Practical foundations for probability: Prediction methods and calibration

Benedikt Höltgen
University of Tübingen

June 30, 2024

Abstract

Although probabilistic statements are ubiquitous, probability is still poorly understood. This shows itself, for example, in the mere stipulation of policies like expected utility maximisation and in disagreements about the correct interpretation of probability. In this work, we provide an account of probabilistic predictions that explains when, how, and why they can be useful for decision-making. We demonstrate that a calibration criterion on finite sets of predictions allows one to anticipate the distribution of utilities that a given policy will yield. Based on this, we specify assumptions under which expected utility maximisation is a sensible decision criterion. We also introduce the notion of prediction methods and argue that all probabilities are outputs of such prediction methods. This helps to explain how the calibration criterion can be satisfied and to show that also supposedly objective probabilities are model-dependent. We compare our account of probability with common interpretations and show that it recovers key intuitions behind the latter. We, thus, provide a novel account of what probabilities are and how they can enable successful decision-making under uncertainty.

1 Introduction

We make probabilistic statements and use probabilistic reasoning all the time: If the predicted 'probability of rain' is sufficiently high, you bring your umbrella or even stay at home. You decide to undergo surgery if this is thought to significantly 'raise your chances' of recovery. Given that such probabilistic statements permeate both science and our everyday lives, it is quite remarkable that it is still an open question what exactly we mean by them and how they are useful: Do they refer to degrees of belief, to relative frequencies of repeated trials, or to physical properties? Beyond philosophical curiosity, the meaning of probability is considered to 'bear at least indirectly, and sometimes directly, upon central scientific, social scientific, and philosophical concerns' (Hájek, 2019). Also in Machine Learning, the meaning of probability is increasingly recognised as a 'pressing question' (Burhanpurkar et al., 2021); as put by Cynthia Dwork, 'without an answer to this definitional question, we don't even know what it is that the ideal algorithm should satisfy' (Dwork, 2022).

The aim of this paper is to provide an account of predictions and prediction methods that explains what probabilities are as well as when, how, and why they are useful for decision-making. It is often simply assumed that probabilities should be action-guiding through (some variant of) expected utility maximisation (EUM): EUM is assumed, for example, as a premise in arguments about rational behaviour in philosophy (Hedden, 2013), as a model

of human behaviour in microeconomics, as a theoretical basis for Machine Learning in the form of expected risk minimisation, and even for the interpretation of evaluation scores in weather forecasting (Palmer and Richardson, 2014). But considering that the predicted event either does or does not occur, it is not well understood why and when decisions based on EUM or other policies lead to desirable outcomes. In this paper, we approach the concept of probability by taking finite sets of (probabilistic) predictions as our starting point. We show that probabilities are useful because they allow us to predict how many events from a given set will actually occur. More specifically, we demonstrate that when predictions are calibrated on relevant sets of events – that is, when the sum of the predictions coincides with the number of occurring events –, then we can predict the distribution of utilities we receive from a specified set of decisions or policy. Based on this, we can explicitly state assumptions under which, for example, EUM actually leads to desired outcomes. Our focus on sets of predictions and relationships between them highlights the importance of the prediction methods that generate them. These methods combine the construction of representations with the application of a predictor (which is basically a mathematical function). We argue that instead of relative frequencies, objective properties, or (models of) degrees of belief, probabilities should generally be seen as outputs of prediction methods. We show that this account not only captures key intuitions behind conventional interpretations of probability but also reveals a generalised version of the reference class problem. While it has been argued that it 'is your problem too' (Hájek, 2007), we suggest that it need not be seen as a problem at all. In sum, our notion of calibrated prediction methods can explain successful decision-making under uncertainty in everyday life rather than in theoretical Dutch books (as used by Bayesians to justify probabilistic reasoning) and provides a novel, useful, and general way of thinking about probability.

The paper is structured as follows. In Section 2, we demonstrate how predictions that are calibrated on relevant sets help us to make actually good decisions (in terms of utility). In Section 3, we introduce a novel notion of prediction methods and show that they cover not only obvious examples like rain forecasts but also supposedly objective probabilities such as relative frequencies and gambling odds. In Section 4, we connect all this to the literature on interpretations of probability and argue that our account captures key intuitions behind other interpretations without inheriting their problems, arguably making them obsolete. Section 5 concludes with a brief summary and an outlook on practical implications for algorithmic predictions.

2 Calibrated sets of predictions

Although it is often assumed that probabilistic predictions are somehow useful for decision-making, it has not been demonstrated in general terms how or under which conditions this is the case. In this slightly technical section, we show how *sets* of predictions are useful to us if they satisfy a form of calibration (Section 2.1): They should allow us to predict how many events from relevant sets of events will actually occur. Although it is almost trivial to specify the calibration assumptions under which we can predict the *distribution of utilities received by specified policies*, this seems to not have been discussed before, let alone its relevance appreciated. We illustrate this theoretical insight with an example involving rain forecasts (Section 2.2). We will show in Section 3.2 why prediction methods should satisfy Kolmogorov's axioms, but for the purpose of this section, it is enough to think of predictions as arbitrary real numbers $p_i \in \mathbb{R}$. Here, a prediction p_i relates to an event A_i

and its label $y_i \in \{0, 1\}$ where $y_i = 1$ or $y_i = 0$ denote that the event does or does not occur, respectively.[1]

2.1 Predicting utility

What is the difference between a prediction of 0.6 and a prediction of 0.9, given that the predicted event either does or does not occur? An important difference surfaces when considering multiple predictions: In general, of 100 events with prediction 0.6, we expect roughly 60 to occur, whereas of 100 events with prediction 0.9, we expect roughly 90 to occur. We can formalise this as a quality criterion for predictions called *calibration*: For a given set of events, the sum of our predictions should coincide with the number of occurring events.

Definition 1 (Calibration)**.**
Predictions $p_1, ..., p_d \in \mathbb{R}$ *are said to be* calibrated *for observations* $y_1, ..., y_d \in \{0, 1\}$ *if they satisfy*

$$\sum_{i=1}^{d} p_i = \sum_{i=1}^{d} y_i. \tag{1}$$

Often, calibration is understood more narrowly as what Dawid (2017) calls 'probability calibration', namely calibration *on sets of equal prediction*. This unnecessarily narrow understanding of calibration may be partly due to historical reasons, as discussed in (Höltgen and Williamson, 2023), but also for its relevance in many settings (Section 2.2). Indeed, we will argue that predicting how many events of certain sets will occur, i.e. calibration in this general sense, is the purpose for having probabilities in the first place. For arbitrary predictions, there is in general no reason to assume that such a criterion would be satisfied. This is already a first hint at the importance of considering the methods that generated the predictions, which we will turn to in Section 3. For now, we focus on implications of calibration rather than how to achieve it.

In order to demonstrate how calibrated predictions are useful for decision-making, we will consider the utilities gained (or losses incurred) in different events. Intuitively, a person's utility is a numerical representation of how much the person values the (non-)occurrence of an event (which is clearly an idealisation). In our setting of binary events, we will use utility functions $u_i : \{0, 1\} \mapsto \mathbb{R}$ where $u_i(0)$ and $u_i(1)$ capture how much I value $y_i = 0$ and $y_i = 1$, respectively. Now calibration can help us to foresee the (non-normalised) distribution of utilities that I will receive: If my predictions are calibrated on sets of equal utility, then I can predict the number of occurring events for each utility, by summing up the relevant predictions. While we will come back to general utility distributions at the end of Section 2.2, we will for now focus on a particularly intuitive property of that distribution which we call cumulative utility.

Definition 2 (Cumulative utility)**.**
My cumulative utility *over a set of outcomes* $\{y_1, ..., y_d\}$ *given utility functions* $u_1, ..., u_d$ *is*

$$\sum_{i=1}^{d} \left[y_i u_i(1) + (1 - y_i) u_i(0) \right]. \tag{2}$$

[1] While this section thus uses a very general notion of prediction, we will argue in Section 3 that predictions ultimately need to be understood in relation to the respective methods that generate them.

As y_i denotes whether event 1 or 0 occurs, the cumulative utility is the *sum over all utility that I actually receive*. This captures how well I will be off overall. We can formally show that with suitably calibrated predictions, it is possible to estimate cumulative utility through a very familiar quantity. The simple idea is that if for each utility value, I correctly predict how many events with this utility will occur, then I will correctly predict my cumulative utility. Note that for our purposes, it would be mathematically equivalent to consider the average utility I get at each time step, i.e. the cumulative utility divided by d.

Proposition 3 (Predicting cumulative utility)**.**
Let there be d predictions $p_i \in \mathbb{R}$ for binary outcomes $y_i \in \{0,1\}$, $i \in \{1,...,d\}$ with utility functions $u_i : \{0,1\} \to \mathbb{R}$ and assume that the predictions are calibrated on sets of equal utility $u_i(0)$ and on sets of equal utility $u_i(1)$ (formalised in (8) below).
Then I can correctly predict my cumulative utility (LHS) via

$$\sum_{i=1}^{d} \left[y_i u_i(1) + (1-y_i)u_i(0) \right] = \sum_{i=1}^{d} \left[p_i u_i(1) + (1-p_i)u_i(0) \right]. \tag{3}$$

Proof.
Let $\mathcal{U}$ denote the set of all values that the u_i can take, i.e. $\mathcal{U} := \bigcup_{1 \le i \le d} \{u_i(0), u_i(1)\} \subset \mathbb{R}$.

$$\sum_{i=1}^{d} \left[y_i u_i(1) + (1-y_i)u_i(0) \right] \tag{4}$$

$$= \sum_{u' \in \mathcal{U}} \left(\sum_{i:u_i(1)=u'} y_i + \sum_{i:u_i(0)=u'} (1-y_i) \right) \cdot u' \tag{5}$$

$$= \sum_{u' \in \mathcal{U}} \left(\sum_{i:u_i(1)=u'} p_i + \sum_{i:u_i(0)=u'} (1-p_i) \right) \cdot u' \tag{6}$$

$$= \sum_{i=1}^{d} \left[p_i u_i(1) + (1-p_i)u_i(0) \right] \tag{7}$$

where for (5) = (6), we use the calibration assumption that $\forall u' \in \mathcal{U}$:

$$\sum_{i:u_i(1)=u'} y_i = \sum_{i:u_i(1)=u'} p_i \qquad \text{and} \qquad \sum_{i:u_i(0)=u'} y_i = \sum_{i:u_i(0)=u'} p_i. \tag{8}$$

□

Given that the assumption of *exact* calibration on all sets of *equal* utility is extremely strong, it is worth noting that approximate calibration (Appendix A.1) and calibration on sets of approximately equal utility (Appendix A.2) suffice for approximately correct predictions of the cumulative utility. In Appendix A.3, we show that a weaker calibration criterion for imprecise predictions makes it possible to also incorporate risk aversion. Also note that in the conventional probabilistic framework, where $\hat{\mathrm{Y}}_i$ is the random variable taking value 1 rather than 0 with probability p_i and P_i denotes its distribution, the RHS of (3) is the sum over my expected utilities (my expected cumulative utility):

$$\sum_{i=1}^{d} \left[p_i u_i(1) + (1-p_i)u_i(0) \right] = \sum_{i=1}^{d} \mathbb{E}_{P_i}[u_i(\hat{\mathrm{Y}}_i)]. \tag{9}$$

This shows that maximising expected utility is the policy that maximises my cumulative utility if my predictions are calibrated! To capture the idea of maximising utility, we need a notion of decisions between acts, which is not yet a part of the setup. For simplicity, we consider d binary decisions, at step i consisting in a choice between (p_i^a, u_i^a) and (p_i^b, u_i^b).

Corollary 4 (Comparing policies by expected utility)**.**
For $i \in \{1, ..., d\}$, let there be predictions $p_i^a, p_i^b \in \mathbb{R}$ for binary outcome $y_i \in \{0, 1\}$ and utility functions $u_i^b, u_i^b : \{0, 1\} \to \mathbb{R}$. For $\pi \in \{a, b\}$, policy π is then given by predictions $p_1^\pi, ..., p_d^\pi$ and utility functions $u_1^\pi, ..., u_d^\pi$. Assume that the predictions of both policies are calibrated on sets of equal utility in the sense of (8) for their respective utility functions. Then, for $\hat{\mathsf{Y}}_i^\pi$ and P_i^π as in (9), the policy with the higher expected cumulative utility $\sum_{i=1}^d \mathbb{E}_{P_i^\pi}[u_i^\pi(\hat{\mathsf{Y}}_i^\pi)]$ will actually provide the higher cumulative utility.

2.2 Example: Rain forecasts

We now illustrate the above with an example. Let there be d days and for $i \in \{1, ..., d\}$, let $p_i \in \{0, 0.1, 0.2, ..., 1\}$ be the daily rain forecast (which I cannot influence) and let $y_i \in \{0, 1\}$ denote whether it actually rains ($y_i = 1$ denoting rain). Now assume that across days, my attitude towards rain does not change over time but that it depends on whether I have an umbrella: Let my utilities be given by $u^a(1) = 0$ and $u^a(0) = -1$ if I brought an umbrella and $u^b(1) = -3$ and $u^b(0) = 0$ if I did not bring one. Then my predicted utility when bringing an umbrella on day i is

$$p_i \cdot u^a(1) + (1 - p_i) \cdot u^a(0) = (1 - p_i) \cdot (-1) = p_i - 1 \tag{10}$$

whereas my predicted utility when not bringing an umbrella is

$$p_i \cdot u^b(1) + (1 - p_i) \cdot u^b(0) = -3p_i. \tag{11}$$

As $p_i - 1 > -3p_i \Leftrightarrow p_i > 0.25$, I maximise predicted utility (per day) if I bring an umbrella on days where $p_i > 0.25$. This policy can be defined by

$$u_i = \begin{cases} u^a & \text{if } p_i > 0.25 \\ u^b & \text{if } p_i < 0.25. \end{cases} \tag{12}$$

The calibration criterion (8) in Proposition 3 for this case amounts to

$$\sum_{i:p_i<0.25} y_i = \sum_{i:p_i<0.25} p_i \qquad \text{and} \qquad \sum_{i:p_i>0.25} y_i = \sum_{i:p_i>0.25} p_i. \tag{13}$$

If this condition is satisfied, my cumulative utility will coincide with the sum of my daily predicted utilities.

To assess the relative merits of this particular policy, we need to compare it with other policies. (13) is one instance of a prediction-dependent threshold policy, where I bring an umbrella whenever the predicted probability of rain is higher than a certain threshold (in this case, 0.25). For Proposition 3 to apply to such a policy with any threshold $t \in [0, 1]$, the calibration criterion is

$$\sum_{i:p_i<t} y_i = \sum_{i:p_i<t} p_i \qquad \text{and} \qquad \sum_{i:p_i>t} y_i = \sum_{i:p_i>t} p_i. \tag{14}$$

Now note that as the possible utility functions are the same at each time step, for this condition to be satisfied for all $t \in [0, 1]$, it is enough to satisfy

$$\forall v \in \{0, 0.1, 0.2, ..., 1\} : \sum_{i:p_i=v} y_i = \sum_{i:p_i=v} p_i. \tag{15}$$

Now this is just the common condition of calibration on sets of equal prediction, which is commonly expected of rain forecasters (Gigerenzer et al., 2005) and 'even inexperienced forecasters are capable of displaying', except for extreme predictions (Sanders, 1963, p. 191), see also (Murphy and Winkler, 1977). Under this fairly benign assumption, choosing 0.25 as my threshold maximises not only my *predicted* utility among all threshold-based policies but also my *actual* cumulative utility due to Proposition 3! Hence, people can tailor their policies to their personal utilities and, thus, their decisions to the forecasts. This also demonstrates why probabilistic forecasts are useful even in a deterministic world without 'real' probabilities. We would like to highlight that calibration on sets of equal prediction thus derives its importance (and prevalence) from their frequent concurrence with the sets of equal utility when considering threshold-based policies. Note, however, that calibration cannot be the only quality criterion for predictions here, as the constant base rate predictor also satisfies (15): more refined predictions allow people to better tailor their decisions to their utilities. Another property of interest is then what is sometimes called sharpness or refinement, which relates to the information content of a predictor (DeGroot and Fienberg, 1983). The Brier score, for example, can be decomposed into two terms measuring probability calibration and sharpness, respectively (Sanders, 1963). These two properties are at odds in the sense that it is more difficult to be calibrated on sharper predictions. In this work, we focus not on the selection or construction of predictors but on how they can be useful in general.[2]

Note that the calibration criterion was fairly benign in our example because the utilities are the same each day and we restricted our comparisons to the 10 threshold-based policies (arguably the only sensible policies here).[3] The story would be more complex if we took the utility in the umbrella problem to also depend e.g. on wind speed or on the day of the week. In general, for d binary decisions, there are 2^d possible combinations of decisions! Accordingly, if we wanted to compare the cumulative utility of all possible choices via Proposition 3, this would lead to a very strong calibration criterion – in fact, it would require perfect binary predictions $p_i = y_i$. This should not be surprising, as the best combination of decisions would be to always bring an umbrella if and only if it rains, which we can only ensure if we can discriminate perfectly between rainy and dry days.

Let us now briefly consider alternatives to the cumulative utility for comparing policies. Recall that predictions calibrated on sets of equal utility not only allow us to predict the cumulative (or average) utility but the distribution of utilities more generally. Let $\mathcal{D} := \{\mu : \mathbb{R} \to \mathbb{N}_{\geq 0}\}$ denote the space of utility distributions, i.e. functions indicating how often different utility values occur.[4] Let $\mathcal{U}_\mu := \{u \in \mathbb{R} : \mu(u) > 0\}$ denote the support of a

[2]The relationship between calibration and point-wise loss functions is an interesting research avenue, see e.g. (Gopalan, Kim, and Reingold, 2023).

[3]An essentially equivalent observation has been made by Zhao et al. (2021) who show that probability calibration is necessary and sufficient for predicting the average future loss if we restrict ourselves to loss functions (our utilities) that only depend on prediction and label and prediction-based policies. Note that for us, this is only a special (albeit common) case.

[4]For $u \in \mathbb{R}$, $\mu(u) = n$ then means that the utility u occurs n times in the distribution described by μ. Instead of such distributions, one may also think of multisets. Note that calibration on sets of equal utility generally only allows us to predict *how often*, but not *when* a specified utility will be received.

utility distribution $\mu \in \mathcal{D}$, i.e. the set of utility values that do occur in the distribution μ. With this notation, we can describe the cumulative utility of a distribution $\mu \in \mathcal{D}$ as $\sum_{u \in \mathcal{U}_\mu} u \cdot \mu(u)$. Another potentially relevant property of utility distributions is the smallest received utility, $\min \mathcal{U}_\mu$. For the umbrella policies, optimising for this property would mean that we should always bring an umbrella when $p_i > 0$, as we will otherwise incur a utility of -3 at some point – assuming that the calibration condition (15) holds. The minimum is quite an extreme property of a distribution as it ignores most information about the distribution (in our case, all events except those with the lowest utility). Optimising other properties of utility distributions will lead to other decision criteria but these questions are not the focus of this paper, interesting as they may be. Instead, we focus on prediction methods and probability.

3 Prediction methods

The previous section demonstrated that predictions are useful for decision-making if they are calibrated on relevant sets of events. We considered predictions simply as numbers that fell into our lap, without reference to any relationship between them that would justify such an assumption (except for the rain forecasting example). In this section, we introduce and discuss a general notion of prediction methods; this provides a basis for discussing relationships between predictions, and calibration in particular. In short, the deployment of a prediction method in any given situation involves the construction of a representation and the application of a predictor (Section 3.1). We show that Kolmogorov's axioms are necessary and sufficient for predictors to preserve calibration between sets of events (Section 3.2) and argue that it is often indeed possible to construct calibrated prediction methods (Section 3.3). Lastly, we demonstrate that prediction methods are much more general than the rain forecast example of the previous section. Indeed, we show that supposedly objective probabilities based on symmetries (Section 3.4) or relative frequencies (Section 3.5) actually rely on prediction methods.

3.1 Predictors and prediction methods

We distinguish the (mathematical) function that takes an input and outputs a prediction (which we call predictor or model) from the more complex prediction method that is applied to an actual situation. The latter involves the construction of a representation (potentially including measurements) before applying the former (Figure 1). We illustrate this with a now familiar example.

$$\left.\begin{array}{l}\text{select predictor } \mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R} \\ \\ \text{construct representation } x \in \mathcal{X}\end{array}\right\} \quad \text{compute } \mathsf{p}(x, A)$$

Figure 1: A prediction method gives a prediction for an event A in some situation by selecting a predictor $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ with $A \in \mathcal{A}$, constructing a representation $x \in \mathcal{X}$ of the situation, and computing $\mathsf{p}(x, A)$.

Definition 5 (Predictor)**.**
A predictor *is a function* $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ *on some set* $\mathcal{X}$ *and algebra* $\mathcal{A}$*.*

Definition 6 (Prediction method)**.**
A prediction method *for an event $A \in \mathcal{A}$ is an implicit or explicit scheme for selecting a predictor $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and constructing a representation $x \in \mathcal{X}$ of the given situation.*

For rain forecasts, the prediction method consists in first taking measurements (of temperatures, air pressure, etc.) and then feeding them to a computer model (the predictor) that outputs a prediction for the probability of rain. As in other settings that we will discuss below, one of the two arguments of the predictor is effectively ignored: For rain forecasts, the algebra of events is only implicit, although it can be formalised as $\{\emptyset, \{rain\}, \{\neg rain\}, \{rain, \neg rain\}\}$.[5] By specifying a prediction p_i for rain, the other events get assigned predictions $0, (1 - p_i)$, and 1, respectively, if Kolmogorov's axioms are obeyed. We will demonstrate in Section 3.2 that they should indeed be obeyed in order to provide calibrated predictions. Our notion of prediction methods thus elucidates the relationship between gambling probabilities and weather predictions by tying together abstraction, probability theory, and successful decision-making. One could argue that outputs of predictors which do not obey these axioms should not be called probabilities, but simply real-valued forecasts or predictions. Note that the specification of events $A \in \mathcal{A}$ also involves choices of definition or measurement. For example, how much rain counts as 'no rain' or in which area it is recorded is more or less implicit – and depends on value judgements (Douglas, 2000). These choices are, however, typically not part of the prediction method so we do not discuss them further.

In our rain forecasting example, the representation constructed by the prediction method consists in taking specific measurements of temperature, air pressure, et cetera. More generally, any sort of prediction requires focusing on a subset of all the information that could be taken into account – a representation of the situation. Any given situation has an enormous amount of potentially relevant information; typically, we decide what to look at based on experience as well as common sense or expert knowledge. For rain forecasts, temperature and air pressure are more interesting quantities than the current GDP.[6] Different models for rain prediction (the predictors) can be based on different information – for example, different granularity, location, and timing of the temperature measurements. These models may also work very differently – they may rely on simple look-up tables or sophisticated simulations. They may even rely on human forecasters who also only use limited information for their forecasts. While it is difficult to speak of human predictors as stable mathematical objects, they can arguably be approximated as such. It is interesting to note that our predictors resemble the confirmation function central to logical accounts of probability such as that of Carnap (1950) or Keynes (1921) (with precursors as early as Leibniz, cf. (Hacking, 1975)). Predictors, however, are neither objective relations nor relations between propositions –they are functions of representations in a set $\mathcal{X}$ and events in an algebra $\mathcal{A}$.

We defined calibration for predictions in Section 2.1. Based on this, we can define calibration for predictors and prediction methods.

Definition 7 (Calibration of predictors and prediction methods)**.**
A predictor (or prediction method) is said to be calibrated on events $A_1, ..., A_d \in \mathcal{A}$ if its predictions $p_1, ..., p_d$ are calibrated for observations $y_1, ..., y_d$.

[5] An *algebra* over a set Ω is a set of subsets of Ω that includes both Ω and the empty set and is closed under complements, finite unions, and finite intersections. Algebras are central to Kolmogorov's axiomatisation of probability (Section 3.2).

[6] Still, even the GDP may be predictive, given that long-term weather phenomena like El Niño can have macroeconomic effects (Cashin, Mohaddes, and Raissi, 2017).

Note that implicit in the enumeration $A_1, ..., A_d \in \mathcal{A}$ are d situations to which the prediction method is applied, using representations $x_1, ..., x_d \in \mathcal{X}$. When we speak of calibration without specifying on which sets, we mean a vague notion of 'sets of interests'. Before providing more examples and arguing that calibration is often achievable, we demonstrate a general property of prediction methods, namely, that it is good for a predictor to adhere to the probability calculus.

3.2 Probabilistic predictors

Even attentive readers probably missed the interesting fact in Proposition 3 that $1 - p_i$ automatically emerged as the prediction for $1 - y_i$, without imposing Kolmogorov's axioms. We now show more generally that predictors need to satisfy these axioms (in their second argument) in order to be calibrated on certain sets.[7] Take a predictor $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and d prediction instances represented by $(x_i, A_i) \in \mathcal{X} \times \mathcal{A}$ with $p_i := \mathsf{p}(x_i, A_i)$ and let $y_i \in \{0, 1\}$ denote whether A_i occurs. Let $y_A, y_B, y_{A\cup B}, y_\Omega \in \{0, 1\}$ denote whether events $A, B, A \cup B, \Omega \in \mathcal{A}$ occur at the last instance d. $\mathcal{A}$ is an algebra over some set Ω where Ω is a sure event: It exhausts all possibilities, that is, for its label it is known that $y_\Omega = 1$.

1. *Non-negativity:* If there is a nontrivial subset $I := \{i : p_i < 0\} \subset \{1, ..., d\}$ where p predicts negative values, then p cannot be calibrated on this subset, regardless of whether the predicted events occur:
$$\sum_{i\in I} p_i < 0 \leq \sum_{i\in I} y_i. \tag{16}$$

2. *Normalisation:* Let $A_d = \Omega$ and p be calibrated on the set $\{1, ..., d-1\}$. Then p is calibrated on $\{1, ..., d\}$ if and only if $\mathsf{p}(x_d, \Omega) = 1$, regardless of x_d.

3. *Additivity:* Let $A, B \in \mathcal{A}$ be disjoint events in the sense that $y_A + y_B \leq 1$ (i.e. it cannot be that both labels are equal to 1 at the same instance); this implies $y_A + y_B = y_{A\cup B}$. Now assume p is calibrated on the set $S := \{(x_1, A_1), ..., (x_{d-1}, A_{d-1}), (x_d, A), (x_d, B)\}$, where p is used for two predictions at instance d. Then
$$\mathsf{p}(x_d, A \cup B) + \sum_{i=1}^{d-1} p_i - y_{A\cup B} - \sum_{C\in J} y_i \tag{17}$$
$$= \mathsf{p}(x_d, A \cup B) + \sum_{i=1}^{d-1} p_i - y_A - y_B - \sum_{i=1}^{d-1} y_i \tag{18}$$
$$= \mathsf{p}(x_d, A \cup B) - \mathsf{p}(x_d, A) - \mathsf{p}(x_d, B) + \mathsf{p}(x_d, A) + \mathsf{p}(x_d, B) + \sum_{i=1}^{d-1} p_i - y_A - y_B - \sum_{i=1}^{d-1} y_i \tag{19}$$
$$= \mathsf{p}(x_d, A \cup B) - \mathsf{p}(x_d, A) - \mathsf{p}(x_d, B) \tag{20}$$
where the last step uses the assumption of calibration on S. So under that assumption, p is calibrated on $\{1, ..., d\}$ with $A_d = A \cup B$ if and only if $\mathsf{p}(x_d, A \cup B) = \mathsf{p}(x_d, A) + \mathsf{p}(x_d, B)$, regardless of x_d.

[7] Related observations for calibration on sets of equal prediction have been made in (van Fraassen, 1983).

The sets that allow if-and-only-if statements are quite specific here – which resembles Dutch book arguments where any single inconsistency can in theory be exploited indefinitely. Here, the implications are more practical: If someone is perfectly calibrated on forecasting 'rain' but does not obey the probability axioms on one 'no rain' forecast, then for some utility functions (in the setting of Section 2.2), they are guaranteed to make a sub-optimal decision due to miscalibration.

We can also motivate the definition of conditional probabilities by the demand for calibration (somewhat analogous to definitions via relative frequencies). Consider the task of predicting events $A, B \in \mathcal{A}$ at d instances. For ease of presentation, assume that all d inputs coincide, i.e. $x_1 = ..., x_d = x \in \mathcal{X}$. This allows us to drop p's dependence on $x \in \mathcal{X}$ and consider a predictor $\mathsf{p} : \mathcal{A} \to [0, 1]$ in the following derivation; a more general version is presented in Appendix B. Now assume p to be calibrated on $A \cap B$ and on B across the d instances, where $y_i^{A\cap B}$ and y_i^B denote whether $A \cap B$ and B occur at instance $i \in \{1, ..., d\}$, respectively. That is, assume $\sum_{i=1}^d \mathsf{p}(A \cap B) = \sum_{i=1}^d y_i^{A\cap B}$ and $\mathsf{p}(B) = \frac{1}{d}\sum_{i=1}^d y_i^B > 0$. Then p is calibrated on $A|B$ for $\{i : y_i^B = 1\}$ (i.e. for the set of steps where B occurs, see first line below) if and only if it satisfies $\mathsf{p}(A|B) = \frac{\mathsf{p}(A\cap B)}{\mathsf{p}(B)}$:

$$
\begin{aligned}
\sum_{i:y_i^B=1} \mathsf{p}(A|B) &= \sum_{i:y_i^B=1} y_i^A \\
\sum_{i:y_i^B=1} \mathsf{p}(A|B) &= \sum_{i:y_i^B=1} y_i^{A\cap B} && \text{(since } y_i^A = y_i^{A\cap B} \text{ when } y_i^B = 1\text{)} \\
\sum_{i:y_i^B=1} \mathsf{p}(A|B) &= \sum_{i=1}^{d} y_i^{A\cap B} && \text{(since } y_i^{A\cap B} = 0 \text{ when } y_i^B = 0\text{)} \\
\sum_{i:y_i^B=1} \mathsf{p}(A|B) &= \sum_{i=1}^{d} \mathsf{p}(A \cap B) && \text{(by calibration of } \mathsf{p} \text{ on } A \cap B\text{)} \\
\mathsf{p}(B) \cdot \sum_{i=1}^{d} \mathsf{p}(A|B) &= \sum_{i=1}^{d} \mathsf{p}(A \cap B) && \text{(by calibration of } \mathsf{p} \text{ on } B\text{)} \\
\mathsf{p}(A|B) &= \frac{\mathsf{p}(A \cap B)}{\mathsf{p}(B)}.
\end{aligned}
$$

Summing up, the probability calculus can be seen as a sound and complete system for generating calibrated predictions on certain sets from calibrated predictions on related sets. While it remains an open question whether this is enough to demand that all predictors follow the probability calculus (i.e. that they are probability measures in their second argument), it does provide a rationale for it.

3.3 Induction and the feasibility of calibration

In general, there is no reason why an arbitrary set of predictions should be calibrated. It is, however, often possible to design prediction methods so that they are (approximately) calibrated on sets of interest, which simply means that they neither systematically over-, nor under-predict. In addition to rain forecasts and later examples in Sections 3.4 and 3.5, we can point to machine learning (ML) models which often aim for calibration on sets of equal prediction. It has been observed that especially modern, over-parameterised ML

models need explicit post-processing, whereas others are automatically calibrated on sets of equal prediction (Guo et al., 2017). One could argue that on a high level, humans also do something like this post-processing: if we are repeatedly over- or under-predicting on sets of interest (i.e. are not calibrated), we will (ideally) notice that; since we do not know on which of the individual events our predictions were too low/high, we systematically increase/decrease our predictions for similar predictions in the future. But why should future predictions then also be calibrated and what can we say about the relevant notion of similarity?

Nelson Goodman (1972, p. 18) already suspected 'that rather than similarity providing any guidelines for inductive practice, inductive practice may provide the basis for some canons of similarity'. Any prediction method, indeed any prediction about the future, relies on an inductive assumption: that the future will be similar to the past in some relevant way. This relevance can be made more precise for our purpose: that a prediction method which has repeatedly proven to be (approximately) calibrated on some sets in the past will be (approximately) calibrated on similar sets in the future. Below, we prove a formal result in support of this particular inductive assumption, similar to the argument for induction made in (Williams, 1947).[8] In contrast to the cited work, we are dealing not only with integers but with real numbers, which is why we draw on an established concentration inequality. In particular, we make use of the combinatorial bound provided by Hoeffding's inequality for drawing without replacement.

Proposition 8 (Calibration on samples from a population)**.**
Take a predictor $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to [0,1]$ *and* N *prediction instances represented by* $(x_i, A_i) \in \mathcal{X} \times \mathcal{A}$*. Now consider drawing a 'sample' of* d *instances from the 'population' of* N *instances. Then the samples* $\{j_1, ..., j_d\} \subset \{1, ..., N\}$ *whose average calibration error differs by more than* ϵ *from the average calibration error of the population, i.e. where*

$$\frac{1}{d}\sum_{i=1}^{d}(p_{j_i} - y_{j_i}) - \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i) \geq \epsilon, \tag{21}$$

(y_i *denoting whether* A_i *occurs and* $p_i := \mathsf{p}(x_i, A_i)$*) make up for less than* $\exp\left(-\frac{1}{2}d\epsilon^2\right)$ *of all possible samples of that size.*

Proof.
Our result follows directly from Hoeffding's inequality for drawing without replacement. We simply insert $z_i := p_i - y_i$, $a = -1$ and $b = +1$ in the below statement taken from Proposition 1.2 of (Bardenet and Maillard, 2015):

Let $Z = (z_1, ..., z_N)$ be a finite population of N points and $Z_1, ..., Z_d$ be a random sample drawn without replacement from Z. Let

$$a := \min_{1 \leq i \leq N} z_i \quad \text{and} \quad b := \max_{1 \leq i \leq N} z_i. \tag{22}$$

Then for all $\epsilon > 0$,

$$\mu\left[\frac{1}{d}\sum_{i=1}^{d} Z_i - \frac{1}{N}\sum_{i=1}^{N} z_i \geq \epsilon\right] \leq \exp\left(-\frac{2d\epsilon^2}{(b-a)^2}\right), \tag{23}$$

[8] Note that we need not mean to provide an a priori justification for induction here: our goal is not to solve any (old or new) riddle of induction but to elucidate the role of probability in it.

where μ measures the proportion of admissible combinations in drawing d of the N points. □

Hence, the average calibration error on large enough samples will mostly be close to the average calibration error of the whole population. In a move analogous to that of (Williams, 1947), we can also infer that if I am approximately calibrated on a large enough sample from a population or set of prediction instances, I will in most cases also be calibrated on the whole set and, thus, on similar sets in the future. Let us illustrate the bound with concrete numbers. If I have an average calibration error of 0.2 on the whole population, then I will get a calibration error of less than 0.05 in less than 10% of possible samples of size $d = 200$; for $d = 500$, this goes down to 0.4% of samples. Hence, the vast majority of possible samples will not mislead me into thinking that I will be well-calibrated in the future in such a setting. Note that the proposition only provides an upper bound and the actual number of non-representative samples will be yet lower. While this result does not prove the possibility of induction, it shows that calibration in the past is an indicator for calibration in the future – on sets that can be thought to be drawn from the same population. Whether this is a sensible model in a given situation depends on whether there is reason to believe that the sample is unbiased.[9] Another interesting implication of the result concerns the mixing of predictions from multiple different calibrated prediction methods. If n prediction methods are calibrated on d events each, then the resulting $n \cdot d$ predictions are clearly also calibrated on the $n \cdot d$ events; Proposition 8 can now also be applied to this larger set of predictions, now inferring from the population to subsets: It shows that most large enough subsets of these $n \cdot d$ predictions will also be approximately calibrated, even though they come from a mix of prediction methods. This is important because it shows that Proposition 3 is not only relevant for predictions from the same prediction method.

3.4 Example: Symmetry-based predictions

In many settings, probabilities are (typically) not thought to be model-dependent, to be based on a particular prediction method; this includes probabilities for symmetrical gambling devices. Assume you go to a casino where they offer a novel game based on a symmetric 8-sided and a symmetrical 20-sided 'die' (an octahedron and an icosahedron). Given that it is an official casino, you assume that the dice are indeed symmetrical. How do you make predictions? You can represent any possible outcome that you wish to predict as the set of admissible combinations of faces, which can e.g. be represented as the algebra $\mathcal{A} = 2^{\{1,...,8\}\times\{1,...,20\}}$. For the prediction, you presumably ignore the name of the croupier, the surface of the table, and so on and only focus on the symmetry of the dice. Each face of the octahedron corresponds to a prediction of $\frac{1}{8}$ and each face of the icosahedron corresponds to a prediction of $\frac{1}{20}$. How exactly this reasoning is captured by a predictor $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ is under-determined; in particular, what counts as an argument versus a part or parameter of the predictor: You may consider a predictor that only takes symmetrical dice, such that the input space $\mathcal{X} = \{x\}$ can be ignored. Alternatively, you may take $\mathcal{X} = \Delta_8 \times \Delta_{20}$ to be the set of all possible combinations of potentially biased 8-sided and 20-sided dice, of which you consider the element $x := (\frac{1}{8}, ..., \frac{1}{8}, \frac{1}{20}, ..., \frac{1}{20}) \in \mathcal{X}$, reflecting your assumption of fairness. You may also take a still larger $\mathcal{X}$ that can also capture dice with other numbers of faces. In any case, you proceed by transforming your representation of fair dice into a

[9] An illuminating analysis of the difference between the means in terms of the bias of the sampling procedure can be found in (Meng, 2018).

number in $[0, 1]$ through a combinatorial model p. You construct a representation and you calculate.

Such symmetry-based prediction methods in gambling situations typically assume that the setup is 'fair'. Indeed, gambling devices are produced in such a way that each outcome should occur equally often, which allows us to make roughly calibrated predictions on large samples.[10] We know from experience that the process of rolling a fair die is so opaque and chaotic that it is practically impossible for us to predict better than uniformly. If we were more proficient in discerning minuscule variations in die-throws and background conditions (as Laplace's demon would be), we might be able to make finer predictions. Indeed, Edward Thorp and Claude Shannon developed a device to better predict roulette outcomes which they successfully deployed in casinos in the 1960s (Thorp, 1998).[11] After all, the assumption of a fair die or roulette wheel is a particular way of representing (your beliefs about) the situation you are in.

3.5 Example: Frequency-based predictions

Predictions that are explicitly based on looking up relative frequencies of similar events also use a very simple type of prediction method. Such a method could be used for gambling instead of (or combined with) symmetry-based considerations. There are also more interesting examples, such as medical risks. Doctors typically base risk predictions on past experience, sometimes by explicitly looking up data about similar people. In an example recently discussed in (Dawid, 2017), which we shall return to later, Angelina Jolie got told that she had a 87% risk of breast cancer – with the number presumably coming from statistical data about women with a particular genetic mutation. Hence, the doctors used the following prediction method: They chose a particular representation of Angelina, as a woman with this gene mutation, and then used a predictor that is basically a look-up table. Presumably, women without that gene mutation get assigned into different categories (or 'reference classes') for which there is enough data for doctors to believe that the relative frequency in this category is stable over time. Different doctors may use different categories, that is, different representations. If the relative frequencies are indeed stable over time, the doctors will be roughly calibrated in their predictions on large enough sets of people.[12] As in the rain example, we can consider this as a case where the implicit 4-element algebra is ignored. Alternatively, we could see it as an application of a more general predictor that can output predictions for different diseases, based on multiple look-up tables. This would make $\mathcal{A}$ more complex by adding more diseases as fundamental events and means that $\mathcal{X}$ needs to be fine enough that all relevant categories for all diseases can be distinguished.

These examples may suggest that induction about calibration always reduces to stable relative frequencies of repeated trials; this is not the case. Consider simulation models in the rain prediction example: If the measurements are fine enough, it may be the case that no input $x_i \in \mathcal{X}$ occurs more than once and no prediction is issued more than once,

[10] Hacking (1975, p. 4) notes that already '[t]he dice in the cabinets of the Cairo Museum of Antiquity, which the guards kindly let me roll for a long afternoon, appear to be exquisitely well balanced.' It also seems that probabilistic calculations were already known to gamblers before mathematicians started to take an interest in it in the 17th century (Garber and Zabell, 1979).

[11] Note that the story does not prove that there is a better way to make predictions *before* the ball is started, but there may well also be such regularities.

[12] To much public as well as scholarly amazement, early 'statisticians' in the 19th century discovered numerous stable relative frequencies at the population level, such as a 'frightening regularity with which the same crimes are reproduced' (Porter (1986, p. 49) citing the highly influential Adolphe Quetelet).

implying that there are no repeated trials. Further examples are given by logistic regression or more complex Machine Learning models used for probabilistic predictions. In general, calibrated predictors may rely on some structure in the relationship between inputs and labels that does not reduce to stable relative frequencies. Otherwise, simulation-based and ML-based predictions could simply be replaced by 'reference class forecasting', as developed in (Flyvbjerg, Glenting, and Rønnest, 2004). We will elaborate on the relation between prediction methods and frequentist reasoning in Section 4.3; more generally, we now turn to the interpretation of our account.

4 An interpretation of probability

> Norms of belief are as remote from empirical claims about nature as is Hume's simpler subjectivism. Propensity theories of probability propose a physical property that cannot be recorded and does not necessitate or preclude any occurrence. [...] any limiting-frequency claim is consistent with any claim about any finite collection of events.
>
> – Clark Glymour (2001)

Even though Bertrand Russell's famous dictum that 'probability is the most important concept in modern science, especially as nobody has the slightest notion what it means' (cited in (Bell, 1945, p. 582)) is almost a century old and there have certainly been many new developments, a thorough understanding is still lacking. The purpose of this section is to argue that probabilities should generally be seen as outputs of prediction methods and used to predict how many events from a set of them will occur. While this has already been hinted at throughout, we now specifically relate this account to the literature on interpretations of probability. This topic is often thought to depend on the question of whether the world is deterministic or not – we briefly argue that this question is indeed not relevant to our account (Section 4.1). It has sometimes been argued that there are two concepts of probability – one relating to degrees of belief, the other to relative frequencies (see e.g. (Hacking, 1975) for a historical account). We show how our account relates to both, providing a comprehensive account of probability (Sections 4.2 and 4.3). Given that *sets* of predictions are at the heart of our account, we discuss what this means for *individual* predictions (Section 4.4). Lastly, we survey other interpretations of probability, highlighting similarities and differences compared to our account (Section 4.5).

4.1 The question of (in-)determinism

> During the nineteenth century it became possible to see that the world might be regular and yet not subject to universal laws of nature. A space was cleared for chance.
>
> – Ian Hacking (1990)

The question of whether the universe is deterministic is sometimes taken to bear directly on how we think about probability. For example, David Lewis (1980, p. 120) thought that objective or physical probabilities rely on indeterminism. Karl Popper (1959) proposed the propensity account of probability (according to which probabilities are physical properties) in the context of Quantum Mechanics (QM). We should briefly note that QM does not imply

that the world is indeterministic – different, empirically indistinguishable interpretations of QM disagree on this question (not even getting into the question of scientific realism). So there is certainly no need to presuppose this. Even if QM came with true 'probabilities', it is unclear that they would be of any relevance to the probabilities we deal with day-to-day, for two reasons: First, QM probabilities need to be described by a more general theory of probability than that axiomatised by Kolmogorov (Streater, 2000). Second and more importantly, even for a coin flip, we have no access to the true QM-based probabilities (assuming it is indeterministic): We are neither able to determine the initial conditions, that is, the complete wave function, nor to take into account the extremely high number of occurring quantum interactions.[13] The resulting values may, thus, vastly differ from any predictions we are able to make – which, as we have shown, are still useful.

Indeed, the intuition that probabilities are objective may depend less on QM and more on the often strong interpersonal agreements about 'correct' prediction methods e.g. for gambling. In the words of Michael Strevens (2006, p. 31), probabilities in such settings 'have attained a certain kind of stability under the impact of additional information. This stability gives them the appearance of objectivity, hence of reality, hence of physicality'. We argued that this is misleading in Section 3.4, as illustrated by the roulette story of Thorp and Shannon. In line with this, the 'erosion of determinism' indeed did not follow the advent of QM but of higher-level statistical regularities discovered during the previous century, as captured by Hacking's epigraph above.[14] On the contrary, this erosion, if anything, facilitated the formulation of QM. In sum, we do not see compelling reasons to either believe in determinism or indeterminism nor to think that indeterminism at the level of QM would contribute much to probabilistic reasoning. It is therefore a strength of our account that, showing how probabilities can be constructed and used, it remains agnostic regarding the question of determinism. Our notion of prediction does not even require that the predicted events lie in the future, it suffices if the observations/labels are not available to the predictor.

4.2 Relation to degrees of belief

> Chances are degrees of belief [...]; not those of any actual person, but in a simplified system to which those of actual people, especially the speaker, in part approximate.
>
> – Frank P. Ramsey (1928/1931)

While probabilities are often thought to have a close connection to degrees of belief, as already reflected in Pascal's wager, we hold that the notion of probability does not depend on degrees of belief. In accordance with earlier chapters, prediction methods can be used in a purely mechanical way, without any involvement of beliefs. A machine that takes measurements and uses a predictor could make decisions based on probabilities without it being plausible to ascribe to it beliefs that come in degrees. However, probabilities *can* also be used to describe human reasoning under uncertainty. In most cases, it is difficult to pin down a particular prediction method, especially as human predictions tend to be qualitative. But humans also take only specific information into account and exploit regularities such

[13] This would require a QM version of Laplace's demon.

[14] The growing popularity of numerical predictions in the 19th century was fuelled both by the broad applicability of statistical methods (Porter, 1986) and the increasing need for decision makers to communicate and justify their judgements (Porter, 1995).

as symmetries or observed relative frequencies.[15] In some cases, the gap between human reasoning and quantitative prediction methods can become fairly small – it seems that some people, modestly described as 'super-forecasters', are particularly good at making calibrated quantitative predictions (Mellers et al., 2015). As shown in Section 3.2, we can get calibrated predictions from calibrated predictions of related events using the probability calculus. This provides at least a pro tanto reason for considering the probability axioms to constitute constraints on rationality. The notion of prediction methods can also shed light on imprecise notions of (subjective) uncertainty. Particularly vague degrees of belief or disagreements between different methods can be represented by imprecise predictions (Appendix A.3)[16] while Knightian uncertainty (Knight, 1921) corresponds to the absence of a trusted prediction method.

While probabilities are not reducible to degrees of belief, the latter, in a sense, 'happen' to roughly behave like outputs of prediction methods. In Section 2.1, we derived that calibrated prediction methods can be harnessed via an expected utility maximisation policy – the same policy that is often taken to be an approximation of human decision-making: We tend to make decisions such that good outcomes seem more likely to us. There are, of course, considerable caveats. The most crucial ones are arguably diminishing marginal utility and risk aversion, already highlighted by Ramsey (1926/1931, p. 172) and analysed e.g. in (Wakker, 1994). Still, expected utility maximisation is useful as a baseline model – and, in contrast to Bayesianism, we derived rather than assumed it under specific assumptions. In Ramsey's words, it is an 'artificial system of psychology, which like Newtonian mechanics can, I think, still be profitably used even though it is known to be false' (Ramsey, 1926/1931, p. 173). In this sense, Section 3 tells us an idealised story of human reasoning and decision-making: Humans happen to implement something close to prediction methods – in that sense, probabilities can model human degrees of belief, as expressed in Ramsey's epigraph. This does not mean that probabilities *only* model degrees of belief (given that probabilities can also be used in a purely mechanical way) – on the contrary, one could say that degrees of belief aspire to be outputs of useful prediction models.

4.3 Relation to frequencies

> If we are asked to find the probability holding for an individual future event, we must first incorporate the case in a suitable reference class. An individual thing or event may be incorporated in many reference classes, from which different probabilities will result. This ambiguity has been called the *problem of the reference class*.
>
> – Hans Reichenbach (1949)

Although our approach implies that probabilities are constructed, it also explains their strong connection to observed relative frequencies. Indeed, if we restricted ourselves to calibration on sets of equal probability (as calibration is sometimes understood), the relationship would be even closer: Then, the calibration condition would be equivalent to the

[15] Note that this is compatible with Hume's position that such processes are not deliberate, that '[a]ll inferences from experience, therefore, are effects of custom, not of reasoning' (Hume, 1777, 5.1/43f). Against Bishop Butler, he held that '[c]ustom, then, is the great guide of human life' (ibid.).

[16] Such interval probabilities were also considered by De Finetti and Savage (1962) as models of degrees of belief in 'the case of a number of decision-makers who have to make a collective decision, and, second, the case of a single individual who experiences a 'kind of personality dissociation" (Feduzi, Runde, and Zappia, 2012, p. 348).

definition of probability in finitary frequentism:

$$\sum_{i=1}^{n} p_i = \sum_{i=1}^{n} y_i \qquad \Leftrightarrow \qquad p_i = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{24}$$

Instead of taking this as a definition, we think it more adequate to see it as a special case of our main quality criterion. Not just because such finitary definitions are problematic (Hájek, 1996)[17], but also because it implies a too narrow evaluation criterion.

The ties between frequentism and our account become particularly clear in reference to the so-called reference class problem. Its metaphysical version is a problem for objectivist theories like frequentism which claim a unique true probability for each event (Hájek, 2007). The epistemic version concerns the question of how a reference class should be chosen for a given event, as different choices would lead to different probabilities. This has led, for example, the cited author to argue that conditional probabilities are actually primitive, as probabilities are always conditional on a certain conceptualisation of events. While the form of our predictors does resemble a conditional probability, they are not actually conditional probabilities as $\mathcal{X}$ is just an arbitrary set rather than an event space on which probabilities are defined. Furthermore, it has been supposed that, rather than a marginal probability, '[v]arious frequentists could tell us the conditional probability that John Smith will live to age 61, *given* that he is a consumptive Englishman aged 50' (ibid., p. 582, original emphasis). This ignores that there might, for example, be different mortality tables resulting in different ratios. More generally, the choice of representation does not yet fix the prediction. This aspect is clear for prediction methods, as different methods may use the same scheme of constructing representations but different predictors, relying e.g. on different mortality tables. In sum, we agree that 'probability is a subtler idea than relative frequency' (Freedman, 1997, p. 23). We would also argue that the reference class problem is not actually a problem. Different prediction methods may be calibrated on different sets, so one can choose a prediction method that promises calibration on sets of interest. Similar observations, without reference to calibration, have been made e.g. in terms of the 'goal-dependence in scientific ontology' (Danks, 2015).

4.4 Individual predictions

> Legends of prediction are common throughout the whole Household of Man. God speaks, spirits speak, computers speak. Oracular ambiguity or statistical probability provides loopholes, and discrepancies are expunged by Faith.
>
> – Ursula K. Le Guin, *The Left Hand of Darkness* (1969)

Our discussions have focused on sets of predictions rather than individual ones. This is not a coincidence as we take probabilities to not be free-floating numbers but to rely on prediction methods which are useful when sets of predictions are calibrated.[18] But what exactly is the relation between a prediction and the corresponding event? And can we evaluate the quality of a single non-trivial prediction? That is, is a prediction of 0.6 better than a prediction of 0.4 if the predicted event occurs? What should we do if we only get a single prediction

[17] Glymour (2001) argues for what he calls an instrumentalist and approximate version of finite frequentism which takes probabilities to be descriptions of frequencies rather than defining the former through the latter. While this is not too far from our somewhat pragmatic approach in spirit, his focus is more on the description of populations through distributions and he rejects the relevance of decision theory.

[18] Of course, singletons can be (approximately) calibrated when predictions are (approximately) 0 or 1.

for some level of utility? The short and perhaps not very satisfying answer is that a single prediction does not have much meaning on its own. Its relationship to the predicted event is mediated through the prediction method and the regularity that it exploits.

The first three questions all relate to the dependence of a prediction on the method that generated it. In particular, they relate to the previous discussion of the generalised reference class problem: Which prediction method is most useful depends on which sets we want to be calibrated on (although the perfect binary predictor is always optimal). In hindsight, we can usually say which prediction would have been good or correct. The validation of prediction methods cannot, however, be thus reduced to comparisons between individual predictions. It is also important to emphasise that the probability is not a property of the event, as it is constructed and depends on the choices of both the representation and the predictor. As discussed in Sections 3.4 and 4.1, gambling setups only appear to have objective probabilities because of their relative stability under additional information. Proper scoring rules also do not help with the issue of evaluating single predictions. They do allow us to put a number on our intuition that 0.6 is somehow a better prediction than 0.4 if the predicted event occurs; but so does any notion of calibration error, such as the ℓ_1 loss deployed in Appendix A.1. After all, proper scoring rules are meant to be 'appropriate for evaluating and comparing forecasters who *repeatedly present their predictions*' (DeGroot and Fienberg, 1983, p. 12, our emphasis). We also mentioned the case of Angelina Jolie who stated 'My doctors estimated that I had an 87 per cent risk of breast cancer', with the number presumably coming from statistical data about women with a particular genetic mutation. Dawid (2017) asks, 'Was Angelina (or her doctors) right to interpret it as her own individual risk?' (p. 3456). On our account, they were – with the qualification that this risk is model-dependent and constructed rather than objective and discovered – as is any other probability.

The fourth and last question about acting on single predictions is similar but more complex. In general, we suggest that policies rather than single actions should be the subject of justification and evaluation. An ex-post evaluation of a decision would ignore the prediction and just consider whether an alternative decision would have been better in hindsight – this is not particularly helpful. Instead, what is familiar also from legal and ethical reasoning (especially deontological, but even rule-consequentialist), is to judge decisions by the reasons or maxims that they were based on.[19] For example, we showed that maximising expected utility is a sensible policy if we can assume calibration on sets of equal utility and wish to maximise cumulative utility (Section 2.1). As noted before, being calibrated for all possible combinations of decisions would require perfect discrimination. In the umbrella example of Section 2.2, we showed that the calibration criterion can be more benign when comparing a more restricted set of sensible policies. But the problem is more difficult e.g. when we only have a few predictions for particularly grave events: If we only make a few high-stakes decisions such as a choice of treatment for breast cancer (where one may even argue that the concept of numerical utility breaks down), it seems too big of an assumption to hope for calibration on such a small set. For such situations, it may then be more sensible to be risk-averse than in low-stakes settings where there are multiple events with comparable utility (cf. Buchak, 2013; Thoma, 2019).[20] A way to model this would be via imprecise calibration as explored in Appendix A.3.

[19]This has been stated in particularly succinct form by Maurice Merleau-Ponty (1955, p. 9): 'Il n'y a pas des décisions justes, il n'y a qu'une politique juste.'

[20]This creates an asymmetry for the doctor-patient relationship, but also for algorithmic predictions, similar to the insurance setting (Fröhlich and Williamson, 2024). CS Peirce (1878) thought that when probabilistic reasoning is confronted with limited trials, 'logicality inexorably requires that our interests [...] must not stop at our own fate, but must embrace the whole community'.

4.5 Comparisons with other interpretations

> Each interpretation that we have canvassed seems to capture some crucial insight into a concept of [probability], yet falls short of doing complete justice to this concept.
>
> – Alan Hájek (2019)

Any new satisfactory account of probability can be expected to make proponents of previous accounts feel vindicated on some aspects that are particularly close to their hearts. We think that this is the case for our notion of probabilities as outputs of prediction methods aiming to predict numbers of occurring events. We already commented on the relation to logical accounts of probability in Section 3.2. In this section, we briefly survey a number of other prominent interpretations and highlight what we take to be the most interesting similarities and differences w.r.t. our account.

Bayesians usually posit that probability and its theory are concerned with degrees of belief and rationality constraints thereon. What we agree with is that probabilities are constructed and that it is misguided to search for true probabilities. However, we ground them in prediction methods rather than degrees of belief (Section 4.2) and highlight that these methods aim to track structure in sets of observations. This makes it possible to replace notions of internal cohesion or rationality with that of calibration, and thereby an actual guide to decision-making. A Bayesian account that is particularly close to ours is that of (Dawid, 2017).[21] On the one hand, his suggestion to arrive at 'probability forecast[s] by assessing the odds at which I would be willing to bet' (p. 3471) is clearly Bayesian in the spirit of (De Finetti, 1937). On the other hand, he also suggests evaluating individual predictions on aggregate data via calibration – although its precise scope and relevance do not yet become entirely clear. In particular, it remains unclear why calibration on future data is important and on which (finite/infinite) sets it matters.[22] In comparison, our notion of prediction methods focuses on (potentially) inter-subjective models and the construction of representations $x \in \mathcal{X}$ which are decoupled from the events $A \in \mathcal{A}$ that we wish to be calibrated on. In a way, then, we posit a variant of Bayesianism without degrees of belief or betting and with a more concrete connection to the world, enabling not only the avoidance of sure loss in Dutch books but successful action in everyday life.

Hypothetical *frequentism* can be defined as the suggestion that 'the probability of an attribute A in a reference class B is the value the limiting relative frequency of occurrences of A within B would be if B were infinite' (Hájek, 2019). This captures the intuition of identifying probabilities with ratios in repeated trials. While this sounds very different to our account at first glance, we already discussed two similarities in Section 4.3: One is the dependence of individual probabilities on other events and a choice of representation (via prediction methods in our case), leading to a generalisation of the reference class problem. Furthermore, probabilities equating relative frequencies is a special case of our notion of calibration, which we consider for finite sets. We do reject the jump to declaring that probabilities themselves are 'out there' in any interesting sense. The account of Glymour (2001), mentioned in Footnote 17, provides, in a sense, an intermediate account.

Karl Popper abandoned frequentism in favour of his *propensity* account because the former could not make sense of sequences with few trials. He thus proposed that frequentists

[21] Dawid's work in statistics more generally, along with the mathematical work of (Shafer and Vovk, 2019) that it partially inspired, is quite close to ours in spirit.

[22] For example, his notion of H-based calibration seems to require calibration on all sets that cannot be further distinguished – which can amount to calibration on individual datapoints.

should alter their theory by letting it 'say that admissible sequences must be either virtual or actual sequences which are *characterised by a set of generating conditions* – by a set of conditions whose repeated realisation produces the elements of the sequence' (Popper, 1959, p. 34, original emphasis). This is still an objectivist theory, dispensing with the reliance on infinite trials but invoking a new sort of mysterious property (especially in the case of a deterministic universe, which Popper did not seem to assume). We argued that relevant probabilities are independent of 'true' probabilities that may or may not be implied by Quantum Mechanics (Section 4.1). Propensity accounts often have a frequentist flavour, highlighting the importance of sets of events in a rather indirect way. It is interesting to note that, as the equivalence classes of generative conditions are idealisations (ignoring background conditions, cf. Section 3.4), they can be seen as the representation constructed by prediction methods. However, propensities are typically thought to be physical rather than model-dependent, which is in stark contrast to our account – although the relevant literature sometimes also invites a reading of model-dependent propensities.

Another interesting interpretation of probability is the *best-systems account* of David Lewis (1994), which also posits objective chances: On this view, 'the chances are what the probabilistic laws of the best system say they are' (p. 480). 'The best system is the one that strikes as good a balance as truth will allow between simplicity and strength. [...] If nature is kind, the best system will be robustly best [...] It's a reasonable hope' (ibid., p. 478f). Now this account presupposes what may seem a tremendous kindness of nature as well as a perhaps weak notion of truth and objectivity – the latter fits well into Lewis' Humean view on laws of nature. What is interesting here about Lewis' account is that it resonates with the hope for a best level of predictive depth expressed in (Dawid, 2017, p. 3465) – similar to the 'primary resolution' of (Li and Meng, 2021). Indeed, Dawid could be seen as linking the best-systems view on probability with our more pragmatic notion of model-based predictions. The clearest differences on the side of Lewis are the integration within a more global systematisation of the universe and the belief in objectivity, hinging on the existence of a privileged description. If there were an objectively best predictor and we assigned to it some notion of truth, these differences would blur.[23] However, this hope for or pretension of objectivity is also what Clark Glymour criticises in typical frequentist takes.

In line with our analysis, Glymour thinks that central problems with Bayesianism and frequentism lie, respectively, in the neglect of empirical claims and the unnecessary stipulation of objectively true probabilistic statements:

> The sometimes bitter debates between those who describe themselves as frequentists and those who describe themselves as subjective Bayesians has often turned on charges by the former that the latter abandon the "objectivity" of science and by the latter that the former dissemble about the "subjectivity" of their probability judgements. My belief is that, among statisticians anyway, the dispute often confuses content with justification. The "objectivity" of the frequentists is in the content of their probability judgements, which, while usually stated as about an unempirical probability, are often really vague empirical claims about finite frequencies. That sort of objectivity is genuinely lost in subjective Bayesian interpretations. The "subjectivity" kept hidden by frequentists is that there is often no explicit justification beyond their own opinion for aspects

[23] This is perhaps not surprising given the subjectivism and pragmatism of Frank Ramsey, whom Lewis credits with a first formulation of a best-systems approach.

> of their empirical claims. That subjectivity can be made entirely explicit without sacrificing the objective–that is empirical–content of frequency claims, and its recognition does not require, or even invite, recourse to subjective probability. Bayesian criticisms do address a confused and uncertain frequentist statistical practice, in which the point of making empirical claims is often forgotten or fudged. (Glymour, 2001, p. 299f)

We have argued that our account avoids these problems by stating that probabilities are constructed rather than discovered while still taking their justification directly from empirical observations. Even more, we connect successful decision-making with empirical evaluation and assumptions about induction through a general notion of calibration, which has not been considered a central concept by any of the conventional accounts.

5 Discussion

Recognising the importance of prediction methods and calibration, we have provided a more or less pragmatic account of probabilities and how they are actually useful for decision-making under uncertainty. We showed that if predictions satisfy an (often sensible) calibration criterion, then it is possible to predict the distribution of utilities that a given policy will yield. In particular, the sum of one's predicted utilities will match the actual cumulative utility, which provides a rationale for expected utility maximisation. However, the larger the pool of considered policies, the stronger the required calibration assumption. A central element of our account is the semi-formal notion of prediction methods which construct representations of given situations and feed them to a model. Arguably, the novelty here consists less in the consideration of predictions than in the connections drawn to abstractions and successful decision-making. This covers not only rain forecasts but also supposedly objective frequency- and symmetry-based probabilities, capturing key intuitions behind other interpretations of probability. While our examples suggest that many prediction methods are approximately calibrated on sets of interest, this is not always the case. In domains where this is difficult, it may be advisable to resort to imprecise predictors or to refrain from probabilistic reasoning altogether. In line with calibration being a criterion for sets of predictions rather than individual ones, prediction methods track structures among sets of events rather than free-floating individual probabilities (even in a potentially indeterministic world). Our discussion of such structures has been very shallow – more substantial insights likely need to be more case-specific, which provides interesting avenues for further research. Given the dependence of causality on probability, the conclusions about model dependence arguably also spill over to the former; we plan to take a deeper dive into this topic in future work.

We believe that this novel account of probability has implications for the thriving area of algorithmic predictions. For example, it highlights the importance of evaluating calibration beyond common notions like the Expected Calibration Error, as explored in (Höltgen and Williamson, 2023). It also highlights that one should be wary of the often-invoked concept of a 'true distribution' from which one can 'sample', especially outside highly controlled gambling settings. Note that our insights about binary events also extend to probability distributions, as the latter are defined by cumulative distributions that specify the probabilities of values falling into certain intervals – that is, for binary events. It has been observed that algorithmic predictions based on Machine Learning tend to convey an air of authority and objectivity, as the many choices involved in data collection (construction of representations)

and model tuning (choice of predictor) often remain beneath the surface (Moss, 2022). This is particularly relevant for predictions about people. Our work highlights that probabilities, e.g. of finding a job, are not properties of people; instead, they depend on the selected representation and model, which, in turn, depends on data about other people. Hence also our answer to Cynthia Dwork's question from the introduction: there is no ideal algorithm, as there are no true probabilities to uncover and different algorithms can be better suited for different goals. The observation that probabilities always reflect structures in the data and model choices, rather than individual properties, also bears on questions concerning the collection (or rather, construction) of datasets: Which representations of people through data are admissible in different situations (Di Bello and O'Neil, 2020)? What effects does the choice of representation, which constructs the joints along which society is cut, have on societal inequalities downstream? While we hope that this work helps to sharpen the view on topics relating to probability, a sea of open questions still calls for further exploration.

Acknowledgements

For helpful feedback on previous versions, I would like to thank Bob Williamson, Jannik Thümmel, Kate Vredenburgh, Konstantin Genin, Rabanus Derr, and Timo Freiesleben.

References

Bardenet, Rémi and Odalric-Ambrym Maillard (2015). "Concentration inequalities for sampling without replacement". In: *Bernoulli* 21.3, pp. 1361–1385.

Bell, Eric Temple (1945). *The development of mathematics*. 2nd ed. McGraw-Hill Book Company.

Buchak, Lara (2013). *Risk and rationality*. Oxford University Press.

Burhanpurkar, Maya et al. (2021). "Scaffolding sets". In: *arXiv preprint arXiv:2111.03135*.

Carnap, Rudolf (1950). *Logical foundations of probability*. Unicersity of Chicago Press.

Cashin, Paul, Kamiar Mohaddes, and Mehdi Raissi (2017). "Fair weather or foul? The macroeconomic effects of El Niño". In: *Journal of International Economics* 106, pp. 37–54.

Danks, David (2015). "Goal-dependence in (scientific) ontology". In: *Synthese* 192, pp. 3601–3616.

Dawid, Philip (2017). "On individual risk". In: *Synthese* 194.9, pp. 3445–3474.

De Cooman, Gert and Jasper De Bock (2022). "Randomness is inherently imprecise". In: *International Journal of Approximate Reasoning* 141, pp. 28–68.

De Finetti, Bruno (1937). "La prévision: ses lois logiques, ses sources subjectives". In: *Annales de l'institut Henri Poincaré*. Vol. 7. 1, pp. 1–68.

De Finetti, Bruno and Leonard J Savage (1962). "Sul modo di scegliere le probabilità iniziali". In: *Biblioteca del Metron, Serie C* 1, pp. 81–154.

DeGroot, Morris H and Stephen E Fienberg (1983). "The comparison and evaluation of forecasters". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 32.1-2, pp. 12–22.

Di Bello, Marcello and Collin O'Neil (2020). "Profile evidence, fairness, and the risks of mistaken convictions". In: *Ethics* 130.2, pp. 147–178.

Douglas, Heather (2000). "Inductive risk and values in science". In: *Philosophy of Science* 67.4, pp. 559–579.

Dwork, Cynthia (2022). "Fairness, randomness, and the crystal ball". In: *Munich AI Lectures*. URL: https://www.youtube.com/watch?v=n4XftI9GOfA (visited on 04/08/2023).
Feduzi, Alberto, Jochen Runde, and Carlo Zappia (2012). "De Finetti on the insurance of risks and uncertainties". In: *The British journal for the philosophy of science*.
Flyvbjerg, Bent, Carsten Glenting, and Arne Rønnest (2004). "Procedures for dealing with optimism bias in transport planning". In: *London: The British Department for Transport, Guidance Document*.
Freedman, David (1997). "Some issues in the foundation of statistics". In: *Topics in the Foundation of Statistics*, pp. 19–39.
Fröhlich, Christian and Robert C Williamson (2024). "Insights from insurance for fair machine learning: Responsibility, performativity and aggregates". In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.
Garber, Daniel and Sandy Zabell (1979). "On the emergence of probability". In: *Archive for History of Exact Sciences*, pp. 33–53.
Gigerenzer, Gerd et al. (2005). ""A 30% chance of rain tomorrow": How does the public understand probabilistic weather forecasts?" In: *Risk Analysis: An International Journal* 25.3, pp. 623–629.
Glymour, Clark (2001). "Instrumental probability". In: *The Monist* 84.2, pp. 284–300.
Goodman, Nelson (1972). "Seven strictures on similarity". In.
Gopalan, Parikshit, Michael P Kim, and Omer Reingold (2023). "Characterizing notions of omniprediction via multicalibration". In: *arXiv preprint arXiv:2302.06726*.
Guo, Chuan et al. (2017). "On calibration of modern neural networks". In: *International Conference on Machine Learning*. PMLR, pp. 1321–1330.
Hacking, Ian (1975). *The emergence of probability*. Cambridge University Press.
— (1990). *The taming of chance*. 17. Cambridge University Press.
Hájek, Alan (1996). ""Mises redux"—redux: Fifteen arguments against finite frequentism". In: *Erkenntnis* 45, pp. 209–227.
— (2007). "The reference class problem is your problem too". In: *Synthese* 156, pp. 563–585.
— (2019). "Interpretations of probability". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University.
Hedden, Brian (2013). "Incoherence without exploitability". In: *Noûs* 47.3, pp. 482–495.
Höltgen, Benedikt and Robert C Williamson (2023). "On the richness of calibration". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1124–1138.
Hume, David (1777). *An enquiry concerning human understanding*.
Keynes, John Maynard (1921). *A treatise on probability*. Macmillan and Co.
Knight, Frank Hyneman (1921). *Risk, uncertainty and profit*. Vol. 31. Houghton Mifflin.
Le Guin, Ursula K (1969). *The left hand of darkness*. Ace Books.
Lewis, David (1980). "A subjectivist's guide to objective chance". In: *Philosophical Papers (1986)* 2, 83—132.
— (1994). "Humean supervenience debugged". In: *Mind* 103.412, pp. 473–490.
Li, Xinran and Xiao-Li Meng (2021). "A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff". In: *Journal of the American Statistical Association*.
Mellers, Barbara et al. (2015). "Identifying and cultivating superforecasters as a method of improving probabilistic predictions". In: *Perspectives on Psychological Science* 10.3, pp. 267–281.

Meng, Xiao-Li (2018). "Statistical paradises and paradoxes in big data (i) law of large populations, big data paradox, and the 2016 us presidential election". In: *The Annals of Applied Statistics* 12.2, pp. 685–726.

Merleau-Ponty, Maurice (1955). *Les aventures de la dialectique*. Gallimard.

Moss, Emanuel (2022). "The objective function: Science and society in the age of machine intelligence". In: *arXiv preprint arXiv:2209.10418*.

Murphy, Allan H and Robert L Winkler (1977). "Reliability of subjective probability forecasts of precipitation and temperature". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 26.1, pp. 41–47.

Palmer, Tim and David Richardson (2014). "Decisions, decisions...!" In: *ECMWF Newsletter*, pp. 12–14. DOI: 10.21957/bychj3cf. URL: https://www.ecmwf.int/node/17333.

Peirce, Charles Sanders (1878). "The doctrine of chances". In: 12, pp. 604–615.

Popper, Karl R (1959). "The propensity interpretation of probability". In: *The British journal for the philosophy of science* 10.37, pp. 25–42.

Porter, Theodore M (1986). *The rise of statistical thinking, 1820-1900*. Princeton University Press.

— (1995). *Trust in numbers*. Princeton University Press.

Ramsey, Frank P (1926/1931). "Truth and probability". In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R.B. Braithwaite. Routledge. Chap. VII, pp. 156–198.

— (1928/1931). "Further considerations". In: *The Foundations of Mathematics and other Logical Essays*. Ed. by R.B. Braithwaite. Routledge. Chap. VIII, pp. 199–211.

Reichenbach, Hans (1949). *The theory of probability*. University of California Press.

Sanders, Frederick (1963). "On subjective probability forecasting". In: *Journal of Applied Meteorology and Climatology* 2.2, pp. 191–201.

Seidenfeld, Teddy, Mark J Schervish, and Joseph B Kadane (2012). "Forecasting with imprecise probabilities". In: *International Journal of Approximate Reasoning* 53.8, pp. 1248–1261.

Shafer, Glenn and Vladimir Vovk (2019). *Game-theoretic foundations for probability and finance*. John Wiley & Sons.

Streater, RF (2000). "Classical and quantum probability". In: *Journal of Mathematical Physics* 41.6, pp. 3556–3603.

Strevens, Michael (2006). "Probability and chance". In: *Encyclopedia of Philosophy, second edition. Macmillan Reference USA, Detroit*.

Thoma, Johanna (2019). "Risk aversion and the long run". In: *Ethics* 129.2, pp. 230–253.

Thorp, Edward O (1998). "The invention of the first wearable computer". In: *Digest of Papers. Second International Symposium on Wearable Computers*. IEEE, pp. 4–8.

van Fraassen, Bas C (1983). "Calibration: A frequency justification for personal probability". In: *Physics, Philosophy and Psychoanalysis: Essays in Honour of Adolf Grünbaum*, pp. 295–319.

Wakker, Peter (1994). "Separating marginal utility and probabilistic risk aversion". In: *Theory and Decision* 36.1, pp. 1–44.

Walley, Peter (1991). *Statistical reasoning with imprecise probabilities*. Chapman-Hall.

Williams, Donald (1947). *The ground of induction*. Harvard University Press.

Zhao, Shengjia et al. (2021). "Calibrating predictions to decisions: A novel approach to multi-class calibration". In: *Advances in Neural Information Processing Systems* 34, pp. 22313–22324.

A Generalising Proposition 3

A.1 Approximate calibration

Here, we generalise Proposition 3 to only require approximate calibration – we give a bound on how large the calibration error on each set of equal utility can be in order to keep the difference between predicted and cumulative utility below some $\epsilon > 0$.

Proposition 9 (Predicting cumulative utility: Approximate calibration)**.**
We assume the same setting as in Proposition 3 except that we now require all utilities to be positive – one may otherwise simply shift the values to a positive domain. If we then replace condition (8) with the assumption that $\forall u' \in \mathcal{U}$*,*

$$\left| \sum_{i:u_i(0)=u'} y_i - \sum_{i:u_i(0)=u'} p_i \right| \leq \frac{\epsilon}{2 \cdot u' \cdot |\mathcal{U}|} \tag{25}$$

and the same for $u_i(1)$*, then*

$$\left| \sum_{i=1}^{d} [y_i u_i(1) + (1-y_i)u_i(0)] - \sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)] \right| \leq \epsilon. \tag{26}$$

Proof.

$$\begin{aligned}
&\left| \sum_{i=1}^{d} [y_i u_i(1) + (1-y_i)u_i(0)] - \sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)] \right| \\
&= \left| \sum_{u' \in \mathcal{U}} \left(\left(\sum_{i:u_i(1)=u'} y_i - \sum_{i:u_i(1)=u'} p_i \right) + \left(\sum_{i:u_i(0)=u'} p_i - \sum_{i:u_i(0)=u'} y_i \right) \right) \cdot u' \right| && (27) \\
&\leq \sum_{u' \in \mathcal{U}} \left(\left| \sum_{i:u_i(1)=u'} y_i - \sum_{i:u_i(1)=u'} p_i \right| + \left| \sum_{i:u_i(0)=u'} p_i - \sum_{i:u_i(0)=u'} y_i \right| \right) \cdot u' && (28) \\
&\leq \sum_{u' \in \mathcal{U}} \left(\frac{\epsilon}{2 \cdot u' \cdot |\mathcal{U}|} + \frac{\epsilon}{2 \cdot u' \cdot |\mathcal{U}|} \right) \cdot u' && (29) \\
&= \epsilon && (30)
\end{aligned}$$

□

While we use a symmetric ℓ^1 loss here, it may be interesting to also look into other measures of error. For example, for settings where under-prediction and over-prediction are valued differently, it may be instructive to look into asymmetric error functions.

A.2 Approximate utility level sets

Here, we generalise Proposition 3 to only require calibration on sets of approximately equal utility: For this, we divide the utility spectrum into bins of some size $\delta > 0$ and bound the resulting difference between predicted and cumulative utility by a term dependent on δ and the number of predictions d.

Proposition 10 (Predicting cumulative utility: Approximate utility)**.**
We assume the same setting as in Proposition 3 except that we now require all utilities to be positive – otherwise, one may simply shift the values to a positive domain. We partition the interval of relevant utilities from the lowest $u_i(a)$ to the highest $u_i(a)$ with $i \in \{1, ..., d\}, a \in \{0, 1\}$ into bins $B_1, ..., B_m$ of size $\leq \delta$. If we then replace condition (8) with the assumption that $\forall k \in \{1, ..., m\}$,

$$\sum_{i:u_i(0)\in B_k} y_i = \sum_{i:u_i(0)\in B_k} p_i \qquad \textit{and} \qquad \sum_{i:u_i(1)\in B_k} y_i = \sum_{i:u_i(1)\in B_k} p_i \tag{31}$$

then

$$\left|\sum_{i=1}^{d} [y_i u_i(1) + (1-y_i)u_i(0)] - \sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)]\right| \leq \delta \cdot d. \tag{32}$$

Proof.
The maximal mismatch occurs when for each bin B_k and each $a \in \{0, 1\}$, one half of the $\{i : u_i(a) \in B_k\}$, we have $(y_i - p_i) = 1$ and $u_i(a) = m_k + \delta/2$ whereas for the other half, $(y_i - p_i) = -1$ and $u_i(a) = m_k - \delta/2$, with m_k denoting the midpoint of B_k. This gives

$$\forall k \in \{1, ..., m\}, a \in \{0, 1\} : \quad \left|\sum_{i:u_i(a)\in B_k} (y_i - p_i) \cdot u_i(a)\right| \leq \left|\sum_{i:u_i(a)\in B_k} \delta/2\right| = b_k^a \cdot \delta/2 \tag{33}$$

where $b_k^a := |\{1 \leq i \leq d \mid u_i(a) \in B_k\}|$. Therefore,

$$\left|\sum_{i=1}^{d} [y_i u_i(1) + (1-y_i)u_i(0)] - \sum_{i=1}^{d} [p_i u_i(1) + (1-p_i)u_i(0)]\right| \tag{34}$$

$$\leq \left|\sum_{i=1}^{d} [y_i \cdot u_i(1) - p_i \cdot u_i(1)]\right| + \left|\sum_{i=1}^{d} [(1-y_i) \cdot u_i(0) - (1-p_i) \cdot u_i(0)]\right| \tag{35}$$

$$= \left|\sum_{i=1}^{d} [(y_i - p_i) \cdot u_i(1)]\right| + \left|\sum_{i=1}^{d} [(y_i - p_i) \cdot u_i(0)]\right| \tag{36}$$

$$\leq \sum_{k=1}^{m} \left(\left|\sum_{i:u_i(1)\in B_k} (y_i - p_i) \cdot u_i(1)\right| + \left|\sum_{i:u_i(0)\in B_k} (y_i - p_i) \cdot u_i(0)\right|\right) \tag{37}$$

$$\leq \sum_{k=1}^{m} \left(b_k^1 \cdot \delta/2 + b_k^0 \cdot \delta/2\right) \tag{38}$$

$$= \delta \cdot d \tag{39}$$

□

A.3 Imprecise calibration

We now consider imprecise forecasts which give interval predictions $[a, b] \subset \mathbb{R}$ and represent them as tuples $p^* = (\underline{p}, \bar{p}) \in \mathbb{R}^2$ of the lower and upper probability. This allows for a weaker calibration criterion where the number of occurring events need not exactly match the sum of predictions but should lie between the sum of the lower and the sum of the higher predictions.

Definition 11 (Imprecise calibration)**.**
Imprecise predictions $p_1^*, ..., p_d^*$ are said to be imprecisely calibrated *for observations $y_1, ..., y_d \in \{0,1\}$ if they satisfy*

$$\sum_{i=1}^{d} \underline{p}_i \leq \sum_{i=1}^{d} y_i \quad \textit{and} \quad \sum_{i=1}^{d} \bar{p}_i \geq \sum_{i=1}^{d} y_i. \tag{40}$$

Note that for our definition, the vacuous forecast that always predicts $(0,1)$ is always imprecisely calibrated.[24] One could also apply the criterion of imprecise calibration to a set of precise predictions, by simply converting every precise prediction p_i into an imprecise forecast $[p_i - \epsilon, p_i + \epsilon]$ for some ϵ – this epsilon may also monotonically decrease in d to account for lower variance on larger sets.

Proposition 12 (Predicting cumulative utility, imprecise version)**.**
Let there be d imprecise predictions $p_i^* \in \mathbb{R}^2$ for binary outcomes $y_i \in \{0,1\}$, $i \in \{1, ..., d\}$ with utility functions $u_i : \{0,1\} \to \mathbb{R}$ and assume that the predictions are imprecisely calibrated on sets of equal utility $u_i(0)$ and on sets of equal utility $u_i(1)$ (formalised in (43) below).
Then I can correctly predict a range for my cumulative utility (LHS) via

$$\sum_{i=1}^{d} [y_i u_i(1) + (1 - y_i)u_i(0)] > \sum_{i=1}^{d} \left[\underline{p}_i u_i(1) + (1 - \underline{p}_i)u_i(0)\right] \tag{41}$$

and

$$\sum_{i=1}^{d} [y_i u_i(1) + (1 - y_i)u_i(0)] < \sum_{i=1}^{d} [\bar{p}_i u_i(1) + (1 - \bar{p}_i)u_i(0)] \tag{42}$$

The proof is analogous to that of Proposition 3, now with the calibration assumptions

$$\begin{aligned} \sum_{i:u_i(1)=u'} y_i > \sum_{i:u_i(1)=u'} \underline{p}_i \quad &\text{and} \quad \sum_{i:u_i(0)=u'} y_i > \sum_{i:u_i(0)=u'} \underline{p}_i, \\ \sum_{i:u_i(1)=u'} y_i < \sum_{i:u_i(1)=u'} \bar{p}_i \quad &\text{and} \quad \sum_{i:u_i(0)=u'} y_i < \sum_{i:u_i(0)=u'} \bar{p}_i. \end{aligned} \tag{43}$$

This allows people to not only optimise their utility but to also take risk-averse or risk-seeking inclinations into account – selecting policies not based on the expected exact cumulative utility but on e.g. the lowest or highest estimation of it. Here, we can see an analogy between the move from deterministic to probabilistic and the move from precise to imprecise predictions: The former allows people to take their (cardinal) preferences into account (Section 2.2) whereas the latter allows them to take their risk aversion into account. If we know that we will be calibrated, risk aversion does not make much sense. Cases where we are less sure of it can be represented by an assumption of imprecise calibration. Note that this notion of risk-aversion also captures unwillingness to bet, for decisions between the utility function of a bet and the constant zero utility function with $u(0) = u(1) = 0$.

[24] Similar issues are discussed in the literature on imprecise probability (Walley, 1991), particularly for Brier-style scoring rules (Seidenfeld, Schervish, and Kadane, 2012) and randomness (De Cooman and De Bock, 2022, Prop. 9).

B Conditional probabilities, generalised

We here generalise the analysis of *conditional probabilities* in Section 3.2. Consider a predictor $\mathsf{p} : \mathcal{X} \times \mathcal{A} \to [0,1]$, events $A_1, ..., A_d, B \in \mathcal{A}$, and inputs $x_1, ..., x_d \in \mathcal{X}$. We assume $\mathsf{p}(x_1, B) = ... = \mathsf{p}(x_d, B)$ and that p is calibrated on $\{(x_i, A_i \cap B) : 1 \leq i \leq d\}$ and $\{(x_i, B) : 1 \leq i \leq d\}$, that is,

$$\sum_{i=1}^{d} \mathsf{p}(x_i, A_i \cap B) = \sum_{i=1}^{d} y_i^{A_i \cap B} \tag{44}$$

and

$$\mathsf{p}(x_1, B) = \frac{1}{d} \sum_{i=1}^{d} y_i^B > 0. \tag{45}$$

Then p is calibrated on $\{(x_i, A_i|B) : y_i^B = 1\}$ (i.e. for the set of steps where B occurs) if and only if it satisfies

$$\sum_{i=1}^{d} \mathsf{p}(x_i, A_i|B) = \sum_{i=1}^{d} \frac{\mathsf{p}(x_i, A_i \cap B)}{\mathsf{p}(x_i, B)}, \tag{46}$$

as we derive below. In particular, a sufficient condition is

$$\mathsf{p}(x_i, A_i|B) = \frac{\mathsf{p}(x_i, A_i \cap B)}{\mathsf{p}(x_i, B)}. \tag{47}$$

Now consider the special case where $A_i = ... = A_d =: A$ and $x_i = ... = x_d =: x$. Here, the familiar definition of conditional probabilities

$$\mathsf{p}(x, A|B) = \frac{\mathsf{p}(x, A \cap B)}{\mathsf{p}(x, B)} \tag{48}$$

is necessary and sufficient for p to be calibrated for predictions of $A|B$ based on inputs x on the set of steps where B occurs. That is, making predictions for $A|B$ rather than A allows us to be calibrated on the set where B occurs (which predictions for A would usually not be).

Now the promised derivation of the characterisation (46):

$$\sum_{i:y_i^B=1} \mathsf{p}(x_i, A_i|B) = \sum_{i:y_i^B=1} y_i^{A_i|B}$$

$$\sum_{i:y_i^B=1} \mathsf{p}(x_i, A_i|B) = \sum_{i:y_i^B=1} y_i^{A_i\cap B} \qquad \text{(since } y_i^{A_i|B} = y_i^{A_i} = y_i^{A_i\cap B} \text{ when } y_i^B = 1\text{)}$$

$$\sum_{i:y_i^B=1} \mathsf{p}(x_i, A_i|B) = \sum_{i=1}^{d} y_i^{A_i\cap B} \qquad \text{(since } y_i^{A_i\cap B} = 0 \text{ when } y_i^B = 0\text{)}$$

$$\sum_{i:y_i^B=1} \mathsf{p}(x_i, A_i|B) = \sum_{i=1}^{d} \mathsf{p}(x_i, A_i \cap B) \qquad \text{(by (44))}$$

$$\mathsf{p}(x_1, B) \cdot \sum_{i=1}^{d} \cdot \mathsf{p}(x_i, A_i|B) = \sum_{i=1}^{d} \mathsf{p}(x_i, A_i \cap B) \qquad \text{(by (45))}$$

$$\sum_{i=1}^{d} \mathsf{p}(x_i, A_i|B) = \sum_{i=1}^{d} \frac{\mathsf{p}(x_i, A_i \cap B)}{\mathsf{p}(x_i, B)}.$$