Fairness and Bounded Rationality

William Brown
June 8, 2018

1 Introduction

Recent years have witnessed an explosion of research concerning questions of fairness in algorithmic decision-making. The widespread availability of powerful computing resources and massive data sets has resulted in the use of algorithmic tools supplanting the role of humans for a growing range of procedures, including including criminal sentencing, applicant screening for jobs and college admissions, and loan approval. Institutions typically design these algorithms with the aim of minimizing (or maximizing) a particular metric relevant to the task at hand, such as the rate of recidivism or loan defaulting. When these algorithms are implemented with this single goal in mind, and the people in consideration come from several populations with different distributional characteristics, concerns about fairness and discrimination can easily arise. For example, an algorithm for determining bail eligibility may end up disproportionally rejecting defendants from a particular minority racial group who would in fact be unlikely to violate the conditions of bail. This is even possible if the attribute we wish to protect (e.g. gender, race) is not present in the dataset, as other present attributes may vary in their predictive power across populations [1].

To avoid or mitigate these issues of discrimination, a range of algorithmic techniques have been developed which can provide guarantees of fairness across different learning settings. This first requires formalizing a mathematical definition of fairness, and there is no universal consensus about what the optimal definition should be. Each of the prominently considered definitions has a connection to some intuitive notion of fairness, as well as its own drawbacks and tradeoffs. Forcing an algorithm to treat similar individuals similarly aligns with meritocratic ideas about fairness, but overlooks discrimination across groups of people with different characteristics [2]. Other algorithms focus on satisfying parity at the group-level, but this can become difficult if we desire to protect a possibly exponential of subgroups defined by any combination of protected features [3]. Further, attempting to satisfy both individual and group parity constraints simultaneously is intractable [1,4]. This suggests that we are unlikely to soon arrive at a unifying definition of fairness for all scenarios, in the way that differential privacy has been broadly applied, and we need to consider the implications of each possible choice when designing a fairness mechanism for a particular situation.

Research progress on algorithmic fairness has resulted in mechanisms with different types of fairness guarantees across a variety of settings. If the end goal of these research trajectories is to result in realizable mechanisms which benefit human agents, we believe that these mechanisms should be realistic in the demands they place on human agents to solve complex computational problems. While in certain circumstances it may be feasible to expect an agent to solve any computationally tractable problem, such as in the case of organizations developing their own selection algorithms, there are also plenty of situations involving the decision-making of individual human agents where this may not be as realistic.

We consider questions of algorithmic fairness from the perspective of bounded rationality. We focus our attention on noisy utility models, where agents’ observations about their utilities. Many mechanisms with fairness guarantees consider online learning problems with
repeated agent interactions, and make use of algorithmic techniques such as Upper Confidence Bounds in the bandit setting, or Follow The Perturbed Leader in the fully observable setting. We show that both of these algorithms are robust to certain types of noisy utility models, which directly implies the preservation of fairness guarantees relying on these algorithms. Additionally, we formalize a problem related to fairness and bounded rationality, where a regulator with limited information wishes to incentivize an organization to develop fair algorithms (This problem was presented by Aaron Roth in a recent discussion). We provide penalty schemes which satisfy certain notions of disparate harm minimization and discuss possible further directions of progress.

2 Noisy Utility Models

There has been a great deal of research showcasing the failure of rational actor assumptions in game-theoretic settings. In experiments with human subjects, participants in simple economic games frequently take actions which are irrational in the sense of maximizing expected value given the assumption that all payers are rational. In Prisoner’s Dilemma games, agents sometimes are willing to trust that other players will not defect, even though defection is strictly dominant [5]. People are often risk-averse even in the case of a higher expected value, but the existence and success of lotteries is evidence that we are often bad at comprehending minuscule probabilities. Several formulations of bounded rationality have been considered to encapsulate this phenomenon, such as prospect theory, bounded depth induction, costly computation, and noisy utility models. We focus on noisy utility models because they encapsulate the phenomenon of being rational on aggregate, while sometimes being wildly optimistic or pessimistic. Additionally, they are mathematically convenient to analyze in the online learning setting with which many fairness mechanisms are concerned.

Random utility models are introduced by Blume in “How Noise Matters” [6], where the effects of noise are studied in a range of game theoretic settings. Upon taking an action, an agent’s observed utility is a perturbation of the true utility of the action from some possibly unknown distribution. Notably, random noise models with skew-symmetric distributions are shown to not affect convergence in coordination games. We consider a formulation of noisy utility models which is semantically similar to Blume’s model, but allows us to relax the assumption that the noise distribution is static over time.

Definition 1. Noisy Utility Model. A noisy utility model \( X(a, i, t) \) is a collection of random variables parameterized by action spaces, agent indices, and time indices. Consider a game in which an agent \( i \) plays an action from \( A_i \) in each of \( T \) rounds. The true utility for \( i \) of playing \( a_i \) when all other players play \( a_{-i} \) is \( u_i(a_i, a_{-i}) \), where \( u_i(\cdot, \cdot) \) is player \( i \)'s utility function. A player playing this game with a noisy utility model \( X \) does not observe their true utility in each round, and rather observes:

\[
\hat{u}_i(a_i, a_{-i}, t) = u_i(a_i, a_{-i}) + X(a_i, i, t)
\]

Definition 2. Symmetric Bounded Noisy Utility Model A symmetric bounded noisy utility model is a noisy utility model \( X \) with mean 0 and values in \([−1, 1]\) for all \( a, i, t \).

These models allow us to encapsulate an appealing notion of bounded rationality. Notably, the degree to which an agent is irrational when considering the utility of their most recent action can vary over time, by agent, and by the actions taken by other agents.

3 Bandit Optimization with Noisy Utility

Bandit algorithms common in existing literature on algorithmic fairness [7, 8, 9], particularly in online optimization settings. We consider the application of symmetric bounded noisy utility models to an agent using a version of the Upper Confidence Bounds algorithm
for bandit optimization, and show stability of the traditional no-regret guarantee. In the standard bandit setting, the assumption is typically made that the distribution of rewards for each action remains fixed over time. We observe that this assumption is not necessary, allowing for the stability of existing results when agents play with a symmetric bounded noisy utility model.

### 3.1 Chernoff-Hoeffding Bound

**Theorem 1.** Let \( X = \sum_{i=1}^{n} X_i \), where each \( X_i \) is an independent random variable taking values in \([0, 1]\). Then:

\[
\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2e^{-2t^2/n}
\]

The absolute value statement follows from union-bounding the standard upper and lower Chernoff-Hoeffding bounds.

**Proof.** This follows from substituting \( t = n\sqrt{\frac{\ln(2/\delta)}{2n}} \) in Theorem 1:

\[
\Pr[|X/n - \mathbb{E}[X]/n| \geq t/n] \leq 2e^{-2t^2/n}
\]

\[
\Pr[|X/n - \mathbb{E}[X]/n| \geq \sqrt{\frac{\ln(2/\delta)}{2n}}] \leq 2e^{-\ln(2/\delta)} = 2\delta
\]

\[
\Pr[|X/n - \mathbb{E}[X]/n| \leq \sqrt{\frac{\ln(2/\delta)}{2n}}] \geq 1 - \delta
\]

\[\square\]

### 3.2 Noisy Upper Confidence Bounds

We present a formulation of the UCB algorithm for bandit optimization where an agent observes their utility perturbed by a (symmetric bounded) noisy utility model \( X \) [10]. In each round \( t \), the agent chooses one of \( k \) actions \( a_i \) and receives a noisy signal \( r_i + x_i (r_i \sim D_i, x_i \sim X(i,t)) \) about the reward. After \( T \) rounds, the agent receives \( R = \sum_{t=1}^{T} r_i^t \) as their true reward. We assume that every \( X(i,t) \) takes a value in \([\frac{1}{2}, \frac{3}{2}]\) and every \( D_i \) has range \([\frac{1}{2}, \frac{3}{2}]\), resulting in every \( r_i + x_i \) value being in \([0, 1]\). Other constant bounded ranges can be mapped to this range with only constant cost factors.

**NoisyUCB(\( \delta, T, X \))**

Define \( w(n) = \sqrt{\frac{\ln(2/\delta)}{2n}} \). Initialize empirical means \( \hat{\mu}_i^0 = \frac{1}{2} \) for each action, as well as upper and lower confidence bounds \( u_i^0 = 1 \) and \( l_i^0 = 0 \) and play counts \( n_i^0 = 0 \).

**for** \( t = 1 \) **to** \( T \) **do**

- Pick an action \( i_t \in \arg \max u_i^{t-1} \). Observe signal \( s_i^t = r_i^t + x_i^t, r_i^t \sim D_{i_t}, x_i^t \sim X(i_t, t) \).

- For each \( i \neq i_t \), set \( (\hat{\mu}_i^t, u_i^t, l_i^t, n_i^t) = (\hat{\mu}_i^{t-1}, u_i^{t-1}, l_i^{t-1}, n_i^{t-1}) \).

- For \( i = i_t \), set \( n_i^t \leftarrow n_i^{t-1} + 1, \hat{\mu}_i^t \leftarrow \frac{n_i^{t-1}\hat{\mu}_i^{t-1} + s_i^t}{n_i^t}, u_i^t \leftarrow \hat{\mu}_i^t + w(n_i^t), l_i^t \leftarrow \hat{\mu}_i^t - w(n_i^t) \).

**end for**

Observe reward \( R = \sum_{t=1}^{T} r_i^t \).
Theorem 3. Let \( \text{OPT} = n \cdot \max D_i \). For any \( k \) actions, any \( 0 \leq \delta \) and any symmetric bounded noisy utility model \( X \), let \( R \) be the reward obtained by playing NoisyUCB(\( \delta, T, X \)) for \( T \) rounds. With probability \( 1 - \delta \):

\[
\text{OPT} - R \leq O\left(\sqrt{k \cdot T \cdot \ln\left(\frac{T}{\delta}\right)}\right)
\]

Equivalently,

\[
\text{Regret}(T) \leq O\left(\sqrt{k \cdot T \cdot \ln\left(\frac{T}{\delta}\right)}\right)
\]

\[
\text{AvgRegret}(T) \leq O\left(\frac{\sqrt{k \cdot T \cdot \ln\left(\frac{T}{\delta}\right)}}{T}\right)
\]

The proof is equivalent to the standard analysis of the UCB algorithm [10], with the primary addition being that we allow the distributional properties of observed rewards for each action to differ in each round. The original proof makes use of a Chernoff-Hoeffding bound with the same guarantees as Theorem 2. The mean of the reward signal perturbation added by the noisy utility process is always zero and bounded, and because the bound only requires the expected reward of each action to remain constant over time, the no-regret results of the algorithm are stable to noisy models of bounded rationality.

Recent work in developing online algorithms with fairness guarantees has made frequent use of the upper confidence bound approach, often with chaining [7,8,9]. The same analysis that we provide can be applied to other variants of the UCB algorithm involving interval chaining, so we can preserve the fairness guarantees when noisy bounded rationality is present for these algorithms as well. While the stability guarantees for noisy utility models essentially follow directly from the analysis of the original UCB algorithm, they offer an appealing semantic guarantee. Namely, if we are sampling rewards from a fixed distribution, it does not matter if our initial beliefs about the received rewards are sometimes wildly optimistic or pessimistic depending on the time or action chosen, as long as we are rational in expectation for every action. Particularly in the case of “Fairness Incentives for Myopic Agents” by Kannan et al [9], where agents are not assumed to care about fairness by default, we are also allowed to assume that these agents are not perfectly rational. In many real-world settings where we wish to incentivize fair behavior, agents will not be tabulating their exact utility from every action and playing according to a well-defined online algorithm. But if we believe that agents are reasonable in expectation absent perfect reward signals at every point in time, and that they will act to maximize future expected return, we can arrive at the same results.

4 An Options Purchasing Detour

We pause our discussion about fairness to observe an example which illustrates the practicality of no-regret learning when bounded rationality is at play. Suppose you are a financial analyst interested in purchasing European sell options for some asset class of size \( k \). For whatever reason, the options broker you are working with only sells options with an expiry date of \( T + 1 \). These options can only be exercised on the expiry date, and they cannot be resold.

At each time step, prices are available for sell options on all \( k \) assets, but we are severely restricted in our ability to evaluate whether these options are worthwhile. We may have to purchase relevant financial data for a vendor, compensate other analysts, or pay Amazon Web Services for enabling any computation we do, and so we assume that we are only able to generate a prediction about the value of purchasing a single option at each time step. Furthermore, we have no guarantees that our predictions are accurate. We will only
assume that they are unbiased, and that there are upper and lower bounds on both true and predicted profits. We will also make the assumption that the relationship between the payoff of an option at time $T + 1$ and the cost of purchasing (both the asset and the right to sell) is independent. This is essentially saying that whether the broker overvalues or undervalues the option at a given time is time-independent. Of course, we will not know the true value of our purchases until the expiry date. Because of this independence assumption, deciding which asset to investigate at a given time is equivalent to pulling one of $k$ arms of a multi-armed bandit, and the price of the asset at time $T + 1$, whether or not we choose to exercise the options we have purchased, will dictate our reward.

Researching and purchasing options in this setting can be viewed as deploying some unbiased prediction algorithm which makes use of a context vector obtained from choosing to research that option at a given time. The option-asset pair will be purchased whenever the predicted profit is positive. Using UCB to choose which assets to research with an unbiased but inaccurate algorithm for price prediction is identical to playing with a perfect algorithm whose predictions are perturbed by a bounded noisy process with mean zero. Thus, if we can generate unbiased predictions for a single option at each time step, we can play according to NoisyUCB. The guarantees of NoisyUCB dictate that our total reward upon exercising the optimal subset of our purchased options at time $T + 1$ will have sublinear regret compared to a strategy with perfect predictions which can choose the optimal asset a priori to research in all rounds. That is, if we could have known the prices at $T + 1$ for all $k$ assets a priori and select a single asset to purchase options for only times when profit will be positive, using NoisyUCB.

The flexibility of noisy utility models to have their distributions be time and action dependent allows us to capture realistic dynamics about researching different assets, namely that some assets may be inherently more difficult to predict, and that the variance of our predictions may change as the time until the expiry date decreases. Some details about this setting are certainly artificial, but we believe there are reasonable arguments to be made for some of the assumptions. The ability to generate unbiased predictions about profit can be justified if we have access to sufficient historical data about option profitability and research contexts, assuming the macro-level market patterns informing the research process are unchanged.

However, to be able to meaningfully profit from this UCB approach, our predictions need to be less biased (or less variant) than those of the option seller, whose predictions will dictate the option price they offer. If we consider the only binary outcome as to whether an option purchase results in a profit or a loss, the means of each bandit arm (say, in $[-1, 1]$) represent the proportion. NoisyUCB will allow us to obtain $\tilde{O}(\sqrt{KT})$ regret, but this will nullify any ability to profit if the prices offered by the broker are properly calibrated by approximately unbiased predictors (i.e. if the mean for all arms is $\tilde{O}(\sqrt{KT/T})$). But if the predictions made by the options broker are sufficiently biased for any one asset such that some arm has mean $\omega(\sqrt{KT/T})$, NoisyUCB will be able to identify such an asset and reliably profit from its sequence of options purchases. We can’t get a free lunch in an unpredictable market via algorithmic tricks alone, but we can use them to exploit the mistakes of the broker if our predictions are better calibrated.

5 Regulating for Fairness

We consider the problem of imposing fairness constraints on an algorithm designer who is responsible for allocating resources to applicants from multiple populations. We view this from the perspective of a regulatory agency who wants to promote fair treatment of applicants and has the ability to impose costs on the algorithm designer if their allocations result in adverse discrimination, but the regulator is restricted in the information it has about how much it costs the algorithm designer to improve fairness.

Suppose a financial institution is responsible for deciding whether to grant loans to
members of two distinct groups, one of which constitutes the majority of the population. The institution will receive an application from an agent whose group membership is public, as well as a vector containing relevant personal data (such as their credit score, employment, payment history, etc.), and will grant the loan if they believe the agent is likely to repay based on an algorithmic classifier. Possibly due to historical patterns of discrimination or other characteristic differences between the groups, identical vectors from members of the two groups do not indicate the same information about their likelihood of repaying, and so using the same classifier for both groups will likely result in many inaccurate predictions. As a result, the institution will make use of a separate classifier for members of each group.

We view the classifiers as machine learning algorithms which the institution can make as accurate as they want with sufficient data and experimental optimization, but increasing accuracy is costly. The institution is only incentivized to maximize their profit, and will choose accuracy levels which maximize the difference between their expected profit and their cost of optimization. If the size or loanworthy proportion differs between groups, the accuracies chosen by the institution to optimize profit may differ substantially between the two groups.

5.1 Formalizing the Problem

- An Evaluator is responsible for allocating Resources (i.e. loans, jobs) to interested agents.
- At each time $t$, the Evaluator receives an application from an agent who belongs to either Class $\alpha$ or Class $\beta$. The class of the applicant is publicly observable, and the probability of the applicant belonging to $\alpha$ in any round is $p$.
- All agents have some real-valued type $d$, and are qualified to receive a Resource if $d > \tau$.
- Types in $\alpha$ are independently distributed according to $D_\alpha = N(\mu_\alpha, \eta_\alpha)$. Types in $\beta$ are independently distributed according to $D_\beta = N(\mu_\beta, \eta_\beta)$. (We use $\eta$ to indicate Gaussian variance, as to avoid confusion with a noise parameter $\sigma^2$. We will later consider relaxing the Gaussian assumption)
- The Evaluator does not see an agent’s true type, and instead observes a noisy signal of the type, e.g. $s_{\alpha,t} = d_{\alpha,t} + k_{\alpha,t}$, where $k_{\alpha,t} \sim N(0, \sigma^2_\alpha)$
- The Evaluator receives reward 1 for correctly determining whether $k > \tau$, and reward 0 for an incorrect determination.
- The Evaluator can select any positive values for $\sigma^2_\alpha$ and $\sigma^2_\beta$ they desire, but they must pay an upfront cost of $c(1/\sigma^2_\alpha) + c(1/\sigma^2_\beta)$, where $c(\cdot)$ is monotonically increasing. The values will be selected to maximize total expected profit over $T$ rounds.
- A Regulator has the goal of minimizing some notion of harm or unfairness for the applicants. The Evaluator cannot operate without approval from the Regulator, and the Regulator may impose arbitrary costs on the Evaluator, as long as the penalty scheme is determined and shared before the sequence of $T$ evaluations.
- The Regulator does not know $c(\cdot)$, but knows $D_\alpha$ and $D_\beta$, as well as the chosen values of $\sigma^2_\alpha$ and $\sigma^2_\beta$.

We can first observe that pure profit optimization can result in unfair outcomes between the groups with respect to parity of false positive and false negative rates. Even if $\mu_\alpha = \mu_\beta$ and $\eta_\alpha = \eta_\beta$, if $p \geq 0.5$, simple cost functions exist where the Evaluator will choose $\sigma^2_\alpha << \sigma^2_\beta$, resulting in higher rates of misclassification for the $\beta$ group.
For clarity, we provide the definitions of the false positive and negative rates for each group, as well as total error rates, which we will consider when designing penalty schemes for the Evaluator:

\[
\begin{align*}
FP_\alpha &= \Pr[\text{accepted} | \text{not qualified, in group } \alpha] \\
FP_\beta &= \Pr[\text{accepted} | \text{not qualified, in group } \beta] \\
FN_\alpha &= \Pr[\text{rejected} | \text{qualified, in group } \alpha] \\
FN_\beta &= \Pr[\text{rejected} | \text{qualified, in group } \beta] \\
Error_\alpha &= FN_\alpha \cdot \Pr[\text{qualified} | \text{in group } \alpha] + FP_\alpha \cdot \Pr[\text{not qualified} | \text{in group } \alpha] \\
Error_\beta &= FN_\beta \cdot \Pr[\text{qualified} | \text{in group } \beta] + FP_\beta \cdot \Pr[\text{not qualified} | \text{in group } \beta]
\end{align*}
\]

Because we have Gaussian distributions for both our population types and classification noise, the Regulator can directly calculate these values by considering the range of signals where the Evaluator will have above 50% confidence that an applicant has a type above \( \tau \).

5.2 Compensation for Discrimination

Our primary strategy will be to consider penalty schemes where the Regulator demands that the Evaluator compensate individuals from the group which is classified with less accuracy. Because the Regulator does not know \( c(\cdot) \), which may be arbitrary, it may be impossible to incentivize the regulator to choose equal error rates. For example, consider:

\[
c(1/\sigma^2) = \begin{cases} 
T/2 & 1/\sigma^2 \geq \epsilon \\
0 & 1/\sigma^2 < \epsilon
\end{cases}
\]

For any \( p > 0.5 \), it will be rational for the Evaluator to make \( \sigma_\alpha^2 \) arbitrarily small and \( \sigma_\beta^2 \) no smaller than \( 1/\epsilon \). Any penalty scheme designed to prompt the regulator to normalize error rates will result in a total cost greater than \( T \), and so it would be more profitable for the Evaluator to either use highly inaccurate classifiers (with no cost) or abstain from classification altogether. This is the case even if \( c(\cdot) \) is known to the Regulator. Even for more natural cost functions, especially if \( p \) is large or if we allow \( c(\cdot) \) to differ for the two groups, the Regulator may be unable to align the incentives of the Evaluator with even approximately achieving error parity.

However, we can make progress if we allow for other methods of providing agents utility other than correct classification. Namely, if the Evaluator is mandated to pay some subsidy to every agent from the group whose error rate is larger (\( \beta \) henceforth), we can avoid disparate harm to the group. We give such approaches corresponding to three different methods for evaluating applicant utility based on their qualification and classification. In all approaches, our goals of normalizing utility will occur regardless of the values of \( \sigma_\alpha^2 \) and \( \sigma_\beta^2 \) which the Evaluator chooses, and so we can be agnostic to the exact form of \( c(\cdot) \).

5.2.1 Normalizing Group Utility

First, we consider the scenario where the utility received by an applicant is identical to the utility received by the Evaluator upon classification. Here, the average utility received by a member of each group is simply equal to the group’s error rate. One desirable goal for the Regulator may be ensuring that the average applicant in each group obtains equal utility, as it may be impossible to gain much useful information about the true type of an applicant in \( \beta \) if \( \sigma_\beta^2 \) is large. In this case, the Regulator can mandate that the Evaluator pay \( Error_\beta - Error_\alpha \) to each applicant from \( \beta \), regardless of the classification result. This results in an expected utility of \( 1 - Error_\beta \) for applicants in both groups.
5.2.2 Normalizing False Negative Harms

While the Evaluator wishes to avoid giving out loans to unqualified applicants, the Regulator may be less concerned with this. Applicants who receive a loan that they are unqualified for may still be better off than if they had been rejected, but the Evaluator wishes to avoid this because of the associated risk of default. Here we assume that unqualified applicants receive a utility of 1 regardless of their application result, and the only agents receiving utility of 0 are qualified applicants who are rejected. To normalize to group utility, the Regulator can force the Evaluator to provide a payment of 

$$(FN_\beta \cdot \Pr[\text{qualified}|\text{in group } \beta] - FN_\alpha \cdot \Pr[\text{qualified}|\text{in group } \alpha])$$

to each agent in $\beta$.

5.2.3 Normalizing Utility with Group Qualification

Finally, our best-case scenario may be to wish that resources are distributed proportionally according to group qualification. We have seen that this may be impossible, but we can achieve a similar result through regulation. Here, we assume that all applicants obtain utility 0 except for qualified applicants who are classified positively, who receive utility 1. The expected utility of an applicant from $\beta$ will be $(1 - FN_\beta) \Pr[\text{qualified}|\text{in group } \beta]$. To equalize the expected utility of an applicant from $\beta$ with $\Pr[\text{qualified}|\text{in group } \beta]$, we can mandate a payment of $FN_\beta \cdot \Pr[\text{qualified}|\text{in group } \beta]$. Likewise, to equalize expected utility for $\alpha$ with $\Pr[\text{qualified}|\text{in group } \alpha]$, we can mandate a payment of $FN_\alpha \cdot \Pr[\text{qualified}|\text{in group } \alpha]$. This is identical to the previous case with an additional payment of $FN_\alpha \cdot \Pr[\text{qualified}|\text{in group } \alpha]$ to all agents. While the Evaluator would prefer to keep this additional profit, achieving the utility distributions of perfect classifiers in this case is impossible without compensating both groups, as the false negative rates for both will always be positive given our assumptions about $c(\cdot)$.

5.2.4 Expanding to More Groups

We may wish to consider scenarios where the number of groups is larger than 2. The first two normalization criteria can be generalized by pairwise comparisons to the group with lowest error when determining payment amounts. The third only involves computing each group’s payment individually, and thus can be generalized as well. A useful property of compensating for discrimination is that if a population is sufficiently small such that the Evaluator can never profit by improving classification accuracy, which may be more common when many groups are involved, we can always demand that they compensate those applicants and achieve the desired normalization guarantees.

5.2.5 Scaling Compensation with Qualification

The approach of awarding the same payment to all applicants from a class gives us guarantees about the utility of the entire class in expectation, but we may be able to maintain these guarantees while awarding proportionally larger subsidies to more qualified applicants. Based on the signal received for an applicant, the Regulator can also calculate their posterior probability of qualification, just as the Evaluator does. We can then demand that the Evaluator pay a subsidy which is not only a function of the group, but also an increasing function of the signal.

5.2.6 Individual Fairness

It is worth additionally considering individual notions of fairness, which dictate that more qualified individuals are always preferred probabilistically over less qualified individuals, which we can interpret here as that applicant’s expected utility. We satisfy this within groups with respect to applicants’ true types, and across groups with respect to the binary attribute of qualification. However, this is not necessarily satisfied with respect to applicants’
types across groups. If $\sigma^2_\alpha$ can be arbitrarily small while $\sigma^2_\beta$ is still quite large (as was the case with our example for $c(\cdot)$), we can essentially learn nothing about the qualification of an applicant in $\beta$. An applicant in $\alpha$ with true type only slightly above $\tau$ will have a higher expected utility than an applicant in $\beta$ with when no regulation is involved. If we mandate payments to correct for this, the expected utility of unqualified applicants in $\beta$ will approach that of their more qualified peers. If we are in the case where unqualified applicants receive utility of 1 from rejection, we can satisfy individual fairness. Individual fairness is impossible if those applicants receive utility 0, as the subsidy provided to a highly unqualified applicant in $\beta$ will raise their expected utility beyond that of an applicant in $\alpha$ whose type is just below $\tau$ and will be correctly classified with high probability.

5.3 Further Directions

• We currently only consider a one-shot training procedure followed by a sequence of predictions. In conjunction with this, we assume that once the classifier is trained, the Evaluator no longer has to pay a continual cost for using it. If the parameters of $D_\alpha$ and $D_\beta$ are unknown to either the Evaluator or the Regulator, a one-shot classifier resulting in a signal threshold for acceptance will be impractical. This could motivate an online learning approach, where the Evaluator maintains estimates of the parameters for each distribution. If the Evaluator pays a cost for each round it uses a classifier, this can be formalized as a bandit problem where they can explore a range of values and eventually obtain a no-regret optimum. If the algorithm pays a cost every time it improves its classifier and cannot recover the cost by later reducing accuracy, . In turn, the Regulator may only have access to data about the classification noise chosen by the Evaluator in previous rounds. Our goal would be to construct algorithmic techniques for navigating these tradeoffs, arriving at convergence to an approximation of any of the aforementioned subsidy schemes.

• Here we primarily consider qualification as a binary feature, where applicants’ true types are either above or below $\tau$. We may wish to consider more complicated utility functions for both the Evaluator and for applicants. This would allow us to view the signal as a measure of likelihood that an agent will default on a loan, and the true outcome of whether they default will dictate utility.

• We may want to relax the Gaussian assumptions about both the type distributions and the classifier noise. As long as the distributions are known, the Evaluator can still determine the posterior probability, albeit with more involved calculations, allowing the regulator to impose payment requirements satisfying our normalization guarantees. This becomes more challenging in combination with other considered relaxations of the problem.

• Recent work by Liu et al. [11] has considered a similar problem involving fairness in allocation of resources such as loans for two groups, with a focus on the long-term implications of unfairness in classification. If agents in a group are misclassified at higher rates and subsequently default on their loans, their credit scores will drop, possibly harming their group’s perceived qualification, feeding into a cycle which perpetuates unfairness. We may consider the impact of regulation schemes in this context to determine whether compensation schemes can provide positive long-term results.

References


