Right Decisions from Wrong Predictions: A Mechanism Design Alternative to Individual Calibration

Shengjia Zhao
Stanford University

Stefano Ermon
Stanford University

Abstract

Decision makers often need to rely on imperfect probabilistic forecasts. While average performance metrics are typically available, it is difficult to assess the quality of individual forecasts and the corresponding utilities. To convey confidence about individual predictions to decision-makers, we propose a compensation mechanism ensuring that the forecasted utility matches the actually accrued utility. While a naive scheme to compensate decision-makers for prediction errors can be exploited and might not be sustainable in the long run, we propose a mechanism based on fair bets and online learning that provably cannot be exploited. We demonstrate an application showing how passengers could confidently optimize individual travel plans based on flight delay probabilities estimated by an airline.

1 Introduction

People and algorithms constantly rely on probabilistic forecasts (about medical treatments, weather, transportation times, etc.) and make potentially high-stake decisions based on them. In most cases, forecasts are not perfect, e.g., the forecasted chance that it will rain tomorrow does not match the true probability exactly. While average performance statistics might be available (accuracy, calibration, etc), it is generally impossible to tell whether any *individual* prediction is reliable (individually calibrated), e.g., about the medical condition of an *specific patient* or the delay of a *particular flight* (Vovk et al., 2005; Barber et al., 2019; Zhao et al., 2020). Intuitively, this is because multiple *identical* datapoints are needed to confidently estimate a probability from empirical frequencies, but identical datapoints are rare in real world applications (e.g. two patients are always different). Given these limitations, we study alternative mechanisms to convey confidence about *individual* predictions to decision-makers.

We consider settings where a single forecaster provides predictions to many decision makers, each facing a potentially different decision making problem. For example, a personalized medicine service could predict whether a product is effective for thousands of individual patients (Ng et al., 2009; Pulley et al., 2012; Bielinski et al., 2014). If the prediction is accurate for 70% of patients, it could be accurate for Alice but not Bob, or vice-versa. Therefore, Alice might be hesitant to make decisions based on the 70% *average* accuracy. In this setting, we propose an insurance-like mechanism that 1) enables each decision maker to confidently make decisions as if the advertised probabilities were individually correct, and 2) is implementable by the forecaster with provably vanishing costs in the long run.

To achieve this, we turn to the classic idea (De Finetti, 1931; Jaynes, 1996) that a probabilistic belief is equivalent to a willingness to take bets. We use the previous example to illustrate that if the forecaster is willing to take bets, a decision maker can bet with the forecaster as an "insurance" against mis-prediction. Suppose Alice is trying to decide whether or not to use a product. If she uses the product, she gains \$10 if the product is effective and loses \$2 otherwise. The personalized medicine service (forecaster) predicts that the product is effective with 50% chance for Alice. Under this probability Alice expects to gain \$4 if she decides to use the product, but she is worried the probability is incorrect. Alice proposes a bet: Alice pays the forecaster \$6 if the product is effective, and the forecaster pays Alice \$6 otherwise. The forecaster should accept the bet because under its own forecasted probability the bet is fair (i.e., the expectation is zero if the forecasted probabilities are true for Alice). Alice gets the guarantee that if she decides to use the product, *ef-*

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

fective or not, she gains \$4 — equal to her expected utility under the forecasted (and possibly incorrect) probability. In general, we show that *Alice has a way of choosing bets for any utility function and forecasted probability, such that her true gain equals her expected gain under the forecasted probability.*

From the forecaster's perspective, if the true probability that Alice's treatment is effective is actually 10%, then the forecaster will lose \$4.8 from this bet in expectation. However, in our setup, the forecaster makes probabilistic forecasts for many different decision makers, each selecting some bet based on their utility function and forecasted probability. The forecaster might gain or lose on *individual* bets, but it only needs to not lose on the entire set of bets *on average* for the approach to be sustainable. Intuitively, our mechanism averages individual decision maker's difference between forecasted gain and true gain so the difficult requirement that *each* difference should be negative has been reduced to an easier requirement that the *average* difference should be negative.

However, this protocol leaves the forecaster vulnerable to exploitation. For example, if Alice already knows that the product will be ineffective; she could still bet with the forecaster for the malicious purpose of gaining \$6. Surprisingly we show that in the online setup (Cesa-Bianchi and Lugosi, 2006), the forecaster has an algorithm to adapt its forecasts and guarantee vanishing loss in the long run, even in the presence of malicious decision makers. This is achieved by first using any existing online prediction algorithm to predict the probabilities, then applying a post processing algorithm to fine-tune these probabilities based on past gains/losses (similar to the idea of recalibration (Kuleshov and Ermon, 2017; Guo et al., 2017)).

As a concrete application of our approach, we simulate the interaction between an airline and passengers with real flight delay data. Risk averse passengers might want to avoid a flight if there is possibility of delay and their loss in case of delay is high. We show if an airline offers to accept bets based on the predicted probability of delay, it can help risk-averse passengers make better decisions, and increase both the airline's revenue (due to increased demand for the flight) and the total utility (airline revenue plus passenger utility).

We further verify our theory with large scale simulations on several datasets and a diverse benchmark of decision tasks. We show that forecasters based on our post-processing algorithm consistently achieve close to zero betting loss (on average) within a small number of time steps. On the other hand, several seemingly reasonable alternative algorithms not only lack theoretical guarantees, but often suffer from positive average betting loss in practice.

2 Background

2.1 Decision Making with Forecasts

This section defines the basic setup of the paper. We represent the decision making process as a multi-player game between nature, a forecaster and a set of (decision making) agents. At every step t nature reveals an input observation x_t to the forecaster (e.g. patient medical records) and selects the hidden probability $\mu_t^* \in [0,1]$ that $\Pr[y_t = 1] = \mu_t^*$ (e.g. probability treatment is successful), We only consider binary variables ($y_t \in \{0,1\} = \mathcal{Y}$) and defer the general case to Appendix B.

The forecaster chooses a forecasted probability $\mu_t \in [0,1]$ to approximate μ_t^*. We also allow the forecaster to represent the lack of knowledge about μ_t^*, i.e. the forecaster outputs a confidence $c_t \in [0,1]$ where the hope is that $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t]$.

At each time step, one or more agents can use the forecast μ_t and c_t to make decisions, i.e. to select an action $a_t \in \mathcal{A}$. However, for simplicity we assume that different agents make decisions at different time steps, so at each time step there is only a single agent, and we can uniquely index the agent by the time step t. The agent knows its own loss (negative utility) function $l_t : \mathcal{A} \times \mathcal{Y} \to [-M, M]$ (the forecaster does not have to know this) where $M \in \mathbb{R}_+$ is the maximum possible loss. This protocol is formalized below.

Protocol 1: Decision Making with Forecasts
For $t = 1, \cdots, T$

1. Nature reveals $x_t \in \mathcal{X}$ to forecaster and chooses $\mu_t^* \in [0,1]$ without revealing it
2. Forecaster reveals $\mu_t, c_t \in (0,1)$ where $(\mu_t - c_t, \mu_t + c_t) \subset (0,1)$
3. Agent t has loss function $l_t : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ and reveals a_t selected according to μ_t, c_t and l_t
4. Nature samples $y_t \sim \text{Bernoulli}(\mu_t^*)$ and reveals y_t; Agent incurs loss $l_t(a_t, y_t)$

We make no assumptions on nature, forecaster, or the agents. They can choose any strategy to generate their actions, as long as they do not look into the future (i.e. their action only depends on variables that have already been revealed). In particular, we make no i.i.d. assumptions on how nature selects y_t and μ_t^*; for example, nature could even select them adversarially to maximize the agent's loss.

2.2 Individual Coverage

Ideally in Protocol 1 the forecaster's prediction μ_t, c_t should satisfy $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t])$ for each individual t (this is often called individual coverage or individual calibration in the literature). However, many existing results show that learning individually calibrated probabilities from past data is often impossible (Vovk et al., 2005; Barber et al., 2019; Zhao et al., 2020) unless the forecast is trivial (i.e. $[\mu_t - c_t, \mu_t + c_t] = [0, 1]$).

One intuitive reason for this impossibility result is that in many practical scenarios for each x_t we only observe a single sample $y_t \sim \mu_t^*$. The forecaster cannot infer μ_t^* from a single sample y_t without relying on unverifiable assumptions.

2.3 Probability as Willingness to Bet

A major justification for probability theory has been that probability can represent willingness to bet (De Finetti, 1931; Halpern, 2017). For example, if you truly believe that a coin is fair, then it would be inconsistent if you are not willing to win \$1 for heads, and lose \$1 for tails (assuming you only care about average gain rather than risk). More specifically a forecaster that holds a probabilistic belief should be willing to accept any bet where it gains a non-negative amount in expectation.

For binary variables, we consider the case where a forecaster believes that a binary event $Y \in \{0, 1\}$ happens with some probability μ^* but does not know the exact value of μ^*. The forecaster only believes that $\mu^* \in [\mu - c, \mu + c] \subset [0, 1]$. The forecaster should be willing to accept any bet with non-negative expected return under *every* $\mu^* \in [\mu - c, \mu + c]$. For example, assume the forecaster believes that a coin comes up heads with at least 40% chance and at most 60% chance. The forecaster should be willing to win \$6 for heads, and lose \$4 for tails; similarly the forecaster should be willing to lose \$4 for heads, and win \$6 for tails.

More generally, according to Lemma 1 (proved in Appendix D), a forecaster believes that the probability of success $\Pr[Y = 1] = \mu^*$ of the binary event Y satisfies $\mu^* \in [\mu - c, \mu + c]$ if and only if she is willing to accept bets where she loses $b(Y - \mu) - |b|c, \forall b \in \mathbb{R}$.

Lemma 1. *Let $\mu, c \in (0, 1)$ such that $[\mu - c, \mu + c] \subset [0, 1]$, then a function $f : \mathcal{Y} \to \mathbb{R}$ satisfies $\forall \tilde{\mu} \in [\mu - c, \mu + c]$, $\mathbb{E}_{Y \sim \tilde{\mu}}[f(Y)] \leq 0$ if and only if for some $b \in \mathbb{R}$ and $\forall y \in \{0, 1\}$, $f(y) \leq b(y - \mu) - |b|c$.*

In words, a forecaster is willing to lose $f(Y)$ if f has non-positive expectation under every probability the forecaster considers possible. However, every such function f are smaller (i.e. forecaster loses less) than $b(Y - \mu) - |b|c$ for some $b \in \mathbb{R}$. Therefore, we only have to consider whether a forecaster is willing to accept bets of the form $b(Y - \mu) - |b|c$.

3 Decisions with Unreliable Forecasts

In Protocol 1, agents could make decisions based on the forecasted probability μ_t, c_t and the agent's loss l_t. For example, the agent could choose

$$a_t := \arg\min_{a \in \mathcal{A}} \mathbb{E}_{Y \sim \mu_t} l_t(a, Y) \tag{1}$$

to minimize the expected loss under the forecasted probability.

However, how can the agent know that this decision has low expected loss under the *true probability* μ_t^*? This can be achieved with two desiderata, which we formalize below:

We denote the agent's maximum / average / minimum expected loss under the forecasted probability as

$$L_t^{\max} = \max_{\tilde{\mu} \in \mu_t \pm c_t} \mathbb{E}_{Y \sim \tilde{\mu}}[l_t(a_t, Y)]$$
$$L_t^{\text{avg}} = \mathbb{E}_{Y \sim \mu_t}[l_t(a_t, Y)]$$
$$L_t^{\min} = \min_{\tilde{\mu} \in \mu_t \pm c_t} \mathbb{E}_{Y \sim \tilde{\mu}}[l_t(a_t, Y)]$$

and true expected loss as $L_t^* = \mathbb{E}_{Y \sim \mu_t^*}[l_t(a_t, Y)]$. If the agent knows that

Desideratum 1 $L_t^* \in [L_t^{\min}, L_t^{\max}]$
Desideratum 2 The interval size c_t is close to 0.

then the agent can infer that the true expected loss L_t^* is not too far off from the forecasted expected loss L_t^{avg}. This is because if c_t is small then $L_t^{\min}$ will be close to $L_t^{\max}$. Both L_t^* and L_t^{avg} will be sandwiched in the small interval $[L_t^{\min}, L_t^{\max}]$.

However, we show that desiderata 1 and 2 often cannot be achieved simultaneously. To guarantee $L_t^* \in [L_t^{\min}, L_t^{\max}]$ the forecaster in general must output individually correct probabilities (i.e. $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t]$), as shown by the following proposition (proof in Appendix D).

Proposition 1. *For any $\mu_t, c_t, \mu_t^* \in (0, 1)$ where $(\mu_t - c_t, \mu_t + c_t) \subset (0, 1)$*

1. If $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t]$ then $\forall l_t : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ we have $L_t^* \in [L_t^{\min}, L_t^{\max}]$

2. If $\mu_t^* \notin [\mu_t - c_t, \mu_t + c_t]$ then $\forall l_t : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, if $\forall a \in \mathcal{A}, \ell_t(a, 0) \neq \ell_t(a, 1)$, then $L_t^* \notin [L_t^{\min}, L_t^{\max}]$

In words, if $\mu_t^* \notin [\mu_t - c_t, \mu_t + c_t]$, we cannot guarantee that $L_t^* \in [L_t^{\min}, L_t^{\max}]$ unless the agent's loss function is trivial (e.g. it is a constant function). However, in

Section 2.2 we argued that it is usually impossible to achieve $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t]$ unless c_t is very large (i.e. $[\mu_t - c_t, \mu_t + c_t] = [0, 1]$). If c_t is too large, the interval $[L_t^{\min}, L_t^{\max}]$ will be large, and the guarantee that $L_t^* \in [L_t^{\min}, L_t^{\max}]$ would be practically useless even if it were true. This means the forecaster cannot convey confidence in individual predictions it makes, and as a result the agent can't be very confident about the expected loss it will incur.

3.1 Insuring against unreliable forecasts

Since it is difficult to satisfy desiderata 1 and 2 simultaneously, we consider relaxing desideratum 1. In particular, we study what guarantees are possible for each individual decision maker even when $\mu_t^* \notin [\mu_t - c_t, \mu_t + c_t]$, i.e., the prediction is wrong.

We consider the setup where each agent can receive some side payment (a form of "insurance" which could depend on the outcome Y, and could be positive or negative) from the forecaster, and we would like to guarantee

Desideratum 1'

$$\underbrace{L_t^* - \mathbb{E}_{Y\sim\mu_t^*}[\text{payment}(Y)]}_{\text{True expected loss w. side payment}} \in \underbrace{[L_t^{\min}, L_t^{\max}]}_{\text{Forecasted expected loss range}}$$

In other words, we would like the expected loss under the true distribution to be predictable *once we incorporate the side payment.*

Note that desideratum 1' can be trivially satisfied if the forecaster is willing to pay any side payment to the decision agent. For example, an agent can choose $\text{payment}(Y) := \mathbb{E}_{Y\sim\mu_t}[l_t(a_t, Y)] - l_t(a_t, Y)$ to satisfy desideratum 1'. However, if the forecaster offers any side payment, it could be subject to exploitation. For example, decision agents could request the forecaster to pay \$1 under any outcome y_t. Such a mechanism cannot be sustainable for the forecaster.

3.2 Insuring with fair bets

Even though the forecaster cannot offer arbitrary payments to the decision agent, we show that the forecaster can offer a sufficiently large set of payments, such that [i] each decision agent can select a payment to satisfy Desideratum 1' and [ii] the forecaster has an algorithm to guarantee vanishing loss in the long run, even when the decision agents tries to exploit the forecaster.

In fact, the "fair bets" in Section 2.3 satisfy our requirement. Specifically, the forecaster can offer the set

$$\{\text{payment}(Y) := b(Y - \mu_t) - |b|c_t, \forall b \in [-M, M]\}$$

as available side payment options. The constant $M \in \mathbb{R}_+$ caps the maximum payment each decision agent can request (in our setup l_t is also upper bounded by M). This set of payments satisfy both [i] (which we show in this section) and [ii] (which we show in the next section).

Before we proceed to show [i] and [ii], for convenience, we formally write down the new protocol. Compared to Protocol 1, the decision agent selects some "stake" $b_t \in [-M, M]$, and receive side payment $b_t(Y - \mu_t) - |b_t|c_t$ from the forecaster.

Protocol 2: Decision Making with Bets For $t = 1, \cdots, T$

1. Nature reveals observation $x_t \in \mathcal{X}$ and chooses $\mu_t^* \in [0, 1]$ without revealing it
2. Forecaster reveals $\mu_t, c_t \in (0, 1)$ where $(\mu_t - c_t, \mu_t + c_t) \subset (0, 1)$
3. Agent t has loss function $l_t : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ and reveals action $a_t \in \mathcal{A}$ and stake $b_t \in [-M, M]$ selected according to μ_t, c_t and l_t
4. Nature samples $y_t \sim \text{Bernoulli}(\mu_t^*)$ and reveals y_t
5. Agent incurs loss $l_t(a_t, y_t) - b_t(y_t - \mu_t) + |b_t|c_t$; forecaster incurs loss $b_t(y_t - \mu_t) - |b_t|c_t$

Denote the agent's true expected loss with side payment as (i.e. the LHS in Desideratum 1')

$$L_t^{\text{pay}} := \underbrace{L_t^*}_{\text{decision loss}} \underbrace{- \mathbb{E}_{Y\sim\mu_t^*}[b_t(Y - \mu_t) + |b_t|c_t]}_{\text{payment from forecaster}} \quad (2)$$

then we have the following guarantee[1] for any choice of μ_t, c_t, μ_t^*, a_t and l_t

Proposition 2. *If the stake* $b_t = l_t(a_t, 1) - l_t(a_t, 0)$ *then* $L_t^{\text{pay}} \in [L_t^{\min}, L_t^{\max}]$

In words, the agent has a choice of stake b_t that only depends on variables known to the agent (l_t and a_t) and does not depend on variables unknown to the agent (μ_t^*, y_t). If the agent chooses this b_t, she can be certain that desideratum 1' is satisfied, regardless of what the forecaster or nature does (they can choose any μ_t, c_t, μ_t^*).

This mechanism allows the agent to **make decisions as if the forecasted probability is correct**, i.e. as if $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t]$. This is because Proposition 2 is true for any choice of action a_t (as long as the agent chooses b_t according to Proposition 2 after selecting a_t). Intuitively, for any action a_t the agent selects, she can guarantee to achieve a total loss close to $\mathbb{E}_{Y\sim\mu_t} l_t(a_t, Y)$ (assuming c_t is small). This is the same guarantee she would get as if $\mu_t^* \in [\mu_t - c_t, \mu_t + c_t]$.

[1] For the more general version of the proposition in the multi-class setup, see Appendix A.1.

In addition, if [ii] is satisfied (i.e. the forecaster has vanishing loss), the forecaster also doesn't lose anything, so should have no incentive to avoid offering these payments. We discuss this in the next section.

Algorithm 1: Post-Processing for Exactness

Invoke Algorithm 2 and 3 with $K = (T/\log T)^{1/4}$
for $t = 1, \cdots, T$ **do**
  Receive $\hat{\mu}_t$ and $\hat{c}_t$ from Algorithm 2
  Receive λ_t from Algorithm 3
  Output $\mu_t = \hat{\mu}_t$, $c_t = \hat{c}_t + \lambda_t$
  Input y_t and b_t
  Set $r_t = (b_t/\sqrt{|b_t|})(\mu_t - y_t) - \sqrt{|b_t|}\hat{c}_t$,
   $s_t = -\sqrt{|b_t|}$, Send (r_t, s_t) to Algorithm 3

Algorithm 2: Online Prediction

Choose any initial value for θ_1, ϕ_1
for $t = 1, \cdots, T$ **do**
  Input x_t and output $\hat{\mu}_t = \mu_{\theta_t}(x_t)$, $\hat{c}_t = c_{\phi_t}(x_t)$
  Input y_t and b_t
  $\theta_{t+1} = \theta_t - \eta \frac{\partial}{\partial \theta}(\mu_{\theta_t}(x_t) - y_t)^2$
  $\phi_{t+1} = \phi_t - \eta \frac{\partial}{\partial \phi}\left(b_t(\hat{\mu}_t - y_t) - |b_t| c_{\phi_t}(x_t)\right)^2$

4 Probability Forecaster Strategy

In this section we study the forecaster's strategy. As motivated in the previous section, the goal of the forecaster (in Protocol 2) is to:

1) have non-positive cumulative loss when T is large, so that the side payments are sustainable
2) output the smallest c_t compatible with 1), so that forecasts are as sharp as possible

Specifically, the forecaster's average cumulative loss (up to time T) in Protocol 2 is

$$\frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t - y_t) - |b_t| c_t \tag{3}$$

Whether Eq.(3) is non-positive or not depends on the actions of all the players: forecaster μ_t, c_t, nature y_t and agent b_t. Our focus is on the forecaster, so we say that a sequence of forecasts $\mu_t, c_t, t = 1, 2, \cdots$ is **asymptotically sound** relative to $y_1, b_1, y_2, b_2, \cdots$ if the forecaster loss in Protocol 2 is non-positive, i.e.

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t - y_t) - |b_t| c_t \le 0 \tag{4}$$

In subsequent development we will use a stronger definition than Eq.(4). We say that a sequence of forecasts $\mu_t, c_t, t = 1, 2, \cdots$ is **asymptotically exact** relative to $y_1, b_1, y_2, b_2, \cdots$ if the forecaster loss in Protocol 2 is exactly zero, i.e.

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t - y_t) - |b_t| c_t = 0 \tag{5}$$

Intuitively asymptotic soundness requires that the forecaster should not lose in the long run; asymptotic exactness requires that the forecaster should neither lose nor win in the long run — a stronger requirement.[2]

The reason we focus on asymptotic exactness is because the forecaster should output the smallest possible c_t to achieve sharp forecasts. Observe that the left hand side of Eq.(4) is increasing if c_t decreases. Therefore, whenever the forecaster is asymptotically sound but not asymptotically exact (i.e. the left hand side in Eq.(4) is strictly negative), there is some room to decrease c_t without violating asymptotic soundness.

Algorithm 3: Swap Regret Minimization

Input: number of discrete interval K
Partition $[-1, 1]$ into equal intervals $[-1 = v_0, v_1)$, $\cdots$, $[v_{K-1}, v_K = 1]$
For each interval init an empty set $\mathcal{D}_k$, set $v^0 = 0$
for $t = 1, \cdots, T$ **do**
  Initialize an empty ordered list $\mathcal{V}^t$
  Initialize $v^t = v^{t-1}$ and **while** $v^t \notin \mathcal{V}^t$ **do**
    $\lambda_t^{v^t} = \inf_{\lambda \in [-1, 1)} \frac{1}{|\mathcal{D}_{v^t}|} \sum_{r_t, s_t \in \mathcal{D}_{v^t}} (r_t + s_t \lambda)^2$
    Append v^t to $\mathcal{V}^t$
    Set v^t as the k that satisfies
     $\lambda_t^{v^t} \in [v_k, v_{k+1})$
  Remove all elements before v^t from $\mathcal{V}^t$
  Select v^t uniform randomly from $\mathcal{V}^t$
  Choose $\lambda_t = \lambda_t^{v^t}$ and send λ_t to Algorithm 1
  Receive (r_t, s_t) from Algorithm 1, add to $\mathcal{D}_{v^t}$

4.1 Online Forecasting Algorithm

We aim to achieve asymptotic exactness with minimal assumptions on $y_t, b_t, t = 1, 2, \cdots$ (we only assume boundedness). This is challenging for two reasons: an adversary could select $y_t, b_t, t = 1, 2, \cdots$ to violate asymptotic exactness as much as possible (e.g. decision agents could try to profit on the forecaster's loss); in Protocol 2 the agent's action b_t is selected *after* the forecaster's prediction μ_t, c_t are revealed, so the agent has last-move advantage.

[2] In mechanism design literature, Eq.(4) and Eq.(5) are typically referred to as weak and strong budget balanced. Here we use the terminology in probability forecasting literature.

Nevertheless asymptotic exactness can be achieved as shown in Theorem 1 (proof in Appendix C). In fact, we design a post-processing algorithm that modifies the prediction of a base algorithm (similar to recalibration (Kuleshov and Ermon, 2017; Guo et al., 2017)). Algorithm 1 can modify any base algorithm (as long as the base algorithm outputs some μ_t, c_t at every time step) to achieve asymptotic exactness, even though the finite time performance could be hurt by a poor base prediction algorithm.

Theorem 1. *Suppose there is a constant $M > 0$ such that $\forall t$, $|b_t| \leq M$, there exists an algorithm to output μ_t, c_t in Protocol 2 that is asymptotically exact for μ_t^*, b_t generated by any strategy of nature and agent. In particular, Algorithm 1 satisfies*

$$\left(\frac{1}{T}\sum_{t=1}^{T} b_t(\mu_t - y_t) - |b_t|c_t\right)^2 = O\left(\sqrt{\frac{\log T}{T}}\right) \quad (6)$$

For this paper we use as our base algorithm a simple online gradient descent algorithm (Zinkevich, 2003) shown in Algorithm 2. Specifically Algorithm 2 learns two regression models (such as neural networks with a single real number as output) μ_θ and c_ϕ. μ_θ is trained to predict μ_t^* by minimizing the standard L_2 loss $\min_\theta \sum_{\tau=1}^{t}(\mu_\theta(x_\tau) - y_\tau)^2$ while c_ϕ is trained to to minimize the squared payoff of each bet $\min_\phi \sum_{\tau=1}^{t}(b_\tau(\hat{\mu}_\tau - y_\tau) - |b_\tau|c_\phi(x_\tau))^2$

Based on Algorithm 2, Algorithm 1 learns an additional "correction" parameter $\lambda_t \in \mathbb{R}$ by invoking Algorithm 3. Intuitively, up to time t, if the forecaster has positive cumulative loss in Protocol 2, then the c_ts have been too small in the past, Algorithm 1 will select a larger λ_t to increase c_t; conversely if the forecaster has negative cumulative loss, then the c_ts have been too large in the past, and Algorithm 1 will select a smaller λ_t to decrease c_t.

Despite the straight-forward intuition, the difficulty comes from ensuring Theorem 1 for *any* sequence of $y_t, b_t, t = 1, \cdots$. In fact, Algorithm 3 needs to be a swap regret minimization algorithm (Blum and Mansour, 2007). Appendix C provides a detailed explanation and proof.

Special Case: Monotonic Loss In general, the forecaster selects c_t carefully to achieve asymptotic exactness and protect itself from exploitation. However, there are special cases where the c_t is not necessary (i.e. the forecaster can always output $c_t \equiv 0$).

In particular, c_t is not necessary whenever the loss function satisfies $\forall t, l_t(1, a_t) \geq l_t(0, a_t)$. Intuitively, $y_t = 1$ is never better (incurs equal or higher loss) than $y_t = 0$. For example, ineffective treatment is never better than effective treatment; delayed flight is never better than on-time flights. Under this assumption and according to Proposition 2, decision makers can choose a non-negative stake $0 \leq b_t := l_t(1, a_t) - l_t(0, a_t)$ to ensure $L_t^{\text{pay}} \in [L_t^{\min}, L_t^{\max}]$. In other words, in Protocol 2 we can restrict $b_t \geq 0$ without losing the guarantee of Proposition 2. In this situation the forecaster can achieve asymptotic exactness even when $c_t \equiv 0$

Corollary 1. *Under the assumptions of Theorem 1 if additionally $b_t \geq 0, \forall t$, there exists an algorithm to output μ_t in Protocol 2 with $c_t \equiv 0$ that is asymptotically exact for μ_t^*, b_t generated by any strategy of nature and agent.*

4.2 Offline Forecasting

Our new definition of asymptotic soundness in Eq.(4) can be extended to the offline setup, where nature's action in Protocol 2 is i.i.d. sampled from random variables X, Y, i.e. $x_t \sim X, \mu_t^* = \mathbb{E}[Y \mid x_t]$. In addition, the agent's action b_t in Protocol 2 is a (fixed) function of x_t i.e. $b_t = b(x_t)$ for some $b : \mathcal{X} \to [-M, M]$. In this setup, the forecaster can also select its actions μ_t, c_t based on fixed functions of x_t.

In Appendix A.2 we formally define soundness in the offline setup, and show that for certain choices of agent's action b we can recover existing notions of calibration (Dawid, 1985; Guo et al., 2017; Kleinberg et al., 2016; Kumar et al., 2019) or multicalibration (Hébert-Johnson et al., 2017). In other words, if a forecaster satisfies the existing notions of calibration, there are some functions $b : \mathcal{X} \to [-M, M]$: as long as the decision making agents limit their actions to $b_t = b(x_t)$, the forecaster will be asymptotically sound. The benefit is that once deployed, the forecaster does not have to be updated (compared to the online setup where the forecaster must continually update via Algorithm 1). However, the short-coming is that we make strong assumptions on how the agents choose bets to insure themselves.

5 Related Work

Calibration: A forecaster is calibrated if among the times the forecaster predicts that an event happens with α probability, the event indeed happens α fraction of the times (Brier, 1950; Murphy, 1973; Dawid, 1984; Platt et al., 1999; Zadrozny and Elkan, 2001; Guo et al., 2017). Calibration can be achieved even when the data is not i.i.d. (Cesa-Bianchi and Lugosi, 2006; Kuleshov and Ermon, 2016). However, calibration is an average performance measurement and provides no guarantee on the correctness of individual probability predictions.

Scoring rule: a (proper) scoring rule is a function $s(y, p_Y)$ that measures the "quality" of a probability forecast p_Y if the outcome y is observed (Brier, 1950; Savage, 1971; Gneiting and Raftery, 2007; Dawid and Musio, 2014). However, achieving a high score only reflects high *average* quality, rather than the quality of individual predictions.

Conformal prediction: Many applications only require a confidence interval (i.e. a subset of $\mathcal{Y}$) instead of the joint probability. A confidence interval forecaster (or conformal forecaster) is δ-exact if $1-\delta$ proportion of the predicted intervals contain the observed outcome. There are algorithms that guarantee exactness for exchangeable data (Vovk et al., 2005; Shafer and Vovk, 2008). However, exactness guarantees the proportion of predictions that contain the observed outcome, rather than any individual prediction.

The above approaches provide no guarantees on the correctness of individual predictions. The classic method that can guarantee individual predictions is **non-parametric learning**. Algorithms such as nearest neighbor or Gaussian processes can produce correct individual probabilities with infinite training data (Bishop, 2006), but have no guarantees when training data is finite or not i.i.d.

In the finite data regime, a notable research direction is **individual calibration**, i.e. calibration on every data sample. Individual calibration is sometimes possible with a randomized forecaster (Zhao et al., 2020). However, for randomized forecasts, calibration cannot be interpreted as forecasting correct probabilities. Without randomization, individual calibration is often impossible (Vovk et al., 2005; Vovk, 2012; Zhao et al., 2020; Barber et al., 2019).

Individual calibration can be relaxed to **group calibration**, i.e. calibration on pre-specified subsets of the data (Kleinberg et al., 2016). Notably, (Hébert-Johnson et al., 2017) achieve calibration for a parametric set of subsets. Several impossibility results (Kleinberg et al., 2016; Pleiss et al., 2017) show that often group calibration cannot be meaningfully achieved.

Our contribution Our approach has the main desired effect of individual calibration (decision makers can confidently use the forecasted probability "as if" it is correct) without actually achieving individual calibration, hence are not limited by the impossibility results above. The key difference that makes our guarantees possible (without i.i.d. assumptions, well specification assumptions, infinite data, or randomization) is that the forecasts depend on downstream decision tasks. Rather than predicting perfect probabilities, we aim for the attainable objective of achieving exactness for actually encountered decision tasks.

6 Case Study on Flight Delays

In this section we study a practical application that could benefit from our proposed mechanism. Compared to other means of transport, flights are often the fastest, but usually the least punctual. Different passengers may have different losses in case of delay. For example, if a passenger needs to attend an important event on-time, the loss from a delay can be very large, and the passenger might want to choose an alternative transportation method. The airline company could predict the probability of delay, and each passenger could use the probability to compute their expected loss before deciding to fly or not. However, as argued in Section 2.2, there is in general no good way to know that these probabilities are correct. Even worse, the airline may have the incentive to underreport the probability of delay to attract passengers.

Instead the airline can use Protocol 2 to convey confidence to the passengers that the delay probability is accurate. In this case, Protocol 2 has a simple form that can be easily explained to passengers as a "delay insurance". In particular, if a passenger buys a ticket, he can choose to insure himself against delay by specifying the amount b_t^1 he would like to get paid if the airplane is delayed. The airlines provides a quote on the cost b_t^0 of the insurance (i.e. the passenger pays b_t^0 if the flight is not delayed). Note that this would be equivalent to Protocol 2 if the airline first predicts the probability of delay μ_t, c_t and then quotes $b_t^0 := \frac{b_t^1(\mu_t+c_t)}{1-\mu_t-c_t}$.

If a passenger buys the right insurance according to Proposition 2, their expected utility (or negative loss) will be fixed — she does not need to worry that the predicted delay probability might be incorrect. In addition, if the airline follows Algorithm 1 the airline is also guaranteed to not lose money from the "delay insurance" in the long run (no matter what the passengers do), so the airline should be incentivized to implement the insurance mechanism to benefit its passengers "for free".

Passenger Model Since the passengers' utility functions are unknown, we model three types of passengers that differ by their assumptions on μ_t^* when they make their decision:

1. Naive passengers don't care about delays and assume the airline doesn't delay.
2. Trustful passengers assume the delay probability forecasted by the airline is correct.
3. Cautious passengers assume the worst (i.e. they choose actions that maximizes their worst case utility)

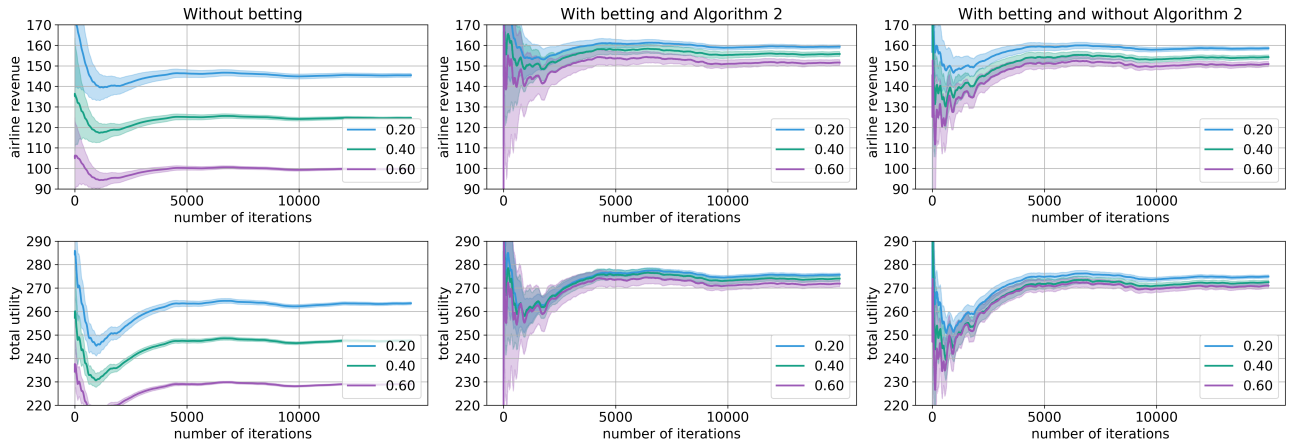

Figure 1: The airline's revenue (**Top**) and total utility (of both airline and passenger, **Bottom**) with and without the betting mechanism. Different colors represent the percentage of cautious passengers. The x-axis represents the number of flights that has happened, and the y-axis represents the average utility per passenger across all past flights. **Left**: Without the betting mechanism that insure passengers against delay **Middle and Right**: With the betting mechanism, the airline revenue increases (because it is able to charge a higher ticket price due to increased demand) and the total utility increases. The middle panel is the utility with both Algorithm 1 and Algorithm 3, while the right panel only uses Algorithm 1 (i.e. it always sets $\lambda_t = 0$). In general the middle panel achieves faster convergence, so with fewer iterations, the utility is better than the right panel.

In this experiment we will vary the proportion of cautious passengers, and equally split the remaining passengers between naive and trustful. The naive and trustful passengers do not care about the risk of misprediction, so they do not buy the delay insurance (i.e. they always choose $b_t^1 = 0$), while cautious passengers always buy insurance that maximize worst case utility.

6.1 Simulation Setup

Dataset We use the flight delay and cancellation dataset (DoT, 2017) from the year 2015, and use flight records of the single biggest airline (WN). As input feature, we convert the source airport, target airport, and scheduled time into one-hot vectors, and binarize the arrival delay into 1 (delay $>$ 20min) and 0 (delay $<$ 20min). We use a two layer neural network with the leaky ReLU activation for prediction.

Passenger Utility Let $y \in \{0, 1\}$ denote whether a delay happens, and $a \in \{0, 1\}$ denote whether the passenger chooses to ride the plane. We model the passenger utility (negative loss) as

$$-l(y, a) = \begin{cases} y = *, a = 0 & r^{\text{alt}} \\ y = 0, a = 1 & r^{\text{trip}} - c^{\text{ticket}} \\ y = 1, a = 1 & r^{\text{trip}} - c^{\text{ticket}} - c^{\text{delay}} \end{cases}$$

where r^{alt} is the utility of the alternative option (e.g. taking another transportation or cancelling the trip). For simplicity we assume that this is a single real number. r^{trip} is the reward of the trip, c^{ticket} is the cost of the ticket, and c^{delay} is the cost of a delayed flight. For each flight we sample 1000 potential passengers by randomly drawing the values $r^{\text{alt}}, r^{\text{trip}}$ and c^{delay} (for details see appendix).

Airline Pricing Based on the passenger type (naive, trustful, cautious) and passenger parameter $r^{\text{alt}}, r^{\text{trip}}$ and c^{delay}, each passenger will have a maximum they are willing to pay for the flight. For simplicity we assume the airline will choose c^{ticket} at the highest price for which it can sell 300 tickets. The passengers who are willing to pay more than c^{ticket} will choose $a = 1$, and other passengers will choose $a = 0$.

6.2 Delay Insurance Improves Total Utility

The simulation results are shown in Figure 1. Using the betting mechanism is strictly better for both the airline's revenue (i.e. ticket price * number of tickets) and the total utility (airline revenue + passenger utility). This is because the cautious passengers always make decisions to maximize their worst case utility. With the betting mechanism, their worst case utility becomes closer to their actual true utility, so their decision ($a = 1$ or $a = 0$) will better maximize their true utility. The airline also benefits because it can charge a higher ticket price due to increased demand (more cautious passengers will choose $a = 1$).

We also consider several alternatives to Algorithm 3.

Code is available at https://github.com/ShengjiaZhao/ForecastingWithBets

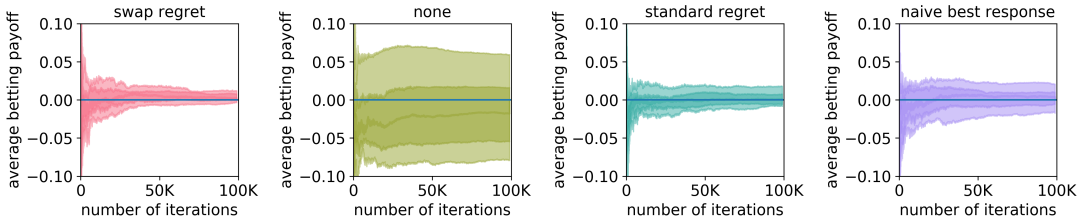

Figure 2: Comparing forecaster loss in Protocol 2 for different forecaster algorithms on MNIST (results for Adult dataset are in appendix B.2). Each plot is an average performance across 20 different decision tasks, where we plot the top 10%, 25%, 50%, 75%, 90% quantile in forecaster loss. If the forecaster achieves asymptotic exactness defined in Eq.(5), then the loss should be close to 0. **Left** panel is Algorithm 1, and the rest are other seemingly reasonable algorithms explained in Section 7. The loss of a forecaster that use Algorithm 1 typically converges to 0 faster, while alternative algorithms often fail to converge.

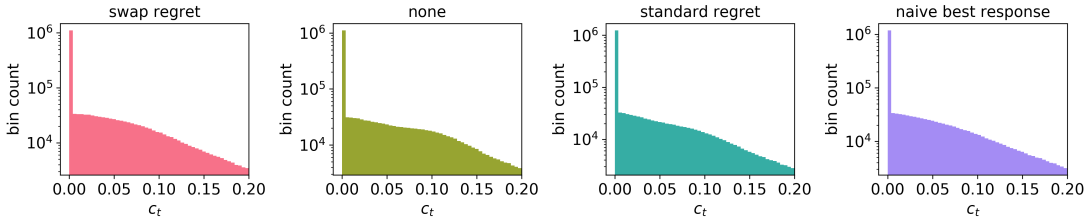

Figure 3: Histogram of the interval size c_t produced by the forecaster algorithm across all the tasks. There is no noticeable difference between the different algorithms. Notably the interval sizes are typically quite small, and big interval size is exponentially less common.

The alternative algorithms do not provide theoretical guarantees; in practice, they also achieve worse convergence to the final utility. This is be a reason to prefer Algorithm 3 if the number of iterations T is small.

7 Additional Experiments

We further verify our theory with simulations a diverse benchmark of decision tasks. We also do ablation study to show that Algorithm 3 is necessary. Several simpler alternatives often fail to achieve asymptotic exactness and have worse empirical performance.

Dataset and Decision Tasks We use the MNIST and UCI Adult (Dua and Graff, 2017) datasets. MNIST is a multi-class classification dataset; we convert it to binary classification by choosing $\Pr[Y = 1 \mid l(x) = i] = (i + 1)/11$ where the $l(x) \in \{0, 1, \cdots, 9\}$ is the digit category. We also generate a benchmark consisting of 20 different decision tasks. For details see Appendix B.2.

Comparison We compare several forecaster algorithms that differ in whether they use Algorithm 3 to adjust the parameter λ_t. In particular, **swap regret** refers to Algorithm 3; **none** does not use λ_t and simply set it to 0; **standard regret** minimizes the standard regret rather than the swap regret; **naive best response** chooses the λ_t that would have been optimal were it counter-factually applied to the past iterations.

Forecaster Model As in the previous experiment, we use a two layer neural network as the forecaster μ_θ and c_ϕ. For the results shown in Figure 6 we also use histogram binning (Naeini et al., 2015) on the entire validation set to recalibrate μ_θ, such that μ_θ satisfies standard calibration (Guo et al., 2017).

Results The results are plotted in Figure 2,3 in the main paper and Figure 5,6 in Appendix B.2. There are three main observations: **1)** Even when a forecaster is calibrated, for individual decision makers, the expected loss under the forecaster probability is almost always incorrect. **2)** Algorithm 1 has good empirical performance. In particular, the guarantees of Theorem 1 can be achieved within a reasonable number of time steps, and the interval size c_t is usually small. **3)** Seemingly reasonable alternatives to Algorithm 1 often empirically fail to be asymptotically exact.

8 Conclusion

In this paper, we propose an alternative solution to address the impossibility of individual calibration based on an insurance between the forecaster and decision makers. Each decision maker can make decisions as if the forecasted probability is correct, while the forecaster can also guarantee not losing in the long run. Future work can explore social issues that arise, such as honesty (Foreh and Grier, 2003), fairness (Dwork et al., 2012), and moral/legal implications.

9 Acknowledgements

This research was supported by AFOSR (FA9550-19-1-0024), NSF (#1651565, #1522054, #1733686), JP Morgan, ONR, FLI, SDSI, Amazon AWS, and SAIL.

References

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2019). The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*.

Bielinski, S. J., Olson, J. E., Pathak, J., Weinshilboum, R. M., Wang, L., Lyke, K. J., Ryu, E., Targonski, P. V., Van Norstrand, M. D., Hathcock, M. A., et al. (2014). Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time—using genomic data to individualize treatment protocol. In *Mayo Clinic Proceedings*, pages 25–33. Elsevier.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blum, A. and Mansour, Y. (2007). From external to internal regret. *Journal of Machine Learning Research*, 8(Jun):1307–1324.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Dawid, A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290.

Dawid, A. P. (1985). Calibration-based empirical probability. *The Annals of Statistics*, pages 1251–1274.

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72(2):169–183.

De Finetti, B. (1931). On the subjective meaning of probability. *Fundamenta mathematicae*, 17(1):298–329.

DoT (2017). 2015 flight delays and cancellations. `https://www.kaggle.com/usdot/flight-delays`.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM.

Foreh, M. R. and Grier, S. (2003). When is honesty the best policy? the effect of stated company intent on consumer skepticism. *Journal of consumer psychology*, 13(3):349–356.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*.

Halpern, J. Y. (2017). *Reasoning about uncertainty*. MIT press.

Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. (2017). Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*.

Jaynes, E. T. (1996). *Probability theory: the logic of science*. Washington University St. Louis, MO.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kuleshov, V. and Ermon, S. (2016). Estimating uncertainty online against an adversary. *arXiv preprint arXiv:1607.03594*.

Kuleshov, V. and Ermon, S. (2017). Estimating uncertainty online against an adversary. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Kumar, A., Liang, P. S., and Ma, T. (2019). Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pages 3792–3803.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, page 2901. NIH Public Access.

Ng, P. C., Murray, S. S., Levy, S., and Venter, J. C. (2009). An agenda for personalized medicine. *Nature*, 461(7265):724–726.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.

Pulley, J. M., Denny, J. C., Peterson, J. F., Bernard, G. R., Vnencak-Jones, C. L., Ramirez, A. H., Delaney, J. T., Bowton, E., Brothers, K., Johnson, K., et al. (2012). Operational implementation of prospective genotyping for personalized medicine: the design of the vanderbilt predict project. *Clinical Pharmacology & Therapeutics*, 92(1):87–95.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pages 475–490.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616. Citeseer.

Zhao, S., Ma, T., and Ermon, S. (2020). Individual calibration with randomized forecasting. *arXiv preprint arXiv:2006.10288*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936.