Individual Calibration with Randomized Forecasting

Shengjia Zhao [1] **Tengyu Ma** [1] **Stefano Ermon** [1]

Abstract

Machine learning applications often require calibrated predictions, e.g. a 90% credible interval should contain the true outcome 90% of the times. However, typical definitions of calibration only require this to hold on average, and offer no guarantees on predictions made on individual samples. Thus, predictions can be systematically over or under confident on certain subgroups, leading to issues of fairness and potential vulnerabilities. We show that calibration for individual samples is possible in the regression setup if the predictions are randomized, i.e. outputting randomized credible intervals. Randomization removes systematic bias by trading off bias with variance. We design a training objective to enforce individual calibration and use it to train randomized regression functions. The resulting models are more calibrated for arbitrarily chosen subgroups of the data, and can achieve higher utility in decision making against adversaries that exploit miscalibrated predictions.

1. Introduction

Many applications of machine learning, such as safety-critical systems and medical diagnosis, require accurate estimation of the uncertainty associated with each prediction. Uncertainty is typically represented using a probability distribution on the possible outcomes. To reflect the underlying uncertainty, these probabilities should be calibrated (Cesa-Bianchi & Lugosi, 2006; Vovk et al., 2005; Guo et al., 2017). In the regression setup, for example, the true outcome should be below the predicted 50% quantile (median) roughly 50% of the times (Kuleshov et al., 2018).

However, even when the probability forecaster is calibrated, it is possible that the true outcome is *always* below the predicted median on a subgroup (e.g., men), and *always* above the predicted median for another (e.g., women). In other words, the standard notion of calibration is a property that needs to hold only *on average across all predictions* made (i.e., on the set of all input data points). The predicted probabilities can still be highly inaccurate for *individual* data samples. These predictions can lead to unfair or otherwise suboptimal decisions. For example, a bank might over-predict credit risk for one gender group and unfairly deny loans, or under-predict credit risk for a group that can then exploit this mistake to their advantage. Group calibration (Kleinberg et al., 2016) partly addresses the short-comings of average calibration by requiring calibration on pre-specified groups (e.g. men and women). In particular, (Hébert-Johnson et al., 2017) achieves calibration on any group that can be computed from input by a small circuit. However, these methods are not applicable when the groups are unknown or difficult to compute from the input. For example, groups can be defined by features that are unobserved e.g. due to personal privacy.

Ideally, we would like forecasters that are calibrated on each individual sample. Our key insight is that individual calibration is possible when the probability forecaster is itself randomized. Intuitively, a randomized forecaster can output random probabilistic forecasts (e.g., quantiles) — it is the predicted quantile that is randomly above or below a fixed true value with the advertised probability (see Figure 1). Randomization can remove systematic miscalibration on any group of data samples. Useful forecasters also need to be sharp — the predicted probability should be concentrated around the true value. We design a concrete learning objective that enforces individual calibration. Combined with an objective that enforces sharpness, such as traditional log-likelihood, we can learn forecasters that trade-off calibration and sharpness *Pareto optimally*. The objective can be used to train any prediction model that takes an additional random source as input, such as deep neural networks.

We assess the benefit of forecasters trained with the new objective on two applications: fairness and decision making under uncertainty against adversaries. Calibration on protected groups traditionally has been a definition for fairness (Kleinberg et al., 2016). On a UCI crime prediction

[1]Computer Science Department, Stanford University. Correspondence to: Shengjia Zhao <sjzhao@stanford.edu>, Tengyu Ma <tengyuma@stanford.edu>, Stefano Ermon <ermon@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

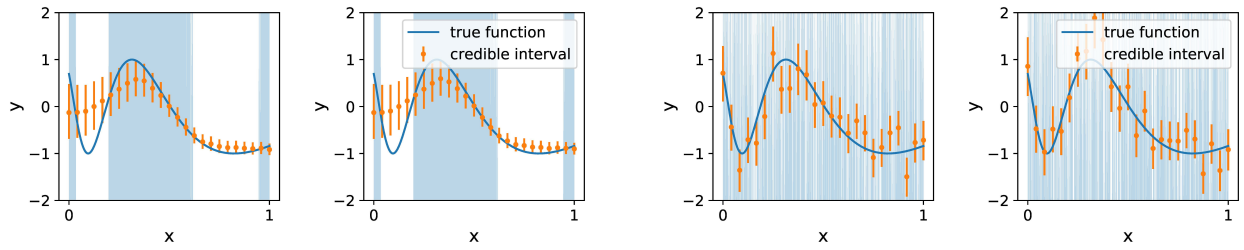

Figure 1: Example of deterministic forecaster that is calibrated on average (**left**) and randomized forecaster that is individually calibrated (**right**). We plot the 80% credible interval and the median (orange dot) the forecaster outputs, and shade in cyan the area where the predicted median is less than the true function value. **Left**: the deterministic forecaster outputs a fixed credible interval (the left 2 plots are identical) and and can be miscalibrated on sub-groups of the data samples (e.g. it is not calibrated for the samples in the shaded area, because the predicted median is always less than the true function value). **Right**: The randomized forecaster outputs a different credible interval each time (the right 2 plots are different), and can remove systematic miscalibration on sub-groups of data samples.

task, we show that forecasters trained for individual calibration achieve lower calibration error on protected groups without knowing these groups in advance.

For decision making we consider the Bayesian decision strategy — choosing actions that minimize expected loss under the predicted probabilities. We prove strong upper bounds on the decision loss when the probability forecaster is calibrated on average or individually. However, when the input data distribution changes, forecasters calibrated on average lose their guarantee, while individually calibrated forecasters are less affected. We support these results by simulating a game between a bank and its customers. The bank approves loans based on predicted credit risk, and customers exploit any mistake of the bank, i.e. the distribution of customers change adversarially for the bank. We observe that the bank incurs lower loss when the credit risk forecaster is trained for individually calibration.

2. Preliminary: Forecaster and Calibration

2.1. Notation

We use bold capital letters $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{H}$ to denote random variables, lower case letters x, y, z, h to denote fixed values, and $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \mathcal{H}$ to denote the set of all possible values they can take.

For a random variable $\mathbf{Z}$ on $\mathcal{Z}$ we will use $\mathbb{F}_{\mathbf{Z}}$ to denote the distribution of $\mathbf{Z}$, and denote this relationship as $\mathbf{Z} \sim \mathbb{F}_{\mathbf{Z}}$. If $\mathcal{Z}$ is an interval in $\mathbb{R}$ we overload $\mathbb{F}_{\mathbf{Z}}$ to denote the cumulative distribution function (CDF) $\mathcal{Z} \to [0, 1]$ of $\mathbf{Z}$.

$\mathbf{X}$ is a random variable on $\mathcal{X}$, if $\mathcal{S} \subset \mathcal{X}$ is a measurable set and $\Pr[\mathbf{X} \in \mathcal{S}] > 0$, we will use the notation $X_{\mathcal{S}}$ as the random variable distributed as $\mathbf{X}$ conditioned on $\mathbf{X} \in \mathcal{S}$.

Let $\mathcal{Y}$ be an interval in $\mathbb{R}$, we use $\mathcal{F}(\mathcal{Y})$ denote the set of all CDFs on $\mathcal{Y}$. We use $d : \mathcal{F}([0, 1]) \times \mathcal{F}([0, 1]) \to \mathbb{R}$ to denote a distance function between two CDFs on $[0, 1]$. For example, the Wasserstein-1 distance is defined for any $\mathbb{F}, \mathbb{F}' \in \mathcal{F}([0, 1])$ as

$$d_{W1}(\mathbb{F}, \mathbb{F}') = \int_{r=0}^{1} |\mathbb{F}(r) - \mathbb{F}'(r)| dr$$

This is the distance we will use throughout the paper. We provide results for other distances in the appendix.

2.2. Problem Setup

Given an input feature vector $x \in \mathcal{X}$ we would like to predict a distribution on the label $y \in \mathcal{Y}$. We consider regression problems where $\mathcal{Y}$ is an interval in $\mathbb{R}$.

Suppose there is a true distribution $\mathbb{F}_{\mathbf{X}}$ on $\mathcal{X}$, and a random variable $\mathbf{X} \sim \mathbb{F}_{\mathbf{X}}$. For each $x \in \mathcal{X}$, we also assume there is some true distribution $\mathbb{F}_{\mathbf{Y}|x}$ on $\mathcal{Y}$, and a random variable $\mathbf{Y} \sim \mathbb{F}_{\mathbf{Y}|x}$. As a convention, the random variable $\mathbf{Y}$ only appears in any expression along side x or $\mathbf{X}$. Its distribution is always defined conditioned on x (or after we have randomly sampled $x \sim \mathbb{F}_{\mathbf{X}}$).

A **probability forecaster** is a function $h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})$ that maps an input $x \in \mathcal{X}$ to a continuous CDF $h[x]$ over $\mathcal{Y}$. Note that $h[x]$ is a CDF, i.e. it is a function that takes in $y \in \mathcal{Y}$ and returns a real number $h[x](y) \in [0, 1]$. We use $[\cdot]$ to denote function evaluation for x and $(\cdot)$ for y.

Let $\mathcal{H} \stackrel{\text{def}}{=} \{h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})\}$ be the set of possible probability forecasters. We consider **randomized** forecasters $\mathbf{H}$ which is a random function taking values in $\mathcal{H}$.

To clarify notation, $\mathbf{H}[\mathbf{X}](\mathbf{Y}), \mathbf{H}[x](\mathbf{Y})$ and $\mathbf{H}[x](y)$ are all random variables taking values in $[0, 1]$, but they are random variables on different sample spaces.

- $\mathbf{H}[\mathbf{X}](\mathbf{Y})$ is a random variable on the sample space $\mathcal{H} \times \mathcal{X} \times \mathcal{Y}$ — All of $\mathbf{H}, \mathbf{X}, \mathbf{Y}$ are random.

	Individual calibration (Definition 1)	$\overset{\text{Thm 2}}{\Longrightarrow}$	Adv group calibration (Definition 4)	$\Longrightarrow$	Group calibration (Definition 3)	$\Longrightarrow$	Average calibration (Definition 2)
R.V. that needs to be uniform	$\mathbf{H}[x](\mathbf{Y})\ \forall x \in \mathcal{X}$		$\mathbf{H}[\mathbf{X}_S](\mathbf{Y})$ $\forall S \subset \mathcal{X}$		$\mathbf{H}[\mathbf{X}_{S_i}](\mathbf{Y})$ $i = 1, \cdots, n$		$\mathbf{H}[\mathbf{X}](\mathbf{Y})$
Deterministic	Not Achievable (Prop 1)		Not achievable (Prop 4)		Achievable		Achievable
Randomized	Achievable* (Thm 1)		Achievable* (Thm 2)		Achievable		Achievable

Figure 2: Relationships between different notions of calibration, ordered from strongest (individual calibration) to the weakest (average calibration). *With caveats in certain situations, see additional discussion in Appendix A.6.

- $\mathbf{H}[x](\mathbf{Y})$ is a random variable on the sample space $\mathcal{H} \times \mathcal{Y}$, while x is just a fixed value in $\mathcal{X}$.
- $\mathbf{H}[x](y)$ is a random variable on the sample space $\mathcal{H}$, while x, y are just fixed values in $\mathcal{X} \times \mathcal{Y}$.

Similarly $h[\mathbf{X}](\mathbf{Y}), h[x](\mathbf{Y})$ are also random variables taking values in $[0, 1]$, while $h[x](y)$ is just a number in $[0, 1]$ (there is no randomness). This difference will be crucial to distinguishing different notions of calibration.

We use $\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}, \mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}$ and $\mathbb{F}_{\mathbf{H}[x](y)}$, etc, to denote the CDF of these $[0, 1]$-valued random variables. These are in general different CDFs because the random variables have different distributions (they are not even defined on the same sample space).

If the $\mathbf{H}$ takes some fixed value in $h \in \mathcal{H}$ with probability 1, we call it a deterministic forecaster. Because deterministic forecasters are a subset of randomized forecasters, we will use $\mathbf{H}$ to denote them as well.

2.3. Background: Perfect Probability Forecast

We consider several criteria that a good probability forecaster $\mathbf{H}$ (randomized or deterministic) should satisfy, and whether they are achievable.

Given some input $x \in \mathcal{X}$, an ideal forecaster should always output the CDF of the true conditional distribution $\mathbb{F}_{\mathbf{Y}|x}$. We call such a forecaster a "perfect forecaster".

However, learning an (approximately) perfect forecaster from training data is almost never possible. Usually each $x \in \mathcal{X}$ appear at most once in the training set (e.g. it is unlikely for the training set to contain identical images). It would be almost impossible to infer the entire CDF $\mathbb{F}_{Y|x}$ from a single sample $y \sim \mathbb{F}_{Y|x}$ (Vovk et al., 2005) without strong assumptions.

2.4. Individual Calibration

Because perfect probability forecasters are difficult to learn, we relax our requirement, and look at which desirable property of a perfect forecaster to emulate.

We first observe that for some x, when the random variable $\mathbf{Y}$ is truly drawn from a continuous CDF $\mathbb{F}_{\mathbf{Y}|x} \in \mathcal{F}(\mathcal{Y})$, by the inverse CDF theorem, $\mathbb{F}_{\mathbf{Y}|x}(\mathbf{Y})$ should be a random variable with uniform distribution in $[0, 1]$ — As a notation reminder, $\mathbb{F}_{\mathbf{Y}|x}$ is a fixed (CDF) function $\mathcal{Y} \to [0, 1]$, $\mathbf{Y}$ is a random variable taking values in $\mathcal{Y}$. Therefore, $\mathbb{F}_{\mathbf{Y}|x}(\mathbf{Y})$ is a random variable taking values in $[0, 1]$. Also recall the convention that whenever $\mathbf{Y}$ appears in an expression, its distribution is always conditioned on x, i.e. $\mathbf{Y} \sim \mathbb{F}_{\mathbf{Y}|x}$.

If $\mathbf{H}$ is indeed a perfect forecaster, then $\forall x \in \mathcal{X}$, $\mathbf{H}[x]$ should always equal the true CDF: $\mathbf{H}[x] = \mathbb{F}_{\mathbf{Y}|x}$. Therefore $\mathbf{H}[x](\mathbf{Y})$ is a uniformly distributed random variable. In other words, $\mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}$ is the CDF function of a uniform random variable. Conversely, we can require this property for any good forecaster.

Formally, let $d : \mathcal{F}([0, 1]) \times \mathcal{F}([0, 1]) \to \mathbb{R}^+$ be any distance function (such as the Wasserstein-1 distance) between CDFs over $[0, 1]$. For convenience we use $\mathbb{F}_{\mathbf{U}}$ to denote the CDF of a uniform random variable in $[0, 1]$. We can measure

$$\mathrm{err}_{\mathbf{H}}(x) \overset{\text{def}}{=} d(\mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}})$$

and if $\mathrm{err}_{\mathbf{H}}(x) = 0$ for all $x \in \mathcal{X}$, we say the forecaster $\mathbf{H}$ satisfies individual calibration.

In practice individual calibration can only be achieved approximately, i.e. $\mathrm{err}_{\mathbf{H}}(x) \approx 0$ for most values of $x \in \mathcal{X}$, which we formalize in the following definition.

Definition 1. *A forecaster $\mathbf{H}$ is (ϵ, δ)-probably approximately individually calibrated (PAIC) (with respect to distance metric d) if*

$$\Pr\left[\mathrm{err}_{\mathbf{H}}(\mathbf{X}) \le \epsilon\right] \ge 1 - \delta$$

This definition is intimately related to a standard definition of calibration for regression (Gneiting et al., 2007; Kuleshov et al., 2018) which we restate slightly differently

Definition 2. *A forecaster $\mathbf{H}$ is ϵ-approximately average calibrated (with respect to distance metric d) if*

$$d(\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}}) \le \epsilon$$

Note that $d(\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}}) = 0$ is equivalent to the original definition of calibrated regression in (Kuleshov et al., 2018)

$$\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}(c) = \Pr[\mathbf{H}[\mathbf{X}](\mathbf{Y}) \leq c] = c = \mathbb{F}_{\mathbf{U}}(c), \forall c \in [0,1]$$

In words, under the ground truth distribution, y should be below the c-th quantile of the predicted CDF exactly c percent of the times. More generally, an ϵ-approximately average calibrated forecaster with respect to Wasserstein-1 distance also has ϵ ECE (expected calibration error) — a metric commonly used to measure calibration error (Guo et al., 2017). For details see Appendix A.1.1.

Despite the similarity, individual calibration is actually much stronger compared to average calibration (Figure 2). Individual calibration requires $\mathbf{H}[x](\mathbf{Y})$ be uniformly distributed for *every* x. average calibration only require this *on average*: $\mathbf{H}[\mathbf{X}](\mathbf{Y})$ is uniformly distributed only if $\mathbf{X} \sim \mathbb{F}_{\mathbf{X}}$ — it may not be uniformly distributed if $\mathbf{X}$ has some other distribution. For example, if $\mathcal{X}$ can be partitioned based on gender, the forecaster can be uncalibrated when $\mathbf{X}$ is restricted to a particular gender.

2.5. Group Calibration

To address the short-coming of average calibration in Definition 2, a stronger notion of calibration has been proposed (Kleinberg et al., 2016; Hébert-Johnson et al., 2017). We choose in advance measurable subsets $\mathcal{S}_1, \cdots, \mathcal{S}_k \subset \mathcal{X}$ such that $\Pr[\mathbf{X} \in \mathcal{S}_i] \neq 0, \forall i \in [k]$ (for Hébert-Johnson et al. (2017) these are sets that can be identified by a small circuit from the input), and define the random variables $\mathbf{X}_{\mathcal{S}_1}, \cdots, \mathbf{X}_{\mathcal{S}_k}$ (Recall that $\mathbf{X}_{\mathcal{S}_i}$ is distributed by $\mathbf{X}$ conditioned on $\mathbf{X} \in \mathcal{S}_i$).

Definition 3 (Group Calibration)**.** *A forecaster* $\mathbf{H}$ *is* ϵ*-approximately group calibrated w.r.t. distance metric* d *and* $\mathcal{S}_1, \cdots, \mathcal{S}_k \subset \mathcal{X}$ *if*

$$\forall i \in [k], d\left(\mathbb{F}_{\mathbf{H}[\mathbf{X}_{\mathcal{S}_i}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}}\right) \leq \epsilon$$

This can alleviate some of the shortcomings of average calibration. However, the groups must be pre-specified or easy to compute from the input features $\mathbf{X}$. A much stronger definition (Figure 2) is group calibration for any subset of $\mathcal{X}$ that is sufficiently large.

Definition 4 (Adversarial Group Calibration)**.** *A forecaster* $\mathbf{H}$ *is* (ϵ, δ)*-adversarial group calibrated (with respect to distance metric* d*) if for* $\forall \mathcal{S} \subset \mathcal{X}$ *such that* $\Pr[\mathbf{X} \in \mathcal{S}] \geq \delta$ *we have*

$$d\left(\mathbb{F}_{\mathbf{H}[\mathbf{X}_{\mathcal{S}}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}}\right) \leq \epsilon \tag{1}$$

2.6. Interpreting Individual Calibration

We remark that when the forecaster is stochastic, the notion of calibration has a different interpretation compared to deterministic forecasters. When the forecaster is deterministic, individual calibration in Definition 1 implies that almost surely, the forecasted distribution must be identical to the true distribution — a much stronger requirement than calibration for stochastic forecasters. However, as we show in the remaining of the paper, individual calibration for deterministic forecasters is stronger but unverifiable, while for stochastic forecasters is weaker but verifiable (thus can also be optimized with a learning algorithm).

We also remark that we extend the definition of regression calibration proposed in (Kuleshov et al., 2018). This is not the unique reasonable definition for regression calibration. For example, we can partition $\mathcal{Y}$ into bins to convert the regression problem into a classification problem, apply the classification calibration definition (Vovk et al., 2005; Guo et al., 2017), and take the number of bins to infinity. This leads to a different definition for calibration than ours (even in the limit of infinitely many bins). Alternative definitions have also been proposed in (Levi et al., 2019). Individual calibration that extends these alternative definitions is beyond the scope of this paper.

3. Impossibility Results for Deterministic Forecasters

We first present results showing that individual calibration and adversarial group calibration are impossible to verify with a deterministic forecaster. Therefore, there is no general method to train a classifier to achieve individual calibration. This motivates the need for randomized forecasters.

Given a finite dataset $\mathcal{D} = \{(x_i, y_i)\}$ and a deterministic forecaster $h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})$, suppose some verifier $T(\mathcal{D}, h) \to \{\text{yes}, \text{no}\}$ aims to verify if h is (ϵ, δ)-PAIC. We claim that no verifier can correctly decide it (unless $\epsilon \geq 1/4$ or $\delta = 1$ which means the calibration is trivially bad). This proposition is for the Wasserstein-1 distance d_{W1}. Examples of other distances are given in the Appendix B.1.

Proposition 1. *For any distribution* $\mathbb{F}$ *on* $\mathcal{X} \times \mathcal{Y}$ *such that* $\mathbb{F}_{\mathbf{X}}$ *assigns zero measure to individual points* $\{x \in \mathcal{X}\}$ *sample* $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\} \overset{\text{i.i.d.}}{\sim} \mathbb{F}$. *For any deterministic forecaster* h *and any function* $T(\mathcal{D}, h) \to \{\text{yes}, \text{no}\}$ *such that*

$$\Pr_{\mathcal{D} \sim \mathbb{F}}[T(\mathcal{D}, h) = \text{yes}] = \kappa > 0\,,$$

there exists a distribution $\mathbb{F}'$ *such that (a)* h *is not* (ϵ, δ)*-PAIC w.r.t* $\mathbb{F}'$ *for any* $\epsilon < 1/4$ *and* $\delta < 1$*, and (b)*

$$\Pr_{\mathcal{D} \sim \mathbb{F}'}[T(\mathcal{D}, h) = \text{yes}] \geq \kappa$$

This additionally implies that no learning algorithm can guarantee to produce an individually calibrated forecaster.

because otherwise the learning algorithm and its guarantee can serve as a verifier. Proofs of the proposition and a similar negative result for adversarial group calibration are in Appendix B.1. In addition to impossibility of verification, it is also known (in the conformal prediction setup) that no forecaster can guarantee individual calibration in a non-trivial way (Barber et al., 2019). For a discussion see related works.

In light of these negative results, there are two options: one can make additional assumptions about the true distribution $\mathbb{F}_{\mathbf{XY}}$, such as Lipschitzness. However, these assumptions are usually hard to verify, and their usefulness diminishes as the dimensionality of $\mathcal{X}$ increases. We propose an alternative: in contrast to deterministic forecasters, there is a sufficient condition for individual calibration for *randomized* forecasters. The sufficient condition is verifiable can be conveniently converted into a training objective.

4. Individual Calibration with Randomized Forecasting

4.1. Reparameterized Randomized Forecaster

We will consider randomized forecasters that are deterministic functions applied on the input feature vector x and some fixed random seeds R. (For readers familiar with variational Bayesian inference (Kingma & Welling, 2013), this is reminiscent of the reparameterization trick.) Concretely, we choose a deterministic function $\bar{h} : \mathcal{X} \times [0,1] \to \mathcal{F}(\mathcal{Y})$, and let $\mathbf{R} \sim \mathbb{F}_{\mathbf{U}}$. Define the randomized forecaster $\mathbf{H}[x]$ as

$$\mathbf{H}[x] = \bar{h}[x, \mathbf{R}] \tag{2}$$

In addition, we would like $\bar{h}[x, r]$ to be a monotonic function of r for all x. Any continuous function that is not monotonic can be transformed into one by shuffling r.

4.2. Sufficient Condition for Individual Calibration

In this subsection, we introduce sufficient (but not necessary) conditions of individual calibration for randomized forecasters defined in Eq. (2). First, recall that the definition of individual calibration requires $\mathbf{H}[x](\mathbf{Y}) := \bar{h}[x, \mathbf{R}](\mathbf{Y})$ to be a uniform distribution for most of $x \in \mathcal{X}$ sampled from $\mathbb{F}_{\mathbf{X}}$. This condition is hard to verify given samples, because for each x in the training (or test) set, we typically only observe a single corresponding label $y \sim \mathbb{F}_{\mathbf{Y}|x}$.

As an alternative, we propose to verify whether $\bar{h}[x, \mathbf{R}](y)$ is uniformly distributed (for the unique sample y). Therefore we introduce a stronger but verifiable condition:

$$\bar{h}[x, \mathbf{R}](y) \text{ is uniformly distributed} \tag{3}$$

for most (random) choices of x, y under $\mathbb{F}_{XY}$.

The benefit is that the condition — $\bar{h}[x, \mathbf{R}](y)$ is uniformly distributed — can be written in an equivalent form when $\bar{h}[x, r]$ is a monotonic function in r (proved in Theorem 1)

$$\bar{h}[x, r](y) = r, \forall r \in [0, 1],$$

We formalize a relaxed version of this condition (allowing for approximation errors) as follows:

Definition 5. *A forecaster $\bar{h}$ is (ϵ, δ)-monotonically probably approximately individually calibrated (mPAIC) if*

$$\Pr\left[|\bar{h}[\mathbf{X}, \mathbf{R}](\mathbf{Y}) - \mathbf{R}| \geq \epsilon\right] \leq \delta$$

Note that even though we want to achieve $\bar{h}[x, r](y) = r, \forall r$, this does not mean that $\bar{h}$ ignores the input x — $\bar{h}[x, r](.)$ has the special form of a CDF, so $\bar{h}[x, r]$ must output a CDF concentrated around the observed label y to satisfy mPAIC.

The following theorem formalizes our previous intuition that monotonic individual calibration (mPAIC) is obtained by imposing additional restrictions on individual calibration (PAIC) — mPAIC is a sufficient condition for PAIC.

Theorem 1. *If $\bar{h}$ is (ϵ, δ)-mPAIC, then for any $\epsilon' > \epsilon$ it is $(\epsilon', \delta(1-\epsilon)/(\epsilon'-\epsilon))$-PAIC with respect to the 1-Wasserstein distance.*

Proof can be found in Appendix B.2. This theorem shows that mPAIC implies PAIC (up to different constants ϵ, δ). In particular, if a forecaster achieves mPAIC perfectly (i.e. $\epsilon = 0, \delta = 0$), it is also perfectly PAIC.

The benefit of Definition 5 is that it can be verified with a finite number of validation samples, and is amenable to uniform convergence bounds, so that we can train it following the standard empirical risk minimization framework.

Proposition 2. *[Concentration] Let $\bar{h}$ be any (ϵ, δ)-mPAIC forecaster, and $(x_1, y_1), \cdots, (x_n, y_n) \overset{i.i.d.}{\sim} \mathbb{F}_{\mathbf{XY}}$, $r_1, \cdots, r_n \overset{i.i.d.}{\sim} \mathbb{F}_{\mathbf{U}}$, then with probability $1 - \gamma$*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(|\bar{h}[x_i, r_i](y_i) - r_i| \geq \epsilon) \leq \delta + \sqrt{\frac{-\log \gamma}{2n}}$$

4.3. Learning Calibrated Randomized Forecasters

Given training data $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{F}_{\mathbf{XY}}$ we would like to learn a forecaster that satisfy mPAIC. We propose a concrete set of forecasters and a learning algorithm we will use in all of our experiments.

We will model uncertainty with Gaussian distributions $\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$. We parameterize $\bar{h} = \bar{h}_\theta$ as a deep neural network with parameters θ. The neural

networks takes in concatenation of x and r and outputs the μ, σ that decides the returned CDF.

Inspired by Proposition 2, we optimize the mPAIC objective on the training data defined as [1]

$$\mathcal{L}_{\text{PAIC}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}|\bar{h}_\theta[x_i, r_i](y_i) - r_i|$$

Practically, this objective enforces the calibration but does not take into account the sharpness of the forecaster. For example, a simple $\bar{h}$ can be constructed that trivially outputs r (Appendix A.1.2). We would also like to minimize the variance σ of the predicted Gaussian distribution. Therefore we also regularize with the log likelihood (i.e. log derivative of the CDF) objective which encourages the sharpness of the prediction

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{d}{dy}\bar{h}[x_i, r_i](y_i)$$

Because we model uncertainty with Gaussian distributions, the $\mathcal{L}_{\text{NLL}}$ objective is equivalent to the standard squared error (MSE) objective typically used in regression (Myers & Myers, 1990) literature. We use a hyper-parameter α to trade-off between the two objectives:

$$\mathcal{L}_\alpha(\theta) = (1-\alpha)\mathcal{L}_{\text{PAIC}}(\theta) + \alpha\mathcal{L}_{\text{NLL}}(\theta) \tag{4}$$

In other words, when $\alpha \approx 0$ the objective is almost exclusively PAIC, while when $\alpha = 1$ we reduce to the standard log likelihood maximization (i.e. MSE) objective.

5. Application I: Fairness

Individual calibration provides guarantees on the performance of a machine learning model on individual samples. As we will show, this has numerous applications. We begin discussing its use in settings where fairness is important.

5.1. From Individual Calibration to Group Calibration

In high-stakes applications of machine learning (e.g., healthcare and criminal justice), it is imperative that predictions are fair. Many definitions of fairness are possible (see related work section), and calibration is a commonly used one. For example, in a healthcare application we would like to prevent a systematic overestimation or underestimation of a predicted risk for different socio-economic groups (Pfohl et al., 2019). One natural requirement is that predictions for every group are calibrated, that is, the true value below the $r\%$ quantile exactly $r\%$ of the times.

If the protected groups are known at training time, we could enforce group calibration as in Definition 3. However, it can be difficult to specify which groups to protect a priori. Some groups are also defined by features that are unobserved e.g. due to personal privacy.

We propose to address the problem by requiring a stronger notion of calibration, adversarial group calibration, where *any group of sufficiently large size needs to be protected*. Moreover, we can achieve this stronger notion of calibration because it is implied by individual calibration.

Theorem 2. *If a forecaster is (ϵ, δ)-PAIC with respect to distance metric $\mathcal{W}_p$, then $\forall \delta' \in [0,1]$, $\delta' > \delta$, it is $(\epsilon + \delta/\delta', \delta')$-adversarial group calibrated with respect to $\mathcal{W}_p$.*

We prove a stronger version of this theorem in Appendix B.3. We know from theory that a forecaster trained on the individual calibration objective Eq.(4) can achieve good individual calibration (and thus group calibration) on the training data. We will now experimentally verify that the benefit generalizes to test data.

5.2. Experiments

Experiment Details. We use the UCI crime and communities dataset (Dua & Graff, 2017) and we predict the crime rate based on features of the neighborhood (such as racial composition). For training details and network architecture see Appendix B.3. [2]

Recalibration. Post training recalibration is a common technique to improve calibration. For completeness we additionally report all results when combined with recalibration by isotonic regression as in (Kuleshov et al., 2018). Post training recalibration improves average calibration, but as we show in the experiments, has limited effect on individual or adversarial group calibration.

5.2.1. Performance Metrics

1) Sharpness metrics: Sharpness measures whether the prediction $\mathbf{H}[x]$ is concentrated around the ground truth label y. We will use two metrics: negative log likelihood on the test data $\mathbb{E}\left[-\log\bar{h}(\mathbf{X},\mathbf{R})(\mathbf{Y})\right]$, and expected standard deviation of the predicted CDFs $\mathbb{E}\left[\sqrt{\text{Var}[\bar{h}(\mathbf{X},\mathbf{R})]}\right]$. Because forecaster outputs Gaussian distributions to represent uncertainty, this is simply the average standard deviation σ predicted by the forecaster.

2) Calibration metrics: we will measure the (ϵ, δ)-adversarial group calibration (with 1-Wasserstein distance) defined in Definition 4 and Eq.(1). In particular, we measure ϵ as a function of δ — for a smaller group (smaller δ) the ϵ should be larger (worse calibration), and vice versa. A better forecaster should have a smaller ϵ for any given value of δ. We show in Appendix A.1.1 that measuring ϵ

[1] In practice we always sample a new r_i for each training step.

[2] https://github.com/ShengjiaZhao/Individual-Calibration

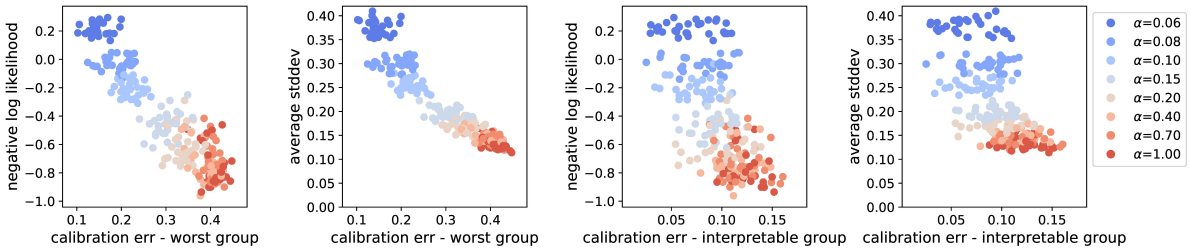

Figure 3: Sharpness (negative log likelihood and average standard deviation) vs. calibration error (on worst group and worst interpretable group) on the UCI crime dataset for different values of α (recall that α is the coefficient to trade-off $\mathcal{L}_{\text{NLL}}$ and $\mathcal{L}_{\text{PAIC}}$, $\alpha = 1$ corresponds to standard training with $\mathcal{L}_{\text{NLL}}$). Each dot represent the performance of a classifier trained with some value of α. It can be seen that there is a trade-off: a smaller value of α (more weight on $\mathcal{L}_{\text{PAIC}}$ and less weight on $\mathcal{L}_{\text{NLL}}$) leads to better calibration and worse sharpness. **Left 2**: Calibration loss on the worst group that contain 20% of the test data. **Right 2**: Calibration loss on the worst interpretable group.

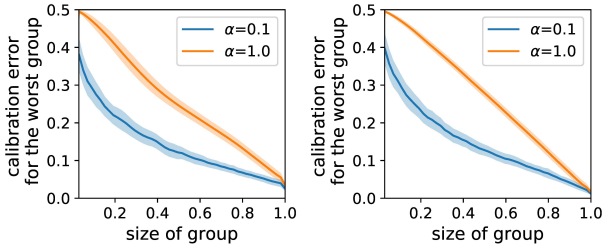

Figure 4: Calibration error as a function of the worst group size. The shaded region represents the one standard deviation error bar on random training/validation partitions. A forecaster trained with the PAIC objective ($\alpha = 0.1$) has better calibration on adversarially chosen groups. **Left**: Without recalibration. **Right** With post training recalibration. Post training recalibration improves average calibration (i.e. group size = 1.0) but does not improve adversarial group calibration (group size < 1.0).

is identical to the commonly used ECE (Guo et al., 2017) metric for miscalibration.

The worst adversarial group may be an uninterpretable set, which is argurably less important than interpretable groups. Therefore, we also measure group calibration with respect to a set of known and interpretable groups. In particular, for each input feature we compute its the median value in the test data, and consider the groups that are above/below the median. For example, if the input feature is income, we considers the group with above median income, and the group with below median income. We also consider the intersection of any two groups.

5.2.2. Results

The results are shown in Figure 3 and 4. We compare different values of α (with $\alpha \approx 0$ we learn with $\mathcal{L}_{\text{PAIC}}$, and with $\alpha \approx 1$ we learn with $\mathcal{L}_{\text{NLL}}$).

The main observation is forecasters learned with smaller α almost always achieve better group calibration (both adversarial group and interpretable group) and worse sharpness (log likelihood and variance of predicted distribution). This shows a trade-off between calibration and sharpness. Depending on the application, a practitioner can adjust α and find the appropriate trade-off, e.g. given some constraint on fairness (maximum calibration error) achieve the best sharpness (log likelihood).

We also observe that post training recalibration improves average calibration (i.e. when the size of adversarial group is 100% in Figure 4). However, we cannot expect recalibration to improve individual or adversarial group calibration — we empirically confirm this in Figure 4.

In Table 1 in Appendix B.3 we also report the worst interpretable groups that are miscalibrated. These do correspond to groups that we might want to protect, such as percent of immigrants, or racial composition.

6. Application II: Decision under Uncertainty

Machine learning predictions are often used to make decisions. With good uncertainty quantification, the agent can consider different plausible outcomes, and pick the best action in expectation.

More formally, suppose there is a set of actions $a \in \mathcal{A}$, and some loss function $l : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. If we had a perfect forecaster (i.e. $\mathbf{H}[x] = \mathbb{F}_{\mathbf{Y}|x}$), then given input x, Bayesian decision theory would suggest to take the action that minimizes expected loss (Fishburn & Kochenberger, 1979) under the predicted probability.

$$l_{\mathbf{H}}(x) \stackrel{\text{def}}{=} \min_a \mathbb{E}_{\tilde{\mathbf{Y}} \sim \mathbf{H}[x]}[l(x, \tilde{\mathbf{Y}}, a)] \tag{5}$$

$$\phi_{\mathbf{H}}(x) \stackrel{\text{def}}{=} \arg\min_a \mathbb{E}_{\tilde{\mathbf{Y}} \sim \mathbf{H}[x]}[l(x, \tilde{\mathbf{Y}}, a)] \tag{6}$$

However, perfect forecaster is almost never possible in

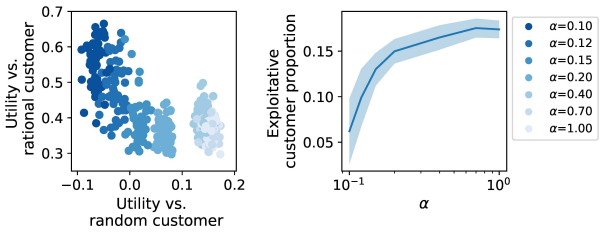

Figure 5: Comparison between individual calibration ($\alpha \approx 0$) and baseline ($\alpha = 1$). **Left**: Each dot represents a random roll-out for different values of α, where we plot the bank's average utility when the customers either decide to apply randomly or rationally. Individually calibrated forecaster perform worse than baseline when the customers are random, but better when customers are rational. **Middle Left**: Proportion of exploitative customers (customers with $y < y_0$ but decide to apply). Random forecasters have less systematic bias and discourages exploitative customers.

practice. Nevertheless, calibration provides some guarantee on the decision rule in Eq.(6) for certain loss functions. In particular, we consider loss functions $l(x, \cdot, a)$ that, for each x, a, are either monotonically non-increasing or non-decreasing in y. We call these loss functions **monotonic**. For example, if y represents stock prices and $a \in \{\text{buy}, \text{sell}\}$, then when $a = \text{buy}$, loss is decreasing in y; when $a = \text{sell}$, loss is increasing in y.

In the following theorem (proof in Appendix B.4) we show that the actual loss cannot exceed the expected loss in Eq.(5) too often. This would be Markov's inequality if Eq.(5) takes expectation under the true distribution $\mathbb{F}_{\mathbf{Y}|x}$. Interestingly the inequality is still true when the expectation is under the predicted distribution $\mathbf{H}[x] \neq \mathbb{F}_{\mathbf{Y}|x}$.

Theorem 3. *Suppose $l : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ is a monotonic non-negative loss, let $\phi_{\mathbf{H}}$ and $l_{\mathbf{H}}$ be defined as in Eq.(6)*

1. If $\mathbf{H}$ is 0-average calibrated, then $\forall k > 0$

$$\Pr[l(\mathbf{X}, \mathbf{Y}, \phi_{\mathbf{H}}(\mathbf{X})) \geq k l_{\mathbf{H}}(\mathbf{X})] \leq 2/k$$

2. If $\mathbf{H}$ is $(0, 0)$-PAIC, then $\forall x \in \mathcal{X}, k > 0$

$$\Pr[l(x, \mathbf{Y}, \phi_{\mathbf{H}}(x)) \geq k l_{\mathbf{H}}(x)] \leq 1/k$$

6.1. Case Study: Credit Prediction

Suppose customers of a financial institution are represented with a feature vector x and a real-valued credit worthiness score $y \in \mathcal{Y}$. The bank has a financial product (e.g. credit card or loan) with a minimum threshold $y_0 \in \mathbb{R}$ for credit worthiness. If a customer chooses to apply for the product, the bank observes x, and uses forecaster $\mathbf{H}$ to predict their true credit worthiness y. There is a positive utility for saying 'yes' to a qualified customer, and a negative utility for saying 'yes' to a disqualified customer. More specifically, the utility (negative loss) for the bank is

	$y \geq y_0$	$y < y_0$
'yes'	1	-3
'no'	0	0

Guarantees from Average Calibration: Suppose the bank uses the Bayesian decision rule in Eq.(6). If $\mathbf{H}$ is average calibrated, Theorem 3 would imply that when the bank says 'yes', at most 25% will be to unqualified customers (when customers truly come from the distribution $\mathbb{F}_{\mathbf{X}}$). For details see Appendix A.4.

The Bayesian decision rule in Eq.(6) is fragile when $\mathbf{H}$ is only average calibrated because the guarantee above is void if the customers do not come from the distribution $\mathbb{F}_{\mathbf{X}}$. For example, suppose some unqualified customers know that their credit scores are overestimated by the bank, then these customers are more likely to apply. More concretely, if only the customers who would be mistakenly approved ended up applying, then the bank is guaranteed to lose (it will suffer a loss of -3 for each customer).

Guarantees from Individual Calibration: If $\mathbf{H}$ is individually calibrated (and hence also calibrated with respect to any adversarial subgroup), the bank cannot be exploited — no matter which subgroup of customers choose to apply, the fraction of unqualified approvals is at most 25%.

6.1.1. Simulation

We perform simulations to verify that individual calibrated forecasters are less exploitable and can achieve higher utility in practice. We model the customers as rational agents that predict their utility and exploit any mistake by the bank. For detailed setup, see Appendix A.4.

The results are shown in Figure 5. We compare different values of $\alpha \in [0.1, 1.0]$ (recall when $\alpha \approx 0$ we almost exclusive optimize $\mathcal{L}_{\text{PAIC}}$, and when $\alpha = 1$ we exclusively optimize $\mathcal{L}_{\text{NLL}}$). We compare the average utility on random customers (customers drawn from $\mathbb{F}_{\mathbf{XY}}$) and rational customers (Appendix A.4).

The main observations is that when customers are random, individual calibrated forecasters perform worse because average calibration is already sufficient. On the other hand, when the customers are rational, and try to exploit any systematic bias with the decision maker, individually calibrated forecasters perform much better.

7. Related Work

Randomized Forecast: Randomized forecast has been used in adversarial environments. In online learning where the true label can be adversarial, randomized forecasters

can achieve low regret (Cesa-Bianchi & Lugosi, 2006; Shalev-Shwartz et al., 2012). In security games / Stackelberg games (Tsai et al., 2010; Trejo et al., 2015), defender needs randomization to play against attackers. (Perdomo et al., 2020) gives a theoretical characterization of prediction in (possibly adversarial) environments when the loss function is strongly convex.

Calibration: Definitions of average calibration first appeared in the statistics literature (Brier, 1950; Murphy, 1973; Dawid, 1984; Cesa-Bianchi & Lugosi, 2006). Recently there has been a surge of interest in recalibrating classifiers (Platt et al., 1999; Zadrozny & Elkan, 2001; 2002; Niculescu-Mizil & Caruana, 2005), especially deep networks (Guo et al., 2017; Lakshminarayanan et al., 2017). Average calibration in the regression setup has been studied by (Gneiting et al., 2007; Kuleshov et al., 2018).

Group calibration for a small number of pre-specified groups has been studied in (Kleinberg et al., 2016). Interestingly (Hébert-Johnson et al., 2017; Kearns et al., 2017) can achieve calibration for any group computable by a small circuit, but is computationally difficult (likely no polynomial time algorithm). Similarly (Barber et al., 2019) achieve calibration for a set of groups, but only has efficient algorithms for special sets of groups. (Kearns et al., 2019) proposes a notion of individual calibration applicable when there are many prediction tasks, and each individual draws multiple prediction tasks. (Joseph et al., 2016) achieves a notion similar to individual calibration for fairness, but needs strong realizability assumptions that are difficult to verify. (Liu et al., 2018) proves an upper bound on calibration error for any group. However, it is unclear how to compute the upper bound if the group labels are not provided.

(Barber et al., 2019) show that if a forecaster *always* outputs an individually calibrated confidence interval (i.e. the true label belong to the interval with the advertised probability), then the size of the interval cannot be smaller than a trivially constructed forecaster (with large interval size). Our algorithm is not bound by this impossibility result because our algorithm do not *always* produce individually calibrated forecasters — success of the algorithm depends on the inductive bias and the data distribution. However, whenever the algorithm succeeds in producing individually calibrated forecasters, we do obtain post-training guarantees by Theorem 2.

Fairness: In addition to calibration (Kleinberg et al., 2016; Hébert-Johnson et al., 2017; Kearns et al., 2017), other definitions of fairness include metric based fairness (Dwork et al., 2012), equalized odds (Hardt et al., 2016), conterfactual fairness (Kusner et al., 2017; Kilbertus et al., 2017), and representation fairness (Zemel et al., 2013; Louizos et al., 2015; Song et al., 2018). The trade-off between these definitions are discussed in (Pleiss et al., 2017; Kleinberg et al., 2016; Friedler et al., 2016; Corbett-Davies et al., 2017).

8. Conclusion and Future Work

In this paper we explore using randomization to achieve individual calibration for regression. We show that these individually calibrated predictions are useful for fairness or decision making under uncertainty. One future direction is extending our results to classification. The challenge is that there is no natural way to define a CDF for a discrete random variables. Another open question is a good theoretical characterization of the trade-off between sharpness and individual calibration.

9. Acknowledgements

This research was supported by AFOSR (FA9550-19-1-0024), NSF (#1651565, #1522054, #1733686), JP Morgan, ONR, TRI, FLI, SDSI, and SAIL. Toyota Research Institute ("TRI") provided funds to assist the authors with their research but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. TM is also supported in part by Lam Research and Google Faculty Award.

We are thankful for valuable feedback from Kunho Kim (Stanford), Ananya Kumar (Stanford) and Wei Chen (MSRA).

References

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

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

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

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM, 2017.

Dawid, A. P. 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, 1984.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

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

Fishburn, P. C. and Kochenberger, G. A. Two-piece von neumann-morgenstern utility functions. *Decision Sciences*, 10(4):503–518, 1979.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

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

Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pp. 325–333, 2016.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Kearns, M., Roth, A., and Sharifi-Malvajerdi, S. Average individual fairness: Algorithms, generalization and experiments. *arXiv preprint arXiv:1905.10607*, 2019.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pp. 656–666, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017.

Levi, D., Gispan, L., Giladi, N., and Fetaya, E. Evaluating and calibrating uncertainty prediction in regression tasks. *arXiv preprint arXiv:1905.11659*, 2019.

Liu, L. T., Simchowitz, M., and Hardt, M. The implicit fairness criterion of unconstrained learning. *arXiv preprint arXiv:1808.10013*, 2018.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Malik, A., Kuleshov, V., Song, J., Nemer, D., Seymour, H., and Ermon, S. Calibrated model-based deep reinforcement learning. *arXiv preprint arXiv:1906.08312*, 2019.

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

Myers, R. H. and Myers, R. H. *Classical and modern regression with applications*, volume 2. Duxbury press Belmont, CA, 1990.

Niculescu-Mizil, A. and Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632. ACM, 2005.

Perdomo, J. C., Zrnic, T., Mendler-Dünner, C., and Hardt, M. Performative prediction. *arXiv preprint arXiv:2002.06673*, 2020.

Pfohl, S., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., and Shah, N. H. Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 271–278, 2019.

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

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

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2018.

Trejo, K. K., Clempner, J. B., and Poznyak, A. S. A stackelberg security game with random strategies based on the extraproximal theoretic approach. *Engineering Applications of Artificial Intelligence*, 37:145–153, 2015.

Tsai, J., Yin, Z., Kwak, J.-y., Kempe, D., Kiekintveld, C., and Tambe, M. Urban security: Game-theoretic resource allocation in networked domains. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

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

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

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699. ACM, 2002.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.