the probability that heads turn up is* θ, where θ is a number between 0 and 1. Models are, for example, *the coin is fair* (ex. model 1), which is a singleton set, or *the coin favours heads* (ex. model 2), which is an uncountable set of hypotheses.

Scenario 2 (Anvil drop)**.** Galileo, author of the definitive guide to Earth's gravity, established that sufficiently heavy objects, when dropped from equal height, hit the ground simultaneously. Now he wants to find out how falling time relates to height. To this end, he has brought an anvil and a stopwatch to the tower of Pisa. He performs an experiment by climbing up to some floor at height h, dropping the anvil, and measuring the amount of time t it takes the anvil to hit the pavement. His hypotheses relate t to h, for example via the equation $h \approx 4t^2$. (We write $\approx$ to signify that we use a fixed zero-mean noise model. In this case, $t = \sqrt{h/4} + Z$, where Z is a normally distributed random variable with mean zero. The outcome of an experiment, the measured time t, is noisy; it cannot be obtained exactly due to limited reaction speed and precision. We approximate the combined influence of such small errors with a normal distribution. Thus, again we use probabilistic hypotheses.)

Models are, for example, *linear gravity* (ex. model 1) or *quadratic gravity* (ex. model 2).

Reality, worlds In this thesis, we use the word *reality* in a technical sense. It designates a data-generating process, which we cannot identify, but on which we can run experiments. We regard the procedural details of the execution of the experiment as part of reality itself. For example, in Scenario 2, reality is the process that generates a stopwatch reading when provided with a height.

Table 1.1 Example scenarios with hypotheses and models.

Scenario	Biased coin	Anvil drop
Input	none	height h
Outcome	H(eads) or T(ails)	time t with Gaussian noise
Ex. hypothesis	$P(\mathrm{H}) = 0.6$	$h \approx 4t^2$
Ex. model 1	$\{P(\mathrm{H}) = 0.5\}$	$\{h \approx \alpha t \mid \alpha \in [1, 10]\}$
Ex. model 2	$\{P(\mathrm{H}) = \theta \mid \theta \in (0.5, 1]\}$	$\{h \approx \alpha t^2 \mid \alpha \in [1, 10]\}$

Reality is not necessarily a probability distribution, although we will require this assumption later when we turn to truth finding.

Reality, which we will also call *the state of nature*, is unknown to us. We can, however, consider a collection of candidate explanations of reality: alternative states of nature that we cannot yet distinguish from the actual state. We adopt the nomenclature of modal reasoning, and call such alternative states *possible worlds*. We refer to the current state, reality, as the *actual world*. Again, in truth finding, we assume that the actual world is a possible world.

Overfitting Intuitively, a hypothesis *over-fits* if it describes past outcomes well, but predicts future outcomes poorly. Such a hypothesis is too specific; it describes the noise instead of the regularity in the past outcomes, hence it misses the general pattern. See Figure 1.1 for a graphical example. Ten data-points were generated from the true dotted curve, with Gaussian noise added. The best fitting polynomials of degree 1, 3 and 9 are shown. The under-fitting curve (a) describes the sampled points poorly, and explains the true curve poorly. The over-fitting curve (c) describes the sampled points *perfectly*, but explains the true curve *extremely* badly. Curve (b) strikes a good balance between descriptive and explanatory quality.

The statistical literature, e.g. [Mit97], is full of theorems showing that to get good predictive quality, you need to take the hypothesis that optimises some trade-off between complexity and goodness of fit.

1.3 Selection tasks

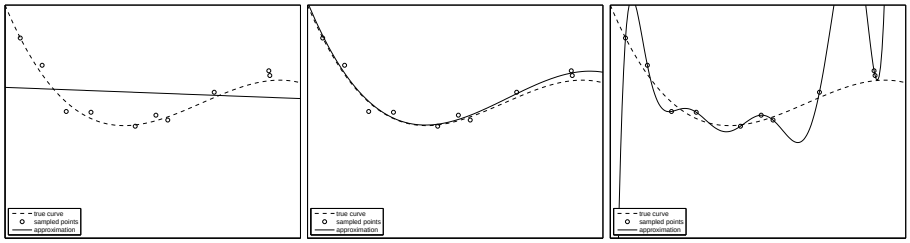
Learning amounts to finding regularity in data. An elegant formalisation of this idea is given by the theory of *Kolmogorov complexity*, where *all* computable regularities are considered. See [LV93] for an introduction to the field. The Kolmogorov complexity itself is not computable.

To obtain a computable notion of regularity, one must restrict the class of regularities under consideration. Such restrictions lie at the heart of minimum description length (MDL) methods, see [GMP05]. As in MDL, we use models to explicitly state which regularities are considered.

The following question is a more precise formulation of Question 1:

Figure 1.1 Example of fit

(a) Degree 1: under-fit **(b)** Degree 3: good fit **(c)** Degree 9: over-fit



Question 3. Given a sequence of data obtained from reality and a list of regularities, how can we infer the best explanation for reality on the list, i.e. learn something about it?

We list three answers to Question 3, that differ in the precise interpretation of "best explanation": hypothesis selection, model selection, and truth finding. But first we make the following observation about selection methods on probabilistic objects in general.

Probabilistic humility The general principle

$$\text{probabilities in} \Rightarrow \text{probabilities out} \tag{PIPO}$$

states that, once a problem has been formalised in probability theory, then reasoning within probability theory can only yield the probability of new events. The goal of model selection is to find the best model. By PIPO, probability theory can only give us a probability distribution on the candidate models. We need an additional criterion, outside of probability theory, to judge the quality of such distributions. The same holds for hypothesis selection and truth finding.

Hypothesis selection The hypothesis-selection problem is stated as follows: *given a collection of hypotheses and data obtained from reality, find the hypothesis that explains reality best.* Of course, one cannot evaluate directly how well a hypothesis explains reality, one can only evaluate how well it describes the data. To guard against over-fitting, one must select a hypothesis that strikes a balance between expressiveness and simplicity. This can be achieved by (1) adopting a measure of complexity for hypotheses, and (2) penalising hypotheses by their complexity. For example, in Figure 1.1, we penalise polynomials according to their order. This ensures that we prefer (b) over (c).

The hypothesis-selection problem between two hypotheses is called *hypothesis testing*. To obtain a true selection, beyond PIPO, one uses a significance level as a selection threshold.

Model selection The model-selection problem is the following: *given a collection of models and data obtained from reality, find the model that explains reality best.* This problem is often solved by reducing it to the hypothesis-selection problem using universal codes. A universal code for a model corresponds to a *single* hypothesis, composed of a weighted average over the partaking hypotheses. Two well-known approaches are Bayesian and MDL model selection, see [GMP05].

Truth finding The truth-finding problem is the following: *given a collection of competing models and data obtained from reality, find the true model.* We regard the hypotheses in the models as possible worlds, and assume that one of them is reality. The model that contains reality is called the *true model*, and we want to obtain as much information about its identity as possible. Selecting a single model with certainty is generally impossible, because we are working with probabilistic hypotheses. We allow a more general answer: a probability distribution on models, which expresses any remaining uncertainty about the true model. The performance of such a distribution is evaluated by the well-known *log loss* measure. The truth-finding problem is discussed in Chapter 3.

Quantifying information The truth-finding problem is different from the preceding two problems, for it makes the additional assumption that reality is in one of the models. The availability of a true model allows a natural measure of *error* or *loss*, namely the amount of information that we, the learner, lack about this true model. To measure this amount, consider the following hypothetical situation. Suppose that there is a helpful external observer that knows which model is true. This observer sends us a message (e.g. an SMS), to tell us which model is true. *The more information we already possess about the true model, the shorter this message needs be.* We equate the amount of information that we lack about the true model with the length of the shortest message that will make us totally informed about the true model.

Log loss The above sending of messages is formalised in information theory using codes. Codes allow us to measure message lengths in bits. Throughout this thesis, we will use probability distributions as mathematical generalisations of codes. A probability distribution can be regarded as a code with idealised (non-integer) code-lengths. This correspondence will be explained in more detail in §2.6.1. Let P be the distribution on models that represents our uncertainty about the true model, i.e. that we use as a code, and let $\mathcal{M}^*$ be the true model. Then the amount of information that we lack about $\mathcal{M}^*$, denoted $L_{\log}$ and called the *log loss*, is given by

$$L_{\log}(P, \mathcal{M}^*) = -\log P(\mathcal{M}^*). \tag{1.1}$$

Minimising the log loss is equivalent to maximising the probability that we assign to the true model. Of course, we — the learner — cannot determine the log loss ourselves, because we do not know the true model.

We stress that the distribution P on models that is learned can *always* be interpreted as a code. Only in special situations can the probabilities that P assigns to the models be interpreted as their relative frequencies of occurrence, or as the learner's subjective degree of belief in their truth. We describe these situations in §3.7.4.

Truth-finding game To analyse the truth-finding problem with log loss, we formulate it as a game. The *truth-finding game*, a strictly competitive game with chance moves, is played between the players Learner, Chance and Nature. The arena of the truth-finding game is a set of competing models. Nature picks the world that generates the data from one of the models, Chance actually generates the data, and Learner tries to gain as much information as possible about the true model using the data. The entire game is shown in Protocol 1.1.

Protocol 1.1 The truth-finding game

Arena: Competing models $\mathbb{M} = \{\mathcal{M}_1, \mathcal{M}_2 \ldots\}$.
Require: Number of outcomes n.
1: Nature covertly chooses a hypothesis θ^*. Say θ^* is in model $\mathcal{M}^*$.
2: Chance samples a sequence of outcomes $y_1, y_2 \ldots y_n$ from θ^*.
3: Learner expresses his belief as a probability distribution P on models.
Loss: Learner suffers -log $P(\mathcal{M}^*)$.

Example 1.1 (Biased coin)**.** The following is a run of the truth-finding game for Scenario 1. We start with two models: *fair coin*, and *coin favours heads*. Formally, we have $\mathbb{M} = \{\mathcal{M}_1, \mathcal{M}_2\}$, where

$$\begin{aligned} \mathcal{M}_1 &= \big\{P(\mathrm{H}) = 0.5\big\} && \text{fair coin} \\ \mathcal{M}_2 &= \big\{P(\mathrm{H}) \in [0.6, 1]\big\} && \text{coin favours heads} \end{aligned}$$

We have slightly altered the definition of the second model in this example for illustrative purposes. This version of the biased coin scenario will be called Reduced Biased Coin in Chapter 3, where we continue this example. Say n, the number of coin flips we will perform, is fourteen. Now Nature starts by choosing a world from either of the models. Say she picks the coin with bias 0.6, from the second model. This means that, for the rest of this game, model two is the true model. (This simple strategy for Nature is not optimal. The best (minimax optimal) strategy is given in Chapter 3.) Then Chance generates 14 outcomes of the coin with bias 0.6. Say these are the outcomes:

$$\mathrm{H}, \mathrm{T}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{T}, \mathrm{T}, \mathrm{T}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{T}, \mathrm{T}, \mathrm{H}$$

Finally, Learner must express his belief about the true model as a probability distribution on models. He does not know which model is true, but he has seen the outcomes. Disregarding the order of the outcomes, he could just count: $6 \times T$ and $8 \times H$. Now, following the worst-case-optimal strategy described in Chapter 3, he produces the following probability distribution on models:

$$P(\mathcal{M}_1) = 0.4701 \qquad P(\mathcal{M}_2) = 0.5299$$

Note that Learner only slightly favours the second model. This a cautious choice, because the data are not very informative. Then the true model $\mathcal{M}_2$ is revealed, and the information that Learner lacks about this model is computed using the log loss. It is given by $-\log P(\mathcal{M}_2) = 0.9162$. Our analysis will show that the expected loss for Learner, using the worst-case-optimal strategy is 0.9040. The current loss is higher, but this is not due to Nature, but to Chance. The outcome, which Chance generated at random, is just not very informative.

Worst-case-optimal strategy In this thesis we analyse the truth-finding game from a worst-case perspective. That is, we search for a learning procedure, a *strategy*, that constructs a probability distribution on models from observations, such that, using this procedure, we gain as much information about the true model as possible, in the worst state of nature for this particular procedure. The motivation for this approach is that it gives the best performance guarantees if we are not prepared to make further assumptions.

To evaluate a strategy, we compute its *risk*. This is the mean loss that the learner obtains using this strategy, where the average is taken over all Chance's moves. The risk of a strategy does depend on Nature's move, but no longer on Chance. Then, taking the worst-case-optimal strategy for Learner, we eliminate dependence on Nature. Therefore, the worst-case-optimal strategy and its risk require no assumptions about Nature. One of the models has to be true, but, in our worst-case analysis, it is immaterial which one.

Worst-case belief Worst-case analysis is quite different from Bayesian analysis. In the latter, it is assumed that the learner can always construct a probability distribution on possible worlds, expressing his prior uncertainty about the actual world. The crucial difference is that the Bayesian learner uses this distribution for two purposes. First, he selects the act that is optimal with respect to this distribution. Second, he assesses his own performance with respect to this distribution, using it as though it were *true*. We think that this approach is essentially circular.

It is interesting that our worst-case analysis also constructs a probability distribution on possible worlds. This probability distribution can be interpreted as the prior belief that the learner should have about the true model, in the sense that it is a worst-case-optimal mixed strategy for Nature. For Learner, believing this distribution is not problematic; if it is not true, then Nature does not play optimally, and the incurred risk for Learner can only decrease. The constructed probability distribution on models depends heavily on the structure of the models, and is often particularly non-uniform. This directly contradicts Bayesian philosophy, which, taken in a weak form, prescribes the assumption of a smooth, fairly uniform, distribution, in the absence of specific prior knowledge.

1.4 Experiments

We now address the task of truth finding when we can perform experiments.

Experiment design The experiment-design problem is the following: *given a collection of competing models, perform the experiments that, in the end, yield most information about the true model*. We allow experiments to be chosen sequentially, this means that we can choose the next experiment based on the data obtained in all previous experiments. Example 1.2 is provided below as an illustration.

Previously, hypotheses were probability distributions on outcomes. In experiment design, hypotheses are conditional probability distributions on outcomes given input. Here, the input is the experiment selected by the learner. The true hypothesis fixes the way experiments work, by dictating the probability of each outcome for each input.

The experiment-design problem can be seen as experiment selection followed by truth finding. The data used for truth finding are the outcomes of the selected experiments. The task of experiment design amounts to choosing an experimentation strategy to maximise the amount of information that truth finding obtains.

To analyse the experiment-design problem, we translate it into a game.

Experimentation game The *experimentation game*, a strictly competitive game with chance moves, is played between the players Experimenter, Chance and Nature. The experimentation game extends the truth-finding game with experiments. This extra ability for Learner licenses his new name: Experimenter. A run of the game proceeds as follows. Nature initially picks the world in which all experiments take place. Then Experimenter chooses an experiment, and Chance responds by generating an outcome according to the actual world and the chosen experiment. These two steps are repeated a predetermined number of

times, allowing Experimenter to base his choice on previous outcomes. Finally, Experimenter is evaluated as in the truth-finding game. He must provide a probability distribution on models, and suffers the log loss, that is, the amount of information this distribution lacks about the true model. The experimentation game is summarised in Protocol 1.2.

Protocol 1.2 The experimentation game

Arena: Competing models $\mathbb{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$.
Require: Number of experiments n.
1: Nature covertly chooses a hypothesis θ^*. Say θ^* is in model $\mathcal{M}^*$.
2: **for** n turns **do**
3: Experimenter chooses an experiment input ξ.
4: Chance generates an outcome as predicted by θ^* on ξ.
5: **end for**
6: Experimenter expresses his belief as a probability distribution P on models.
Loss: Experimenter suffers $-\log P(\mathcal{M}^*)$.

Example 1.2 (Anvil drop)**.** The following is a run of the experimentation game for Scenario 2. We start with two models: *linear gravity* and *quadratic gravity*. Formally, we have $\mathbb{M} = \{\mathcal{M}_1, \mathcal{M}_2\}$, where

$$\begin{aligned} \mathcal{M}_1 &= \left\{h \approx \alpha t \mid \alpha \in [1, 10]\right\} \quad \text{linear gravity} \\ \mathcal{M}_2 &= \left\{h \approx \alpha t^2 \mid \alpha \in [1, 10]\right\} \quad \text{quadratic gravity} \end{aligned}$$

The tower of Pisa has six galleries. Correcting for inclination, the first one is located at the height of four metres above the tower base. Each next one adds another four metres of height. Experimenter, Galileo, drops an anvil by gently pushing it over the edge of a loggia.

First Nature chooses the actual world; suppose she chooses $h \approx 5t^2$. More precisely, this means $t \sim \sqrt{h/5} + \epsilon$, where ϵ is Gaussian noise with variance 0.1. This fixes quadratic gravity as the true model for the rest of the game. Historical fiction[1] tells us that Galileo dropped two heavy objects from the tower. We adopt this number of experiments. Say the first anvil is dropped from the topmost gallery, $h = 6 \cdot 4 = 24$. Then Chance generates an outcome according to $h \approx 5t^2$, say $t = 2.2$ seconds. Now Experimenter can choose another experiment, say he drops the second anvil from $h = 3 \cdot 4 = 12$, and Chance generates $t = 1.6$.

It now remains to perform the evaluation step, which continues exactly as in the truth-finding game. The data, the concatenation of successive input/outcome pairs, are

$$\langle 24, 2.2\rangle, \langle 12, 1.6\rangle$$

[1]This story, although reported by Galileo's own student, is widely considered to be a legend according to [Wik06].

1.5 Contribution of the thesis

Truth-finding[2] is, to the best of our knowledge, a new way to formalise the learning problem. Its prime motivations are the following.

- Models are the interesting level of abstraction for learning.
- Worst-case analysis is a good answer to the absence of prior information, as it provides rigorous bounds without further assumptions.

We provide the worst-case solution of the truth-finding game, in the form of a procedure to find the worst-case-optimal strategy. We give an algorithm, and prove that it finds the worst case optimal strategy. As already described, this procedure constructs a probability distribution on possible worlds, and then acts optimally with respect to this distribution. One might say that this probability distribution is *objective*, as it is induced by the structure of the game, and not based on the learner's judgement.

The second half of this thesis is devoted to experiment design. We use truth finding as a building block, and hence adopt its motivations. We add the following

- Experiments are performed sequentially.

There is considerable literature on Bayesian experiment design, but most of the literature covers a setting in which all experiments are performed simultaneously. See [CV95] for an overview. The sequential setting is, of course, more powerful.

There also literature on frequentist experiment design, for example [Puk93]. Also here, there is a focus on performing experiments simultaneously. Other than that, it is not clear how this approach is related.

Theorems 2.30, 2.31 and 3.41 are more minor contributions. The last theorem is both interesting and simple to prove; we suspect it is not new, but we could not find a published statement of this result.

Related work

Information theory is covered in [CT90], game theory in [Bin91]. Decision theory is covered in the classical [Fer67]. Non-Bayesian experiment design is covered in [Puk93], which focuses on linear models. [CV95] reviews the current state of Bayesian experiment design, while [SW00] compares Bayesian experiment design to Maximum Entropy Selection. The relation between Bayes acts and the Maximum Entropy Principle is treated in detail in [GD04]. Minimum Description Length model selection is covered in [GMP05].

Organisation of the thesis

Chapter 2 provides notions and results that will be used later. It also serves to introduce our notation and conventions. Among other things, Chapter 2

[2]The term *truth finding* already has a legal meaning, and it has also been defined as a desirable quality of learning frameworks. Our usage refers to the interpretation/definition that we present in the first half of this thesis.

introduces the relevant strategic game theory that allows us to analyse and solve the games in later chapters.

Chapter 3 addresses the problem of truth finding. It provides an analysis and solution of the truth-finding game. It also addresses the problem that there often is no analytical solution, and provides a numerical solution to a simple example.

Experiment design is covered in Chapter 4, where we analyse and solve the experimentation game. Chapter 5 concludes and provides a list of directions for future research.

Acknowledgements

I would like to express my gratitude to the people that made this work possible, both the research and this document. I want to thank Sirée Koolen-Wijkstra, my lovely wife, for all her care and love, and for the huge amount of time she invested in my thesis. And I have to say that her massage is unsurpassed.

I was encouraged, inspired challenged and coached by my colleagues at CWI: Peter Grünwald, Steven de Rooij and Tim van Erven. Their sustained enthusiasm and support made doing research and writing this thesis an extremely pleasant experience.

I am grateful to Tikitu de Jager for his thoroughly pedantic proofreading. A native English speaker with his LaTeX expertise, understanding of Dutch, technical background, linguistic knowledge and typographical zeal is hard to find.

Chapter 2

Preliminaries

This chapter covers the notions and results needed for the development of the theory of truth finding and experiment design in Chapters 3 and 4. This chapter does not contain new material, with the exception of §2.6.3; it is intended as a reference, and serves to introduce notation. Readers that are familiar with some of the areas of research described herein may skip these without difficulty, because standard notation has been used wherever possible.

2.1 General notation

We denote by $\mathbb{N}$ and $\mathbb{R}$ the sets of natural and real numbers. Both contain 0. The extended real numbers are defined by $\overline{\mathbb{R}} := [-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\}$, and they are endowed with the intuitive order and the corresponding order topology. $\mathbb{R}_+$ is the set of non-negative real numbers. $\mathbb{R}_{++}$ is the set of positive real numbers. As usual, $\mathbb{R}^n$ is the n-fold Cartesian product of $\mathbb{R}$. We use log for the binary logarithm. It is convenient to regard log as a function from $\mathbb{R}_+$ to $\overline{\mathbb{R}}$, by defining $\log 0 := -\infty$.

2.2 Set theory

Notation 2.1. Let Φ and Ω be sets. We denote the power set of Ω by $\wp(\Omega)$. The identity function on Ω is denoted by $\mathbf{1}_\Omega$. We denote by $[\Omega \to \Phi]$ the set of all functions from Ω to Φ. We abbreviate $f \in [\Omega \to \Phi]$ to $f : \Omega \to \Phi$. We write $f : \Omega \twoheadrightarrow \Phi$ if f is a surjective function from Ω to Φ.

Notation 2.2. Let Φ be a set, I a well-ordered set of indices, and $\Omega_i \subseteq \Phi$ for each $i \in I$, with duplicates allowed. We denote the function $i \mapsto \Omega_i$ by $\langle \Omega_i \rangle_{i \in I}$. We call a function of this form an *I-family* or an *I-sequence*.

Definition 2.3. Let Φ be a set. We denote by Φ^* and $\Phi^{<\omega}$ the set of all finite sequences over Φ. We denote by Φ^ω the set of all infinite sequences over Φ.

Definition 2.4. Let $f : \Phi \times \Omega \to \Psi$. For all $x \in \Phi$, we denote by $f(x, \cdot) := \big\{(y, f(x, y)) \mid y \in \Omega\big\}$. Clearly, $f(x, \cdot) : \Omega \to \Psi$. The function $f(x, \cdot)$, viewed as a function of x, is called the *Schönfinkelisation* or *Currying* of f. We analogously define $f(\cdot, y)$.

Definition 2.5. Let Ω be a set. A set $\Phi \subseteq \wp(\Omega)$ is called a *partition* of Ω if

1. $\bigcup \Phi = \Omega$. (Φ covers Ω.)
2. $\emptyset \notin \Phi$.
3. $\Psi \cap \Theta = \emptyset$ for all different $\Psi, \Theta \in \Phi$. (The elements of Φ are pairwise disjoint.)

2.3 Linear algebra

Whenever n is clear from the context, we denote by $\mathbf{0}$ and $\mathbf{1}$ the zero and unity vectors in $\mathbb{R}^n$, i.e. the vectors that have all entries set to either zero or one. For $1 \leq i \leq n$, $\mathbf{e}_i$ is the unit vector of dimension i. We denote the transpose of a vector p by p^T.

Definition 2.6. The *unit n-simplex* is the set given by

$$\Delta_n := \left\{ p \in \mathbb{R}^n_+ \mid p^T \mathbf{1} = 1 \right\}.$$

It is also called the *standard n-simplex* or *probability n-simplex*. One can equivalently define Δ_n as the convex hull of $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$. Note that we number the unit simplices by the number of partaking unit vectors, whereas some other authors number by dimension (which is $n-1$), or equivalently include the zero vector in the convex hull definition.

Each discrete probability distribution on a set of n outcomes can be represented by a point in Δ_n and vice versa. The unit 2- and 3-simplices are shown in Figure 2.1.

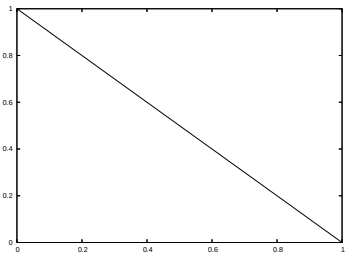
2.4 Convex analysis

All sets in this section are subsets of $\mathbb{R}^n$. A set C is convex if it is closed under linear interpolation. We denote by $\mathrm{conv}(\Omega)$ the convex hull of the set Ω, i.e. the intersection of all convex sets that contain Ω. The convex hull operation preserves openness, closedness and boundedness, hence also compactness.

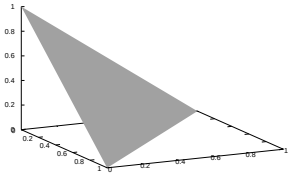
The following results can be found in a standard textbook on convex optimisation, for example in [BV04].

Figure 2.1 Examples of unit simplices

(a) unit 2-simplex: a line segment in the plane



(b) unit 3-simplex: an equilateral triangle in 3-space



Theorem 2.7 (Supporting Hyperplane Theorem)**.** For every convex set C, and point x on the border of C, there is a hyperplane P through x, such that C is contained in one of the half-spaces of P.

Theorem 2.8 (Separating Hyperplane Theorem)**.** Let H and K be convex sets in $\mathbb{R}^n$ with disjoint interior. Then there exists a hyperplane $\{x \mid a^T x = b\}$ that separates H and K.

Theorem 2.9. If $f : \mathbb{R}^2 \to \mathbb{R}$ is convex in (x, y), and C a convex non-empty set, then the function

$$g(x) = \inf_{y \in C} f(x, y)$$

is convex in x, provided $g(x) > -\infty$ for some x.

2.5 Probability theory

Probability theory deals with probabilities of events, that is, sets of outcomes. A rigid formalisation of probability theory using measure theory is given in Appendix A. For the current exposition, it suffices to define a *probability distribution* as a function that assigns probabilities, i.e. numbers from $[0, 1]$, to events, obeying certain conditions. We abbreviate probability distribution to distribution whenever convenient.

Throughout this thesis, we will use three standard types of event sets, depending on the type of the set of outcomes $\mathcal{X}$ as follows:

- If $\mathcal{X}$ is finite, we use the events $\wp(\mathcal{X})$.
- If $\mathcal{X} \subseteq \mathbb{R}$, we use the events $\mathrm{Bor}(\mathbb{R})$. This is the Borel σ-algebra on $\mathbb{R}$, i.e. the smallest set of events that contains all open sets of $\mathbb{R}$, and that is closed under complements and countable unions. If $\mathcal{X} \subseteq \mathbb{R}^n$, we use the events $\mathrm{Bor}(\mathbb{R}^n)$.
- If $\mathcal{X}$ is a set distributions on a finite set of size n, we identify it with $\Delta_n \subset \mathbb{R}^n$, and use the Borel σ-algebra on the latter.

Note that in all cases, the singleton sets of $\mathcal{X}$ appear in the event set.

Definition 2.10. A pair $\langle \mathcal{X}, \Sigma \rangle$, where $\mathcal{X}$ is a set of one of the above categories, and Σ is the corresponding set of events, is called a *sample space*. Because Σ is always clear from the context, we identify a sample space with its carrier $\mathcal{X}$.

Definition 2.11. For any sample space $\mathcal{X}$, we denote by $\mathfrak{D}(\mathcal{X})$ the set of all probability distributions on $\mathcal{X}$.

So far, we have not assumed any structure on the set of outcomes $\mathcal{X}$. The most straightforward way to obtain structure is to use the available structures on $\mathbb{R}$ by assigning real numbers to outcomes, thereby transforming the set of outcomes into a subset of $\mathbb{R}$.

Definition 2.12. A function $X : \mathcal{X} \to \mathbb{R}$ is called a *random variable*. We say that X is finite, countable or uncountable if its range is.

Sometimes, it is useful to translate the set of outcomes into some set different from $\mathbb{R}$. We call such transformations *pseudo random variables*.

As was just stated, a random variable transforms outcomes into real numbers. Via this transformation, we can forget about the original distribution, and consider the induced distribution on $\mathbb{R}$.

Notation 2.13. If a random variable X is distributed according to P, we write $X \sim P$.

Definition 2.14. Let X be a random variable defined on the sample space $\mathcal{X}$ with distribution P. We define the *expected value* or *expectation* of X by

$$\mathrm{E}\left[X\right] := \int_{\mathcal{X}} X \,\mathrm{d}P$$

Definition 2.15. Let X be a random variable on the sample space $\mathcal{X}$. We say that X is *constant* if $\exists c \forall x \in \mathcal{X} : X(x) = c$. We call X *almost surely constant* if $\exists c : P(X = c) = 1$. This implies $P(X = \mathrm{E}\left[X\right]) = 1$.

Remark 2.16. A random variable X on $\mathcal{X}$ is almost surely constant if all measure of P is assigned to a region where X is constant. This can be solely due to X, namely when X is constant, or solely due to P, namely when P puts all measure on a single point, or partially due to both.

Theorem 2.17 (Jensen's Inequality [Wil91, Theorem, p. 61])**.** Let $\mathcal{X}$ be a convex set, P a probability distribution on $\mathcal{X}$. Then for any convex function $f : \mathcal{X} \to \mathbb{R}$,

$$\mathrm{E}_P\left[f(X)\right] \geq f\left(\mathrm{E}_P\left[X\right]\right) \tag{2.1}$$

Moreover, if f is strictly convex, then equality in (2.1) implies that X is an almost surely constant random variable.

One level of abstraction higher, we work with a meta-distribution on sets of probability distributions. We can interpret such a meta-distribution as a prior probability; one first samples a distribution according to this meta-distribution, and then generates an outcome according to the sampled distribution. For more detail, see [GD04, Section 9.2]. Such a meta-distribution can be collapsed into a single distribution on outcomes as follows.

Definition 2.18. Let $\mathcal{X}$ be a set, $\mathcal{Q}$ a convex set of distributions on $\mathcal{X}$, and $\mathbb{Q}$ a distribution on $\mathcal{Q}$. We define $\mathrm{E}_{\mathbb{Q}}\left[Q\right]$, the *expected distribution of* $\mathbb{Q}$ *on* $\mathcal{X}$, by

$$\mathrm{E}_{\mathbb{Q}}\left[Q\right](A) := \mathrm{E}_{\mathbb{Q}}\left[Q(A)\right] = \int_{\mathcal{Q}} Q(A) \,\mathrm{d}\mathbb{Q}$$

where $Q = \mathbf{1}_{\mathcal{Q}}$ is a pseudo random variable, and A any event of $\mathcal{X}$.

Definition 2.19. Let X, Y be pseudo random variables with ranges $\mathcal{X}$ and $\mathcal{Y}$. The distribution P that gives the distribution of the pair $\langle X, Y\rangle$ is called the *joint distribution* of X and Y. The *marginal distributions* of X, Y are given by

$$\begin{aligned} P_X(E_X) &:= P(E_X \times \mathcal{Y}) \\ P_Y(E_Y) &:= P(\mathcal{X} \times E_Y), \end{aligned} \tag{2.2}$$

for events E_X and E_Y over $\mathcal{X}$ and $\mathcal{Y}$. We write $P(E_X)$ for $P_X(E_X)$ and $P(E_Y)$ for $P_Y(E_Y)$ whenever convenient.

Definition 2.20. Let n be the number of outcomes, and let P be a probability distribution on a countable space. The *n-fold product distribution* of P, denoted P^n, is given by

$$P^n(y_1, \ldots, y_n) := \prod_{i=1}^{n} P(y_i)$$

2.6 Information theory

Information theory exploits the relationship between probability distributions and codes. In this section, we restrict attention to probability distributions on a finite or countable set. The proofs of the theorems that we merely state below can be found in [CT90].

2.6.1 Quantifying information

Bits and codes

A *bit* is a variable that ranges over two values. These values can have many interpretations, for example: on/off, high/low, 0/1 and true/false. In general, a single bit allows one to distinguish between two arbitrary possibilities. To distinguish between more than two possibilities, one uses a *code*: a collection of *code words*, sequences of bits, with an interpretation for each code word. We restrict attention to codes with the following properties:

- *Prefix-free.* A code is prefix free if no codeword is a prefix of another. A non-prefix-free code has two code words a, b such that b extends a. Such a code uses both the *content* and the *length* of code words to convey information. Such a code is ambiguous; upon receiving, bit by bit, code word a, one cannot tell whether the message has ended and is a, or whether it will continue, actually being b.

- *Irredundant.* A code is irredundant if no two code words have the same interpretation. A code that has two code words with the same interpretation is obviously inefficient, because code words can only be used one at a time.

- *Complete.* A code is complete if addition of any new code word renders it non-prefix-free. A code that remains prefix-free when a new code word is added does not use its full potential.

We will henceforth simply use the word code for an irredundant complete prefix-free code.

Codes are used to quantify the information content of objects: the amount of information a certain object contains, with respect to a given code, is given by the length of the shortest code word that is interpreted as this object.

Idealised bits and probability distributions

There is a strong correspondence between codes and certain probability distributions, as shown by the following theorem.

Theorem 2.21 (Kraft Inequality [LV93, p.74])**.** Let $\ell_1, \ell_2, \ldots$ be a finite or infinite sequence of natural numbers. There is a prefix-free code with this sequence as lengths of its binary code words iff

$$\sum_i 2^{-\ell_i} \leq 1.$$

Moreover, a code is complete iff this holds with equality.

Consequently, a code on $\mathcal{X}$ with code words w_1 for x_1, w_2 for x_2, etc. corresponds to a probability distribution P with $P(x_i) = 2^{-\ell_i}$, where $\ell_i = |w_i|$ is the length, in bits, of code word w_i. Note that this correspondence does not use the content of code words, only their lengths.

The inverse of this transformation transforms an arbitrary probability distribution on $\mathcal{X}$ into a list of code word lengths as follows

$$\ell_1 = -\log P(x_1), \quad \ell_2 = -\log P(x_2), \quad \ldots$$

These code word lengths can be non-integral. We call such numbers *idealised code lengths*, and their unit *idealised bits*.[1]

This mathematical generalisation gives us a much more fine-grained way to measure the information content of objects: the amount of information that a certain object x contains, with respect to a probability distribution P, is given by $-\log P(x)$. Conversely, we say that this is the amount of information that P lacks about x.

2.6.2 Basic coding

In the following definitions we use the convention, based on continuity arguments, that $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$ for $p > 0$ and all q, see [CT90]. It follows that $p \log \frac{p}{q} = p \log p - p \log q$ even when $q = 0$.

Entropy

Definition 2.22. Let $X \sim P$. The *entropy of* X is defined by

$$\mathcal{H}(X) := \mathrm{E}_P \left[-\log P(X) \right] \tag{2.3}$$

The entropy of X equals the expected codelength, when the outcome X is encoded using its source distribution P as the code. When P is a distribution on $\mathcal{X}$, we abbreviate $\mathcal{H}(\mathbf{1}_{\mathcal{X}})$ to $\mathcal{H}(P)$.

Definition 2.23. The *conditional entropy of* Y *given* $X = x$ is defined by

$$\mathcal{H}(Y|X = x) := \mathop{\mathrm{E}}_{Y \sim P(\cdot|X=x)} \left[-\log P(Y|X = x) \right]. \tag{2.4}$$

The *expected conditional entropy of* Y *given* X[2] is given by

$$\mathrm{E}_X \left[\mathcal{H}(Y|X) \right] = \mathrm{E}_{X,Y} \left[-\log P(Y|X) \right]. \tag{2.5}$$

[1] An important technique in information theory called the *Shannon-Fano code* allows us to find code words $w_1, w_2, \ldots$ such that $\lceil \ell_i \rceil$, ℓ_i rounded up, equals $|w_i|$. A second technique, called *arithmetic coding*, allows us to actually achieve these idealised code lengths when we code a sequence of objects, even when a different probability distribution is used for each object.

[2] In [CT90], the expected conditional entropy is abbreviated to $\mathcal{H}(Y|X)$. We cannot adopt this shorthand, as it would cause ambiguity in the sequel.

Theorem 2.24 (Chain Rule of Entropy)**.**

$$\mathcal{H}(X,Y) = \mathcal{H}(X) + \mathrm{E}_X\left[\mathcal{H}(Y|X)\right]. \tag{2.6}$$

KL-divergence

Definition 2.25. For distributions P, Q, we define the *Kullback-Leibler divergence* of Q from P by

$$\mathcal{D}\left(P\|Q\right) := \mathrm{E}_P\left[\log\frac{P(X)}{Q(X)}\right] \tag{2.7}$$

$$= \mathrm{E}_P\left[-\log Q(X)\right] - \mathcal{H}(P). \tag{2.8}$$

The Kullback-Leibler divergence of Q from P is the number of additional bits one expects to use when coding an outcome from P using Q instead of P.

Note that $\mathcal{D}$ is not symmetric in P and Q, so it is not a distance. It does however have the following important property.

Theorem 2.26 (Information Inequality)**.** Let P, Q be probability distributions. Then

$$\mathcal{D}\left(P\|Q\right) \geq 0 \tag{2.9}$$

with equality iff $P = Q$.

This means that, in expectation, the best code for outcomes that are generated from P, is P itself.

Theorem 2.27. Let P^n, Q^n be product distributions. Then

$$\mathcal{D}(P^n\|Q^n) = n\mathcal{D}(P\|Q) \quad \text{and} \quad \mathcal{H}(P^n) = n\mathcal{H}(P).$$

Proof. It suffices to show

$$\mathop{\mathrm{E}}_{Y^n\sim P^n}\left[\log Q^n(Y^n)\right] = \mathop{\mathrm{E}}_{Y_1\sim P}\cdots\mathop{\mathrm{E}}_{Y_n\sim P}\left[\log Q(Y_1)+\cdots+\log Q(Y_n)\right] \tag{2.10}$$

$$= \mathop{\mathrm{E}}_{Y_1\sim P}\left[\log Q(Y_1)\right] + \cdots + \mathop{\mathrm{E}}_{Y_n\sim P}\left[\log Q(Y_n)\right] \tag{2.11}$$

$$= n\mathop{\mathrm{E}}_{Y\sim P}\left[\log Q(Y)\right], \tag{2.12}$$

the rest is just definition chasing. □

Theorem 2.28. $\mathcal{D}\left(P\|Q\right)$ is convex in the pair (P, Q), that is, for all distributions P_1, P_2, Q_1, Q_2 and for all $0 \leq \lambda \leq 1$ we have

$$\mathcal{D}\left(\lambda P_1 + (1-\lambda)P_2\|\lambda Q_1 + (1-\lambda)Q_2\right) \;\leq\; \lambda\mathcal{D}\left(P_1\|Q_1\right) + (1-\lambda)\mathcal{D}\left(P_2\|Q_2\right)$$

Mutual information

Definition 2.29. For random variables X and Y with joint distribution P and marginal distributions P_X and P_Y, we define the *mutual information* between X and Y by

$$\mathcal{I}(X;Y) := \mathrm{E}_P\left[\log\frac{P(X,Y)}{P_X(X)P_Y(Y)}\right] \tag{2.13}$$

$$= \mathcal{D}(P\|P_XP_Y) \tag{2.14}$$

$$= \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y) \tag{2.15}$$

Note that this quantity is symmetric in X and Y.

2.6.3 Advanced coding

We analyse three somewhat advanced coding scenarios.

Conditional coding

The following theorem is a generalisation of the information inequality. Suppose that a pair of outcomes X, Y is generated from P, and we are first told $X = x$, and then need to encode Y. Then the best code for this, again in expectation, is the conditional distribution $P(Y|X = x)$. Of course, this conditional probability is not defined when $P(X = x) = 0$, but then simultaneously, the probability that we observe x in the first place is zero.

Theorem 2.30 (Generalized Information Inequality)**.** Let $\mathcal{X}$ and $\mathcal{Y}$ be sample spaces, and let P, Q be probability distributions over $\mathcal{X} \times \mathcal{Y}$. Then

$$\mathrm{E}_P\left[-\log P(Y|X)\right] \leq \mathrm{E}_P\left[-\log Q(Y|X)\right] \tag{2.16}$$

with equality if and only if $P(Y|X) = Q(Y|X)$ almost surely. Here, almost surely means whenever $P(X) > 0$.

Proof. For all x s.t. $P(x) > 0$, we have, by the Information Inequality,

$$\mathrm{E}_{Y|x}\left[-\log P(Y|x)\right] \leq \mathrm{E}_{Y|x}\left[-\log Q(Y|x)\right], \tag{2.17}$$

with equality iff $Q(Y|x) = P(Y|x)$. Taking the expectation over $P(X)$ in (2.17) yields (2.16), observing that the x where $P(x) = 0$ do not contribute to the expectation at all. This immediately shows that equality holds iff $P(Y|X) = Q(Y|X)$ almost surely. □

Meta-coding

Suppose we have a meta-distribution on codes, that we want to use to encode an outcome from P. We can either sample a code from our meta-distribution, and then use that to encode the outcome, or we can encode the outcome with the expected code. The following theorem proves that, in expectation, the latter is better.

Theorem 2.31. Let $\mathcal{X}$ be a set, and $\mathcal{Q}$ a convex set of distributions on $\mathcal{X}$. Then for all distributions $\mathbb{Q}$ on $\mathcal{Q}$ and P on $\mathcal{X}$:

$$\mathrm{E}_{\mathbb{Q}}\left[\mathcal{D}(P\|Q)\right] \geq \mathcal{D}(P\|\,\mathrm{E}_{\mathbb{Q}}\left[Q\right]). \tag{2.18}$$

Proof. First, observe that (2.18) holds iff

$$\mathrm{E}_{\mathbb{Q}}\,\mathrm{E}_P\left[\log\frac{P(X)}{Q(X)}\right] \geq \mathrm{E}_P\left[\log\frac{P(X)}{\mathrm{E}_{\mathbb{Q}}\left[Q(X)\right]}\right] \tag{2.19}$$

$$\text{iff} \qquad \mathrm{E}_P\,\mathrm{E}_{\mathbb{Q}}\left[-\log Q(X)\right] \geq \mathrm{E}_P\left[-\log \mathrm{E}_{\mathbb{Q}}\left[Q(X)\right]\right] \tag{2.20}$$

Second, note that $f_x : \mathcal{Q} \to [0, \infty]$ defined by $f_x(Q) := -\log Q(x)$ is a convex function (of Q) for all $x \in \mathcal{X}$. Application of Jensen's Inequality (2.1) yields:

$$\mathrm{E}_{\mathbb{Q}}\left[f_x(Q)\right] \geq f_x(\mathrm{E}_{\mathbb{Q}}\left[Q\right]) \quad \text{so} \quad \mathrm{E}_{\mathbb{Q}}\left[-\log Q(x)\right] \geq -\log \mathrm{E}_{\mathbb{Q}}\left[Q(x)\right]$$

As this holds for all x, it also holds in expectation, which proves (2.20). □

Conditional meta-coding

The next theorem generalises the previous theorem to the case where P is a joint distribution on X, Y, and we need to encode Y given that $X = x$.

Theorem 2.32. Let $\mathcal{X}$, $\mathcal{Y}$ be sets, and $\mathcal{Q}$ a convex set of conditional distributions on $\mathcal{Y}$ given $\mathcal{X}$. Then for all distributions $\mathbb{Q}$ on $\mathcal{Q}$ and P on $\mathcal{X} \times \mathcal{Y}$:

$$\mathrm{E}_P \, \mathrm{E}_{\mathbb{Q}} \left[-\log Q(Y|X) \right] \geq \mathrm{E}_P \left[-\log \mathrm{E}_{\mathbb{Q}} \left[Q(Y|X) \right] \right]. \tag{2.21}$$

Proof. The proof is analogous to that of Theorem 2.31, using the convex function $f_{x,y}(Q) := -\log Q(y|x)$ instead. □

2.7 Game theory

Definition 2.33. A triple $\mathcal{G} = \langle S_{\mathrm{A}}, S_{\mathrm{B}}, \pi \rangle$ is called a *matrix game* if $\pi : S_{\mathrm{A}} \times S_{\mathrm{B}} \to \mathbb{R}$. A matrix game is a two-player zero-sum game in strategic form. We call the elements of S_{A} and S_{B} *pure strategies* for players A and B, and π the *payoff*.

Definition 2.34. We define the *minimax* value $\overline{\mathrm{V}}$ and *maximin* value $\underline{\mathrm{V}}$ by

$$\overline{\mathrm{V}} := \inf_{s_{\mathrm{B}} \in S_{\mathrm{B}}} \sup_{s_{\mathrm{A}} \in S_{\mathrm{A}}} \pi(s_{\mathrm{A}}, s_{\mathrm{B}}) \tag{2.22}$$

$$\underline{\mathrm{V}} := \sup_{s_{\mathrm{A}} \in S_{\mathrm{A}}} \inf_{s_{\mathrm{B}} \in S_{\mathrm{B}}} \pi(s_{\mathrm{A}}, s_{\mathrm{B}}) \tag{2.23}$$

These value can be interpreted as follows. The maximin value (supposing that the supremum is attained) is the highest payoff that player A can guarantee, when player B chooses her move after learning the move of player A. Similarly, the minimax value is the least payoff that player B can guarantee when her move is reported to player A before he has to choose his move.

If $\overline{\mathrm{V}} = \underline{\mathrm{V}}$, we call this quantity the *value* of $\mathcal{G}$ and denote it by just V. If a game has a value, then *playing second provides no advantage*.

Remark 2.35. Always $\underline{\mathrm{V}} \leq \overline{\mathrm{V}}$, but not necessarily $\underline{\mathrm{V}} = \overline{\mathrm{V}}$.

Definition 2.36. We call $(\tilde{s}_{\mathrm{A}}, \tilde{s}_{\mathrm{B}})$ a *saddle-point* of π if for all $s_{\mathrm{A}}, s_{\mathrm{B}}$

$$\pi(\tilde{s}_{\mathrm{A}}, s_{\mathrm{B}}) \geq \pi(\tilde{s}_{\mathrm{A}}, \tilde{s}_{\mathrm{B}}) \geq \pi(s_{\mathrm{A}}, \tilde{s}_{\mathrm{B}}).$$

The existence of a saddle-point guarantees that $\mathcal{G}$ has a value. On the other hand, $\mathcal{G}$ may have a value but no saddle point. This occurs when the infimum is not a minimum in (2.22), or the supremum is not a maximum in (2.23).

Definition 2.37. Let $\mathcal{G}$ be a matrix game. We call $\Sigma_{\mathrm{A}} := \mathfrak{D}(S_{\mathrm{A}})$ and $\Sigma_{\mathrm{B}} := \mathfrak{D}(S_{\mathrm{B}})$ the *mixed strategies* for players A and B. We lift the pure strategy payoff π to the mixed strategy payoff $\Pi : \Sigma_{\mathrm{A}} \times \Sigma_{\mathrm{B}} \to \mathbb{R}$ by

$$\Pi(\sigma_{\mathrm{A}}, \sigma_{\mathrm{B}}) := \mathop{\mathrm{E}}_{\substack{s_{\mathrm{A}} \sim \sigma_{\mathrm{A}} \\ s_{\mathrm{B}} \sim \sigma_{\mathrm{B}}}} \left[\pi(s_{\mathrm{A}}, s_{\mathrm{B}}) \right] := \int_{S_{\mathrm{A}} \times S_{\mathrm{B}}} \pi(s_{\mathrm{A}}, s_{\mathrm{B}}) \, \mathrm{d}\sigma_{\mathrm{A}} \, \mathrm{d}\sigma_{\mathrm{B}}.$$

We call $\widetilde{\mathcal{G}} := \langle \Sigma_{\mathrm{A}}, \Sigma_{\mathrm{B}}, \Pi \rangle$ the *mixed matrix game generated by* $\mathcal{G}$. Its minimax value $\overline{\mathrm{V}}$ and maximin value $\underline{\mathrm{V}}$ are given by

$$\overline{\mathrm{V}} := \inf_{\sigma_{\mathrm{B}} \in \Sigma_{\mathrm{B}}} \sup_{\sigma_{\mathrm{A}} \in \Sigma_{\mathrm{A}}} \Pi(\sigma_{\mathrm{A}}, \sigma_{\mathrm{B}}) \tag{2.24}$$

$$\underline{\mathrm{V}} := \sup_{\sigma_{\mathrm{A}} \in \Sigma_{\mathrm{A}}} \inf_{\sigma_{\mathrm{B}} \in \Sigma_{\mathrm{B}}} \Pi(\sigma_{\mathrm{A}}, \sigma_{\mathrm{B}}) \tag{2.25}$$

Theorem 2.38 (Minimax Theorem [Bin91, Theorem 6.4.4])**.** If $\mathcal{G} = \langle S_{\mathrm{A}}, S_{\mathrm{B}}, \pi \rangle$ is a matrix game with compact sets S_{A} and S_{B} and continuous payoff π, then $\underline{\mathrm{V}} = \overline{\mathrm{V}}$ for $\widetilde{\mathcal{G}}$.

Corollary 2.39. Let $\mathcal{G} = \langle S_{\mathrm{A}}, S_{\mathrm{B}}, \pi \rangle$ be a matrix game, with S_{A} and S_{B} finite, then $\underline{\mathrm{V}} = \overline{\mathrm{V}}$ for $\widetilde{\mathcal{G}}$.

Definition 2.40. A pair $\langle T, < \rangle$ is a *tree* if

- $<$ is a strict partial order on T, i.e. irreflexive, antisymmetric, transitive,
- there is a unique $<$-least element, and
- for each $x \in T$, $\{ y \in T \mid y < x \}$ is well-ordered by $<$.

As usual, we shall use T to refer to $\langle T, < \rangle$. The $<$-maximal elements are called *terminals* or *leaves* of T, the other elements are called *non-terminals*. The sets of leaves and non-terminals are denoted $\mathrm{lf}(T)$ and $\mathrm{nt}(T)$. The unique $<$-least element of T is called the *root* of T.

Definition 2.41. A *game tree* is a sextuple

$$\left\langle T, <, A, m, \langle M_t \rangle_{t \in T}, \langle I_a \rangle_{a \in A}, \langle \pi_a \rangle_{a \in A} \right\rangle$$

where the following conditions hold:

- $\langle T, < \rangle$ is a finite tree. We call the elements of T *positions*.
- A is a set. We refer to the elements of A as *players*.
- For each position $t \in T$, M_t is a set of moves. We say that a move $m \in M_t$ can be *performed* at t. We define $M := \bigcup_{t \in T} M_t$.
- $m : \mathrm{nt}(T) \to A$. We call m the *move function*. It specifies which player has to move at each position.
- For each player $a \in A$, I_a is a partition of $m^{-1}(a)$. The elements of I_a are called the *information sets* for player a. We use I_a as an equivalence relation on the positions where player a has to move. Positions that are related by I_a, i.e. that are in the same information set, are called *indistinguishable* to player a.
- For each player $a \in A$, $\pi_a : \mathrm{lf}(T) \to \mathbb{R}$. π_a is called the *payoff function* for player a.
- $\forall a \in A \forall i \in I_a \forall t \in i \forall t' \in i \, (M_t = M_{t'})$. That is, for each player, the same moves are available at indistinguishable positions. We lift the move function to information sets by defining $M_i := \{ m \mid m \in M_t \wedge t \in i \}$ for each $i \in I_a$.

Definition 2.42. Given a game tree T, a function

$$s : I_a \to M$$

is called a strategy for player a if for each $i \in I_a$, $s(i) \in M_i$. That is, a strategy prescribes a move for player a, selected from the moves that are available, under the constraint that the move only depends on what player a can actually distinguish.

Chapter 3

Truth finding

In this chapter, we discuss *truth finding*: obtaining information about the true model from data. The setting of truth finding is the following. We possess a list of candidate explanations for reality, and assume that reality is on this list. The candidate explanations on the list are grouped into models. We suppose that the effects of our actions depend on reality only indirectly; they only depend on the model that contains reality. Hence it is important to know which model is true. We approach this learning task with a blank slate. The only available information is the outcome of a fixed experiment. This leads to the *truth-finding problem*: how to use data to obtain as much information about the true model as possible, in the worst case. The worst case is taken over all possible choices for the state of nature.

A related problem which, using our terminology, could be called *true world finding* or *hypotheses selection*, is covered in great detail in [Fer67]. Our setup is quite similar, allowing us to reuse several of the general theorems. Of course, some work is needed to prove that the preconditions hold in our setup. The subtle difference of focusing on finding the true model instead of true world admits a new decomposition of strategies, using the convex hulls of the models.

In §3.1 we formally state the truth-finding problem. The analysis of the truth-finding problem is performed in terms of the Nature vs Learner truth-finding game in §3.2. We give a representation of the strategies in §3.3. We subsequently prove that the truth-finding game has a value, and that there is a minimax strategy for Learner in §3.4. We then turn to the computation of the minimax strategy in §3.5. We show that it equals the best response to the optimal mixed strategy for Nature. We show that the latter strategy itself is easy to compute. At the end of this chapter, in §3.7, we relate truth finding to two other interpretations of learning from data: prediction and compression. We conclude and summarise in §3.8. All theorems given without proof are not original to this thesis, and are referenced accordingly.

3.1 Formalisation

We first formalise the setting described in the introduction to this chapter. Then we give the formal statement of the truth-finding problem.

Notational conventions

We typographically distinguish sets, elements and random variables. We use Roman and Greek lowercase symbols (y) for elements, and the corresponding uppercase symbols (Y) for random variables. We use calligraphic script ($\mathcal{Y}$) for sets, and the blackboard bold font ($\mathbb{M}$) for sets of sets. We use lowercase boldface symbols ($\mathbf{m}$), to denote probability mass functions. We have chosen symbols that are mnemonics for the underlying sample space.

For sets we write $\mathcal{Y}^n$ for the n-fold Cartesian product. For elements (and for random variables ranging over elements) we use the following convention: we use y^n as a shorthand for $\langle y_1, \ldots, y_n \rangle$. Hence the typical element of $\mathcal{Y}^n$ is y^n.

3.1.1 Truth-finding frames

Definition 3.1. A quadruple

$$\mathfrak{F} = \left\langle \mathcal{Y}, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}} \right\rangle$$

is called a *truth-finding frame*, or *frame* for short, if the following conditions hold:

- $\mathcal{Y}$ is a sample space, called the *outcome space*. We refer to the elements of $\mathcal{Y}$ as *outcomes*.
- $\mathcal{T}$ is a sample space, called the *possible-world space*. We refer to the elements of $\mathcal{T}$ as *possible worlds*.
- $\mathbb{M}$ is a partition of $\mathcal{T}$, called the set of *models*. The function $M : \mathcal{T} \to \mathbb{M}$ assigns to each $\theta \in \mathcal{T}$ the unique $\mathcal{M} \in \mathbb{M}$ that contains it.
- $\mathbf{p}_\theta$ is a probability distribution on $\mathcal{Y}$ for each $\theta \in \mathcal{T}$. We write $\mathbf{p}(y|\theta)$ for $\mathbf{p}_\theta(y)$. We use the term *mechanics* to refer collectively to $\langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}}$.

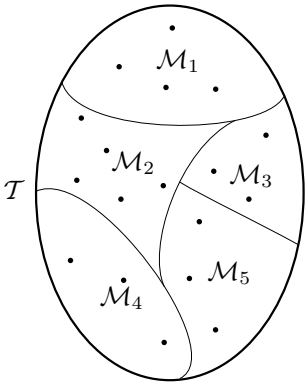
A truth-finding frame incorporates both the available information about reality, and the objective of the learning task. The available information about reality is represented by a set of options, the possible worlds, and a conditional distribution that specifies how each possible world works. The objective of the learning task is given by the partition, which divides the possible worlds into models, clusters of similar worlds. The set of models $\mathbb{M}$ can be regarded as a discretisation of the set of possible worlds $\mathcal{T}$. Figure 3.1 depicts a typical partition of possible worlds into models. We want to obtain as much information as possible about the true model, i.e. the model that contains the true state of nature.

3.1.2 Examples

Throughout this chapter we use three running examples as an illustration. Let $\mathcal{Y} = \{\mathrm{H}, \mathrm{T}\}$ be the sample space of outcomes of a coin flip. We denote by $\mathbf{p}_\theta$ the distribution that assigns probability θ to H, and hence $1 - \theta$ to T. We use the following three variants of the biased coin scenario:

- Biased Coin (BC). As in the introduction to this thesis, we use models *the coin is fair* and *the coin favours heads*.

Figure 3.1 Partition of possible worlds



- Reduced Biased Coin (RBC). In this simplified version, we replace the second model of BC by *the coin favours heads considerably.*
- Binary Biased Coin. (BBC). In this even more simplified version, we use two singleton models: *the coin mildly favours tails* and *the coin favours heads a lot.*

The models for each of the variants are formally specified in Table 3.1. In each case, we have models $\mathbb{M} = \{\mathcal{M}_1, \mathcal{M}_2\}$, possible worlds $\mathcal{T} = \bigcup \mathbb{M}$, and truth-finding frame

$$\mathfrak{F} = \left\langle \mathcal{T}, \mathcal{Y}, \mathbb{M}, \langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}} \right\rangle .$$

3.1.3 Truth-finding problem

Now we can formally state the worst-case truth-finding problem. Let $\mathfrak{F}$ be a frame. The unknown actual world $\theta^* \in \mathcal{T}$ must be classified according to $\mathbb{M}$, based on data generated from $\mathbf{p}_{\theta^*}$. Ideally, we would like to identify the true model $\mathcal{M}^* := M(\theta^*)$, but in our probabilistic framework we cannot do this with certainty. Instead we use the data to construct a probability distribution $\mathbf{m}$ on models, where $\mathbf{m}(\mathcal{M})$ represents our degree of belief that $\mathcal{M}^* = \mathcal{M}$.

An outsider knowing $\mathcal{M}^*$ can evaluate $\mathbf{m}$ by computing the log loss, which is given by $\mathrm{L}(\mathcal{M}^*, \mathbf{m})$, where

$$\mathrm{L}(\mathcal{M}, \mathbf{m}) := -\log \mathbf{m}(\mathcal{M}).$$

Recall that this is the amount of information that we, using $\mathbf{m}$, lack about the true model. Namely, it is the number of bits that the outsider has to transmit

Table 3.1 Biased coin example models

	$\mathcal{M}_1$	$\mathcal{M}_2$
BC	$\{1/2\}$	$(1/2, 1]$
RBC	$\{1/2\}$	$(0.6, 1]$
BBC	$\{1/3\}$	$\{5/6\}$

to us to allow us to identify $\mathcal{M}^*$ with certainty. For convenience, we define

$$\mathrm{L}(\theta, \mathbf{m}) := \mathrm{L}(M(\theta), \mathbf{m}).$$

We do not know $\mathcal{M}^*$, so we cannot evaluate the performance of $\mathbf{m}$. This impasse can be overcome by considering strategies. A strategy f is a function that assigns a distribution on models to *each* outcome, i.e. a conditional distribution on models given outcomes. A strategy f can be pitted against a possible world θ, yielding the following expected loss, or *risk*

$$\mathrm{R}(\theta, f) := \mathop{\mathrm{E}}_{Y \sim \mathbf{p}_\theta} \Big[\mathrm{L}\left(\theta, f(Y)\right) \Big]. \tag{3.1}$$

Hence the problem becomes this: *find a strategy f that minimises the worst-case risk*, i.e. attains

$$\overline{\mathrm{V}} = \inf_f \sup_{\theta \in \mathcal{T}} \mathrm{R}(\theta, f).$$

We will show in §3.4.2 that there is always an f that attains $\overline{\mathrm{V}}$.

3.1.4 Assumptions

We must assume that $|\mathbb{M}| \geq 2$ for there to be a truth-finding *problem*, and we must assume $|\mathcal{Y}| \geq 2$ for there to be a basis for a solution. To simplify analysis, we make the following additional assumptions.

Assumption 3.2. The set of outcomes $\mathcal{Y}$ is finite.

Assumption 3.3. The set of models $\mathbb{M}$ is finite.

These assumptions, albeit restricting, capture many practically interesting cases, for instance the biased coin example. In the anvil drop example, the outcomes are time measurements. The set of time measurements is in principle uncountable, but it is naturally discretised by contemporary stopwatch manufacturers, who supply a fixed number of decimal digits.

These assumptions allow us to identify distributions on $\mathcal{Y}$ and $\mathbb{M}$ with points in the unit $|\mathcal{Y}|$- and $|\mathbb{M}|$-simplices. Recall from Definition 2.6 that the unit n-simplex is given by

$$\Delta_n := \left\{ p \in \mathbb{R}^n_+ \mid p^T \mathbf{1} = 1 \right\}.$$

This identification allows us to regard each model $\mathcal{M} \in \mathbb{M}$ as a subset of $\Delta_{|\mathcal{Y}|}$.

We do not have to place any restrictions on the set of possible worlds, nor on the mechanics.[1] We see this as a passed sanity check; these are the two places where the complexity of real-world applications will be reflected.

3.2 Truth-finding game

Worst-case expected-value optimisation problems like the truth-finding problem are naturally thought of as two-player games with chance moves, and this viewpoint proves fruitful in our case. The players in the truth-finding game are

[1] Of course, there are the standard measurability conditions. We list all the conditions in Appendix A.

called Learner, Nature, and Chance. A run of the *truth-finding game* consists of three steps. First, Nature, the antagonist, chooses the actual world from the list of possible worlds. Second, Chance, an independent and impartial player, generates data from the actual world. Third, Learner, the protagonist, uses these data to update his information about the label of reality, and express his beliefs about the label of the actual world in terms of a probability distribution. Learner's loss is given by the log loss. The complete game is shown in Protocol 3.1.

Protocol 3.1 The truth-finding game

Arena: Truth-finding frame $\mathfrak{F} = \langle \mathcal{Y}, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}} \rangle$.

1: Nature covertly chooses a hypothesis $\theta^* \in \mathcal{T}$. Say $M(\theta^*) = \mathcal{M}^*$.
2: Chance generates an outcome y according to $\mathbf{p}_{\theta^*}$.
3: Learner expresses his belief about the true model as a distribution $\mathbf{m}$ on models.

Loss: Learner suffers $-\log \mathbf{m}(\mathcal{M}^*)$.

3.2.1 Many outcomes

Protocol 3.1 states that Chance generates a single outcome Y from $\mathcal{Y}$. If the outcome space is "small" compared to the possible world space, then a single outcome is not very informative. We can of course have Chance repeat the data-generating process n times. The n outcomes are then, by construction, independent identically distributed (i.i.d.). Sequences of n i.i.d. outcomes can be modelled by a single outcome in a product frame.

Definition 3.4. Let $\mathfrak{F} = \langle \mathcal{Y}, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}} \rangle$ be a frame, and let n be a given number of repetitions. Then the *n-fold product frame* is given by

$$\mathfrak{F}^n := \langle \mathcal{Y}^n, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}_\theta^n \rangle_{\theta \in \mathcal{T}} \rangle , \tag{3.2}$$

where

$$\mathbf{p}_\theta^n(y^n) := \prod_{i=1}^{n} \mathbf{p}_\theta(y_i).$$

With this reduction in place, it suffices to consider only single-outcome truth-finding games.

3.2.2 Extensive form game

Protocol 3.1 specifies a two-player game of imperfect information with chance moves. The extensive form representation of the Binary Biased Coin example is shown in Figure 3.2. This particular example has been chosen for simplicity, but it is typical. The game tree always has the same structure. For different frames, only the probabilities along the edges and the fanout at each level of the game tree change. Nature's move is performed covertly, i.e., Learner is not informed of her choice of world. Hence Learner cannot distinguish the situations after Nature's move. A collection of indistinguishable situations for a particular player is called an *information set* for that player. Information sets are properties of game trees. The information set for Learner after Nature's move is indicated

in Figure 3.2 by a dotted line. After Chance's move, which is made publicly, Learner generally remains uninformed about the exact position in the game tree. This is indicated by the two lower dotted lines in Figure 3.2, which connect indistinguishable states for Learner, i.e. positions in which Chance has selected the same outcome.

The truth-finding game in Protocol 3.1 and its extensive form are essentially sequential. First Nature chooses the world. Second, Chance generates an outcome, according to a distribution that depends on the world chosen by Nature's. Third, Learner chooses a distribution on models, which depends on the outcome that Chance generated. Games with such dependencies are complicated to analyse. We can remove these dependencies by transforming the game into normal form, at the cost of increasing the complexity of the moves.

3.2.3 Normal form game

Each extensive form game has an associated normal form game. The former is represented by a game tree, whereas the latter is a game in the sense of Definition 2.33. In the normal form game, the moves for a player are given by his strategies in the extensive form game. Recall from Definition 2.42 that a strategy for a player in an extensive form game is a function that specifies what that player will do at each position in the game tree she can encounter. A game in normal form is parallel instead of sequential. Both players can choose their moves independently, as all dependencies are now captured by the moves, which internalise the strategies.

We first consider pure strategies, that is, strategies that deterministically choose moves in the extensive form game. Then we turn to mixed strategies: probability distributions over pure strategies.

3.2.4 Pure strategies

A pure strategy for a player assigns a legal move to each information set for that player where it is this player's turn. Nature makes a single move, at the start of a run of the game, hence a pure strategy for Nature is a choice of reality from the possible worlds. Strategies of Nature are identified with elements of $\mathcal{T}$.

When it is Learner's turn, the data have already been generated. A move for Learner consists of choosing a distribution on models. A pure strategy for Learner assigns such a distribution to *each* possible realisation of the data. So Learner's strategies are elements of $\mathcal{F} := \left[\mathcal{Y} \to \mathfrak{D}(\mathbb{M})\right]$. Note that $\mathcal{F}$ is not finite, although by our assumptions $\mathcal{Y}$ and $\mathbb{M}$ are.

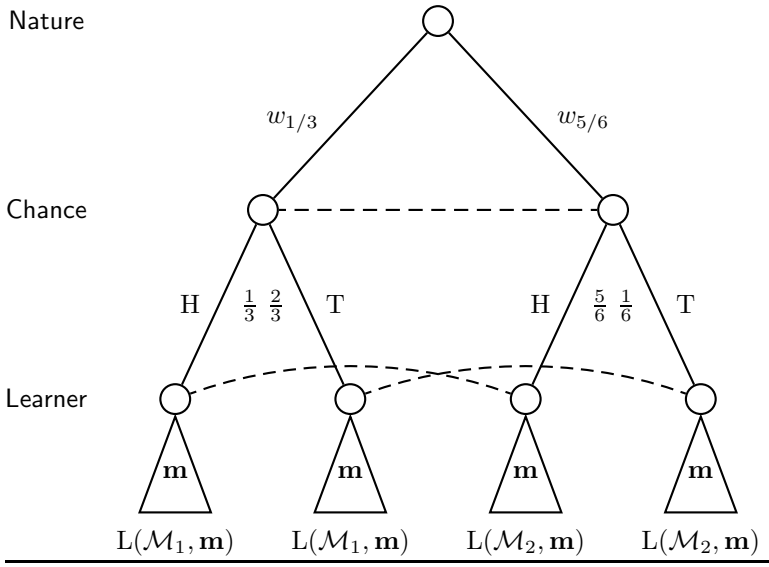
Definition 3.5. When Nature plays $\theta \in \mathcal{T}$ and Learner plays $f \in \mathcal{F}$, then the risk of Learner is given by

$$\mathrm{R}(\theta, f) := \mathop{\mathrm{E}}_{Y \sim \mathbf{p}_\theta} \Big[\mathrm{L}\left(\theta, f(Y)\right) \Big]. \tag{3.3}$$

The expected loss eliminates Chance by taking an expectation over outcomes.

Definition 3.6. Given a truth-finding frame $\mathfrak{F}$, the triple $\mathfrak{G}_\mathfrak{F} := \langle \mathcal{T}, \mathcal{F}, \mathrm{R} \rangle$ is called the *normal form* of the truth-finding game. It is a game in the sense of Definition 2.33.

Figure 3.2 Truth-finding game tree for the BBC example. The information sets of Learner after Nature's and Chance's moves are indicated by dotted lines. We denote the possible worlds by $w_{1/3}$ and $w_{5/6}$ to disambiguate world names and probabilities.



3.2.5 Mixed strategies

We now consider mixed strategies in the game $\mathfrak{G}_{\mathfrak{F}}$. Recall (Definition 2.37) that a mixed strategy is a probability distribution on pure strategies. In some situations, playing a mixed strategy can, in expectation, yield strictly lower risk worst-case risk than playing any single pure strategy. As a fortunate side effect, mixed strategies are also easier to analyse. The set of mixed strategies for Nature is $\mathfrak{D}(\mathcal{T})$. The set of mixed strategies for Learner is $\mathfrak{D}(\mathcal{F})$. $\mathfrak{D}(\mathcal{F})$ is a set of distributions on a set of conditional distributions, which may seem excessively complicated. Luckily, we can restrict attention to a much simpler set. The following is included for completeness.

Definition 3.7. When Nature plays $\mathsf{P}_\Theta \in \mathfrak{D}(\mathcal{T})$ and Learner plays $\mathsf{P}_F \in \mathfrak{D}(\mathcal{F})$, then the expected risk of Learner is given by

$$\mathrm{R}(\mathsf{P}_\Theta, \mathsf{P}_F) = \mathop{\mathrm{E}}_{\Theta \sim \mathsf{P}_\Theta} \mathop{\mathrm{E}}_{F \sim \mathsf{P}_F} \mathop{\mathrm{E}}_{Y \sim \mathbf{p}_\Theta} \Big[\mathrm{L}\left(\Theta, F(Y)\right) \Big]. \tag{3.4}$$

We use Θ, F and Y as random variables ranging over the possible worlds $\mathcal{T}$, Learner's pure strategies $\mathcal{F}$ and the observations $\mathcal{Y}$. Application of the risk function to a probability distribution on either argument will always be interpreted as an expected value, and we will refer to it as just *risk*.

Theorem 2.32 shows that it is never strictly beneficial for Learner to use a mixed strategy. For any mixed strategy of Nature P_Θ, and for all mixed strategies of Learner P_F:

$$\mathrm{R}(\mathsf{P}_\Theta, \mathsf{P}_F) \geq \mathrm{R}\left(\mathsf{P}_\Theta, \mathrm{E}\left[\mathsf{P}_F\right]\right)$$

Pure strategies for Learner already encompass the full power of randomisation. We will not consider mixed strategies for Learner any further. Summarising,

Definition 3.8. When Nature plays $\mathsf{P}_\Theta \in \mathfrak{D}(\mathcal{T})$ and Learner plays $f \in \mathcal{F}$, then the expected risk of Learner is given by

$$\mathrm{R}(\mathsf{P}_\Theta, f) = \mathop{\mathrm{E}}_{\Theta\sim\mathsf{P}_\Theta} \mathop{\mathrm{E}}_{Y\sim\mathbf{p}_\Theta} \Big[\mathrm{L}\left(\Theta, f(Y)\right)\Big]. \tag{3.5}$$

We emphasise that mixed strategies cannot be disregarded for Nature. In particular, in later analysis, we will require Nature to disclose her strategy to Learner before Learner chooses his. Now, if Nature plays a pure strategy, then Learner can always achieve zero loss. (Knowing the actual world, Learner can place all probability on the true model.) On the other hand, when Nature uses a randomised strategy in this scenario, then Learner cannot do this anymore.

Example 3.9 (BBC ctd.)**.** A mixed strategy for Nature is a distribution on $\mathcal{T} = \{1/3, 5/6\}$. Assume Nature chooses P_Θ to be the uniform distribution on the two possible worlds. A pure strategy for Learner is a distribution on $\mathbb{M}$ for each outcome. Learner could for example play the strategy f, given by

f	$\mathcal{M}_1$	$\mathcal{M}_2$
T	3/4	1/4
H	1/4	3/4

The expected risk of Learner is given by

$$\begin{aligned}\mathrm{R}(\mathsf{P}_\Theta, f) &= -\frac{1}{2}\cdot\left(\frac{2}{3}\cdot\log\frac{3}{4} + \frac{1}{3}\cdot\log\frac{1}{4}\right) - \frac{1}{2}\cdot\left(\frac{1}{6}\cdot\log\frac{1}{4} + \frac{5}{6}\cdot\log\frac{3}{4}\right) \\ &\approx 0.8113\end{aligned} \tag{3.6}$$

This strategy attains quite a respectable risk in this particular case. We will show in Example 3.20 that the optimal strategy always attains risk 0.8044.

3.2.6 The joint space

A truth finding frame $\mathfrak{F}$, and a mixed strategy P_Θ for Nature together induce a joint distribution P on possible worlds and outcomes. P_Θ gives a distribution on worlds, and for each world θ, $\mathbf{p}_\theta$ gives a conditional distribution on outcomes. If the set of worlds is countable, then this distribution is specified by

$$\mathsf{P}(\theta, y) = \mathsf{P}_\Theta(\theta)\mathbf{p}_\theta(y).$$

For general distributions on worlds, we have the following. The sample space of P is $\mathcal{T}\times\mathcal{Y}$. Its events are given by the product of the partaking σ-algebras. The joint distribution P is given by

$$\mathsf{P}(E_\mathcal{T}\times E_\mathcal{Y}) := \int_{E_\mathcal{T}} \mathbf{p}(E_\mathcal{Y}|\theta)\mathsf{P}_\Theta(\mathrm{d}\theta). \tag{3.7}$$

We use the following random variables on the joint space:

$$\begin{aligned}\Theta &: (\mathcal{T}\times\mathcal{Y}) \to \mathcal{T} \quad \langle\theta, y\rangle \mapsto \theta \\ Y &: (\mathcal{T}\times\mathcal{Y}) \to \mathcal{Y} \quad \langle\theta, y\rangle \mapsto y \\ M &: (\mathcal{T}\times\mathcal{Y}) \to \mathbb{M} \quad \langle\theta, y\rangle \mapsto M(\theta).\end{aligned} \tag{3.8}$$

These random variables should technically be called *pseudo random variables*, as they map into sample spaces that are not $\langle \mathbb{R}, \mathcal{B} \rangle$. We will only use them as building blocks to construct proper random variables.

These definitions allow us to write the risk as

$$\mathrm{R}(\mathsf{P}_\Theta, f) = \mathrm{E}_{\mathsf{P}} \left[\mathrm{L}\left(M, f(Y)\right) \right]. \tag{3.9}$$

We will henceforth omit P whenever it is clear from the context.

3.3 Representing strategies

In this section, we provide representations for strategies of both Learner and Nature. The representation of Learner's strategies is quite straightforward: his strategies correspond to stochastic matrices. For Nature, we reap the benefits of focusing on finding the true model, instead of finding the true world. In our framework, Nature's strategies admit a simple representation. We show that a strategy for Nature can be decomposed into a distribution on models, and, for each model, a distribution on outcomes, which must be chosen from the convex hull of that model.

3.3.1 Learner's strategies

We showed that Learner does not need mixed strategies, and this simplifies representation considerably. Let $n = |\mathcal{Y}|$ and $k = |\mathbb{M}|$. A pure strategy f for Learner corresponds to a $n \times k$ matrix A, defined by

$$A_{y,\mathcal{M}} = f(\mathcal{M}|y).$$

The matrix A is a *stochastic matrix*, i.e. all entries are from $[0, 1]$, and rows sum to one. This perspective allows a clear visualisation of Learner's strategies.

Definition 3.10. An *equaliser strategy* for Learner is a strategy f, such that $\mathrm{R}(\cdot, f)$ is constant. That is, no matter what Nature does, the risk is always the same.

As the set of models is finite, there is an equaliser strategy for Learner, namely the strategy *just guessing*. This strategy assigns the uniform probability distribution on $\mathbb{M}$ to each outcome. It is easily seen that the risk of this strategy is given by $\log |\mathbb{M}|$.

3.3.2 Nature's strategies

A mixed strategy for Nature is a distribution P_Θ on possible worlds. Such a distribution induces the distribution P on the joint space $\mathcal{T} \times \mathcal{Y}$. Observe, for example in (3.9), that Learner's risk depends on the actual world that Nature chooses only via the true model and the generated outcome. Therefore it is useful to characterise the distributions on $\mathbb{M} \times \mathcal{Y}$ that Nature can realise.

For ease of exposition, we pretend that $\mathcal{T}$ is countable, to be able to write sums instead of integrals. The reader may check that the reasoning extends to the uncountable case. We have

$$\mathsf{P}(\mathcal{M}, y) = \mathsf{P}(y|\mathcal{M})\mathsf{P}(\mathcal{M}), \quad \text{where} \quad \mathsf{P}(y|\mathcal{M}) = \sum_{\theta \in \mathcal{M}} \mathbf{p}(y|\theta)\mathsf{P}(\theta|\mathcal{M}). \tag{3.10}$$

In this formula, $\mathbf{p}$ is fixed by the truth-finding frame. Both $\mathsf{P}(\theta|\mathcal{M})$ and $\mathsf{P}(\mathcal{M})$ are determined by Nature's strategy, from the joint

$$\mathsf{P}(\theta, \mathcal{M}) = \begin{cases} \mathsf{P}(\theta) & \text{if } \theta \in \mathcal{M}, \\ 0 & \text{otherwise.} \end{cases}$$

We give names to three important quantities in (3.10):

- The distribution $\mathsf{P}(M)$ on models is called the *prior on models*. We use the fact that models are events over $\mathcal{T}$.
- For each model $\mathcal{M}$, the distribution $\mathsf{P}(\Theta|M = \mathcal{M})$ on the worlds in $\mathcal{M}$ is called the *within-model prior* of $\mathcal{M}$.
- The term $\sum_{\theta\in\mathcal{M}} \mathbf{p}(y|\theta)\mathsf{P}(\theta|\mathcal{M})$, which we will subsequently abbreviate to $\mathsf{P}(y|\mathcal{M})$ is called the *within-model marginal*.

In (3.10), we separate $\mathsf{P}(\theta)$ into a prior distribution $\mathsf{P}(\mathcal{M})$ and a within-model prior distribution $\mathsf{P}(\theta|\mathcal{M})$, but we can also reverse this process, and obtain any distribution on $\mathcal{T}$ from a prior and a within-model prior.

The within-model marginal is a distribution on outcomes for each model $\mathcal{M}$. It is obtained by weighing the distributions $\mathbf{p}_{[\mathcal{M}]} := \langle \mathbf{p}_\theta \rangle_{\theta\in\mathcal{M}}$ according to the within-model prior weights. It is an expected distribution, where the partaking distributions are fixed (they are specified by the truth-finding frame), while the relative contribution of each possible world in the model is controlled by Nature. The set of all distributions that can be obtained in this way is, by definition, equal to the convex hull of $\mathbf{p}_{[\mathcal{M}]}$.

For each model, Nature can choose the within-model prior *independently* of the other within-model priors. So, by varying her strategy, Nature can achieve each combination of within-model marginals from the convex hull of each of the models. Finally, by choosing the prior on models, Nature fixes $\mathsf{P}(\mathcal{M}, y)$ in (3.10). The prior can also be chosen arbitrarily, independent of the within-model priors. This leads to the following characterisation:

> Each possible strategy for Nature can, for the purpose of truth finding, be equivalently represented by
>
> - a distribution P_Θ on possible worlds; or
> - a prior on models $\mathsf{P}(M)$, and a within-model prior on possible worlds $\mathsf{P}(\Theta|M = \mathcal{M})$ for each model $\mathcal{M}$; or
> - a prior on models, $\mathsf{P}(M)$, and, for each model $\mathcal{M}$, a distribution $\mathbf{p}_\mathcal{M}$ on outcomes, chosen from the convex hull of $\mathbf{p}_{[\mathcal{M}]}$.

The third representation is important enough to receive a name.

Definition 3.11. A pair $\left\langle \mathsf{P}, \langle \mathbf{p}_\mathcal{M} \rangle_{\mathcal{M}\in\mathbb{M}} \right\rangle$ is called *a collapsed strategy for* Nature if P is a probability distribution on $\mathbb{M}$, and $\mathbf{p}_\mathcal{M} \in \text{conv}(\mathbf{p}_{[\mathcal{M}]})$ for each $\mathcal{M} \in \mathbb{M}$.

Example 3.12. Consider the instance of the truth-finding problem given in Figure 3.3. In this case $|\mathcal{Y}| = 3$, $|\mathcal{T}| = 23$ and $|\mathbb{M}| = 2$. Finiteness of $\mathcal{T}$ implies that each model is also finite, thus each model corresponds to a finite set of points in $\mathbb{R}^3$. The convex hull of a finite set is a polyhedron, and in Δ_3 these are polygons.

3.4 Solution of the truth-finding game

The previous sections introduced the truth-finding game. Now we show that it can be solved. We first prove that the truth-finding game has a value, and that Learner always has a minimax strategy. Then we provide an algorithm to find the value. We start by eliminating a pathological case, in which the value can be obtained immediately.

3.4.1 Triviality

We showed in the previous section that a strategy for Nature can be decomposed into a probability distribution on models, and, per model, a within-model distribution on possible worlds in that model. The per-model distribution on possible worlds can be summarised (by taking the expectation) into a per-model distribution on outcomes, that is, a single distribution. This distribution necessarily lies in the convex hull of the model. A strategy for Nature can thus be reduced to a finite mixture of distributions on outcomes.

When the convex hulls of all models share a common world θ, we call the truth-finding problem trivial. This is because the best strategy for nature is to choose the prior over models uniform, and all within-model priors such that they put all probability on θ. Then no outcome provides information that helps differentiating the models, and Learner can perform no better than just guessing, suffering loss equal to $\log |\mathbb{M}|$.

3.4.2 Value

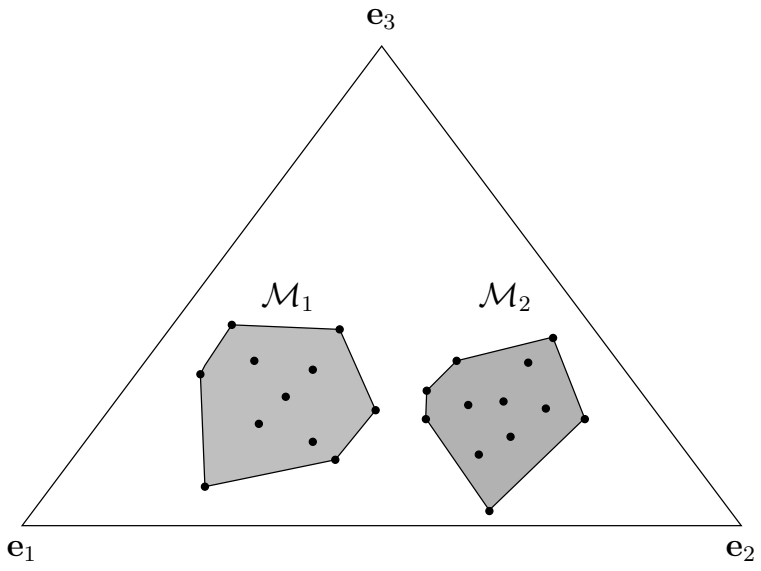
The truth-finding game has a value (Definition 2.34) if

$$\sup_{\mathsf{P}_\Theta} \inf_{f} \mathrm{R}(\mathsf{P}_\Theta, f) = \inf_{f} \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, f).$$

The right hand expression is the lowest risk that Learner can guarantee in the game where he moves first, and Nature, after learning Learner's choice, moves second. Analogously, the left hand expression is the highest risk that Nature can guarantee when she has to play first, and Learner chooses his move after learning her move. We will show that the truth-finding game has a value, hence there is no advantage to playing second. We denote by $\mathrm{V}(\mathfrak{F})$ the *value of the truth-finding frame on the arena* $\mathfrak{F}$.

Note that in general it is not necessarily the case that strategies exist that witness this value. A distribution P_Θ that attains the left hand side is called *least favourable*. A strategy f that attains the right hand side is called *minimax*. We will show that a minimax strategy always exists in the truth-finding game.

Figure 3.3 Convex hull of models. The triangle shows the projection of Δ_3 onto the plane. Each model is a set of points in $\mathbb{R}^3$. The shaded areas are the convex hulls of the models.



Definition 3.13. A set $C \subseteq \mathfrak{D}(\mathcal{F})$ is *essentially complete* [Fer67] if given any $\mathsf{P}_F \in \mathfrak{D}(\mathcal{F})$, there is a $\mathsf{P}'_F \in C$ such that

$$\forall \theta : \mathrm{R}(\theta, \mathsf{P}'_F) \leq \mathrm{R}(\mathsf{P}_F, \theta).$$

Exposing an essentially complete class simplifies matters, for there is no reason to consider strategies outside of that class anymore. One can always find strategies within the class that are just as good. In our case, the set of pure strategies for Learner is essentially complete. In the following theorems, we interpret each pure strategy $f \in \mathcal{F}$ as a mixed strategy that puts all probability on f.

Theorem 3.14. $\mathcal{F}$ is essentially complete.

Proof. Theorem 2.30. □

The following theorem will allow us to conclude that the truth-finding game has a value and that a minimax strategy exists. The remainder of this section is spent verifying the preconditions of this theorem for the truth-finding game.

Theorem 3.15 ([Fer67, Theorem 2.9.2])**.** Let C be essentially complete for the game $\langle \mathcal{T}, \mathcal{F}, \mathrm{R} \rangle$. Assume there is a topology on C such that

- C is compact and
- R is lower semi-continuous in $f \in C$ for all $\theta \in \mathcal{T}$.

Then the game has a value, i.e.,

$$\sup_{\mathsf{P}_\Theta} \inf_f \mathrm{R}(\mathsf{P}_\Theta, f) = \inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, f).$$

Moreover, a minimax f attaining $\inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, f)$ exists in C.

Theorem 3.16. $\mathcal{F}$ is compact.

Proof. Using Assumption 3.2 and Assumption 3.3, let $n = |\mathcal{Y}|$ and $k = |\mathbb{M}|$. Then

$$\mathcal{F} = \left[\mathcal{Y} \to \mathfrak{D}(\mathbb{M})\right] \cong (\Delta_k)^n \tag{3.11}$$

is closed and bounded, hence compact, in $\mathbb{R}^{nk}$. □

Definition 3.17. The extended logarithm function $\underline{\log}(x) : \mathbb{R} \to \overline{\mathbb{R}}$ is defined by

$$\underline{\log}(x) = \begin{cases} \log(x) & \text{if } x > 0, \\ -\infty & \text{otherwise.} \end{cases}$$

Lemma 3.18. $\underline{\log}$ is continuous.

Proof. We need to show that $\underline{\log}$ is both upper and lower semi-continuous. The intervals of type $[-\infty, r)$ and $(r, +\infty]$, with $r \in \mathbb{R}$, are a subbasis for the order topology on $\overline{\mathbb{R}}$. We need to show that the following sets are open for all $r \in \mathbb{R}$

$$\text{(lower)} \qquad \underline{\log}^{-1}\big((r, +\infty]\big) = \big\{x \in \mathbb{R} \mid \underline{\log}(x) > r\big\} = (2^r, +\infty) \tag{3.12}$$

$$\text{(upper)} \qquad \underline{\log}^{-1}\big([-\infty, r)\big) = \big\{x \in \mathbb{R} \mid \underline{\log}(x) < r\big\} = (-\infty, 2^r) \tag{3.13}$$

Both sets are open intervals, hence open sets in the order topology on $\mathbb{R}$, for each $r \in \mathbb{R}$. We conclude that $\underline{\log}$ is continuous. □

Theorem 3.19. $\mathrm{R}(\theta, f)$ is continuous in f for all $\theta \in \mathcal{T}$.

Proof. Fix θ. We have

$$\mathrm{R}(\theta, f) = \mathrm{E}_{Y \sim \mathbf{p}_\theta}\left[-\log f(M(\theta)|y)\right] \tag{3.14}$$

$$= -\sum_y \mathbf{p}_\theta(y) \log f(M(\theta)|y) \tag{3.15}$$

A weighted sum of finitely many continuous functions is continuous. It remains to show that for given $y, \mathcal{M}$, the function $f \mapsto \log f(\mathcal{M}|y)$ is continuous. We already proved that the (extended) logarithm function is continuous on $\mathbb{R}$ in a single argument, so it is also continuous on the subspace topology on $[0, 1]$. Continuous functions are closed under composition, hence it suffices to show that the function $f \mapsto f(\mathcal{M}|\theta)$ is continuous. The preimage of a basic open set $[0, r)$ or $(r, 1]$ under this function is the intersection of $\mathcal{F}$ with an open half-space, hence it is open. □

We have now achieved the central result of this chapter. We proved in Theorems 3.16 and 3.19 that the preconditions of Theorem 3.15 hold for the truth-finding game, hence we can conclude:

The truth-finding game has a value, and furthermore, Learner has a minimax strategy.

Example 3.20 (BBC ctd.)**.** The following tables show the minimax strategies for Learner in the binary biased coin model, for $n = 1, 2, 3$ outcomes.

$n=1$, $V=0.8044$			$n=2$, $V=0.6565$			$n=3$, $V=0.5399$		
#H	$\mathbf{m}(\mathcal{M}_1)$	$\mathbf{m}(\mathcal{M}_2)$	#H	$\mathbf{m}(\mathcal{M}_1)$	$\mathbf{m}(\mathcal{M}_2)$	#H	$\mathbf{m}(\mathcal{M}_1)$	$\mathbf{m}(\mathcal{M}_2)$
0	0.8038	0.1962	0	0.9427	0.0573	0	0.9852	0.0148
1	0.2906	0.7094	1	0.6220	0.3780	1	0.8690	0.1310
			2	0.1413	0.8587	2	0.3988	0.6012
						3	0.0622	0.9378

As expected, the more heads Learner observes, the more he should prefer the model saying that the coin favours heads. Also, the more outcomes, the more information is achieved in expectation.

3.5 Computing the minimax strategy

We have established that the truth-finding game $\langle \mathfrak{D}(\mathcal{T}), \mathcal{F}, \mathrm{R} \rangle$ has a value, and a minimax strategy for Learner exists. More precisely,

$$\sup_{\mathsf{P}_\Theta} \inf_{f} \mathrm{R}(\mathsf{P}_\Theta, f) = \mathrm{V} = \inf_{f} \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, f),$$

where the right-hand side is attained for some strategy f. We now consider the problem of computing that strategy f. This is important, because a straightforward numerical approach, even for a moderate number of outcomes, quickly becomes infeasible.

In this section, we will first show in §3.5.1 that we can find a minimax strategy for Learner in a quire restricted set, namely in the set of strategies that are the optimal response to a strategy of Nature. This is very helpful, as strategies for Nature have far lower complexity than strategies for Learner.

Subsequently, we will show in §3.5.2 that we can directly optimise the left-hand side of the preceding equation. This yields the least favourable distribution, if it exists. Then we apply an important insight from Game Theory, namely, that the minimax strategy is the optimal response to the least favourable distribution. This leads to the somewhat surprising conclusion that Learner, to play optimally, *must assume* that Nature will play according to the least favourable distribution.

If the least favourable distribution does not exist, then there is a chain of less and less favourable distributions, whose limit is not a strategy for Nature. In this case, we can apply the same reasoning. We will conjecture that the limit of the corresponding chain of optimal response strategies is always a strategy for Learner, and that this strategy is minimax.

3.5.1 Extended Bayes

We have shown that the pure strategies for Learner form an essentially complete class. The following allows us to restrict attention even further, namely, to the class of strategies that are an optimal response to some mixed strategy for Nature. The notion of optimal response is formalised by the following two definitions.

Definition 3.21. A strategy $f \in \mathcal{F}$ is called ϵ-*Bayes* with respect to a distribution P_Θ on $\mathcal{T}$ if

$$\mathrm{R}(\mathsf{P}_\Theta, f) \leq \inf_{f \in \mathcal{F}} \mathrm{R}(\mathsf{P}_\Theta, f) + \epsilon.$$

We abbreviate 0-Bayes to just *Bayes*.

A Bayes strategy for Learner is the optimal response to a particular strategy for Nature. When dealing with chains of less and less favourable distributions, we need the following extension.

Definition 3.22. A strategy $f \in \mathcal{F}$ is called *extended Bayes* for a collection D of distributions on $\mathcal{T}$ if f is ϵ-Bayes with respect to some $\mathsf{P}_\Theta \in D$ for each $\epsilon > 0$.

Theorem 3.23 (Complete Class Theorem, [Fer67, Theorem 2.10.3])**.** Let C be essentially complete for the game $\langle \mathcal{T}, \mathcal{F}, \mathrm{R} \rangle$ satisfying the preconditions of Theorem 3.15. Then the set of extended Bayes strategies in C is essentially complete.

So we have that (a) there is a minimax strategy for Learner in the truth-finding game, and (b) if there is a minimax strategy, then there is an extended Bayes minimax strategy. So we only need to consider extended Bayes strategies to find a minimax strategy.

Example 3.24 (BC ctd.)**.** We can now formalise the intuition that the Biased Coin problem is trivial. Recall that in the Biased Coin problem, we have

$$\mathcal{T} = \left[\tfrac{1}{2}, 1\right], \quad \mathcal{M}_1 = \left\{\tfrac{1}{2}\right\}, \quad \text{and} \quad \mathcal{M}_2 = \left(\tfrac{1}{2}, 1\right]$$

Now consider a distribution P_Θ on possible worlds. By §3.3.2, we can represent P_Θ by a distribution $\mathsf{P}(M)$ on models, and for each model, a distribution on outcomes from within the convex hull of that model. In our example, the models equal their convex hulls. We can find two numbers $\alpha \in [0, 1]$ and $\beta \in \mathcal{M}_2$, such that upon defining

$$\begin{array}{lll} \mathsf{P}(\mathcal{M}_1) = \alpha & \mathbf{p}_{\mathcal{M}_1}(\mathrm{H}) = \frac{1}{2} & \mathbf{p}_{\mathcal{M}_2}(\mathrm{H}) = \beta \\ \mathsf{P}(\mathcal{M}_2) = 1 - \alpha & \mathbf{p}_{\mathcal{M}_1}(\mathrm{T}) = \frac{1}{2} & \mathbf{p}_{\mathcal{M}_2}(\mathrm{T}) = 1 - \beta \end{array} \tag{3.16}$$

we have

$$\mathsf{P}(y, \mathcal{M}_i) = \mathsf{P}(\mathcal{M}_i)\mathbf{p}_{\mathcal{M}_i}(y).$$

The distribution in the previous line is explicitly spelled out below, together with the best response for Learner.

$\mathsf{P}(y, \mathcal{M}_i)$

	$\mathcal{M}_1$	$\mathcal{M}_2$
H	$\alpha/2$	$(1-\alpha)\beta$
T	$\alpha/2$	$(1-\alpha)(1-\beta)$

$\mathsf{P}(\mathcal{M}_i|y)$

	$\mathcal{M}_1$	$\mathcal{M}_2$
H	$\frac{\alpha/2}{\alpha/2+(1-\alpha)\beta}$	$\frac{(1-\alpha)\beta}{\alpha/2+(1-\alpha)\beta}$
T	$\frac{\alpha/2}{\alpha/2+(1-\alpha)(1-\beta)}$	$\frac{(1-\alpha)(1-\beta)}{\alpha/2+(1-\alpha)(1-\beta)}$

Now consider what happens when $\alpha = 1/2$ while $\beta \to 1/2$. Then all entries of the above right table tend to $1/2$. This means that Nature has strategies to which the best response for Learner lies arbitrarily close to just guessing. The actual value $\beta = 1/2$ is not a valid choice for Nature, as it is not present in $\mathcal{M}_2$.

Let D be the collection of distributions that Nature can achieve with $\alpha = 1/2$ and $\beta \in \left(1/2, 1\right]$. Learner's strategy of just guessing is extended Bayes with respect to D, while it is not Bayes with respect to any mixed strategy for Nature. The following section will allow us to conclude that just guessing is the minimax strategy for Learner.

3.5.2 Generalised entropy

To find the minimax strategy for Learner, we look at the *best-response-value function*, or *generalised entropy* [GD04], which is given by

$$\begin{aligned}\mathfrak{H}(\mathsf{P}_\Theta) &:= \inf_f \mathrm{R}(\mathsf{P}_\Theta, f) \\ &= \inf_f \mathop{\mathrm{E}}_{M,Y\sim\mathsf{P}} \left[-\log f(M|Y)\right].\end{aligned} \tag{3.17}$$

Recall that here P depends on Nature's strategy P_Θ. The generalised entropy function yields the risk as a function of Nature's strategy, when the optimal response for Learner is used. We will show that it is concave, hence it has no local maxima. The maximum can hence easily be found, at least numerically, using methods from [BV04]. The maximum is achieved for the least favourable distribution (the maximin strategy for Nature). This in turn will help us find the minimax strategy for Learner.

Application of Theorem 2.30 yields that the Bayes response f is given by

$$f(\mathcal{M}|y) = \mathsf{P}(M = \mathcal{M}|Y = y) \tag{3.18}$$

From this, one can easily derive

$$\mathfrak{H}(\mathsf{P}_\Theta) = \mathrm{E}_Y\left[\mathcal{H}(M|Y)\right] \tag{3.19}$$
$$= \mathcal{H}(Y, M) - \mathcal{H}(Y) \tag{3.20}$$
$$= \mathcal{H}(M) - \mathcal{I}(M; Y). \tag{3.21}$$

This last line is particularly telling. If Nature's mixed strategy is publicly announced before Learner has to choose his strategy, then Nature must try to minimise the amount of information $\mathcal{I}(M;Y)$ that the outcome contains about the generating model, while on the other hand maximising the uncertainty $\mathcal{H}(M)$ about the generative model inherent in her strategy.

Theorem 3.25 ([Fer67, p.90])**.** If P_Θ attains maximum generalised entropy, then f, defined in (3.18) as the best response to P_Θ, is a minimax strategy for Learner.

Theorem 3.26. The generalised entropy is bounded. More specifically,

$$0 \leq \mathfrak{H}(\mathsf{P}_\Theta) \leq \log|\mathbb{M}|.$$

Proof. We obtain non-negativity by (3.19), using the facts that entropies of finite random variables are nonnegative, and that expectations preserve this. For the upper bound, consider the (equaliser) strategy of Learner that assigns the uniform distribution on models to each outcome. This strategy has risk $\log|\mathbb{M}|$ irrespective of Nature's move. Now observe that this strategy partakes in the infimum of (3.17). □

Theorem 3.27. $\mathfrak{H}(\mathsf{P}_\Theta)$ is concave.

Proof. Given in Appendix B. The proof is analogous to the proof of the concavity of the normal entropy, which can be found in [CT90]. □

Corollary 3.28. If the convex hull of each model is closed, then there is a least favourable distribution P^*_Θ for Nature, i.e. a distribution for which

$$\mathfrak{H}(\mathsf{P}^*_\Theta) = \sup_{\mathsf{P}_\Theta} \mathfrak{H}(\mathsf{P}_\Theta).$$

We note that if the models are closed to begin with, then their convex hull is closed too.

Proof. The generalised entropy function $\mathfrak{H}$ is concave on $\mathfrak{D}(\mathcal{T})$. A concave function obtains its supremum on a closed set. Closedness of the convex hull of each model implies closedness of their finite union, $\mathcal{T}$, which in turn implies closedness of $\mathfrak{D}(\mathcal{T})$. □

Putting together Theorems 3.25, 3.27 and Corollary 3.28, we obtain the main result of this section:

> If the convex hulls of the models are closed, then a least favourable distribution exists. Moreover, this distribution can easily be found by maximising the concave generalised entropy function. The minimax strategy for Learner is the optimal response to it.

Example 3.29 (BBC ctd.)**.** We analyse the behaviour of the binary biased coin model for arbitrary biases. Let

$$P_1(\mathrm{H}) = \alpha \qquad\qquad P_2(\mathrm{H}) = \beta \tag{3.22}$$

A strategy for Nature specifies the mixing weights $\mathbf{w}$ over P_1 and P_2. To compute the optimal mixing weights, we consider the generalised entropy function. For n outcomes, we have

$$\mathfrak{H}(\mathbf{w}) = \mathrm{E}_{Y^n}\left[\mathcal{H}(M|Y^n)\right] \tag{3.23}$$

Define $n_\mathrm{H}(y^n) := \left|\{i \mid y_i = \mathrm{H}\}\right|$ and $n_\mathrm{T}(y^n) := \left|\{i \mid y_i = \mathrm{T}\}\right|$. When clear from the context, we will omit the argument y^n and just write $n_\mathrm{H}, n_\mathrm{T}$. Clearly $n_\mathrm{H} + n_\mathrm{T} = n$. For n i.i.d. outcomes, we have

$$P_1(y^n) = f_1(n_\mathrm{H}) \quad := \alpha^{n_\mathrm{H}}(1-\alpha)^{n_\mathrm{T}}, \tag{3.24}$$
$$P_2(y^n) = f_2(n_\mathrm{H}) \quad := \beta^{n_\mathrm{H}}(1-\beta)^{n_\mathrm{T}}, \tag{3.25}$$
$$P(y^n, j) = f(n_\mathrm{H}, j) := f_j(n_\mathrm{H})\mathbf{w}(j). \tag{3.26}$$

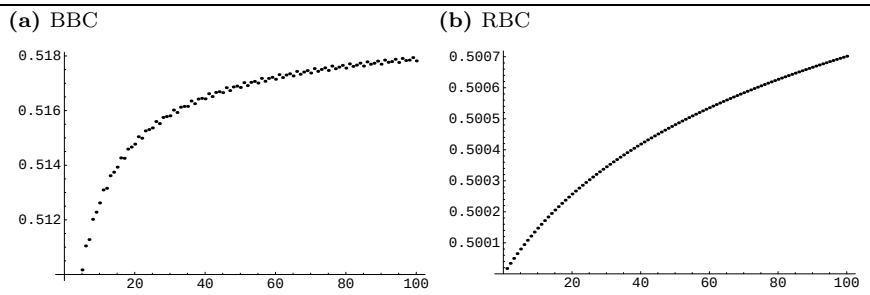
We expand the generalised entropy, and then combine all terms with the same number of heads.

$$\mathfrak{H}_n(\mathbf{w}) = -\sum_{y^n}\sum_{j=1}^{2} P(y^n, j) \log \frac{P(y^n, j)}{P(y^n, 1) + P(y^n, 2)} \tag{3.27}$$

$$= -\sum_{i=0}^{n} \binom{n}{i} \sum_{j=1}^{2} f(i, j) \log \frac{f(i, j)}{f(i, 1) + f(i, 2)} \tag{3.28}$$

In Figure 3.4a, we have plotted $\arg\max_{\mathbf{w}} \mathfrak{H}_n(\mathbf{w})$ as a function of n for $\alpha = 1/3$ and $\beta = 5/6$, i.e. the BBC example. Figure 3.4b shows the same function for $\alpha = 1/2$ and $\beta = 0.6$, i.e. the RBC example. Both graphs have been obtained by applying the Newton method to the concave function $\mathfrak{H}_n(\mathbf{w})$. We cannot explain the small-scale fluctuations in Figure 3.4a. They are most probably due to machine precision issues.

Figure 3.4 Least favourable distribution. Both plots show the least favourable prior weight of P_1 as a function of n.



Arbitrary models

It is strongly suggested by the results of this section that our main result can be extended to the case where the convex hulls of the models are not closed. The convex hull of a set is not necessarily closed, e.g. in the BC example, the convex hull of the second model, the fair coin model, is a half-open interval. There is, however, a standard way to obtain closed convex hulls.

Definition 3.30. Let $C \subseteq \mathbb{R}^n$ be a set. The *closed convex hull* of C is defined as the intersection of all closed convex sets containing C. The closed convex hull of C is the closure of the convex hull of C.

The current state of research only licenses a conjecture.

Conjecture 3.31. The minimax strategy f is the Bayes response to the least favourable distribution for Nature over the closed convex hulls of the models. In particular, when the closed convex hulls have a common intersection, then the truth-finding problem is trivial and Learner can do no better than just guessing.

Regarding Theorem 3.27, one might wonder whether the generalised entropy is strictly concave. In general, this is unfortunately not true. For example, take some truth-finding problem with finite models, and take some possible world θ that is in the support of the least favourable distribution. Now *duplicate* this world, i.e. add a fresh world θ' to the model $M(\theta)$, and set $\mathbf{p}_{\theta'} := \mathbf{p}_\theta$. It is easy to see that in the truth-finding problem on this new frame, the original least favourable distribution is still least favourable. By transferring all probability from θ to θ' we obtain a different least favourable distribution. Any mixture of these is least favourable too. We have a continuum of distributions with identical generalised entropy.

We believe that the representation of §3.3.2 circumvents this problem. The distributions on worlds that are used in the above construction have identical associated collapsed strategies.

Conjecture 3.32. $\mathfrak{H}$, regarded as a function of collapsed strategies for Nature, is strictly concave.

3.6 Similarity

By §3.3.2, each mixed strategy for Nature corresponds to a set of distributions on outcomes, one from the convex hull of each model, with associated mixing weights. Intuitively, for the least favourable distribution, these partaking distributions must be similar. Indeed, the more similar they are, the harder it will be for Learner to tell them apart on the basis of an outcome. The truth-finding problem defines a similarity measure, which we develop in this section. We have formulated our notion of similarity in general terms, as we believe that it is of independent interest.

The central question is the following: given an outcome, drawn from a mixture over probability distributions, what do we know about its origin? In other words, how hard is it to tell which probability distribution in the mixture was used to generated that outcome?

Definition 3.33. Let $\mathcal{Y}$ be a sample space, $\mathbf{P} = \langle P_j \rangle_{j \in \mathcal{J}}$ a family of distributions on $\mathcal{Y}$, and Q a distribution on $\mathcal{J}$. Let P denote the joint distribution on $\mathcal{J} \times \mathcal{Y}$ defined by $\mathsf{P}(j, y) := Q(j) \cdot P_j(y)$. We define the *$Q$-similarity* of $\mathbf{P}$ by

$$\mathrm{S}(\mathbf{P}, Q) = \mathop{\mathrm{E}}_{\langle J,Y \rangle \sim \mathsf{P}} \left[-\log \mathsf{P}(J|Y) \right] \tag{3.29}$$

$$= \mathop{\mathrm{E}}_{Y \sim \mathsf{P}} \left[\mathcal{H}(J|Y) \right] \tag{3.30}$$

This definition has the following interpretation. If nature generates an outcome (Y, J) from P, but only discloses Y, then our uncertainty about J, the actual distribution that generated Y, is measured by $\mathcal{H}(J|Y)$. Hence, $\mathrm{S}(\mathbf{P}, Q)$ is the average amount of information we obtain about J from an outcome generated according to Q.

Lemma 3.34. For all $\mathbf{P}, Q$,

$$0 \leq \mathrm{S}(\mathbf{P}, Q) \leq \log |\mathcal{J}| .$$

Proof. For all y, we have $0 \leq \mathcal{H}(J|Y = y) \leq \log |\mathcal{J}|$. S is defined as an expectation of such entropies, and expectations preserve bounds. □

Lemma 3.35. If Q is a point-distribution then

$$\mathrm{S}(\mathbf{P}, Q) = 0.$$

Proof. If Q is a point-distribution, then the random variable J is a constant. The entropy of a constant is zero, and the expectation of zero is zero. □

In the definition of Q-similarity, the mixing weights over the codes are given by Q. To obtain a general measure of similarity, we take the worst-case Q-similarity. That is, we maximise the similarity over all possible mixing weights.

Definition 3.36. Let $\mathbf{P} = \langle P_j \rangle_{j \in \mathcal{J}}$ be as before. We define the *similarity* of $\mathbf{P}$ by

$$\mathrm{S}(\mathbf{P}) = \max_{Q} \mathrm{S}(\mathbf{P}, Q) \tag{3.31}$$

We believe that there is a connection to the *information channel capacity* (see [CT90, p.184]), defined by

$$C := \max_{Q \text{ on } \mathcal{J}} \mathcal{I}(J;Y) \tag{3.32}$$

$$= \max_{Q \text{ on } \mathcal{J}} \Big(\mathcal{H}(J) - \mathrm{E}_Y \left[\mathcal{H}(J|Y) \right] \Big) \tag{3.33}$$

Intuitively, the distribution Q that attains the maximum in (3.31) *minimises* the average amount of information that the outcome Y transmits about its origin J. The exact relation is a matter for future research.

3.6.1 Koolen distance

We obtain a binary measure of similarity by instantiating (3.31) with $\mathcal{J} = \{1, 2\}$. By Lemma 3.34, the binary similarity takes values between 0 and 1. We now consider its opposite, which we baptise the *Koolen distance*, and abbreviate to *K-distance*. It is given by

$$\begin{aligned} \mathrm{d_K}(P_1, P_2) &:= 1 - \mathrm{S}\left(\{P_1, P_2\}\right) \\ &= 1 - \max_{Q \text{ on } \{1,2\}} \mathrm{E}_Y \left[\mathcal{H}(J|Y) \right]. \end{aligned} \tag{3.34}$$

This definition raises two natural questions. First: is $\mathrm{d_K}$ a metric? We will show that it satisfies minimality and symmetry, but that it violates the triangle inequality. Second: how does the K-distance relate to the Kullback-Leibler divergence (Definition 2.25), a well-known distance on probability distributions? We provide a partial answer in the form of graphical examples.

3.6.2 Metric

Definition 3.37. A function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ is called a *metric* if for all $x, y, z \in \mathcal{X}$ the following conditions hold:

$$d(x, y) = 0 \text{ iff } x = y \qquad \text{(minimality)} \tag{3.35}$$

$$d(x, y) = d(y, x) \qquad \text{(symmetry)} \tag{3.36}$$

$$d(x, z) \leq d(x, y) + d(y, z) \qquad \text{(triangle inequality)} \tag{3.37}$$

We show that $\mathrm{d_K}$ satisfies minimality and symmetry, but violates the triangle inequality.

Theorem 3.38. By definition, $\mathrm{d_K}$ is symmetric.

Theorem 3.39. $\mathrm{d_K}(P_1, P_2) = 0$ iff $P_1 = P_2$.

Proof. Suppose $\mathrm{d_K}(P_1, P_2) = 0$. Then $\max_Q \mathrm{E}_Y \left[\mathcal{H}(J|Y) \right] = 1$. Let Q be a distribution on $\{1, 2\}$ that achieves Q-similarity one. By Lemma 3.35, $Q(1), Q(2) > 0$. One is the maximum achievable entropy on a 2 element set, hence for each y in the support of either P_1 or P_2, we have $\mathcal{H}(J|Y = y) = 1$. Only the uniform distribution on $\{1, 2\}$ has entropy 1, hence $Q(1)P_1(y) = Q(2)P_2(y)$. As this holds for all y in the support of P_1 or P_2, we conclude $P_1 = P_2$ and $Q(1) = Q(2) = 1/2$. □

Theorem 3.40. $\mathrm{d_K}(P_1, P_2)$ violates the triangle inequality.

Proof. Consider the following counterexample. Let $\mathcal{Y}$ be a two-element set. Let P_1 place all probability on the first element, P_2 be uniform, and P_3 place all probability on the second element. Then we have

$$P_1 = \langle 1, 0 \rangle \qquad \mathrm{d_K}(P_1, P_2) \approx 0.3058 \tag{3.38}$$
$$P_2 = \left\langle \tfrac{1}{2}, \tfrac{1}{2} \right\rangle \qquad \mathrm{d_K}(P_2, P_3) \approx 0.3058 \tag{3.39}$$
$$P_3 = \langle 0, 1 \rangle \qquad \mathrm{d_K}(P_1, P_3) = 1 \tag{3.40}$$

Clearly, $\mathrm{d_K}(P_1, P_3) > \mathrm{d_K}(P_1, P_2) + \mathrm{d_K}(P_2, P_3)$. □

As an illustration, Figure 3.5 shows the iso-similarity curves for some distributions on three outcomes. Obtaining the least favourable distribution now is equivalent to finding the distributions within the convex hulls of the models that are most similar. Of course, for truth finding, the distribution Q that attains the maximum similarity is also of importance, as it is used as the prior on models.

3.6.3 K-distance and KL-divergence

The Kullback-Leibler divergence, $\mathcal{D}$, is also not a metric. It violates both symmetry and the triangle inequality. The asymmetry of the KL-divergence can be extreme. When P_2 is a point-distribution while P_1 is not, then $\mathcal{D}(P_1 \| P_2) = \infty$, while $0 < \mathcal{D}(P_2 \| P_1) < 1$.

We show two graphical examples, Figure 3.6 and Figure 3.7. In both we use the Bernoulli model, because specification of a Bernoulli distribution requires a single parameter. In Figure 3.6, we show a contour plot of the K-distance and the KL-divergence. The asymmetry of the KL-divergence is clearly visible. Then, in Figure 3.7, we show both K-distance and KL-divergence in a single graph, for several P_1, as a function of P_2.

3.7 Discussion

This section collects various observations about the truth-finding problem.

3.7.1 Equaliser strategies

Consider the case when $\mathcal{T}$ is finite. Then the convex hull of each model is closed. We have shown that there is a minimax strategy f for Learner, and a least favourable distribution P_Θ for Nature. Let V be the value of the truth-finding game. Then for all θ we have

$$R(\theta, f) \leq V,$$

with equality if θ is in the support of P_Θ. Hence f is almost an equaliser strategy. It guarantees risk exactly V for the possible worlds that Nature uses in the least favourable distribution, and at most V for all other worlds.

Figure 3.5 Iso-similarity graphs

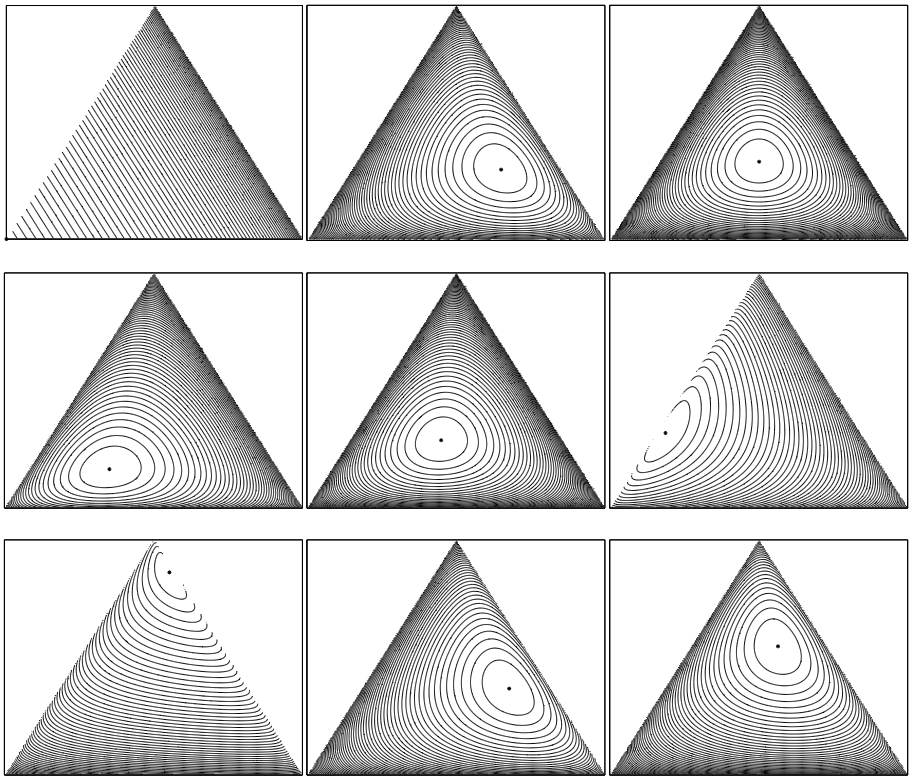


Figure 3.6 K-distance vs KL-divergence I. Contour lines of both loss measures on the Bernoulli model. The vertical axis shows $P_1(\mathrm{H})$, the horizontal axis $P_2(\mathrm{H})$.

(a) K-distance $\mathrm{d_K}(P_1, P_2)$ **(b)** KL-divergence $\mathcal{D}(P_1 \| P_2)$

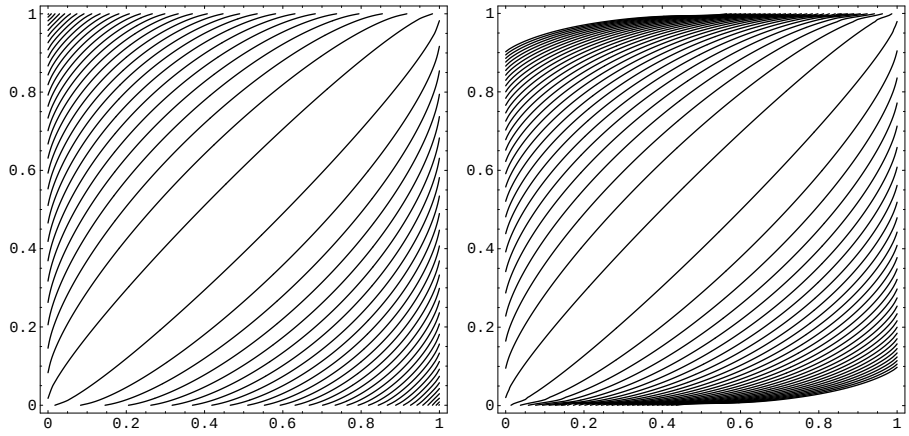
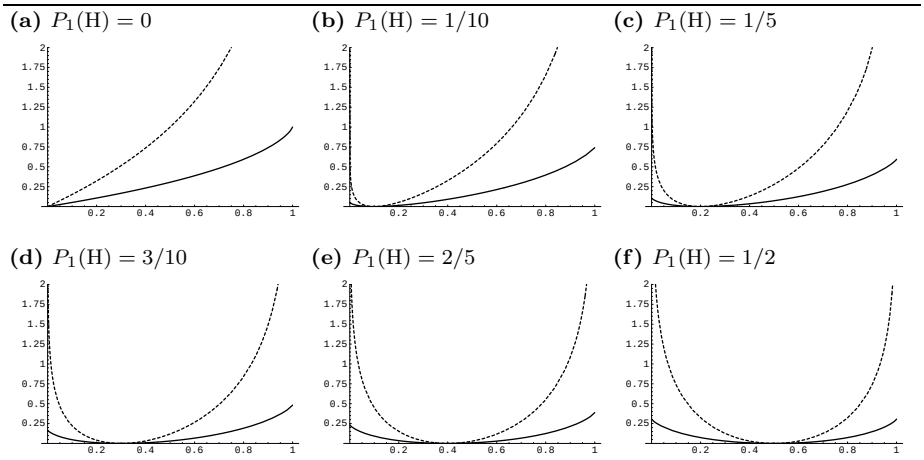


Figure 3.7 K-distance vs KL-divergence II. Graphs of $\mathrm{d_K}(P_1, P_2)$ (solid) and $\mathcal{D}(P_1 \| P_2)$ (dotted) as a function of $P_2(\mathrm{H})$.



3.7.2 Log loss non-decomposability

We have used the log loss as an integral part of our exposition. In particular, we have regarded the probability distribution that Learner plays at the end of the truth-finding game as part of a *pure strategy*. It may seem that equivalently, we can use a loss function that is defined directly on models, and let Learner choose models instead of distributions, and then play a mixed strategy. This is not the case. Recall that the log loss is defined by

$$\mathrm{L}(\mathcal{M}, \mathbf{m}) := -\log \mathbf{m}(\mathcal{M}).$$

We claim that there is no loss function $d : \mathbb{M} \times \mathbb{M} \to \mathbb{R}$ such that

$$\forall \mathbf{m} \forall \mathcal{M} : \mathrm{L}(\mathcal{M}, \mathbf{m}) = \mathrm{E}_{M \sim \mathbf{m}} \left[d(\mathcal{M}, M) \right].$$

Consider the following two simplified *belief games* on a finite set $\mathbb{M}$:

$$\mathcal{G}_1 = \left\langle \mathbb{M}, \mathfrak{D}(\mathbb{M}), \mathrm{L} \right\rangle \qquad \text{and} \qquad \mathcal{G}_2 = \langle \mathbb{M}, \mathbb{M}, d \rangle .$$

In both games, Nature picks an element $\mathcal{M} \in \mathbb{M}$, which we call *the truth*.

1. Learner picks, as a pure strategy, a distribution $\mathbf{m} \in \mathfrak{D}(\mathbb{M})$. We say that Learner *puts his belief on the table*. The loss for Learner is given by the log loss.

2. Learner picks, as a pure strategy, a model $\mathcal{M} \in \mathbb{M}$. We say that Learner *makes a guess*. The loss for Learner is given by $d(\mathcal{M}, \mathcal{M}')$. Of course, Learner can also play a mixed strategy, that is, a distribution $\mathbf{m} \in \mathfrak{D}(\mathbb{M})$. His risk is then given by

$$\mathrm{R}(\mathcal{M}, \mathbf{m}) = \mathrm{E}_{M \sim \mathbf{m}} \left[d(\mathcal{M}, M) \right].$$

Theorem 3.41. There is no function d such that game 1 is identical to game 2. Formally, for every function $d : \mathbb{M} \times \mathbb{M} \to \mathbb{R}$, and for every strategy $\mathcal{M} \in \mathbb{M}$ for Nature, there is a strategy $\mathbf{m}_1$ for Learner in game 1 such that for all strategies $\mathbf{m}_2$ for Learner in game 2 we have $\mathrm{L}(\mathcal{M}, \mathbf{m}_1) \neq \mathrm{R}(\mathcal{M}, \mathbf{m}_2)$.

Proof. Let d be a function as above, and pick an arbitrary strategy $\mathcal{M} \in \mathbb{M}$ for Nature. Let $m = \max_{\mathcal{M}'} d(\mathcal{M}, \mathcal{M}')$ and $m' = \max(0, m)$. Then define

$$\mathbf{m}_1(\mathcal{M}') = \begin{cases} 2^{-m'-1} & \text{if } \mathcal{M}' = \mathcal{M}, \\ \frac{1-2^{-m'-1}}{|\mathbb{M}|-1} & \text{otherwise.} \end{cases} \tag{3.41}$$

It is clear that for all $\mathbf{m}_2 \in \mathfrak{D}(\mathbb{M})$, $\mathrm{R}(\mathcal{M}, \mathbf{m}_2) \leq m'$, but on the other hand our definition of $\mathbf{m}_1$ ensures $\mathrm{L}(\mathcal{M}, \mathbf{m}_1) = -\log \mathbf{m}(\mathcal{M}) = -\log 2^{-m'-1} = m'+1$. □

3.7.3 Truth-finding in context

In this chapter, we have introduced truth finding as a framework of learning. Two other important frameworks for learning are prediction and compression. As an important first step toward understanding the relations between these three frameworks, we provide an exposition of assumptions and performance criteria of each.

Let $\mathcal{X}$ be a finite alphabet. We write $\mathcal{X}^\omega$ for the set of infinite sequences over $\mathcal{X}$. Let let P^* be a probability measure on $\mathcal{X}^\omega$, called *the truth*. Let $\langle X_i \rangle_{i<\omega}$ be discrete random variables, X_i giving the i-th outcome. We will be concerned with prefixes of outcomes. We define the random variable X^n by $X^n(x) := \big\langle X_i(x) \big\rangle_{i<n}$ and abbreviate $x^n := \langle x_i \rangle_{i<n}$. We assume for simplicity that n is fixed. The above three frameworks are formalised as follows.

Prediction Given $X^n = x^n$, produce a probability measure P on the next outcome X_{n+1}. Then observe x_{n+1}, the next outcome. The loss of P is given by $-\log P(X_{n+1} = x_{n+1})$. The risk of P, after observing x^n but before observing x_{n+1}, is given by $\mathrm{E}_{P^*}\big[-\log P(X_{n+1}) \mid X^n = x^n\big]$. More generally, let $P(\cdot|X^n = x^n)$, with abuse of notation, be a function that assigns a probability measure on $\mathcal{X}$ to each n-sequence. Then the overall risk of P is given by

$$\mathrm{E}_{P^*}\left[-\log P(X_{n+1}|X^n)\right].$$

Compression Produce a probability measure P on infinite sequences, then, given $X^n = x^n$, the loss of P is given by the compressed length of x^n, that is, $-\log P(X^n = x^n)$. The risk of P is given by

$$\mathrm{E}_{P^*}\left[-\log P(X^n)\right].$$

Truth finding Assume that $\mathcal{T}$ is a measurable space, with $P^* \in \mathcal{T}$. Let $\langle \mathcal{M}_i \rangle_{i<m}$ be a family of disjoint events that cover $\mathcal{T}$, and let $\mathcal{M}^*$ be the unique such event that contains P^*. Given $X^n = x^n$, produce a probability measure P on $\mathcal{T}$. The loss of P is given by $-\log P(\mathcal{M}^*)$. Again more general, let $P(\cdot|X^n = x^n)$ be a function that assigns a probability measure on $\mathcal{T}$ to each n-sequence. Then the overall risk of P is given by

$$\mathrm{E}_{P^*}\left[-\log P(\mathcal{M}^*|X^n)\right].$$

Figure 3.8 Three types of learning

Goal	Type	Risk
Prediction	$P : \mathcal{X}^n \to \mathfrak{D}(\mathcal{X})$	$\mathrm{E}_{P^*}\left[-\log P(X_{n+1}\vert X^n)\right]$
Compression	$P \in \mathfrak{D}(\mathcal{X}^n)$	$\mathrm{E}_{P^*}\left[-\log P(X^n)\right]$
Truth finding	$P : \mathcal{X}^n \to \mathfrak{D}(\mathcal{T})$	$\mathrm{E}_{P^*}\left[-\log P(\mathcal{M}^*\vert X^n)\right]$

Comparison Note that each framework expresses losses and risks in idealised bits, the unit of uncertainty. This is to be expected from a statistical learning framework. Prediction and compression are most easily related; prediction equals compression of the next outcome, with complete knowledge of the past. In other words, the loss of prediction is the discrete derivative of the loss of compression. Prediction can be based on a probability measure P' on $\mathcal{X}^\omega$ by using $P(X = x|X^n = x^n) = P'(X_{n+1} = x|X^n = x^n)$, where the latter is a proper conditional probability. The best compressors are not necessarily the best predictors and vice versa, as is shown in [vE06].

In practice, prediction and compression are often based on Bayesian universal models. The following section discusses the relation between truth finding and Bayes.

3.7.4 Truth-finding and Bayes

Procedurally, a Bayesian approach to any inference (i.e. learning) problem consists of three steps.

1. Obtaining a prior distribution on possible states of nature. In the case of truth finding, these are the possible worlds.

2. Updating the prior distribution using data, obtaining the so-called posterior distribution.

3. Using the posterior distribution to make the required inferences. In truth finding, this means producing a distribution on models.

Our solution to the worst-case truth-finding problem is Bayesian, according to this definition. This is surprising, as we perform a worst-case analysis, which explicitly disregards any prior belief. Of course, philosophically, our truth-finding solution is not Bayesian.

A Bayesian constructs a prior distribution that reflects his prior belief about the actual world. Consequently, his choice will not depend on the number of outcomes that he is going to observe. Moreover, in the absence of specific prior knowledge, one would consider most worlds possible, and would not have any strong preference for one world over another, which a Bayesian would reflect in a smooth, fairly uniform prior.

The prior distribution on possible worlds that we obtain for truth finding is the least favourable distribution. This distribution is a mixed strategy for Nature. We use this prior distribution, because the minimax strategy for Learner happens to coincide with the optimal response to the least favourable distribution. The least favourable distribution depends on all components of the truth-finding frame, and, in particular on the number of observations that Learner will

receive from Chance. Furthermore, the least-favourable distribution is particularly non-uniform and non-smooth. For closed convex models, we have shown that its support contains exactly one (!) point per model, the most similar points of §3.6.

Our worst-case analysis of the truth-finding problem provides ample motivation for using the least favourable distribution. In particular, it provides the best guaranteed bounds on loss, *even when it is not true*. Our method constructs a distribution on models from data, using the least-favourable distribution as a prior. The probabilities that we assign to models can always be interpreted as code lengths, expressing our uncertainty about the true model. If Nature does play worst-case optimal, i.e. using the least favourable distribution, then the probability that we assign to each model is its conditional probability given the data. Only if the least favourable distribution corresponds to the prior belief of a true Bayesian can our distribution on models be regarded as the Bayesian posterior distribution.

3.8 Conclusion

In this chapter, we introduced truth finding. We formalised it as the truth-finding problem: *given a list of models, use data sampled from reality to obtain as much information as possible about the true model in the worst case*. We reformulated the truth-finding problem as a game between Nature and Learner. We first gave the extensive form of this game, and then transformed it into normal form. We showed how strategies for both players can be represented. Strategies for Learner equal stochastic matrices. Strategies for Nature can be identified with distributions from the convex hull of the models, weighted according to a prior on models. We proved that the truth-finding game has a value, and that Learner has a minimax strategy. Then we showed that this minimax strategy can be obtained from the least favourable distribution for Nature, which can be obtained by optimising the generalised entropy function. We proved that this function is concave, hence the optimisation problem is easy. To find the least favourable distribution for Nature, we need to find similar distributions, one from the convex hull of each model. We extracted a formal definition of similarity from the truth finding setting, which seems very natural. We formulated it in general terms for future research. We related truth finding to prediction and compression, regarded truth finding from a Bayesian perspective. Although conforming to Bayesian practice, we showed that truth finding violates Bayesian philosophy in several ways.

3.8.1 Open questions

- Can a distribution on models, obtained by truth finding, be used for prediction? If so, how, and is it any good?
- What is the exact relation between the closed convex hull and extended Bayes distributions?
- What happens when there is no common world in all models, but a subset of the models does intersect? How trivial is that problem?

- We have analysed truth finding with the log loss. Are there other natural loss measures for truth finding, and if so, how do the obtained distributions on models differ?
- How exactly is similarity related to channel capacity?

Chapter 4

Experiment design

The previous chapter introduced truth finding. In truth finding, the data are considered given. This chapter discusses the extension of truth finding with experiments, called *experiment design*. Here, the data are obtained as outcomes of experiments, and the learner can select which experiment is performed.

We show that, once the selection of experiments has been fixed, the remaining problem is an instance of the truth-finding problem. We covered the truth-finding problem in Chapter 3. In this chapter, we will be concerned with the selection of experiments. For simplicity, we first cover the case where only a single experiment can be performed. We will show that performing many experiments *in parallel* is covered by this case. Then we will turn to the more interesting case of truly *sequential* experiment design.

The structure of this chapter parallels the development of Chapter 3. In §4.1, we introduce the two examples that motivated this research: polynomial regression and the balance scale. Then, in §4.2 we formalise the arena of the experiment-design problem as the experimentation frame. We introduce the experimentation game in §4.4. We solve the experiment-design problem, i.e. obtain the minimax strategy for Experimenter, in two stages. First, in §4.5 we consider pure and mixed single-experiment strategies. Then, in §4.6 we turn to pure and mixed sequential strategies. In each of these four cases, we show that an experimentation strategy for Experimenter induces a truth-finding subgame, and explicitly extract its arena, the induced truth-finding frame. We conclude and summarise in §4.8, and give directions for future research.

4.1 Examples

We start by introducing two examples. These examples will be revisited in §4.7. Each example contains the following components.

- A set of outcomes
- A set of possible worlds
- A classification of possible worlds into models.
- A set of experiments

- Experimental mechanics, conveniently represented here by
  - A regression function
  - Noise

The first three are also used in truth finding. The set of experiments is the new ingredient for experiment design, and the experimental mechanics replace the (ordinary) mechanics of truth finding. Experimental mechanics specify the distributions of outcomes given experiments.

4.1.1 Polynomials

We are given an interval, say $[-1, 1]$, and a (possibly infinite) set of polynomials. We assume that a certain polynomial from this list is the true (but unknown) polynomial, and that it is our task to learn the answer to a specific question about this polynomial. We consider the following natural questions.

- Identity: which polynomial is it?
- Degree: what is the highest exponent occurring with non-zero coefficient?
- Parity: is the degree even or odd?

Each question induces an equivalence relation on the set of polynomials, relating the polynomials with identical answers. The equivalence classes form a partition of the set of polynomials; in our setting we call them models. Hence answering such a question amounts to finding the true model.

To be able to discover the answer to one of these questions, we can perform experiments thus: first we pick a point ξ in the given interval, and then we receive the value Y of the polynomial θ at the given point, perturbed by normally distributed noise. This statement is formalised by the following *regression relation*:

$$Y|\theta, \xi \sim \theta(\xi) + \mathcal{N}(0, \sigma^2). \tag{4.1}$$

Here $\mathcal{N}(\mu, \sigma^2)$ is the standard normal distribution with mean μ and variance σ^2. Note that under this noise function — whichever polynomial is the truth — any outcome is possible, though large deviations from the mean are very unlikely. This means that there will always be uncertainty in our inferences.

4.1.2 Balance scale

This example is based on a well-known riddle, where, using a balance scale, one has to find the odd ball among twelve indistinguishable balls. We are given 12 numbered balls and a classical balance scale, like the one shown in Figure 4.1. We are given for a fact that exactly one of the balls is heavier or lighter than all the others. We are to the answer to one of the following questions through experiment:

- Index: which ball is the odd ball?
- Weight: is the odd ball heavier or lighter?
- Both: which ball is odd, and is it heavier or lighter?

An experiment is performed by placing some balls in the left scale pan, and placing equally many different balls in the right scale pan. The observed outcome is one of LEFT, EVEN and RIGHT. For example, LEFT means that the scale indicates that the content of the left scale pan is heavier than that of the right scale pan.

Let L, R be disjoint equinumerous sets of balls, and let i and w be the index and the weight of the odd ball. Then the true outcome of the balance scale, when the balls in L and R are placed in the left and right scale pan, is given by the following regression function:

$$r_{i,w}(L,R) := \begin{cases} \text{LEFT} & \text{if } i \in L \text{ and } w = \text{HEAVY}, \\ \text{LEFT} & \text{if } i \in R \text{ and } w = \text{LIGHT}, \\ \text{RIGHT} & \text{if } i \in R \text{ and } w = \text{HEAVY}, \\ \text{RIGHT} & \text{if } i \in L \text{ and } w = \text{LIGHT}, \\ \text{EVEN} & \text{otherwise.} \end{cases} \tag{4.2}$$

It is a classical puzzle problem to find both the index and the weight of the odd ball using three sequential experiments.[1] To make the problem harder, we assume that the observations are noisy. This might be caused, for example, by wind. We assume that the probability of observing each outcome is related to the truth by an error matrix, a 3×3 matrix that specifies a probabilistic function from true outcomes to observed outcomes. Examples of error matrices are shown in Figure 4.2.

4.2 Formalisation

We formalise the setting described in the introduction to this chapter. Then we give the formal statement of the experiment-design problem.

4.2.1 Frames

Definition 4.1. A quintuple

$$\mathfrak{F} = \left\langle \mathcal{Y}, \mathcal{T}, \mathbb{M}, \Xi, \langle \mathbf{p}_{\theta,\xi} \rangle_{\theta \in \mathcal{T}, \xi \in \Xi} \right\rangle$$

is called an *experimentation frame*, or *frame* for short, if the following conditions hold:

Figure 4.1 Balance scale

[1] An online statment of and solution to this puzzle is given `http://www.iwriteiam.nl/Ha12coins.html`.

Figure 4.2 Two error matrix examples

(a) Uniform error

truth \ outcome	LEFT	EVEN	RIGHT
LEFT	0.8	0.1	0.1
EVEN	0.1	0.8	0.1
RIGHT	0.1	0.1	0.8

(b) Neighbour error

truth \ outcome	LEFT	EVEN	RIGHT
LEFT	0.9	0.1	0
EVEN	0.1	0.8	0.1
RIGHT	0	0.1	0.9

- $\mathcal{Y}$ is a sample space, called the *outcome space*. We refer to the elements of $\mathcal{Y}$ as *outcomes*.
- $\mathcal{T}$ is a sample space, called the *possible world space.*. We refer to the elements of $\mathcal{T}$ as *possible worlds*.
- $\mathbb{M}$ is a partition of $\mathcal{T}$, called the set of *models*.
- Ξ is a set. We refer to the elements of Ξ as *experiments*.
- $\mathbf{p}_{\theta,\xi}$ is a probability distribution on $\mathcal{Y}$ for each possible world θ and experiment ξ. We write $\mathbf{p}(y|\theta,\xi)$ for $\mathbf{p}_{\theta,\xi}(y)$. We call $\langle \mathbf{p}_{\theta,\xi} \rangle_{\theta\in\mathcal{T},\xi\in\Xi}$ the *experimental mechanics*.

The requirements on the first three components are exactly the same as for truth-finding frames, which are defined in §3.1.1. The set of possible worlds contains all the different states of nature that we consider. We do not know what reality is, but we assume that we can always perform any of the given experiments. When performing an experiment, the outcome depends on chance, but moreover, the generating distribution depends on the actual world through the experimental mechanics. This dependence, which is modelled by the experimental mechanics $\mathbf{p}_{\theta,\xi}$, is key to discovery.

4.2.2 Experiment-design problem

The experiment-design problem can now be succinctly stated thus: *given an experimentation frame $\mathfrak{F}$, choose experiments intelligently, to obtain as much information about the true model as possible.* In more detail, the actual world $\theta^* \in \mathcal{T}$ must be classified according to $\mathbb{M}$. We assume no prior knowledge about reality, but, by performing an experiment and observing its outcome, we can obtain information about the true model. The experiment-design problem then consists of two subproblems:

1. Choosing the experiment to perform.
2. Obtaining information about the true model from the observed outcome.

The second subproblem is solved in Chapter 3. Recall that to measure the amount of information that the learner has about the true model, we use the log loss. The first subproblem is the central problem of this chapter.

4.2.3 Formalisation of the examples

A concise summary of the examples is given in Table 4.1. We explicitly list all the components of the experimentation frames for the polynomial example, and for the balance scale example. In §4.7.1 we describe how an experimentation frame is obtained from a regression function and noise.

4.3 Assumptions

The assumptions that are relevant to truth finding, as given in §3.1.4, remain relevant for experiment design. In addition, we must assume $|\Xi| \geq 2$ for there to be an experiment-design *problem*. We add the following assumption to simplify analysis.

Assumption 4.2. The set of experiments Ξ is finite.

This assumption is satisfied by Example 1.2, the anvil drop, from the introduction. Here the number of experiments is six, the number of floors of the Tower of Pisa. The balance scale example also has a finite number of experiments, namely

$$\sum_{i=0}^{6} \binom{12}{i, i, 12-2i} = 73789.$$

In the case of polynomials, the number of experiments is uncountable. There, we need to approximate the interval $[-1, 1]$ using finitely many points to satisfy the above assumption.

4.4 Experimentation game

The experiment-design problem is a worst-case optimisation task. To analyse it, we model it as a game, which we call the *experimentation game*. This allows us to find the solution in terms of a worst-case-optimal learning strategy. The players in the experimentation game are called Experimenter, Nature and Chance. Nature initially chooses the actual world, thereby fixing the true model. Then Experimenter tries to find out as much as possible about the true model by intelligently selecting experiments. We use the impartial player Chance to model the generation of an outcome for each experiments. Experimenter performs several experiments sequentially. This means that his choice of the next

Table 4.1 Overview of examples

	Polynomials	Balance scale
Outcomes	$\mathcal{Y} = \mathbb{R}$	$\mathcal{Y} = \{\text{LEFT}, \text{EVEN}, \text{RIGHT}\}$
Possible worlds	$\mathcal{T} \subseteq \mathbb{R}^*$	$\mathcal{T} = [1 \dots 12] \times \{\text{HEAVY}, \text{LIGHT}\}$
Models	identity, order parity	identity, index, weight, colour
Experiments	$\Xi = [0, 1]$	$\Xi = \{\langle L, R\rangle \mid L \cap R = \emptyset \wedge |L| = |R|\}$
Regression fn.	$\theta \in \mathcal{T}$	$r_{i,w}$
Noise	$\mathcal{N}(0, \sigma^2)$	error matrix

experiment may depend on all previous data: experiments and outcomes. After all experiments have been performed, we measure the amount of information that Experimenter lacks about the true model using the log loss.

The protocol of the experimentation game is shown in Protocol 4.1. Its extensive form is illustrated in Figure 4.3, where we show the game tree for the case where a single experiment is performed. Note that Experimenter makes a move twice. On the second level of the tree, Experimenter chooses an experiment. On the fourth level, he expresses his belief about the true model as a distribution on models. Information sets, i.e. clusters of positions thar are indistinguishable to Experimenter, are indicated by dotted lines.

To solve for the worst-case-optimal strategy for Experimenter, we transform the experimentation game into normal form. We first discuss the pure strategies that both players have at their disposition. Our approach is similar to that of [GD04] and [Fer67], although our games have much richer internal structure.

We proceed as follows. In §4.5 we cover the simple case $n = 1$, where Experimenter performs a single experiment. We then turn to truly sequential experimentation in §4.6. In each case, a strategy for Experimenter consists of two parts. We call the first part the *experimentation strategy*, and the second part the *learning strategy*. The experimentation strategy dictates which experiments will be performed in each execution of step 3 of Protocol 4.1.

We will show that, once the experimentation strategy has been fixed, the remaining subgame is an instance of the truth-finding game. Experimenter's learning strategy is the strategy that he wields on this subgame, and his loss on the expermentation game as a whole coincides with his loss on the induced truth-finding subgame. In §3.4.2 we proved that the truth-finding game has a value, and showed how to compute a minimax strategy. This perspective allows us to choose the optimal learning strategy for Experimenter, once the experimentation strategy has been fixed.

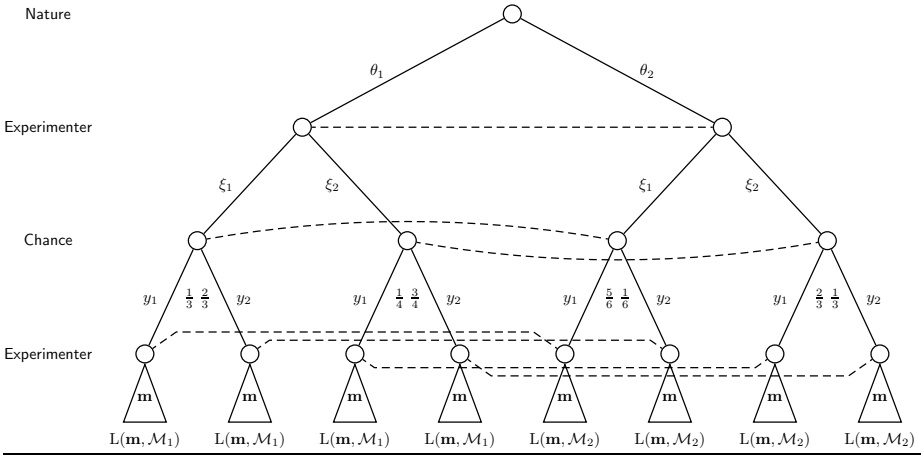
Protocol 4.1 The experimentation game

Arena: Experimentation frame $\mathfrak{F} = \left\langle \mathcal{Y}, \mathcal{T}, \mathbb{M}, \Xi, \langle \mathbf{p}_{\theta,\xi} \rangle_{\theta \in \mathcal{T}, \xi \in \Xi} \right\rangle$.
Require: Number of experiments n.
1: Nature covertly chooses a hypothesis θ^*. Say $M(\theta^*) = \mathcal{M}^*$.
2: **for** n turns **do**
3: Experimenter chooses an experiment ξ.
4: Chance generates an outcome y according to $\mathbf{p}_{\theta^*,\xi}$.
5: **end for**
6: Experimenter expresses his belief as a probability distribution $\mathbf{m}$ on models.
Loss: Experimenter suffers $-\log \mathbf{m}(\mathcal{M}^*)$.

4.5 Single experiment

The basis of game-theoretic analysis is the pure strategy. The strategies for Nature in the experimentation game are the same as those in the truth-finding game: Nature plays a possible world. In this section, we analyse the experimentation game in case only a single experiment is allowed. We first describe pure strategies for Experimenter, and provide a method to find the minimax strat-

Figure 4.3 Experimentation game tree for $n = 1$



egy. We then show that Experimenter must consider mixed strategies, by giving an example where Experimenter can, using mixed strategies, obtain substantially lower risk than when using pure strategies. Finally, we describe mixed strategies for Experimenter, and give references to known methods to obtain the minimax mixed strategy.

4.5.1 Pure strategies for Experimenter

In the experimentation game, Experimenter moves twice. In his first turn, he chooses an experiment. Then, for each possible outcome, he must produce a distribution on models in his second turn.

Definition 4.3. A pair $\langle \xi, f \rangle$ is called a *pure strategy for* Experimenter if

- $\xi \in \Xi$. We call ξ the *experimentation strategy.*
- $f : \mathcal{Y} \to \mathfrak{D}(\mathbb{M})$. We call f the *learning strategy.*

The pure strategy $\langle \xi, f \rangle$ instructs Experimenter to perform experiment ξ, and to subsequently, on observing outcome y, report the distribution $f(y)$ on models.

We can now define Experimenter's risk for the play θ vs $\langle \xi, f \rangle$.

Definition 4.4. Let θ and $\langle \xi, f \rangle$ be pure strategies for Nature and Experimenter. The *risk of the play* $\theta, \langle \xi, f \rangle$ is given by

$$\mathrm{R}\left(\theta, \langle \xi, f \rangle\right) := \mathop{\mathrm{E}}_{Y \sim \mathbf{p}_{\theta,\xi}} \Big[\mathrm{L}\left(\theta, f(Y)\right) \Big]. \tag{4.3}$$

This risk has a clear interpretation: it is the expected amount of information about the true model, expressed in bits, that Experimenter still lacks after observing the outcome of experiment ξ.

The difference between the truth-finding risk (3.3) and the experiment-design risk (4.3) is the presence of the experiment parameter in the distribution over which the expectation is taken. If we fix ξ, then this difference disappears.

Definition 4.5. Let $\mathfrak{F}$ be an experimentation frame, and fix $\xi \in \Xi$. Then the *ξ-subframe* of $\mathfrak{F}$ is given by

$$\mathfrak{F}_\xi := \left\langle \mathcal{Y}, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}_{\theta,\xi} \rangle_{\theta \in \mathcal{T}} \right\rangle. \tag{4.4}$$

Obviously, $\mathfrak{F}_\xi$ is a truth-finding frame, where the outcome that the learner obtains is actually the outcome of experiment ξ.

Recall from §3.4.2 that the truth-finding game on $\mathfrak{F}$ has a value, which we denote by $\mathrm{V}(\mathfrak{F})$. This means the following (using R_{tf} for the truth-finding risk)

$$\sup_{\mathsf{P}_\Theta} \inf_f \mathrm{R}_{\mathrm{tf}}(\mathsf{P}_\Theta, f) = \mathrm{V}(\mathfrak{F}) = \inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}_{\mathrm{tf}}(\mathsf{P}_\Theta, f).$$

The following theorem reduces the experiment-design problem to a number of truth-finding problems on subframes. Note that we allow Nature to play mixed strategies, and we use the convention that the risk function, when applied to a probability distribution, is interpreted as the expected risk.

Theorem 4.6. For all $\xi \in \Xi$,

$$\inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, \langle \xi, f \rangle) = \mathrm{V}(\mathfrak{F}_\xi).$$

Proof. By definition. □

For each experiment $\xi \in \Xi$ we can, using Chapter 3, compute the corresponding minimax strategy f_ξ in the truth-finding game on $\mathfrak{F}_\xi$. Hence the experiment-design problem with pure strategies for Experimenter is a finite optimisation problem. Experimenter is faced with $|\Xi|$ many choices, each choice leading to an essentially unrelated[2] instance of the truth-finding game on a subframe. To find the overall minimax strategy, Experimenter computes the value of the truth-finding game on each subframe, and chooses an experiment for which the resulting truth-finding risk is least.

The algorithm that is sketched in the previous paragraph computes the minimax pure strategy for Experimenter, thereby solving the experiment-design problem restricted to pure strategies. Unfortunately, this is of little use, as playing mixed strategies can be significantly better for Experimenter.

4.5.2 The necessity of mixed strategies

The following example illustrates that Experimenter can reduce the worst-case risk by using mixed strategies, even when only a single experiment is performed. Mixed strategies will be worked out in detail in the next section.

Example 4.7 (Bribed jury)**.** Consider a judge — our Experimenter — that wants to know whether the jury has been bribed. There are two shady members within the jury, call them A and B. To fix the jury's verdict, it suffices to bribe a single member. The judge has the power to command a search on one of them. A search can either yield incriminating evidence, say a lot of cash, or no

[2] In theory, the experimental mechanics can be completely unrelated for all possible worlds and all experiments. In practical applications, however, the experimental mechanics are often not arbitrary, and relations between different subframes abound.

such evidence. We set $\mathcal{Y} = \{\text{CASH}, \text{NOTHING}\}$. The possible worlds are given by $\mathcal{T} = \{\text{A}, \text{B}, \emptyset\}$, where we use $\emptyset$ for no bribe. The models are given by

$$\mathcal{M}_1 = \{\text{A}, \text{B}\} \quad \text{bribed jury} \tag{4.5}$$

$$\mathcal{M}_2 = \{\emptyset\} \quad \text{fair jury} \tag{4.6}$$

We call the experiments $\mathcal{A}$, and $\mathcal{B}$. The experimental mechanics are given by

p		CASH	NOTHING
A	$\mathcal{A}$	1	0
	$\mathcal{B}$	0	1
B	$\mathcal{A}$	0	1
	$\mathcal{B}$	1	0
$\emptyset$	$\mathcal{A}$	0	1
	$\mathcal{B}$	0	1

Now say the judge commits herself to $\mathcal{A}$, that is, she will have jury member A searched. Now the worst that can happen is that Nature plays the uniform distribution on $\{\text{B}, \emptyset\}$. The judge's experiment will yield no information, and the best she can do is just guess blindly between $\mathcal{M}_1$ (bribed jury) and $\mathcal{M}_2$ (fair jury), incurring loss 1. Analogously, if she commits to performing $\mathcal{B}$, she will obtain no information in the worst case.

On the other hand, if the judge uses the uniform distribution on $\{\mathcal{A}, \mathcal{B}\}$, then the risk of her minimax truth-finding strategy is only 0.6942. By symmetry of the experimental mechanics, the uniform distribution on experiments must be optimal for Experimenter. The least favourable distribution for Nature, and Experimenter's best response are shown in Figure 4.4. Figure 4.5 shows the shape of the generalised entropy surface, which will be explained in the next section.

4.5.3 Mixed strategies for Experimenter

A mixed strategy is a probability distribution on pure strategies. Let P_F be a mixed strategy for Experimenter. Such a strategy has the following interpretation. Take $\langle \xi, f \rangle$ from the support of P_F. Then with probability $\mathsf{P}_F(\langle \xi, f \rangle)$,

Figure 4.4 Bribed jury example optimal strategies. (a) shows Nature's worst-case-optimal distribution on possible worlds. Note that both the prior on worlds and the derived prior on models are non-uniform. (b) shows the optimal experimentation strategy for Experimenter. (c) shows the optimal learning strategy for Experimenter.

(a) Nature

θ	$\mathsf{P}_\Theta(\theta)$
A	0.2765
B	0.2765
$\emptyset$	0.4470

(b) Experimenter

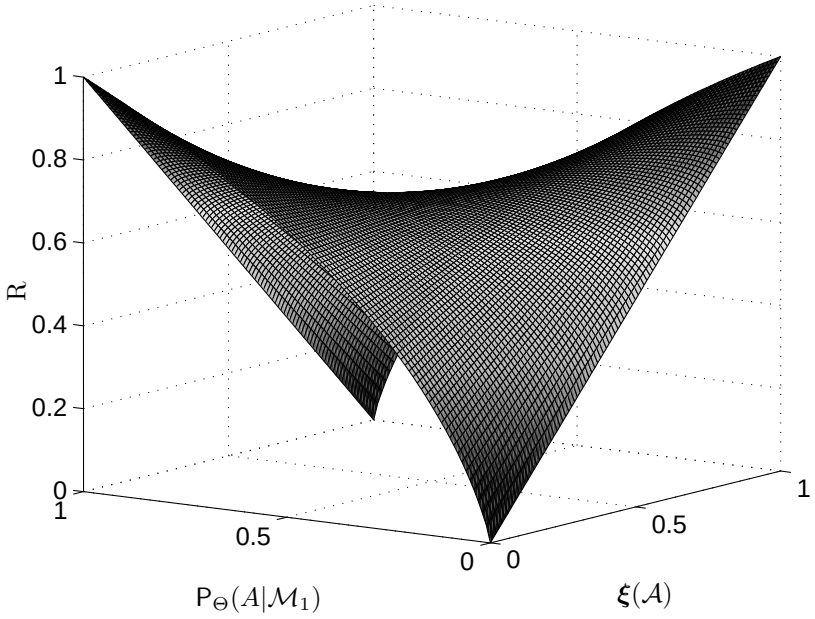
ξ	$\boldsymbol{\xi}(\xi)$
$\mathcal{A}$	1/2
$\mathcal{B}$	1/2

(c) Experimenter

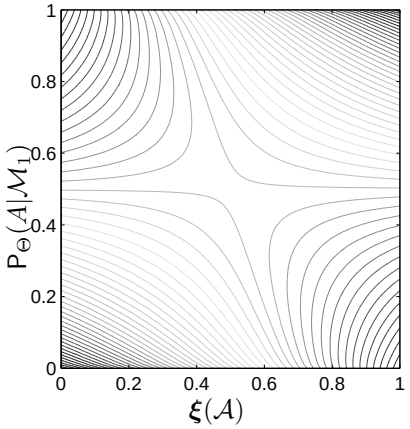
ξ	y	$\mathbf{m}(\mathcal{M}_1)$	$\mathbf{m}(\mathcal{M}_2)$
$\mathcal{A}$	CASH	1	0
$\mathcal{A}$	NOTHING	0.3822	0.6178
$\mathcal{B}$	CASH	1	0
$\mathcal{B}$	NOTHING	0.3822	0.6178

Figure 4.5 Generalised entropy for Bribed Jury. To reduce dimensionality, we fixed $\mathsf{P}_\Theta(\mathcal{M}_1) = \mathsf{P}_\Theta(\mathcal{M}_2) = 1/2$. Both (a) and (b) show $\mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi})$ as a function of $\mathsf{P}_\Theta(A|\mathcal{M}_1)$ and $\boldsymbol{\xi}(\mathcal{A})$. Linearity of $\mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi})$ in $\boldsymbol{\xi}$ for fixed P_Θ can be seen in (a), because for each P_Θ, the graph is a straight line. It can also be seen in (b) where the intersection of the contour lines with any horizontal line yields a set of equidistant points. (c) shows the point-wise maximum for Nature (solid) and minimum for Experimenter (dotted).

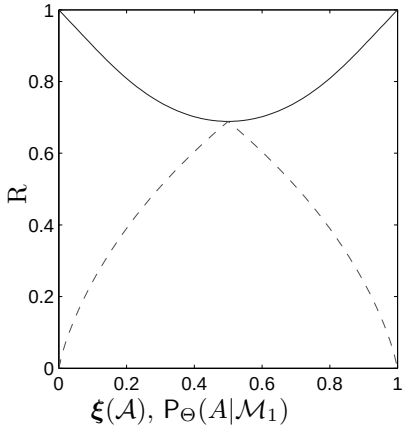
(a) Surface plot



(b) Contour plot



(c) Min-max plot



Experimenter will perform experiment ξ, and subsequently use f in the truth-finding problem on $\mathfrak{F}_\xi$. The risk for mixed strategies is obtained by taking expectations over (4.3).

Definition 4.8. Let P_Θ, P_F be mixed strategies for Nature and Experimenter. The *risk of the play* $\mathsf{P}_\Theta, \mathsf{P}_F$ is given by

$$\mathrm{R}(\mathsf{P}_\Theta, \mathsf{P}_F) := \mathop{\mathrm{E}}_{\Theta\sim\mathsf{P}_\Theta} \mathop{\mathrm{E}}_{\langle X,F\rangle\sim\mathsf{P}_F} \mathop{\mathrm{E}}_{Y\sim\mathbf{p}_{\Theta,X}} \left[\mathrm{L}(\Theta, F(Y))\right] \tag{4.7}$$

Here, $\langle X, F\rangle$ is a random variable, ranging over strategies for Experimenter, that is distributed according to Experimenter's mixed strategy P_F. By Definition 4.3, X is an experiment, and F is a learning strategy, i.e. a function that, given data, produces a distribution on models. Θ, the actual world, is distributed according to Nature's mixed strategy P_Θ.

In §3.2.5, we showed that it is never beneficial for Learner to use a mixed strategy in the truth-finding game. So without loss of generality, we can restrict attention to mixed strategies P_F that have functional support. A mixed strategy P_F has functional support if for each experiment ξ with positive marginal probability, the conditional distribution $\mathsf{P}_F(F|X = \xi)$ assigns all its probability to a single learning strategy. We call the unique such learning strategy f_ξ (When ξ has zero marginal probability, we can choose any f_ξ). Using this, the risk simplifies to

$$\mathrm{R}(\mathsf{P}_\Theta, \mathsf{P}_F) = \mathop{\mathrm{E}}_{\Theta\sim\mathsf{P}_\Theta} \mathop{\mathrm{E}}_{X\sim\mathsf{P}_F} \mathop{\mathrm{E}}_{Y\sim\mathbf{p}_{\Theta,X}} \left[\mathrm{L}(\Theta, f_X(Y))\right].$$

We have decomposed a mixed strategy in a probabilistic component, the generation of the experiment, and a deterministic component, the selection of the distribution on models. The distribution on models is chosen as a function of *both* the experiment and the outcome. The following definition makes this decomposition explicit.

Definition 4.9. A *mixed strategy for* Experimenter *in simple form* is a pair $\langle \boldsymbol{\xi}, f\rangle$, where $\boldsymbol{\xi} \in \mathfrak{D}(\Xi)$, and $f : \Xi \times \mathcal{Y} \to \mathfrak{D}(\mathbb{M})$.

A mixed strategy P_F can be simplified thus:

$$\boldsymbol{\xi}(\xi) := \mathsf{P}_F(\xi), \tag{4.8}$$

$$f(\xi, y) := f_\xi(y). \tag{4.9}$$

Our final simplification of the loss can now be given.

Definition 4.10. When P_Θ is a mixed strategy for Nature, and $\langle \boldsymbol{\xi}, f\rangle$ is a mixed strategy for Experimenter in simple form, then the *risk of the play* $\mathsf{P}_\Theta, \langle \boldsymbol{\xi}, f\rangle$ is given by

$$\mathrm{R}(\mathsf{P}_\Theta, \langle \boldsymbol{\xi}, f\rangle) = \mathop{\mathrm{E}}_{\Theta\sim\mathsf{P}_\Theta} \mathop{\mathrm{E}}_{X\sim\boldsymbol{\xi}} \mathop{\mathrm{E}}_{Y\sim\mathbf{p}_{\Theta,X}} \left[\mathrm{L}(\Theta, f(X, Y))\right]. \tag{4.10}$$

For fixed $\boldsymbol{\xi}$, we arrive at the following crucial observation: the experiment-design risk (4.10) is equal to the truth-finding risk (3.5), *when we regard the experiment as part of the data*. The innermost two expectations of (4.10) specify that Chance draws a pair $\langle X, Y\rangle$, and this pair is subsequently provided to f to determine the distribution on models. The following definition integrates the experiment into the data.

Definition 4.11. Let $\mathfrak{F}$ be an experimentation frame, and $\boldsymbol{\xi}$ a distribution on Ξ. Then the $\boldsymbol{\xi}$*-reduct* of $\mathfrak{F}$ is given by

$$\mathfrak{F}_{\boldsymbol{\xi}} := \left\langle \mathcal{Y} \times \Xi, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}_\theta^{\boldsymbol{\xi}} \rangle_{\theta \in \mathcal{T}} \right\rangle, \tag{4.11}$$

where for each possible world θ, the distribution $\mathbf{p}_\theta^{\boldsymbol{\xi}}$ on $\mathcal{Y} \times \Xi$ is defined by

$$\mathbf{p}_\theta^{\boldsymbol{\xi}}(y, \xi) := \mathbf{p}_{\xi,\theta}(y)\boldsymbol{\xi}(\xi).$$

Again, $\mathfrak{F}_{\boldsymbol{\xi}}$ is a truth-finding frame. This definition generalises Definition 4.5 in the following way. If we take a distribution $\boldsymbol{\xi}$ that concentrates all probability on a single experiment ξ, then for all $\theta \in \mathcal{T}$ and $y \in \mathcal{Y}$ we have $\mathbf{p}_{\theta,\xi}(y) = \mathbf{p}_\theta^{\boldsymbol{\xi}}(\xi, y)$. The data in the $\boldsymbol{\xi}$-reduct $\mathfrak{F}_{\boldsymbol{\xi}}$ are experiment/outcome pairs, while the data in the ξ-subframe $\mathfrak{F}_\xi$ are just outcomes. Other than that, their behaviour is identical for point distributions.

We put this definition to use in the following generalisation of Theorem 4.6.

Theorem 4.12. For all $\boldsymbol{\xi}$,

$$\inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, \langle \boldsymbol{\xi}, f \rangle) = \mathrm{V}(\mathfrak{F}_{\boldsymbol{\xi}}).$$

Proof. As the reader may check, this is purely a matter of definition. □

We are interested in finding the minimax experimentation strategy for Experimenter. Using the above simplification, we need to find the $\boldsymbol{\xi}$ that attains

$$\inf_{\boldsymbol{\xi}} \mathrm{V}(\mathfrak{F}_{\boldsymbol{\xi}}) = \inf_{\boldsymbol{\xi}} \inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, \langle \boldsymbol{\xi}, f \rangle) \tag{4.12}$$

We now apply the two results of Chapter 3: the minimax theorem for truth-finding (see §3.4.2), and the derivation of the game value in terms of the generalised entropy function (see §3.5.2). We obtain:

$$\inf_{\boldsymbol{\xi}} \mathrm{V}(\mathfrak{F}_{\boldsymbol{\xi}}) = \inf_{\boldsymbol{\xi}} \sup_{\mathsf{P}_\Theta} \mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi}), \tag{4.13}$$

where we denote by $\mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi})$ the *doubly generalised entropy*, which, for each $\boldsymbol{\xi}$, is defined as the generalised entropy in the truth-finding game on the $\boldsymbol{\xi}$-reduct $\mathfrak{F}_{\boldsymbol{\xi}}$. One can easily adapt the formula that we obtained for the generalised entropy (3.19) to:

$$\mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi}) = \mathop{\mathrm{E}}_{X,Y} \left[\mathcal{H}(M|X, Y) \right] \tag{4.14}$$

$$= \mathcal{H}(M) - \mathcal{I}(M; X, Y), \tag{4.15}$$

where the discrete random variables M, X and Y are jointly distributed according to

$$\mathsf{P}(\mathcal{M}, \xi, y) = \boldsymbol{\xi}(\xi) \int_{\mathcal{M}} \mathbf{p}(y|\theta, \xi) \mathsf{P}_\Theta(\mathrm{d}\theta).$$

The previous lines cannot be reduced further; the doubly generalised entropy function inextricably intertwines the mixed strategies for Nature and Experimenter. In Chapter 3 we used concavity of the generalised entropy function to obtain the minimax strategy for Learner. Unfortunately, the doubly generalised entropy function is not concave. Still, it is of a sufficiently manageable kind, as shown by the following theorem.

Theorem 4.13. The generalised entropy $\mathfrak{H}$ is *concave-linear*, i.e. $\mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi})$ is

1. a concave function of P_Θ for fixed $\boldsymbol{\xi}$, and
2. a linear function of $\boldsymbol{\xi}$ for fixed P_Θ.

Proof. We cover each claim separately.

1. Theorem 3.27.
2. Rewrite the generalised entropy as

$$\mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi}) = \mathop{\mathrm{E}}_{X} \mathop{\mathrm{E}}_{M,Y|X} \left[-\log \frac{\mathsf{P}(Y, M|X)}{\mathsf{P}(Y|X)} \right] \tag{4.16}$$

and observe that only the outermost expectation depends on $\boldsymbol{\xi}$. □

A linear function is both concave and convex, so $\mathfrak{H}$ is both *concave-convex* and *concave-concave*, the latter also being called *separately concave*. Theorem 4.13 does not imply that $\mathfrak{H}$ is concave. In all but the most trivial cases it is not; see Figure 4.5 for a counterexample.

By (4.13), to find the minimax experimentation strategy we need to find the saddle-point of the doubly generalised entropy function. There is a respectable amount of literature on concave-convex functions. An overview is given in [BV04]. This book contains conditions under which a concave-convex function has a saddle-point. The authors describe two classes of algorithms to find this saddle-point, called *Newton methods* and *Barrier methods*. The method that can be applied to our setting depends on the structure of the models. We did not explicitly try this; deeper research into this subject matter is planned for future work.

4.5.4 Bayesian Maximum Entropy Selection

By (4.14), the minimax optimal experimentation strategy is the strategy $\boldsymbol{\xi}$ that attains

$$\inf_{\boldsymbol{\xi}} \sup_{\mathsf{P}_\Theta} \mathfrak{H}(\mathsf{P}_\Theta, \boldsymbol{\xi}) = \inf_{\boldsymbol{\xi}} \sup_{\mathsf{P}_\Theta} \mathop{\mathrm{E}}_{X,Y} \left[\mathcal{H}(M|X, Y) \right].$$

A Bayesian experimenter has her own means of establishing a prior distribution P_Θ for Nature. Fixing this distribution simplifies matters considerably. The doubly generalised entropy function is a linear function of $\boldsymbol{\xi}$ for fixed P_Θ by Theorem 4.13. This implies that the minimum will be achieved for a point-distribution. Hence, a Bayesian can restrict attention to pure strategies, selecting an experiment (there could be more than one) that attains

$$\min_{\xi} \mathrm{E}_Y \left[\mathcal{H}(M|\xi, Y) \right].$$

This criterion for experiment selection is known in the literature on Bayesian experiment design by the name of Maximum Entropy Selection (MES). See for example [SW00]. The term MES stems from the following application of the chain rule of entropy (2.6):

$$\mathcal{H}(Y, M|\xi) = \mathcal{H}(Y|\xi) + \mathrm{E}_Y \left[\mathcal{H}(M|Y, \xi) \right].$$

Under the common assumption that $\mathcal{H}(Y, M|\xi)$ does not depend on ξ, minimising $\mathrm{E}_Y \left[\mathcal{H}(M|Y, \xi) \right]$ is equivalent to maximising $\mathcal{H}(Y|\xi)$.

4.5.5 Multiple independent experiments

The case where multiple experiments are performed *independently* can easily be reduced to the single experiment case using the following experimentation frame construction. In most general form, n outcomes are generated by n specified, possibly different experiments.

Definition 4.14. Let $\mathfrak{F}$ be an experimentation frame, and n a number of outcomes. The *n-fold product frame* is given by

$$\mathfrak{F}^n := \left\langle \mathcal{Y}^n, \mathcal{T}, \mathbb{M}, \Xi^n, \langle \mathbf{p}^n_{\theta,\xi^n} \rangle_{\theta\in\mathcal{T},\xi^n\in\Xi^n} \right\rangle, \tag{4.17}$$

where

$$\mathbf{p}^n_{\theta,\xi^n}(y^n) := \prod_{i=1}^{n} \mathbf{p}_{\theta,\xi_i}(y_i).$$

A product frame models the situation where we perform n experiments simultaneously. Naturally, these experiments are all performed on the same world, as the state of nature is constant. The n outcomes are generated independently. A product frame models experimentation *in parallel*. We now turn to sequential experimentation.

4.6 Sequential experimentation

In the following, let $n \geq 2$ be the number of sequential experiments. In sequential experimentation, as specified by steps 3 and 4 of Protocol 4.1, Experimenter and Chance alternately choose an experiment and an outcome. We call the sequence of their combined choices the *data*. At each position in the experimentation game where Experimenter is to move, we refer to the data that have been generated so far as *the history*. There is a one-to-one correspondence between histories and information sets for Experimenter in the extensive form of the experimentation game.

4.6.1 Pure strategies for Experimenter

A pure strategy for Experimenter assigns a move — an experiment — to each nonterminal history. Taking this literally, we obtain a viable but slightly baroque definition of strategy, because such a strategy must assign experiments to histories that can never be reached. For example, if the strategy dictates performance of ξ_1 in the empty history $\langle\rangle$, then any history that commences with ξ_2 is unreachable. Without loss of generality, we use the following compact representation of strategies instead.

Definition 4.15. A pair $\langle s, f \rangle$ is called a pure strategy for Experimenter if

- $s : \mathcal{Y}^{<n} \to \Xi$. We call s the *experimentation strategy*.
- $f : \mathcal{Y}^n \to \mathfrak{D}(\mathbb{M})$. We call f the *learning strategy*.

Substituting $n = 1$, we obtain Definition 4.3.

In this definition of strategy we use only the outcomes that Chance generated, instead of full histories. This causes no ambiguity as, given an experimentation strategy s and Chance's moves y^n, one can reconstruct the conducted experiments $\xi^n \in \Xi^n$ and hence the entire history $h^n \in (\Xi \times \mathcal{Y})^n$ thus:

$$\xi_i := s\big(y^{i-1}\big), \qquad h_i := \big\langle \xi_i, y_i \big\rangle . \tag{4.18}$$

Definition 4.16. When θ and $\langle s, f\rangle$ are pure strategies for Nature and Experimenter, then the *risk of the play* $\theta, \langle s, f\rangle$ is given by

$$\mathrm{R}(\theta, \langle s, f\rangle) = \mathop{\mathrm{E}}_{Y^n \sim \mathbf{p}^s_\theta} \big[\mathrm{L}(\theta, f(Y^n))\big] \tag{4.19}$$

where

$$\mathbf{p}^s_\theta(y^n) := \prod_{i=1}^{n} \mathbf{p}\big(y_i \big| \theta, \xi_i\big).$$

As in the single experiment case, a pure strategy for Experimenter induces a truth-finding frame.

Definition 4.17. Let $\mathfrak{F}$ be an experimentation frame, and s be a pure strategy for Experimenter. The *s-subframe of* $\mathfrak{F}$ is given by

$$\mathfrak{F}_s := \Big\langle \mathcal{Y}^n, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}^s_\theta \rangle_{\theta \in \mathcal{T}} \Big\rangle . \tag{4.20}$$

Theorem 4.18. For all learning strategies s,

$$\inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, \langle s, f\rangle) = \mathrm{V}(\mathfrak{F}_s).$$

Proof. By definition. □

Theoretically, to solve the sequential experiment-design problem with pure strategies for Experimenter, we simply solve the truth-finding game on $\mathfrak{F}_s$ for each learning strategy s. The set of learning strategies is given by $\big[\mathcal{Y}^{<n} \to \Xi\big]$, and consequently, the number of learning strategies is

$$\Big|\big[\mathcal{Y}^{<n} \to \Xi\big]\Big| = |\Xi|^{\left(\frac{|\mathcal{Y}|^n - 1}{|\mathcal{Y}| - 1}\right)} \leq |\Xi|^{\left(2|\mathcal{Y}|^{n-1} - 1\right)} \leq |\Xi|^{|\mathcal{Y}|^n} .$$

In practice, considering a doubly exponential number of strategies is infeasible, and we need to use (a) the tree structure of each strategy, (b) the independence of the outcomes given the experiments. We will not look into this further, as we already know that mixed strategies for Experimenter are more powerful.

4.6.2 Mixed strategies for Experimenter

A mixed strategy for Experimenter is a probability distribution on all his pure strategies. In the coming exposition, we prefer to use *probabilistic strategies* instead of mixed strategies. A probabilistic strategy probabilistically assigns an experiment to each history. It can easily be shown that, over finite sets, probabilistic strategies and mixed strategies generate the same behaviour. See for example [Fer67, p.26]. Probabilistic strategies are also called *behavioural strategies*.

Definition 4.19. A pair $\langle \boldsymbol{s}, f \rangle$ is called a *probabilistic strategy for* Experimenter if

- $\boldsymbol{s} : (\Xi \times \mathcal{Y})^{<n} \to \mathfrak{D}(\Xi)$. We call $\boldsymbol{s}$ the *experimentation strategy.*
- $f : (\Xi \times \mathcal{Y})^n \to \mathfrak{D}(\mathbb{M})$. We call f the *learning strategy.*

Probabilistic strategies take the entire history (both experiments and outcomes) into account. This in contrast to sequential pure strategies (Definition 4.15), that use only the past outcomes. Here one can no longer reconstruct the experiments from $\boldsymbol{s}$ and y^n alone, because experiments are chosen probabilistically. Note that for $n = 1$ we obtain a strategy for Experimenter in simple form (Definition 4.9).

Definition 4.20. When P_Θ is a mixed strategy for Nature, and $\langle \boldsymbol{s}, f \rangle$ is a probabilistic strategy for Experimenter, then the *risk of the play* $\mathsf{P}_\Theta, \langle \boldsymbol{s}, f \rangle$ is given by

$$\mathrm{R}(\mathsf{P}_\Theta, \langle \boldsymbol{s}, f \rangle) = \mathop{\mathrm{E}}_{\Theta \sim \mathsf{P}_\Theta} \mathop{\mathrm{E}}_{H^n \sim \mathbf{p}^{\boldsymbol{s}}_\Theta} \left[\mathrm{L}(\Theta, f(H^n)) \right], \tag{4.21}$$

where $H_i := \langle X_i, Y_i \rangle$, and

$$\mathbf{p}^{\boldsymbol{s}}_\theta(h^n) := \prod_{i=1}^{n} \mathbf{p}\big(y_i \big| \theta, \xi_i\big) \boldsymbol{s}(\xi_i | h^{i-1}).$$

Analogously to Definitions 4.5, 4.11 and 4.17, a probabilistic strategy $\boldsymbol{s}$ induces a truth-finding frame.

Definition 4.21. Let $\mathfrak{F}$ be an experimentation frame, and $\boldsymbol{s}$ be a probabilistic strategy for Experimenter. The $\boldsymbol{s}$*-reduct of* $\mathfrak{F}$ is given by

$$\mathfrak{F}_{\boldsymbol{s}} := \left\langle (\Xi \times \mathcal{Y})^n, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}^{\boldsymbol{s}}_\theta \rangle_{\theta \in \mathcal{T}} \right\rangle. \tag{4.22}$$

Theorem 4.22. For all probabilistic strategies for Experimenter $\boldsymbol{s}$,

$$\inf_f \sup_{\mathsf{P}_\Theta} \mathrm{R}(\mathsf{P}_\Theta, \langle \boldsymbol{s}, f \rangle) = \mathrm{V}(\mathfrak{F}_{\boldsymbol{s}}).$$

Proof. By definition. □

4.7 Examples revisited

This section discusses the examples that were introduced in §4.1. We first discuss the polynomial example, then the balance scale example.

4.7.1 Polynomials

In the polynomial example, we want to find the degree of the true polynomial from noisy outcomes of probes. Formulated like this, this problem might seem hard to formalise as an expermentation frame. In this section, we show how regression problems can be modelled in our framework in general. In regression problems (see e.g. [GCSR04]), experimental mechanics arise from a combination of two components: a regression function and noise.

Definition 4.23. A function $r : \Xi \to \mathcal{Y}$ is called a *regression function*.

A regression function deterministically assigns a *true outcome* to each experiment. To generate the *observed outcome*, the true outcome is perturbed in a way that is independent of the actual world. Formally,

Definition 4.24. A function $\epsilon : \Xi \times \mathcal{Y} \to \mathfrak{D}(\mathcal{Y})$ is called *noise*. We write $\epsilon(y'|\xi, y)$ for $\epsilon(\xi, y)(y')$, that is, the probability that we observe y' while y was the true outcome of the experiment ξ.

From a collection of regression functions, grouped into models, and noise, we construct an experimentation frame thus:

Definition 4.25. Let $\mathcal{R} \subseteq [\Xi \to \mathcal{Y}]$ be a collection of regression functions, $\mathbb{M}$ a partition of $\mathcal{R}$, and ϵ noise. The quintuple

$$\mathfrak{F}_{\mathcal{R},\epsilon} := \left\langle \mathcal{Y}, \mathcal{R}, \mathbb{M}, \Xi, \langle \mathbf{p}_{r,\xi} \rangle_{r\in\mathcal{R},\xi\in\Xi} \right\rangle, \tag{4.23}$$

where

$$\mathbf{p}_{r,\xi}(y) = \epsilon\left(y|\xi, r(\xi)\right)$$

is called the *regression frame* generated by $\mathcal{R}, \epsilon$.

4.7.2 Balance scale

In the balance scale problem, for symmetrical error matrices, the experiment mechanics are entirely symmetrical. This suggests that the least favourable distribution is the uniform distribution on worlds. The minimax strategy for Experimenter is hard to compute. In this section, we take a Bayesian approach by assuming the uniform distribution as a prior, and compute the optimal experimentation strategy with respect to it. As explained in §4.5.4, fixing the strategy for Nature brings us within the framework of Bayesian Maximum Entropy Selection. Consequently, we can restrict attention to pure experimentation strategies.

In order to reduce the size of the figures of strategies we restrict the problem to 6 balls. Hence, a world is an element of $[1 \ldots 6] \times \{\text{HEAVY}, \text{LIGHT}\}$. We have analysed the balance scale problem for three different sets of models. First, the minimax strategy for learning the weight of the odd ball is shown in Figure 4.6. Second, the minimax strategy for learning the index of the odd ball is shown in Figure 4.7. Finally, the minimax strategy for learning both the weight and the index of the odd ball is shown in Figure 4.8. These figures depict an annotated subtree of the game tree, which should be read as follows:

- Ellipses represent positions where Experimenter is to move. A number shown in an ellipse is the binary entropy of the distribution on models at that position, *before* any further experiments are performed.
- Rectangles represent positions where Chance is to move. A number on the bottom line of a rectangle is the expected binary entropy of the distribution on models, *after* all future experiments will have been performed.
- Experimenter's moves are labelled by experiments. For example $\{2, 4\} - \{3, 5\}$ indicates that balls 2 and 4 are placed in the left scale pan, whereas balls 3 and 5 are placed in the right scale pan. We only show the optimal move for Experimenter, to reduce the size of the figure.

- Chance's moves are labelled with an outcome and its conditional probability. We show all moves for Chance.
- At the top of rectangles, we show a partition of the balls. Balls within the same equivalence class are indistinguishable to Experimenter, they have always played the same role within each past experiment (i.e. left, right or non-participating). This partition was used to prune the search tree during the computation of the optimal strategy, but it is also interesting by itself.

As an example, consider Figure 4.6. Learner is trying to learn the weight of the odd ball. Before performing any experiments, Experimenter considers both the model "heavy" and the model "light" equally likely, hence the entropy of the true model is one. This is indicated within the topmost circle. The bold arrow going downward indicates that Experimenter will perform, as his first experiment, the weighing of balls 0 and 1 against balls 2 and 3. As indicated at the bottom of the topmost rectangle, Experimenter now expects to have entropy 0.245 at the end of the experimentation. The top line in the topmost rectangle indicates that learner can now distinguish certain balls, as they have assumed different roles in his first experiment. Below the topmost rectangle, there are three outgoing arrows: one for each possible outcome. The labels on these arrows give the conditional probability. In this case EVEN is the most likely outcome, with probability 0.367. The remaining subtree can be interpreted analogously.

4.8 Conclusion

We extended truth finding to experiment design by allowing the learner to choose experiments. We introduced experimentation frames to formalise the arena of experiment design. In experiment design, we face the experiment-design problem: *given an experimentation frame $\mathfrak{F}$, choose experiments intelligently, to obtain as much information about the true model as possible.* In order to solve the experiment-design problem, we formulated it as the experimentation game. A solution then takes the form of a minimax strategy for Experimenter in this game.

We showed that strategies for Experimenter, both pure and mixed, can be decomposed into two parts. The first part, called the experimentation strategy, tells Experimenter which experiments to perform. We showed that each experimentation strategy induces a truth-finding frame. For pure experimentation strategies, we call the induced truth-finding frame a subframe, because the truth-finding mechanics are a slice through the experimental mechanics. Mixed experimentation strategies, on the other hand, require inclusion of the randomly generated experiments in the data. We call the truth-finding frame induced by a mixed strategy a reduct. The second part of a strategy for Experimenter is called the learning strategy. The learning strategy is a strategy for the truth-finding game on the truth-finding frame induced by the experimentation strategy.

Using the Bribed Jury example, we show that, for Experimenter, mixed strategies are more powerful than pure strategies. We concluded that this is due to the entanglement of strategies that takes place within the risk function.

We introduced two approaches to solving the truth-finding game. Minimax pure strategies can be found by solving the induced truth-finding game for each

possible experimentation strategy. For a single experiment, this can be done in practice. Unfortunately, the number of experimentation strategies grows doubly exponential in the number of experiments. For sequential experimentation, this approach is infeasible.

Minimax mixed strategies can be found, because the risk function is concave-linear. General purpose convex optimisation methods, like the Newton method and the Barrier method can be applied under certain conditions on the structure models. This is a matter for future research.

4.8.1 Open questions

- We considered selecting experiments in worst-case-optimal fashion. One could also consider generating a sequence of experiments according to some product distribution, say uniformly at random. What is the relation between the minimax experiment-design risk, and the minimax truth-finding risk on the reduct where experiments are chosen uniformly at random?

- We introduced the balance scale example in §4.1, and discussed it in §4.7. Due to the symmetry of the possible world space and the experimental mechanics, we suspect that the least favourable distribution for Nature is the uniform distribution. Can we introduce a slight asymmetry (preferably by introducing a different partition of the possible worlds into models), such that the worst-case-optimal experiment and the optimal experiment with respect to the uniform distribution are different.

Figure 4.6 Balance scale: minimax strategy for learning weight

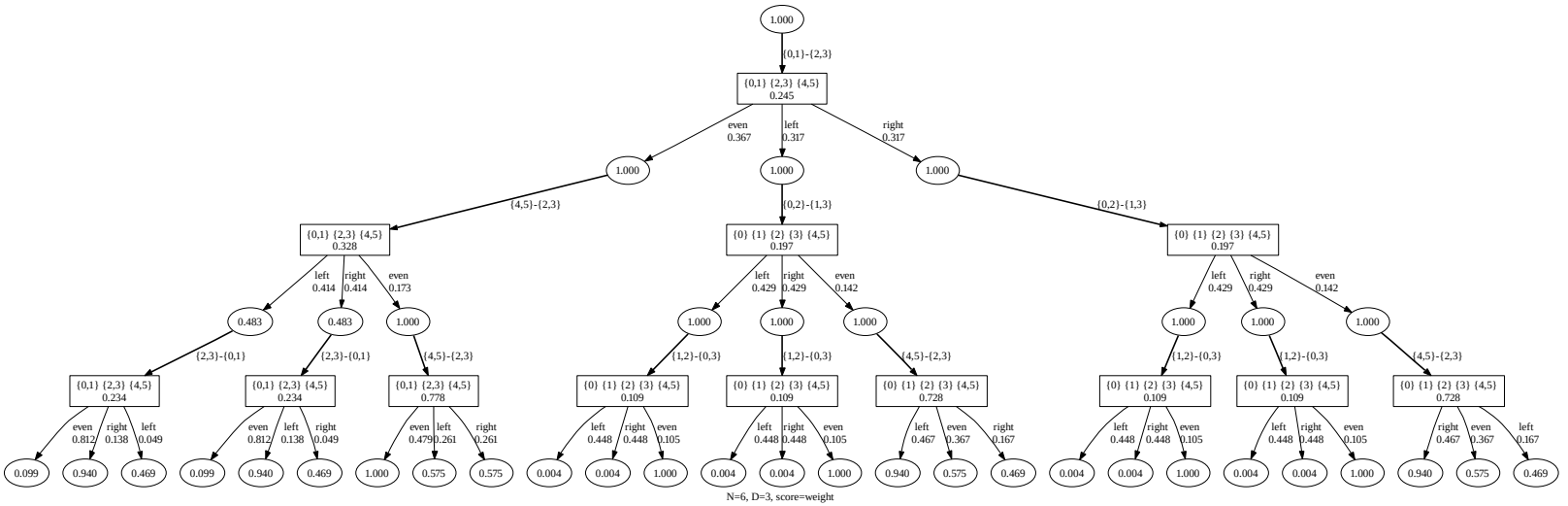


Figure 4.7 Balance scale: minimax strategy for learning index

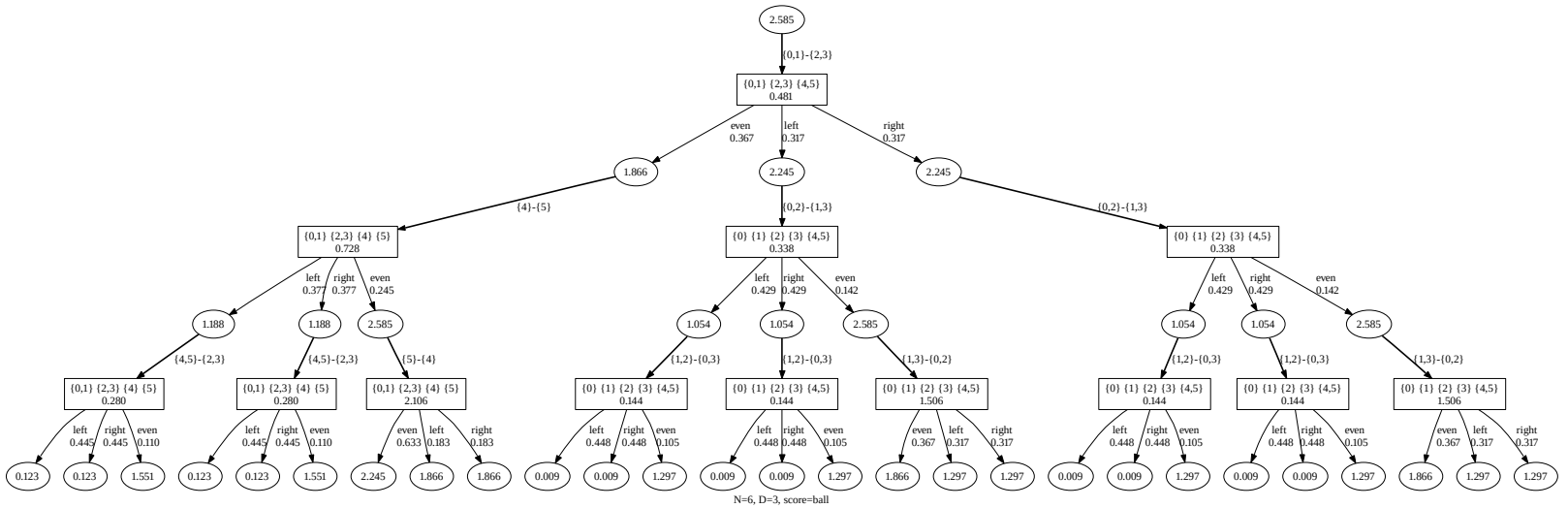
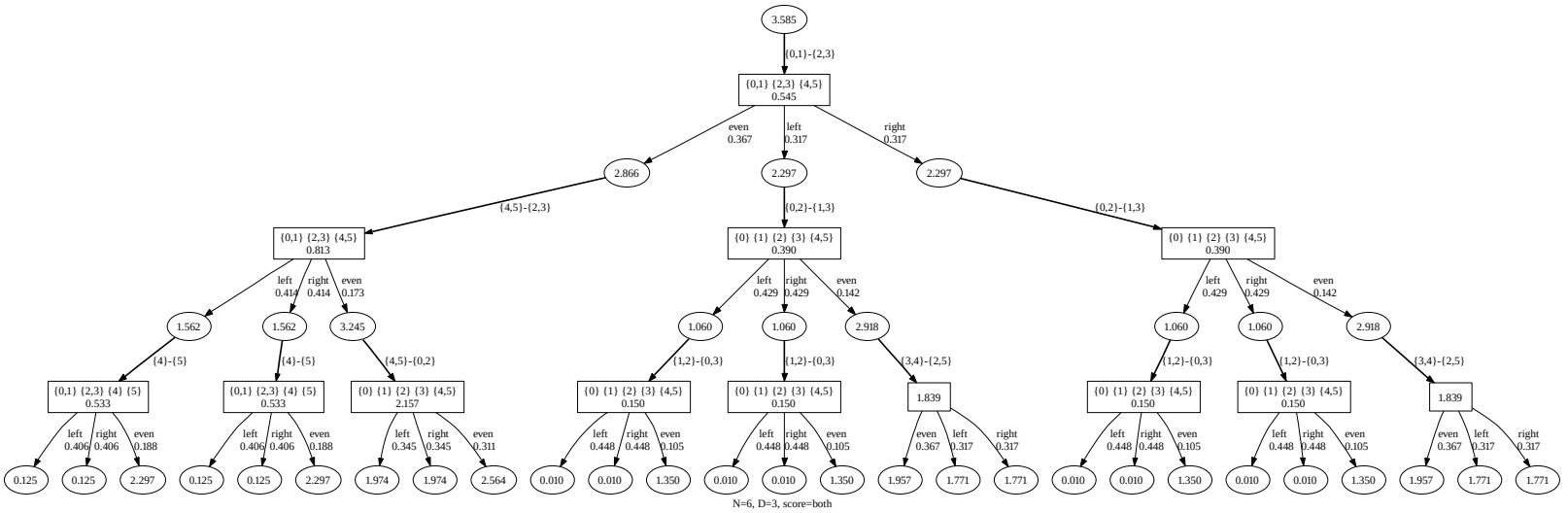


Figure 4.8 Balance scale: minimax strategy for learning both



Chapter 5

Conclusion

Research in the field of statistical machine learning is concerned with general methods for learning from data. Likewise, research in the field of experiment design is concerned with general methods for choosing experiments, with the goal of learning from their outcomes. In this thesis we focus on the following learning problem: *how to obtain as much information as possible about the true model in the worst case?* Here, the worst case is taken over all possible states of nature that the learner considers. We analysed this problem within the setting of statistical machine learning, and within the setting of experiment design.

Setting

In practice, statistical learning problems are often specified using models, collections of similar hypotheses. In this thesis we call hypotheses possible worlds, and assume that one of the hypotheses is true. We call this hypothesis reality and the model that contains it the true model. Models are cognitive devices, introduced by the learner to structure the learning problem at hand. They typically arise by collecting all possible worlds in which the learner wants to take the same action.

The amount of information that the learner lacks about the true model is given by the log loss of the distribution on models that expresses his uncertainty about the true model. We desire worst-case-optimal procedures, to simultaneously obtain the best performance guarantees and circumvent the inherent circularity in the use of subjective priors for both decision-making and performance assessment.

Truth finding

In Chapter 3 we developed, solved and discussed the machine learning version of the above problem, which we coined the truth-finding problem. A natural way to regard the truth-finding problem is as the truth-finding game: a two-player strategic game with chance moves. We proved that the truth-finding game has a value, and that a minimax mixed strategy for the learner always exists.

In the case that a maximin mixed strategy (a least favourable distribution) for nature exists, we showed that the learner's minimax strategy is the optimal response to it. Hence this least favourable distribution may be thought of as

a prior. This result is surprising; we performed a worst-case analysis to factor out all prior knowledge, and yet we obtained a prior distribution on possible worlds. We proved that a least favourable distribution exists when the convex hull of the models is closed. We showed that the generalised entropy function is concave, which implies that we can easily find its optimum, the least favourable distribution, using convex optimisation methods. We showed that any mixed strategy for nature can be decomposed into a prior distribution on models and, for each model, a point from the convex hull of the distributions on outcomes within that model. We conjectured that, by choosing the points from the closed convex hull instead, a least favourable distribution for nature always exists. The distribution on models that the learner obtains using truth finding can always be interpreted as a code. When interpreted as a code, it is the code that minimises the expected message length for encoding the true model. Moreover, if nature plays worst-case-optimally then the learner's distribution on models coincides with the conditional probability distribution on models given the data. Only in this special case does the learner's distribution have a standard probabilistic interpretation. We gave several examples that show that the least favourable distribution is particularly non-uniform, and depends on the number of observations the learner will make. Hence, although the solution to the truth-finding problem is Bayesian in form, we conclude that it is essentially non-Bayesian in philosophy.

The truth-finding problem gives rise to a natural notion of similarity between sequences of distributions. We provide a formal definition in general terms, and hint at its relation to the information channel capacity. We define the Koolen distance between two distributions as the opposite binary similarity, prove that it satisfies minimality and symmetry but violates the triangle inequality, and compare it graphically to the Kullback-Leibler divergence.

Experiment design

In Chapter 4 we turned to the experiment design version of the above problem. We extended truth-finding with experiments, obtaining the experiment-design problem. Again, this problem is naturally viewed as a game. We showed that strategies for the experimenter consist of two parts: an experimentation strategy and a learning strategy. The experimentation strategy is used for experiment selection. We showed that after fixing the experimentation strategy, a truth-finding subgame remains. The learning strategy is then a strategy for this subgame. We showed in Chapter 3 that the optimal learning strategy can easily be found. Hence only a simpler game remains, where the experimenter chooses an experimentation strategy, and nature chooses the actual world. We gave a simple example that showed that the experimenter must consider mixed strategies. The minimax mixed experimentation strategy can be obtained by finding the saddle point of the doubly generalised entropy function. We proved that this function is concave-linear. Concave-linear functions are of relatively low complexity. The saddle point can be found numerically using methods from the convex optimisation literature. There is still much interesting work to be done in this area.

Future work

We return to the problem of learning polynomials. In this problem, the degree of an unknown polynomial has to be learned by sequential probing, where each probe returns the function value, perturbed by Gaussian noise. We feel that we have made an important first step towards a general solution. The following points must still be addressed:

- The truth-finding and experiment-design problem are defined for finite outcome, model and experiment sets. Can these be generalised to uncountable sample spaces?
- The truth-finding problem is defined for finite sets of models. Can these be generalised to countable sets?
- The minimax optimal pure strategy for experimenter in the sequential experiment-design problem can theoretically be obtained by solving a number of truth-finding instances that grows doubly exponential in n. Can we design smarter algorithms for this particular problem?
- In our analysis of the experiment-design problem, we have assumed that the number of experiments n is fixed and known to the experimenter. What can be said in cases where n is not fixed? Can we, for example, fix the amount of information that must be obtained, and then try to minimise the number of experiments that need to be performed?

Ultimately, solutions to these questions will allow us to solve sequential experiment-design problems of the level of complexity of the polynomial degree selection problem.

Appendix A

Measure theory

A.1 Preliminaries

Probability theory deals with probabilities of events, sets of outcomes. Probability is inextricably intertwined with countability. To preclude paradox we use measure theory, which has the following formalisation of events and their probability.

Definition A.1. Let S be a set. A set $\Sigma \subseteq \wp(S)$ is called a *σ-algebra over S* if the following conditions hold:

1. $S \in \Sigma$.
2. If $A \in \Sigma$ then $S \setminus A \in \Sigma$.
3. If $\langle A_n \rangle_{n\in\mathbb{N}}$ is a sequence of elements of Σ then $\bigcup_{n\in\mathbb{N}} A_n \in \Sigma$.

A structure $\mathbb{M} = \langle S, \Sigma \rangle$ is called a *measurable space* if Σ is a σ-algebra over S. We call the elements of Σ *events*. An important measurable space is $\langle \mathbb{R}, \mathcal{B} \rangle$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$, i.e. the smallest σ-algebra on $\mathbb{R}$ containing all open sets.

Definition A.2. Let $\mathbb{M} = \langle S, \Sigma \rangle$ be a measurable space. A function $\mu : \Sigma \to [0, \infty]$ is called a *measure on* $\mathbb{M}$ if it satisfies the properties

1. $\mu(\emptyset) = 0$.
2. If $\langle A_n \rangle_{n\in\mathbb{N}}$ is a sequence of pairwise disjoint elements of Σ, then $\mu\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} \mu(A_n)$. ($\sigma$-additivity)

Additionally, μ is called a *probability measure* if $\mu(S) = 1$.

Definition A.3. A triple $\langle S, \Sigma, \mu \rangle$ is called a *measure space (probability space)* if μ is a measure (probability measure) on $\langle S, \Sigma \rangle$.

An important measurable space is $\langle \mathbb{R}, \mathcal{B}, \mathrm{Leb} \rangle$ where the measure Leb (Lebesgue measure) is generated by closing the following function under countable additivity

$$\mathrm{Leb}([a, b]) := b - a.$$

Lebesgue measure is the uniform measure on $\mathbb{R}$; the measure of an event is given by its length.

Definition A.4. Let $\langle S, \Sigma \rangle$ and $\langle S', \Sigma' \rangle$ be measurable spaces. A function $f : S \to S'$ is called Σ, Σ'-*measurable* if

$$\forall \Phi \in \Sigma' : f^{-1}[\Phi] \in \Sigma.$$

If the second measurable space is $\langle \mathbb{R}, \mathcal{B}, \mathrm{Leb} \rangle$, we say that f is Σ-measurable.

Definition A.5. Let $\mathbb{M} = \langle S, \Sigma, \mu \rangle$ be a measure space. A function $f : S \to \mathbb{R}$ is called a *probability density function* on $\mathbb{M}$ if

- f is Σ-measurable,
- f is non-negative μ-almost everywhere, and
- $\int_S f(x) \, \mathrm{d}\mu = 1$.

A probability density function f on $\mathbb{M}$ generates a probability measure P on $\langle S, \Sigma \rangle$ defined by

$$P(\Omega) := \int_\Omega f(x) \, \mathrm{d}\mu.$$

Definition A.6. For any measurable space $\mathbb{M} = \langle S, \Sigma \rangle$, we denote by $\mathfrak{D}(\mathbb{M})$ the set of all probability measures on $\mathbb{M}$. If $\mathbb{M}$ is clear from the context, we write $\mathfrak{D}(S)$ for $\mathfrak{D}(\mathbb{M})$.

Definition A.7. We denote by $\mathcal{N}_n(\mu, \Sigma)$ the *normal distribution* on $\mathbb{R}^n$ with mean μ and $n \times n$ covariance matrix Σ. When $n = 1$, we write $\mathcal{N}(\mu, \sigma^2)$ instead.

Definition A.8. Fix a probability space $\mathbb{P} = \langle S, \Sigma, P \rangle$. A measurable function $X : S \to \mathbb{R}$ is called a *random variable*.

As stated before, a random variable transforms outcomes into real numbers. Via this transformation, we can forget about the original measure, and consider the induced measure on $\mathbb{R}$. We reserve the term *probability distribution* for a probability measure that specifies the measure of a random variable on $\mathbb{R}$.

Sometimes, it is useful to translate the set of outcomes into some set different from $\mathbb{R}$. We call such transformations *pseudo random variables* if the obvious measurability condition obtains.

Definition A.9. Let X be a random variable defined on a probability space $\mathbb{P} = \langle S, \Sigma, P \rangle$. We define the *expected value* or *expectation* of X by

$$\mathrm{E}[X] := \int_S X \, \mathrm{d}P$$

Definition A.10. Let X be a random variable on a probability space $\mathbb{P}$. We say that X is *constant* if $\exists c \forall x \in S : X(x) = c$. We call X *almost surely constant* if $\exists c : P(X = c) = 1$. This implies $P(X = \mathrm{E}[X]) = 1$.

Remark A.11. A random variable X on $\mathbb{P}$ is almost surely constant if all measure of P is assigned to a region where X is constant. This can be solely due to X, namely when X is constant, or solely due to P, namely when P puts all measure on a single point, or due to both.

Theorem A.12 (Jensen's Inequality)**.** [Wil91, Theorem, p. 61] Let $\mathcal{X}$ be a convex set, P a probability distribution on $\mathcal{X}$. Then for any convex function $f : \mathcal{X} \to \mathbb{R}$,

$$\mathrm{E}_P\left[f(X)\right] \geq f\left(\mathrm{E}_P\left[X\right]\right) \tag{A.1}$$

Moreover, if f is strictly convex, then equality in (A.1) implies that X is an almost surely constant random variable.

One level of abstraction higher, we work with a meta-distribution on sets of probability distributions. We can interpret such a meta-distribution as a prior probability; one first samples a distribution according to this meta-distribution, and then generates an outcome according to the sampled distribution. For more detail, see [GD04, Section 9.2]. Such a meta-distribution can be collapsed into a single distribution on outcomes as follows.

Definition A.13. Let $\mathcal{X}$ be a set, $\mathcal{Q}$ a convex set of distributions on $\mathcal{X}$ with σ-algebra $\Sigma_{\mathcal{X}}$, and $\mathbb{Q}$ a distribution on $\mathcal{Q}$. We define $\mathrm{E}_{\mathbb{Q}}\left[Q\right] : \Sigma_{\mathcal{X}} \to \mathbb{R}$, the *expected distribution of* $\mathbb{Q}$, by

$$\mathrm{E}_{\mathbb{Q}}\left[Q\right](X) := \mathrm{E}_{\mathbb{Q}}\left[Q(X)\right] = \int_{\mathcal{Q}} Q(X)\,\mathrm{d}\mathbb{Q}$$

where $Q = \mathbf{1}_{\mathcal{Q}}$ is a pseudo random variable.

Definition A.14. Let X, Y be pseudo random variables with range $\mathcal{X}$ and $\mathcal{Y}$. The distribution P that gives the distribution of the pair $\langle X, Y\rangle$ is called the *joint distribution* of X and Y. The *marginal distributions* of X, Y are given by

$$P_X(X = x) := \int_{\mathcal{Y}} P(X = x, Y = y)\,\mathrm{d}y \tag{A.2}$$

$$P_Y(Y = y) := \int_{\mathcal{X}} P(X = x, Y = y)\,\mathrm{d}x \tag{A.3}$$

A.2 Truth-finding frame

Definition A.15. A quadruple

$$\mathfrak{F} = \left\langle \mathfrak{Y}, \mathfrak{T}, \mathbb{M}, \langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}} \right\rangle$$

is called a *truth-finding frame*, or *frame* for short, if the following conditions hold:

- $\mathfrak{Y} = \langle \mathcal{Y}, \Sigma_{\mathcal{Y}} \rangle$ is a measurable space. We refer to $\mathfrak{Y}$ as the *outcome space*, and call the elements of $\mathcal{Y}$ *outcomes*.
- $\mathfrak{T} = \langle \mathcal{T}, \Sigma_{\mathcal{T}} \rangle$ is a measurable space. We refer to $\mathfrak{T}$ as the *possible-world space*, and call the elements of $\mathcal{T}$ *possible worlds*.
- $\mathbb{M}$ is a finite partition of $\mathcal{T}$, with the additional demand that $\mathbb{M}$ is a sub σ-algebra of $\Sigma_{\mathcal{T}}$.
- $\langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}} : \mathcal{Y} \times \mathcal{T} \to \mathbb{R}_+$ is $\Sigma_{\mathcal{Y}} \times \Sigma_{\mathcal{T}}$-measurable, and $\mathbf{p}_{\mathcal{T}}$ is a probability density function on $\mathfrak{Y}$ for each $\theta \in \mathcal{T}$.

When $\mathfrak{Y}$ and $\mathfrak{T}$ are clear from the context, we write $\mathfrak{F} = \left\langle \mathcal{Y}, \mathcal{T}, \mathbb{M}, \langle \mathbf{p}_\theta \rangle_{\theta \in \mathcal{T}} \right\rangle$ instead.

Appendix B

Concavity of generalised entropy

The generalised entropy function $\mathfrak{H}$ is concave. This important result will be shown after the following theorem, which is rather useful in proving convexity results in information theory.

Theorem B.1 (Log sum inequality, [CT90, Theorem 2.7.1])**.** For non-negative $a_1, \dots a_n$ and $b_1, \dots b_n$

$$\sum_{i=1} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1} a_i \right) \log \frac{\sum_{i=1} a_i}{\sum_{i=1} b_i}$$

Theorem B.2. $\mathfrak{H}(\mathsf{P}_\Theta)$ is concave.

Proof. Recall from Remark 3.2.6 that a measure P_Θ induces a measure on $\mathcal{T} \times \mathcal{Y}$. This measure can be transformed using M into a measure on $\mathbb{M} \times \mathcal{Y}$. We rewrite (3.19) to

$$\mathfrak{H}(\mathsf{P}_\Theta) = \mathrm{E}_{M,Y} \left[-\log \frac{P(M,Y)}{P(Y)} \right].$$

For arbitrary distributions P and Q on $\mathcal{T}$, and a real number $0 \leq \lambda \leq 1$, define $P^\lambda := \lambda P + (1-\lambda)Q$. We need to show

$$\mathfrak{H}(P^\lambda) \geq \lambda \mathfrak{H}(P) + (1-\lambda)\mathfrak{H}(Q). \tag{B.1}$$

For each $y \in \mathcal{Y}$ and $\mathcal{M} \in \mathbb{M}$, we apply Theorem B.1 with substitutions

$$\begin{array}{lll} n = 2 & a_1 = \lambda P(y, \mathcal{M}) & b_1 = \lambda P(y) \\ & a_2 = (1-\lambda) Q(y, \mathcal{M}) & b_2 = (1-\lambda) Q(y), \end{array}$$

obtaining

$$\begin{aligned} &\lambda P(\mathcal{M}, y) \log \frac{\lambda P(\mathcal{M}, y)}{\lambda P(y)} + (1-\lambda) Q(\mathcal{M}, y) \log \frac{(1-\lambda) Q(\mathcal{M}, y)}{(1-\lambda) Q(y)} \\ &\quad \geq \big(\lambda P(\mathcal{M}, y) + (1-\lambda) Q(\mathcal{M}, y) \big) \log \frac{\lambda P(\mathcal{M}, y) + (1-\lambda) Q(\mathcal{M}, y)}{\lambda P(y) + (1-\lambda) Q(y)} \end{aligned} \tag{B.2}$$

which by observing

$$\begin{aligned}\lambda P(y) + (1-\lambda)Q(y) &= \lambda \sum_{\mathcal{M}} P(\mathcal{M}, y) + (1-\lambda) \sum_{\mathcal{M}} Q(\mathcal{M}, y) \\ &= \sum_{\mathcal{M}} \big(\lambda P(\mathcal{M}, y) + (1-\lambda) Q(\mathcal{M}, y)\big) \\ &= \sum_{\mathcal{M}} P^{\lambda}(\mathcal{M}, y) \\ &= P^{\lambda}(y)\end{aligned} \tag{B.3}$$

reduces to

$$\begin{aligned}\lambda P(\mathcal{M}, y) \log P(\mathcal{M}|y) + (1-\lambda) Q(\mathcal{M}, y) \log Q(\mathcal{M}|y) \\ \geq P^{\lambda}(\mathcal{M}, y) \log P^{\lambda}(\mathcal{M}|y).\end{aligned} \tag{B.4}$$

Summation of the opposite of both sides over $\mathbb{M}$ and $\mathcal{Y}$ yields

$$\mathfrak{H}(P^{\lambda}) = -\sum_{\mathcal{M}, y} P^{\lambda}(\mathcal{M}, y) \log P^{\lambda}(\mathcal{M}|y) \tag{B.5}$$

$$\geq -\lambda \sum_{\mathcal{M}, y} P(\mathcal{M}, y) \log P(\mathcal{M}|y) - (1-\lambda) \sum_{\mathcal{M}, y} Q(\mathcal{M}, y) \log Q(\mathcal{M}|y) \tag{B.6}$$

$$= \lambda \mathfrak{H}(P) + (1-\lambda) \mathfrak{H}(Q) \tag{B.7}$$

This proves (B.1), completing the proof. □

We can decompose a joint distribution P on $\mathbb{M} \times \mathcal{Y}$ into P_M and $\mathsf{P}_{Y|\mathcal{M}}$ for each $\mathcal{M} \in \mathbb{M}$, the latter jointly denoted by $\mathsf{P}_{Y|M}$.

Corollary B.3. $\mathfrak{H}(\mathsf{P}_M, \mathsf{P}_{Y|M}) := \mathfrak{H}(P)$ is concave in all its arguments.

Proof. $\mathfrak{H}\Big(\lambda \mathsf{P}_M + (1-\lambda) Q_M, \lambda \mathsf{P}_{Y|M} + (1-\lambda) Q_{Y|M}\Big) = \mathfrak{H}(\lambda P + (1-\lambda) Q)$. □

Corollary B.4. By Theorem 2.9, noting that $\mathfrak{D}(\mathbb{M})$ is convex, the function

$$G(\mathsf{P}_{Y|M}) := \sup_{\mathsf{P}_M} \mathfrak{H}(\mathsf{P}_M, \mathsf{P}_{Y|M})$$

is concave.

Appendix C

Notation table

Table C.1 Notation for sets, pseudo-random variables and elements

Sort	Set	P.R.V.	Element	Distribution
Possible world	$\mathcal{T}$	Θ	θ	P_Θ
Experiment	Ξ	X	ξ	$\boldsymbol{\xi}$
Outcome	$\mathcal{Y}$	Y	y	$\mathbf{p}$
Model	$\mathbb{M}$	M	$\mathcal{M}$	$\mathbf{m}$
Learning strategy	$\mathcal{F}$	F	f	P_F
History	$\mathcal{H}$	H	h	$-$
Ball weight	$\mathcal{W}$	W	w	$-$
Ball index	$\mathcal{I}$	I	i	$-$
Code index	$\mathcal{J}$	J	j	Q

Table C.2 Notation and type of important functions

Function	Sym	Domain		Range
Learning strategy	f	$\mathcal{Y}^n$	$\to$	$\mathfrak{D}(\mathbb{M})$
Classification	M	$\mathcal{T}$	$\twoheadrightarrow$	$\mathbb{M}$
Mechanics	$\mathbf{p}$	$\mathcal{T}$	$\to$	$\mathfrak{D}(\mathcal{Y})$
Experimental mechanics	$\mathbf{p}$	$\mathcal{T} \times \Xi$	$\to$	$\mathfrak{D}(\mathcal{Y})$

Table C.3 Players

Name	Role
Nature	Chooses the actual world
Chance	Impartial player, used to sample outcomes
Learner	Chooses a distribution on models (given outcomes)
Experimenter	Sequentially chooses experiments, then chooses a distribution on models

Bibliography

[Bin91] Ken Binmore, *Fun and games — a text on game theory*, D. C. Heath & Co., 1991.

[BV04] Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, New York, NY, USA, 2004.

[CT90] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Wiley series in telecommunications, John Wiley, New York, 1990.

[CV95] Kathryn Chaloner and Isabella Verdinelli, *Bayesian experimental design: A review*, Statistical Science **10** (1995), 273–304.

[Fer67] Thomas S. Ferguson, *Mathematical statistics: A decision theoretic approach*, Probability and mathematical statistics, Academic press, 1967.

[GCSR04] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin, *Bayesian data analysis*, second ed., Texts in statistical science, Chapman & Hall/CRC, 2004.

[GD04] Peter D. Grünwald and A. Philip Dawid, *Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory*, The Annals of Statistics **32** (2004), no. 4, 1367–1433.

[GMP05] Peter D. Grünwald, I.J. Myung, and M. Pitt (eds.), *Advances in minimum description length: Theory and applications*, ch. 1 and 2, MIT Press, April 2005.

[LV93] Ming Li and Paul M.B. Vitányi, *An introduction to Kolmogorov complexity and its applications*, Springer-Verlag, Berlin, 1993.

[Mit97] Tom Mitchell, *Machine learning*, McGraw Hill, 1997.

[Puk93] Friedrich Pukelsheim, *Optimal design of experiments*, Wiley series in probability and mathematical statistics, John Wiley & Sons, Inc., 1993.

[SW00] Paola Sebastiani and Henry P. Wynn, *Maximum entropy sampling and optimal Bayesian experiment design*, Journal of the Royal Statistical Society B **62** (2000), no. 1, 145–157.

[vE06] Tim van Erven, *The momentum problem in MDL and Bayesian prediction*, Master's thesis, Universiteit van Amsterdam, May 2006.

[Wik06] Wikipedia, *Leaning tower of Pisa — Wikipedia, the free encyclopedia*, 2006, [Online; accessed 30-October-2006].

[Wil91] David Williams, *Probability with Martingales*, Cambridge university press, 1991.

Index