Fair Performance Metric Elicitation

Gaurush Hiranandani
UIUC
gaurush2@illinois.edu

Harikrishna Narasimhan
Google Research
hnarasimhan@google.com

Oluwasanmi Koyejo
UIUC & Google Research Accra
sanmi@illinois.edu

October 9, 2020

Abstract

What is a fair performance metric? We consider the choice of fairness metrics through the lens of metric elicitation – a principled framework for selecting performance metrics that best reflect implicit preferences. The use of metric elicitation enables a practitioner to tune the performance and fairness metrics to the task, context, and population at hand. Specifically, we propose a novel strategy to elicit fair performance metrics for multiclass classification problems with multiple sensitive groups that also includes selecting the trade-off between performance and fairness. The proposed elicitation strategy requires only relative preference feedback and is robust to both finite sample and feedback noise.

1 Introduction

Machine learning models are increasingly employed for critical decision-making tasks such as hiring and sentencing [38, 3, 9, 12, 26]. Yet, it is increasingly evident that automated decision-making is susceptible to bias, whereby decisions made by the algorithm are unfair to certain subgroups [5, 3, 8, 7, 26]. To this end, a wide variety of group fairness metrics have been proposed – all to reduce discrimination and bias from automated decision-making [21, 11, 14, 24, 41, 27]. However, a dearth of formal principles for selecting the most appropriate metric has highlighted the confusion of experts, practitioners, and end users in deciding which group fairness metric to employ [45]. This is further exacerbated by the observation that common metrics often lead to contradictory outcomes [24].

While the problem of selecting an appropriate fairness metric has gained prominence in recent years [14, 27, 45], it perhaps best understood as a special case of the task of choosing evaluation metrics in machine learning. For instance, when a cost-sensitive predictive model classifies patients into cancer categories [42] even without considering fairness, it is often unclear how the cost-tradeoffs be chosen so that they reflect the expert's decision-making, i.e., replacing

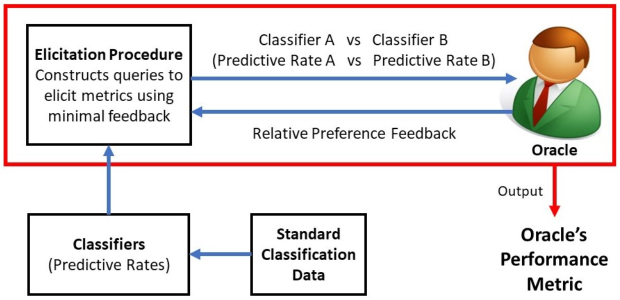

Figure 1: Framework of Metric Elicitation [16].

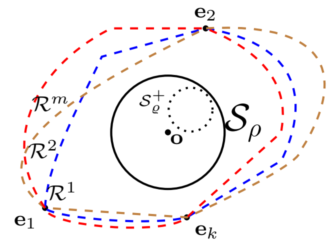

Figure 2: $\mathcal{R}^1 \times \cdots \times \mathcal{R}^m$ (best seen in colors); $\mathcal{R}^u \, \forall \, u \in [m]$ are convex sets with common vertices $\mathbf{e}_i \, \forall \, i \in [k]$ and enclose the sphere $\mathcal{S}_\rho$.

expert intuition by quantifiable metrics. The recently proposed Metric Elicitation (ME) framework [16, 17] provides a solution. ME is a principled framework for eliciting performance metrics using only relative preference feedback over classifiers from an end user (oracle). The motivation behind ME is that employing the performance metrics which reflect user tradeoffs will enable learning models that best capture user preferences [16]. Figure 1 (reproduced from [16]) illustrates the ME framework.

Existing research suggests a fundamental trade-off between algorithmic fairness and performance [21, 43, 9, 6, 27, 45], where in addition to appropriate metrics, the practitioner or policymaker must choose a trade-off operating point between the competing objectives [45]. To this end, we extend the ME framework from eliciting multiclass classification metrics [17] to the task of eliciting *fair* performance metrics from pairwise preference feedback in the presence of multiple sensitive groups. In particular, we elicit metrics that reflect, jointly, the (i) predictive performance evaluated as a weighting of classifier's overall predictive rates, (ii) fairness violation assessed as the discrepancy in predictive rates among groups, and (iii) a trade-off between the predictive performance and fairness violation. Importantly, the elicited metrics are sufficiently flexible to encapsulate and generalize many existing predictive performance and fairness violation measures.

In eliciting group-fair performance metrics, we tackle three challenges as compared to existing work on ME. First, from preference query perspective, the predictive performance and fairness violations are correlated, thus increasing the complexity of joint elicitation. Second, we find that in order to measure both positive and negative violations, the fair metrics are necessarily non-linear functions of the predictive rates, thus existing results on linear ME [17] cannot be applied directly. Finally, as we show, the number of groups directly impacts query complexity. We overcome these challenges by proposing a novel query efficient procedure that exploits the geometric properties of the set of rates.

Contributions. We consider metrics for algorithmically group-fair classification, and propose a novel approach for eliciting predictive performance, fairness violations, and their trade-off point, from expert pairwise feedback. Our procedure uses binary-search based subroutines and recovers the metric with linear query complexity. Moreover, the procedure is robust to both finite sample and oracle feedback noise thus is useful in practice. Lastly, our method can be applied either by querying preferences over classifiers or rates. Such an equivalence is crucial for practical applications [16, 17].

Notation. Matrices and vectors are denoted by bold upper case and bold lower case letters, respectively. We denote the inner product of two vectors by $\langle \cdot, \cdot \rangle$ and the Hadamard product by $\odot$. The ℓ_∞-norm and ℓ_2-norm are denoted by $\|\cdot\|_\infty$ and $\|\cdot\|_2$, respectively. For $k \in \mathbb{Z}_+$, we represent the index set $\{1, 2, \cdots, k\}$ by $[k]$, and the $(k-1)$-dimensional simplex by Δ_k. Given a matrix $\mathbf{A}$, *off-diag*$(\mathbf{A})$ returns a vector of off-diagonal elements of $\mathbf{A}$ in row-major form. The group membership is denoted by superscripts and coordinates of vectors, matrices, and tuples are denoted by subscripts.

2 Background

The standard multiclass, multigroup classification setting comprises k classes and m groups with $X \in \mathcal{X}$, $G \in [m]$ and $Y \in [k]$ representing the input, group membership, and output random variables, respectively. The groups are assumed to be disjoint. We have access to a dataset of size n denoted by $\{(\mathbf{x}, g, y)_i\}_{i=1}^n$, generated *iid* from a distribution $\mathbb{P}(X, G, Y)$. For each group $g \in [m]$, let $\zeta_i^g := \mathbb{P}(Y = i | G = g)$ for $i \in [k]$ be the group-conditional probability of each of the k classes, and $\kappa^g := \mathbb{P}(G = g)$ denote the probability that any instance belongs to the group g. The unconditional probability of class i is given by $\zeta_i := \mathbb{P}(Y = i)$, where clearly $\zeta_i = \sum_g \kappa^g \zeta_i^g$.

Predictive rates: We consider a (randomized) classifier $h^g : \mathcal{X} \to \Delta_k$ for each group g, and use $\mathcal{H}^g = \{h^g : \mathcal{X} \to \Delta_k\}$ to denote the set of all classifiers for group g. The group-conditional rate matrix for a classifier h^g is denoted by $\mathbf{R}^g(h^g, \mathbb{P}) \in \mathbb{R}^{k \times k}$, where its elements are defined as:

$$R_{ij}^g(h^g, \mathbb{P}) := \mathbb{P}(h^g = j | Y = i, G = g) \quad \text{for } i, j \in [k]. \tag{1}$$

We also define the overall classifier $h : (\mathcal{X}, [m]) \to \Delta_k$ by $h(\mathbf{x}, g) := h^g(\mathbf{x})$ and represent the set of all overall classifiers by $\mathcal{H}$. The rate matrix for the overall classifier h is then given by:

$$R_{ij} := \mathbb{P}(h = j | Y = i) = \sum\nolimits_{g=1}^m T_{ij}^g R_{ij}^g, \tag{2}$$

where, for a group g, we have defined a group contribution matrix $\mathbf{T}^g$ as $T_{ij}^g := \kappa^g \zeta_i^g / \zeta_i \; \forall \, i, j \in [k]$. Notice that the predictive rates satisfy the following useful decomposition:

$$R_{ii}^g(h^g, \mathbb{P}) = 1 - \sum\nolimits_{j=1, j \neq i}^k R_{ij}^g(h^g, \mathbb{P}). \tag{3}$$

Using this any rate matrix is uniquely represented by its $q := (k^2 - k)$ off-diagonal elements as a vector $\mathbf{r}^g(h^g, \mathbb{P}) =$ *off-diag*$(\mathbf{R}^g(h^g, \mathbb{P}))$. So we will interchangeably refer to the rate matrix as a *'vector of rates'*. The set of rates associated with a group g is denoted by $\mathcal{R}^g = \{\mathbf{r}^g(h^g, \mathbb{P}) : h^g \in \mathcal{H}^g\}$. For clarity, we will suppress the dependence on $\mathbb{P}$ and h^g if it is clear from the context.

Lastly, let $\mathbf{r}^{1:m} := (\mathbf{r}^1, \ldots, \mathbf{r}^m) \in \mathcal{R}^1 \times \cdots \times \mathcal{R}^m =: \mathcal{R}^{1:m}$ denote the tuple of rates for the m groups for the overall classifier h, then its rates from (2) can now be succinctly written as $\mathbf{r} = \sum_{g=1}^m \boldsymbol{\tau}^g \odot \mathbf{r}^g$, where $\boldsymbol{\tau}^g :=$ *off-diag*$(\mathbf{T}^g)$. Note that $\sum_{g=1}^m \boldsymbol{\tau}^g = \mathbf{1}$, where $\mathbf{1}$ is a vector of ones.

Fairness violation measure: The (approximate) fairness of a classifier is often determined by the 'discrepancy' in rates across different groups e.g. *equalized odds* [14, 4]. So given two groups $u, v \in [m]$, we define the discrepancy in their rates as:

$$\mathbf{d}^{uv} := |\mathbf{r}^u - \mathbf{r}^v|. \tag{4}$$

Since there are m groups, the number of *discrepancy vectors* are $M := \binom{m}{2}$.

2.1 Fair Performance Metric

We aim to elicit a general class of metrics, which recovers and generalizes existing fairness measures, based on trade-off between predictive performance and fairness violation [21, 14, 8, 6, 27]. Let $\phi : [0,1]^q \to \mathbb{R}$ be the cost of overall misclassification (aka. predictive performance) and $\varphi : [0,1]^{m\times q} \to \mathbb{R}$ be the fairness violation cost for a classifier h determined by the overall rates $\mathbf{r}(h)$ and group discrepancies $\{\mathbf{d}^{uv}(h)\}_{u,v=1,v>u}^{m}$, respectively. Without loss of generality (wlog), we assume the metrics ϕ and φ are costs. Moreover, the metrics are scale invariant as global scale does not affect the learning problem [31]; hence let $\phi : [0,1]^q \to [0,1]$ and $\varphi : [0,1]^{m\times q} \to [0,1]$.

Definition 1. *Fair Performance Metric: Let ϕ and φ be monotonically increasing linear functions of overall rates and group discrepancies, respectively. The fair metric Ψ is a trade-off between ϕ and φ. In particular, given $\mathbf{a} \in \mathbb{R}^q, \mathbf{a} \geq 0$ (misclassification weights), a set of vectors $\mathbf{B} \coloneqq \{\mathbf{b}^{uv} \in \mathbb{R}^q, \mathbf{b}^{uv} \geq 0\}_{u,v=1,v>u}^{m}$ (fairness violation weights), and a scalar λ (trade-off) with*

$$\|\mathbf{a}\|_2 = 1, \qquad \sum\nolimits_{u,v=1,v>u}^{m} \|\mathbf{b}^{uv}\|_2 = 1, \qquad 0 \leq \lambda \leq 1, \tag{5}$$

(wlog., due to scale invariance), we define the metric Ψ as:

$$\Psi(\mathbf{r}^{1:m}\,;\,\mathbf{a},\mathbf{B},\lambda) \coloneqq \underbrace{(1-\lambda)}_{\textit{trade-off}}\underbrace{\langle \mathbf{a},\mathbf{r}\rangle}_{\phi(\mathbf{r})} + \lambda \underbrace{\left(\sum\nolimits_{u,v=1,v>u}^{m} \langle \mathbf{b}^{uv}, \mathbf{d}^{uv}\rangle\right)}_{\varphi(\mathbf{r}^{1:m})}. \tag{6}$$

Examples of the misclassification cost $\phi(\mathbf{r})$ include cost-sensitive linear metrics [1]. Many existing fairness metrics for two classes and two groups such as *equal opportunity* [14], *balance for the negative class* [24] *error-rate balance* (i.e., $0.5|r_1^1 - r_1^2| + 0.5|r_2^1 - r_2^2|$) [8], *weighted equalized odds* (i.e., $b_1|r_1^1 - r_1^2| + b_2|r_2^1 - r_2^2|$) [14, 6], etc. correspond to fairness violations of the form $\varphi(\mathbf{r}^{1:m})$ considered above. The combination of $\phi(\mathbf{r})$ and $\varphi(\mathbf{r}^{1:m})$ as defined in $\Psi(\mathbf{r}^{1:m})$ appears regularly in prior work [21, 6, 27]. Notice that the metric is flexible to allow different fairness violation costs for different pairs of groups thus capable of enabling reverse discrimination [33]. Lastly, while the metric is linear with respect to (wrt.) the discrepancies, it is non-linear wrt. the group-wise rates. Hence, standard linear ME algorithm [17] cannot be trivially applied for eliciting the metric in Definition 1.

2.2 Fair Performance Metric Elicitation; Problem Statement

We now state the problem of *Fair Performance Metric Elicitation (FPME)* and define the associated *oracle query*. The broad definitions follow from Hiranandani et al. [16, 17], extended so the rates and the performance metrics correspond to the multiclass multigroup-fair classification setting.

Definition 2 (Oracle Query)**.** *Given two classifiers h_1, h_2 (equivalent to a tuple of rates $\mathbf{r}_1^{1:m}, \mathbf{r}_2^{1:m}$ respectively), a query to the Oracle (with metric Ψ) is represented by:*

$$\Gamma(h_1,h_2\,;\,\Psi) = \Omega\left(\mathbf{r}_1^{1:m},\mathbf{r}_2^{1:m}\,;\,\Psi\right) = \mathbb{1}[\Psi(\mathbf{r}_1^{1:m}) > \Psi(\mathbf{r}_2^{1:m})], \tag{7}$$

where $\Gamma : \mathcal{H}\times\mathcal{H} \to \{0,1\}$ and $\Omega : \mathcal{R}^{1:m}\times\mathcal{R}^{1:m} \to \{0,1\}$. In words, the query asks whether h_1 is preferred to h_2 (equivalent to whether $\mathbf{r}_1^{1:m}$ is preferred to $\mathbf{r}_2^{1:m}$), as measured by Ψ.

Notice that the oracle responds by comparing the arguments of Ψ, which is unknown to us and can be accessed only through queries to the oracle $\Omega(\cdot, \cdot\,;\,\Psi)$. The metrics we consider are functions of rates, thus comparison queries using classifiers are indistinguishable from comparison queries using rates. Henceforth, we denote any query as a rate-based query. Next, we formally state the FPME problem.

Definition 3 (Fair Performance Metric Elicitation with Pairwise Queries (given $\{(\mathbf{x}, g, y)_i\}_{i=1}^n$))**.** *Suppose that the oracle's (unknown) performance metric is Ψ. Using oracle queries of the form $\Omega(\hat{\mathbf{r}}_1^{1:m}, \hat{\mathbf{r}}_2^{1:m}\,;\,\Psi)$, where $\hat{\mathbf{r}}_1^{1:m}, \hat{\mathbf{r}}_2^{1:m}$ are the estimated rates from samples, recover a metric $\hat{\Psi}$ such that $\|\Psi - \hat{\Psi}\| < \omega$ under a suitable norm $\|\cdot\|$ for sufficiently small error tolerance $\omega > 0$.*

Similar to the standard metric elicitation problems [16, 17], the performance of FPME is evaluated both by the fidelity of the recovered metric and the query complexity. As done in decision theory literature [25, 16], we present our FPME solution by first assuming access to population quantities such as the population rates $\mathbf{r}^{1:m}(h, \mathbb{P})$, and then discuss how elicitation can be performed from finite samples, e.g., with empirical rates $\hat{\mathbf{r}}^{1:m}(h, \{(\mathbf{x}, g, y)_i\}_{i=1}^n)$.

2.3 Standard Linear Performance Metric Elicitation – Warmup

We revisit the linear metric elicitation procedure proposed in [17]. The procedure assumes an enclosed sphere $\mathcal{S} \subset \mathcal{Z}$, where $\mathcal{Z}$ is the q-dimensional space of classifier statistics that are feasible, i.e., can be achieved by some classifier. We will use this linear metric elicitation as a subroutine when we consider the more complicated setting defined in Definition 1. Let the oracle's scale invariant, monotonically increasing metric be $\xi(\mathbf{z}) \coloneqq \langle \mathbf{a}, \mathbf{z} \rangle$, such that $\|\mathbf{a}\|_2 = 1, \mathbf{a} \geq 0$, analogous to the misclassification cost in Definition 1. Analogously, the oracle queries are $\Omega\left(\mathbf{z}_1, \mathbf{z}_2\,;\,\xi\right) \coloneqq \mathbb{1}[\xi(\mathbf{z}_1) > \xi(\mathbf{z}_2)]$.

Algorithm 2 in [17] is a coordinate-wise binary search algorithm that takes the query space $\mathcal{S}$, binary-search tolerance ϵ, and the oracle $\Omega(\cdot, \cdot\,;\,\xi)$ as input, and by querying $O(q \log(\pi/2\epsilon))$ queries recovers a metric $\hat{\mathbf{a}}$ with $\|\hat{\mathbf{a}}\|_2 = 1, \hat{\mathbf{a}} \geq 0$ such that $\|\mathbf{a} - \hat{\mathbf{a}}\|_2 \leq O(\sqrt{q}\epsilon)$ (Theorem 2 in [17]). In other words, the algorithm estimates the direction of the gradient $\nabla\xi$, i.e., the slope but not the magnitude. Notice that Algorithm 2 in [17] considers query arguments lying only in one orthant of the sphere $\mathcal{S}$ since it is assumed $\mathbf{a} \geq 0$. However, it is easy to extend the algorithm so that any scale invariant linear metric $\mathbf{a}$ can be elicited, i.e., without the monotonicity condition. This is achieved by asking q additional queries to determine the search orthant, thus this does not affect the query complexity. We refer to this modified algorithm as Standard Linear Performance Metric Elicitation (SLPME) (see Appendix A for details) and summarize the discussion with the following remark.

Remark 1. *Given a q-dimensional space $\mathcal{Z}$ enclosing a sphere $\mathcal{S} \subset \mathcal{Z}$, a metric $\xi(\mathbf{z}) \coloneqq \langle \mathbf{a}, \mathbf{z} \rangle$, and an oracle $\Omega(\cdot, \cdot; \xi)$, the SLPME algorithm (Algorithm 2, Appendix A) provides an estimate $\hat{\mathbf{a}}$ with $\|\hat{\mathbf{a}}\|_2 = 1$ such that the estimated slope is close to the true slope, i.e., ${}^{a_i}/_{a_j} \approx {}^{\hat{a}_i}/_{\hat{a}_j}\ \forall\, i, j \in [q]$.*

Next, we show the existence of a feasible sphere in $\mathcal{R}^g$ so that SLPME can be used to elicit linear rate metrics. We then show how SLPME is used as a subroutine to elicit the fair metric in Definition 1.

3 Geometry of the product set $\mathcal{R}^{1:m}$

Let $\mathbf{e}_i \in \{0,1\}^q$ for $i \in [k]$ be the rates achieved by trivial classifiers, i.e., classifiers predicting only class i on the entire space $\mathcal{X}$. Notice that the rates $\mathbf{e}_i$'s are identical for all groups.

Proposition 1 (Geometry of $\mathcal{R}^{1:m}$; Figure 2)**.** *For any group $g \in [m]$, the set of confusion rates $\mathcal{R}^g$ is convex, bounded in $[0,1]^q$, and has vertices $\{\mathbf{e}_i\}_{i=1}^k$. The intersection of group rate sets $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ is convex and always contains the non-vertex rate $\mathbf{o} = \frac{1}{k}\sum_{i=1}^k \mathbf{e}_i$ associated with the trivial classifier predicting each class with equal probability on the entire space $\mathcal{X}$.*

Since $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ is convex and always contains a non-vertex point $\mathbf{o}$ in the interior, we may make the following assumption (see Figure 2 for an illustration).

Assumption 1. *$\exists$ a q-dimensional sphere $\mathcal{S}_\rho \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ of radius $\rho > 0$ centered at $\mathbf{o}$.*

Such a sphere always exists as long as the class-conditional distributions are not identical, i.e., there is some signal for non-trivial classification conditioned on each group [17]. A method to obtain $\mathcal{S}_\rho$ from [17] is discussed in Appendix B.1. We end this section by outlining a trivial yet useful remark.

Remark 2. *Since the sphere $\mathcal{S}_\rho \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$, any rate $\mathbf{s} \in \mathcal{S}_\rho$ is feasible for all groups, i.e., $\mathbf{s}$ is achievable by some classifier conditioned on each group $g \in [m]$. Moreover, if two groups $u, v \in [m]$ achieve the same rate $\mathbf{s} \in \mathcal{S}_\rho$, then the associated discrepancy vector $\mathbf{d}^{uv} = 0$.*

4 Metric Elicitation

Suppose that the oracle's fair performance metric is $\overline{\Psi}$ parametrized by $(\bar{\mathbf{a}}, \overline{\mathbf{B}}, \bar{\lambda})$ as in Definition 1. We now propose an efficient FPME procedure to elicit the oracle's metric. The overall FPME framework is presented in Figure 3 and is summarized in Algorithm 1. As shown in Figure 3, the FPME procedure has three parts which it follows in sequence: (a) eliciting the misclassification cost $\bar{\phi}(\mathbf{r})$ (i.e., $\bar{\mathbf{a}}$), (b) eliciting the fairness violation $\bar{\varphi}(\mathbf{r}^{1:m})$ (i.e., $\overline{\mathbf{B}}$), and (c) eliciting the trade-off between the misclassification cost and fairness violation (i.e., $\bar{\lambda}$). The framework exploits the sphere $\mathcal{S}_\rho \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ and uses the SLPME procedure (Algorithm 2, Appendix A) and another binary-search based Algorithm 4 (Appendix C.3) as subroutines. We will discuss each step in detail and provide proof of correctness as we progress. We first discuss FPME with no *feedback* noise from the oracle and later show robustness to noisy feedback and query complexity guarantees in Section 5.

4.1 Eliciting the Misclassification Cost $\bar{\phi}(\mathbf{r})$; Part 1 in Figure 3

The response to a query combines the effect of both misclassification cost and fairness violation. Thus, the key to eliciting the misclassification cost $\bar{\phi}$ is to query rates for which the fairness violation is zero, so the oracle's responses are unaffected by the fairness term. Specifically, consider a parametrization $\nu : \mathcal{S}_\rho \to \mathcal{R}^{1:m}$:

$$\nu(\mathbf{s}) \coloneqq (\mathbf{s}, \mathbf{s}, \ldots, \mathbf{s}), \tag{8}$$

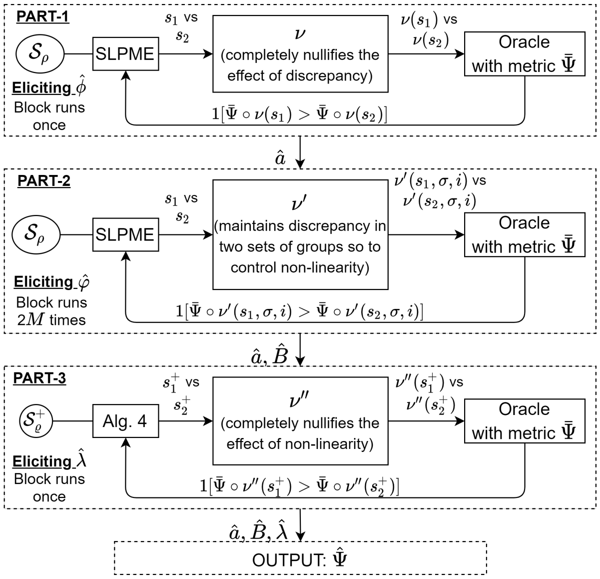

Figure 3: Workflow of the FPME procedure.

Algorithm 1: FPM Elicitation

Input: Query spaces $\mathcal{S}_\rho$, $\mathcal{S}_\varrho^+$, parametrizations ν, ν', ν'', search tolerance $\epsilon > 0$, oracle $\Omega(\cdot,\cdot;\overline{\Psi})$ with metric $\overline{\Psi}$

1: $\hat{\mathbf{a}} \leftarrow \text{SLPME}(\mathcal{S}_\rho, \epsilon, \Omega(\cdot,\cdot\,;\,\overline{\Psi}\circ\nu(\cdot))$
2: **If** $m == 2$
3: $\breve{\mathbf{f}} \leftarrow \text{SLPME}(\mathcal{S}_\rho, \epsilon, \Omega(\cdot,\cdot\,;\,\overline{\Psi}\circ\nu'(\cdot,1))$
4: $\tilde{\mathbf{f}} \leftarrow \text{SLPME}(\mathcal{S}_\rho, \epsilon, \Omega(\cdot,\cdot\,;\,\overline{\Psi}\circ\nu'(\cdot,k))$
5: $\hat{\mathbf{b}}^{12} \leftarrow$ normalized solution from (12)
6: **Else** Let $\mathcal{L} \leftarrow \varnothing$
7: **For** $\sigma \in \mathcal{M}$ **do**
8: $\breve{\mathbf{f}}^\sigma \leftarrow \text{SLPME}(\mathcal{S}_\rho, \epsilon, \Omega(\cdot,\cdot\,;\,\overline{\Psi}\circ\nu'(\cdot,\sigma,1))$
9: $\tilde{\mathbf{f}}^\sigma \leftarrow \text{SLPME}(\mathcal{S}_\rho, \epsilon, \Omega(\cdot,\cdot\,;\,\overline{\Psi}\circ\nu'(\cdot,\sigma,k))$
10: Let ℓ^σ be Eq. (14), extend $\mathcal{L} \leftarrow \mathcal{L} \cup \{\ell^\sigma\}$
11: $\hat{\mathbf{B}} \leftarrow$ normalized solution from (15) using $\mathcal{L}$
12: $\hat{\lambda} \leftarrow$ Algorithm 4 $(\mathcal{S}_\varrho^+, \epsilon, \Omega(\cdot,\cdot\,;\,\overline{\Psi}\circ\nu''(\cdot))$

Output: $\hat{\mathbf{a}}, \hat{\mathbf{B}}, \hat{\lambda}$

i.e., ν assigns all group rates to be $\mathbf{s} \in \mathcal{S}_\rho$. For any element $\nu(\mathbf{s}) \in \mathcal{R}^{1:m}$, the associated discrepancy terms $\mathbf{d}^{uv} = 0$ for all $u, v \in [m]$ (Remark 2). Thus, for query elements constrained to lie in the image of ν, the metric in Definition 1 reduces to a linear metric in $\mathbf{s}$, i.e.:

$$\overline{\Psi}(\nu(\mathbf{s})\,;\,\bar{\mathbf{a}}, \overline{\mathbf{B}}, \bar{\lambda}) = (1-\bar{\lambda})\langle\bar{\mathbf{a}}, \mathbf{s}\rangle.$$

Observe that $\Omega(\nu(\mathbf{s}_1), \nu(\mathbf{s}_2)\,;\,\overline{\Psi}) \equiv \Omega(\mathbf{s}_1, \mathbf{s}_2\,;\,\overline{\Psi}\circ\nu)$; hence, when the query space of the original oracle with metric $\overline{\Psi}$ is restricted to the image of the parametrization ν, it is equivalent to an oracle with metric $\overline{\Psi}\circ\nu$ for the SLPME procedure. Thus line 1 in Algorithm 1 applies the SLPME subroutine with query space $\mathcal{S}_\rho$, binary search tolerance ϵ, and the equivalent oracle $\Omega(\cdot,\cdot\,;\,\overline{\Psi}\circ\nu)$. From Remark 1, this subroutine returns a slope $\mathbf{f}$ with $\|\mathbf{f}\|_2 = 1$ such that:

$$\frac{(1-\bar{\lambda})\bar{a}_i}{(1-\bar{\lambda})\bar{a}_j} = \frac{f_i}{f_j} \implies \frac{\bar{a}_i}{\bar{a}_j} = \frac{f_i}{f_j}. \tag{9}$$

Thus, we set $\hat{\mathbf{a}} := \mathbf{f}$ (line 1, Algorithm 1). See part 1 in Figure 3 for further illustration. Notice that this procedure recovers $\hat{\mathbf{a}}$ independent of the trade-off $\bar{\lambda}$ and fairness violation $\bar{\varphi}$.

4.2 Eliciting the Fairness Violation $\overline{\varphi}(\mathbf{r}^{1:m})$; Part 2 in Figure 3

For the ease of understanding, we divide this subsection into two parts. We first discuss eliciting $\overline{\varphi}(\mathbf{r}^{1:m})$ in the special case of $m = 2$ groups and then extend the proposed procedure for $m > 2$.

4.2.1 Eliciting the Fairness Violation $\overline{\varphi}(\mathbf{r}^{1:m})$ for $m = 2$; lines 2-5 in Algorithm 1

Note that the fairness violation in Definition 1 is non-linear in the rates. So for eliciting the fairness violation weights $\bar{\mathbf{b}}^{12}$, we selectively query rates containing non-zero discrepancy between the two groups such that the non-linearity in the metric is controllable in the oracle's responses. In particular, recall from Section 3 that the rates $\mathbf{e}_i \in \{0,1\}^q \,\forall\, i \in [k]$ associated with trivial classifiers are feasible for all groups $g \in [m]$. Since $\mathbf{e}_i$'s are binary vectors, fixing trivial classifiers for one group and non-trivial for the other reveals the sign of the absolute function in $\overline{\varphi}$. In turn, we are able to linearize the metric. Specifically, consider a parametrization $\nu' : (\mathcal{S}_\rho, [k]) \to \mathcal{R}^{1:2}$ defined as:

$$\nu'(\mathbf{s}, i) \coloneqq (\mathbf{s}, \mathbf{e}_i), \tag{10}$$

i.e., ν' assigns rates $\mathbf{s} \in \mathcal{S}_\rho$ to group 1 and trivial rates $\mathbf{e}_i$ to group 2 (associated with fixing trivial classifier $h^2(\mathbf{x}) = i \,\forall\, \mathbf{x} \in \mathcal{X}$). Since the rates $\mathbf{e}_i$ are known binary vectors and coordinates of s are between 0 and 1, the sign of the absolute function in (6) wrt. each coordinate of $\mathbf{s}$ can be recovered. Thus, for elements constrained to lie in the image of $\nu'(\cdot, i)$, the metric in Definition 1 reduces to:

$$\overline{\Psi}(\nu'(\mathbf{s}, i)\,;\,\bar{\mathbf{a}}, \overline{\mathbf{b}}^{12}, \bar{\lambda}) = \langle (1-\bar{\lambda})\bar{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^2) + \bar{\lambda}\mathbf{w}_i \odot \overline{\mathbf{b}}^{12}, \mathbf{s}\rangle + c_i, \tag{11}$$

where $\mathbf{w}_i \coloneqq 1 - 2\mathbf{e}_i$ and c_i is a constant not affecting the responses. This is again a linear metric in $\mathbf{s}$. Similar to Section 4.1, notice that $\Omega(\nu'(\mathbf{s}_1, i), \nu'(\mathbf{s}_2, i)\,;\,\overline{\Psi}) \equiv \Omega(\mathbf{s}_1, \mathbf{s}_2\,;\,\overline{\Psi} \circ \nu'(\cdot, i))$. Hence, when the query space of the original oracle with metric $\overline{\Psi}$ is restricted to lie in the image of $\nu'(\cdot, i)$, it is equivalent to an oracle with metric $\overline{\Psi} \circ \nu'(\cdot, i)$. Thus we may again employ the SLPME subroutine with the query space $\mathcal{S}_\rho$, binary search tolerance ϵ, and the equivalent oracle $\Omega(\cdot, \cdot\,;\,\overline{\Psi} \circ \nu'(\cdot, i))$.

However, unlike Section 4.1, we must run the SLPME procedure twice to get q independent equations to elicit the q dimensional vector $\overline{\mathbf{b}}^{12}$. This is achieved by running SLPME after fixing $h^2(\mathbf{x}) = 1$ (line 3, Algorithm 1) and then again running SLPME after fixing $h^2(\mathbf{x}) = k$ (line 4, Algorithm 1) for group 2. This results in two slopes $\breve{\mathbf{f}}, \tilde{\mathbf{f}}$ such that $\|\breve{\mathbf{f}}\|_2 = \|\tilde{\mathbf{f}}\|_2 = 1$ from which it is easy to obtain the fairness violation weights:

$$\hat{\mathbf{b}}^{12} = \frac{\tilde{\mathbf{b}}^{12}}{\|\tilde{\mathbf{b}}^{12}\|_2}, \quad \text{where} \quad \tilde{\mathbf{b}}^{12} = \mathbf{w}_1 \odot \left[\delta\breve{\mathbf{f}} - \hat{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^2)\right], \tag{12}$$

where δ is a scalar depending on the known entities $\boldsymbol{\tau}^{12}, \hat{\mathbf{a}}, \breve{\mathbf{f}}^{12}, \tilde{\mathbf{f}}^{12}$. The derivation is provided in Appendix C.2.1. Notice that, due to the scale invariance in $\overline{\varphi}$ (Definition 1), the normalized solution $\hat{\mathbf{b}}^{12}$ is independent of the true trade-off $\bar{\lambda}$ but depends on the elicited vector $\hat{\mathbf{a}}$ from Section 4.1.

4.2.2 Eliciting the Fairness Violation $\overline{\varphi}(\mathbf{r}^{1:m})$ for $m > 2$; line 6-11 in Algorithm 1

We briefly outline the elicitation procedure for $m > 2$ groups, with details in Appendix C.2.2. Let $\mathcal{M} \subset 2^{[m]} \setminus \{\varnothing, [m]\}$ be a non-empty set of sets of groups defined by m. As in the $m = 2$ case, we assign trivial rates to one set of groups $\sigma \in \mathcal{M}$ and rates $\mathbf{s} \in \mathcal{S}_\rho$ to the remaining groups in $[m] \setminus \sigma$. We then follow a similar process by first defining a parametrization $\nu' : (\mathcal{S}_\rho, \mathcal{M}, [k]) \to \mathcal{R}^{1:m}$ as:

$$\nu'(\mathbf{s}, \sigma, i) \coloneqq \mathbf{r}^{1:m} \quad \text{such that} \quad \mathbf{r}^g = \begin{cases} \mathbf{e}_i & \text{if } g \in \sigma \\ \mathbf{s} & \text{o.w.} \end{cases} \tag{13}$$

and then using the SLPME subroutine with the equivalent oracles $\Omega(\cdot, \cdot\,;\,\overline{\Psi} \circ \nu'(\cdot, \sigma, 1))$ and $\Omega(\cdot, \cdot\,;\,\overline{\Psi} \circ \nu'(\cdot, \sigma, k))$ in line 8 and line 9 of Algorithm 1, respectively. See part 2 in Figure 3 for illustration. Thus, for a partitioning of groups defined by $\sigma \in \mathcal{M}$, we get the following generalized version of (12):

$$\sum\nolimits_{u,v} \mathbb{1}\left[|\{u,v\} \cap \sigma| = 1\right] \tilde{\mathbf{b}}^{uv} = \mathbf{w}_1 \odot \left[\delta^\sigma \breve{\mathbf{f}}^\sigma - \hat{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^\sigma)\right], \tag{14}$$

where $\boldsymbol{\tau}^\sigma = \sum_{g\in\sigma} \boldsymbol{\tau}^g$ and $\tilde{\mathbf{b}}^{uv} := \overline{\lambda}\overline{\mathbf{b}}^{uv}/(1-\overline{\lambda})$ is a scaled version of the true (unknown) $\overline{\mathbf{b}}$, which nonetheless can be computed from (14). Since we want to elicit M fairness violation weight vectors, we require M ways of partitioning the groups into two sets so that we construct M independent equations similar to (14). Let $\mathcal{M}$ be those set of sets. $\mathcal{M}$ provides the required system of equations $\mathcal{L}$ in line 10 of Algorithm 1. From the solution of system of equations $\mathcal{L}$ (equation (35), Appendix C.2.2), we recover $\{\tilde{\mathbf{b}}^{uv}\}_{u,v=1,v>u}^m$, which when normalized returns the final fairness violation weights

$$\hat{\mathbf{b}}^{uv} = \frac{\tilde{\mathbf{b}}^{uv}}{\sum_{u,v=1,v>u}^m \|\tilde{\mathbf{b}}^{uv}\|_2} \quad \text{for} \quad u,v \in [m], v > u. \tag{15}$$

Due to the normalization, the elicited fairness violation weights are independent of the trade-off $\overline{\lambda}$.

4.3 Eliciting Trade-off $\overline{\lambda}$; Part 3 in Figure 3

The final step is to elicit the trade-off $\overline{\lambda}$ between the misclassification cost $\overline{\phi}$ and fairness violation $\overline{\varphi}$, given $\hat{\mathbf{a}}$ and $\hat{\mathbf{B}}$. Here, the key insight is to completely remove the non-linearity posed by the absolute function in Definition 1, which then reduces the problem to a one-dimensional binary search. To this end, we generate queries from a q-dimensional sphere $\mathcal{S}_\varrho^+ \subset \mathcal{S}_\rho$ where $\varrho < \rho$, as shown in Figure 2. Clearly, $\mathbf{s}^+ \geq \mathbf{o}$, $\forall\, \mathbf{s}^+ \in \mathcal{S}_\varrho^+$. Specifically, we construct a parametrization $\nu'' : \mathcal{S}_\varrho^+ \to \mathcal{R}^{1:m}$:

$$\nu''(\mathbf{s}^+) := (\mathbf{s}^+, \mathbf{o}, \ldots, \mathbf{o}), \tag{16}$$

where ν'' assigns $\mathbf{s}^+ \in \mathcal{S}_\varrho^+$ to the first group and assigns trivial rates $\mathbf{o}$ to the rest of the groups. For any element $\nu''(\mathbf{s}^+) \in \mathcal{R}^{1:m}$, the associated discrepancy terms $\mathbf{d}^{uv} = 0$ for $u, v \neq 1$ (Remark 2). For the remaining discrepancy terms, the sign of the absolute function in (6) is known since $\mathbf{s}^+ \geq \mathbf{o}$. Thus for elements in the image of ν'', the metric in Definition 1 reduces to a linear metric in $\mathbf{s}^+$, i.e:

$$\overline{\Psi}(\nu''(\mathbf{s}^+)\,;\, \overline{\mathbf{a}}, \overline{\mathbf{B}}, \overline{\lambda}) = \langle (1-\overline{\lambda})\boldsymbol{\tau}^1 \odot \overline{\mathbf{a}} + \overline{\lambda} \sum\nolimits_{v=2}^m \overline{\mathbf{b}}^{1v}, \mathbf{s}^+ \rangle + c, \tag{17}$$

where the constant c does not affect the oracle responses. The following lemma shows that the metric in (17) is quasiconcave in trade-off $\overline{\lambda}$ and is the basis of the proposed subroutine to elicit $\overline{\lambda}$.

Lemma 1. *Under the regularity assumption that* $\langle \mathbf{a}, \sum_{v=2}^m \mathbf{b}^{1v} \rangle \neq 1$, *the function*

$$\vartheta(\lambda) := \max_{\mathbf{s}^+ \in \mathcal{S}_\varrho^+} \Psi(\nu''(\mathbf{s}^+)\,;\, \mathbf{a}, \mathbf{B}, \lambda) \tag{18}$$

is strictly quasiconcave (and therefore unimodal) in λ.

The unimodality of $\vartheta(\lambda)$ allows us to perform the one-dimensional binary search in Algorithm 4 (Appendix C.3) using the query space $\mathcal{S}_\varrho^+$, tolerance ϵ, and the equivalent oracle $\Omega(\cdot\,;\, \overline{\Psi} \circ \nu''(\cdot))$. See Part 3 in Figure 3 for illustration. Algorithm 1 calls Algorithm 4 as a subroutine in line 12 to recover $\hat{\lambda}$. Combining part 1, part 2, and part 3 of the workflow in Figure 3 completes the FPME procedure.

5 Guarantees

We discuss elicitation guarantees under the following feedback model, which is useful in practice.

Definition 4 (Oracle Feedback Noise: $\epsilon_\Omega \geq 0$)**.** *For two rates* $\mathbf{r}_1^{1:m}, \mathbf{r}_2^{1:m} \in \mathcal{R}^{1:m}$*, the oracle responds correctly as long as* $|\overline{\Psi}(\mathbf{r}_1^{1:m}) - \overline{\Psi}(\mathbf{r}_2^{1:m})| > \epsilon_\Omega$*. Otherwise, it may be incorrect.*

In words, the oracle may respond incorrectly if the rates are very close as measured by the metric $\overline{\Psi}$. Since deriving the final metric involves offline computations including certain ratios, we discuss guarantees under a regularity assumption that ensures all components are well defined.

Assumption 2. *We assume that* $1 > c_1 > \overline{\lambda} > c_2 > 0$*,* $\|\bar{\mathbf{a}}\|_\infty > c_3$*,* $\|(1-\overline{\lambda})\bar{\mathbf{a}} \odot \boldsymbol{\tau}^\sigma - \overline{\lambda}\mathbf{w}_i \odot \overline{\mathbf{b}}^\sigma\|_\infty > c_4 \, \forall \, i \in [q], \sigma \in \mathcal{M}$*, for some* $c_1, c_2, c_3, c_4 > 0$*,* $\rho > \varrho \gg \epsilon_\Omega$*, and* $\langle \bar{\mathbf{a}}, \sum_{v=2}^m \overline{\mathbf{b}}^{1v} \rangle \neq 1$*.*

Theorem 1. *Given* $\epsilon, \epsilon_\Omega \geq 0$*, and a 1-Lipschitz fair performance metric* $\overline{\Psi}$ *parametrized by* $\bar{\mathbf{a}}, \overline{\mathbf{B}}, \overline{\lambda}$*, under Assumption 2, Algorithm 1 returns a metric* $\hat{\Psi}$ *with parameters:*

- $\hat{\mathbf{a}}$: *after* $O\left(q \log \frac{\pi}{2\epsilon}\right)$ *queries such that* $\|\bar{\mathbf{a}} - \hat{\mathbf{a}}\|_2 \leq O\left(\sqrt{q}(\epsilon + \sqrt{\epsilon_\Omega/\rho})\right)$.
- $\hat{\mathbf{B}}$: *after* $O\left(Mq \log \frac{\pi}{2\epsilon}\right)$ *queries such that* $\|vec(\overline{\mathbf{B}}) - vec(\hat{\mathbf{B}})\|_2 \leq O\left(mq(\epsilon + \sqrt{\epsilon_\Omega/\rho})\right)$*, where* $vec(\cdot)$ *vectorizes the matrix.*
- $\hat{\lambda}$: *after* $O(\log(\frac{1}{\epsilon}))$ *queries, with error* $|\overline{\lambda} - \hat{\lambda}| \leq O\left(\epsilon + \sqrt{\epsilon_\Omega/\varrho} + \sqrt{mq(\epsilon + \sqrt{\epsilon_\Omega/\rho})/\varrho}\right)$.

We see that the proposed FPME procedure is robust to noise, and its query complexity depends linearly in the number of unknown entities. For instance, line 1 in Algorithm 1 elicits $\hat{\mathbf{a}} \in \mathbf{R}^q$ by posing $\tilde{O}(q)$ queries, the 'for' loop in line 7 of Algorithm 1 runs for M iterations, where each iteration requires $\tilde{O}(2q)$ queries, and finally line 12 in Algorithm 1 is a simple binary search requiring $\tilde{O}(1)$ queries. Previous work suggests that standard linear multiclass elicitation (SLPME) elicits misclassification costs (ϕ) with linear query complexity [17]. Surprisingly, our proposed FPME procedure elicits a more complex (nonlinear) metric without increasing the query complexity order. Furthermore, since sample estimates of rates are consistent estimators, and the metrics discussed are 1-Lipschitz wrt. rates, with high probability, we gather correct oracle feedback from querying with finite sample estimates $\Omega(\hat{\mathbf{r}}_1^{1:m}, \hat{\mathbf{r}}_2^{1:m} \,;\, \Psi)$ instead of querying with population statistics $\Omega(\mathbf{r}_1^{1:m}, \mathbf{r}_2^{1:m} \,;\, \Psi)$, as long as we have sufficient samples. Apart from this, Algorithm 1 is agnostic to finite sample errors as long as the sphere $\mathcal{S}_\rho$ is contained within the feasible region $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$.

6 Experiments

We first empirically validate the FPME procedure and recovery guarantees in Section 6.1 and then highlight the utility of FPME in evaluating real-world classifiers in Section 6.2 .

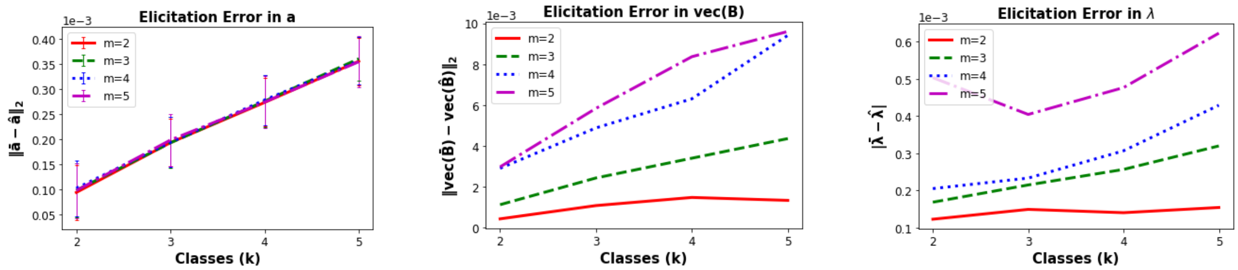

Figure 4: Elicitation error in recovering the oracle's metric.

6.1 Recovering the Oracle's Fair Performance Metric

Recall that a sphere $\mathcal{S}_\rho \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ as long as there is a non-trivial classification signal within each group (Assumption 1). Thus for experiments, we assume access to a feasible sphere $\mathcal{S}_\rho$ with $\rho = 0.2$. We randomly generate 100 oracle metrics each for $k, m \in \{2, 3, 4, 5\}$ parametrized by $\{\bar{\mathbf{a}}, \overline{\mathbf{B}}, \bar{\lambda}\}$. This specifies the query outputs by the oracle for each metric in Algorithm 1. We then use Algorithm 1 with tolerance $\epsilon = 10^{-3}$ to elicit corresponding metrics parametrized by $\{\hat{\mathbf{a}}, \hat{\mathbf{B}}, \hat{\lambda}\}$. Algorithm 1 makes $1 + 2M$ subroutine calls to SLPME procedure and 1 call to Algorithm 4. SLPME subroutine requires exactly $16(q-1)\log(\pi/2\epsilon)$ queries, where we use 4 queries to shrink the interval in the binary search loop and fix 4 cycles for the coordinate-wise search. Also, Algorithm 4 requires $4\log(1/\epsilon)$ queries.

In Figure 4, we report the mean of the ℓ_2-norm between the oracle's metric and the elicited metric. Clearly, we elicit metrics that are very close to the true metrics. Moreover, this holds true across a range of m and k values indicating robustness of the proposed approach. Figure 4(a) shows that the error $\|\bar{\mathbf{a}} - \hat{\mathbf{a}}\|_2$ increases only with the number of classes k and not groups m. This is expected since $\hat{\mathbf{a}}$ is elicited by querying rates that nullify fairness violation (Section 4.1). Figure 5(a) verifies Theorem 1 by showing that $\|\text{vec}(\overline{\mathbf{B}}) - \text{vec}(\hat{\mathbf{B}})\|_2$ increases with both number of classes k and groups m. In accord with Theorem 1, Figure 5(b) shows that the elicited trade-off $\hat{\lambda}$ is also close to the true $\bar{\lambda}$. However, the elicitation error increases consistently with groups m but not with classes k. A possible reason may be the cancellation of errors from eliciting $\hat{\mathbf{a}}$ and $\hat{\mathbf{B}}$ separately.

6.2 Ranking of Classifiers for Real-world Datasets

One of the most important applications of performance metrics is evaluating classifiers, i.e., providing a quantitative score for their quality which then allows us to choose the best (or best set of) classifier(s). In this section, we discuss how the ranking of plausible classifiers is affected when a practitioner employs default metrics to rank (fair) classifiers instead of the oracle's metric or our elicited approximation.

We take four real-world classification datasets with $k, m \in \{2, 3\}$ (see Table 1). 60% of each dataset is used for training and the rest for testing. We create a pool of 100 classifiers for each dataset by tweaking hyperparameters under logistic regression models [23], multi-layer perceptron models [34], support vector machines [19], LightGBM models [22], and fairness constrained optimization based models [30]. We compute the group wise confusion rates on the test data for each model for each dataset. We will compare the ranking of these classifiers

Table 1: Dataset statistics; the real-valued regressor in *wine* and *crime* datasets is binned to three classes.

Dataset	k	m	**#samples**	**#features**	**group.feat**
default	2	2	30000	33	gender
adult	2	3	43156	74	race
wine	3	2	6497	13	color
crime	3	3	1907	99	race

Table 2: Common (baseline) metrics usually deployed to rank classifiers.

Name $\rightarrow$	$\hat{\phi}\hat{\varphi}\hat{\lambda}$_a	$\hat{\phi}\hat{\varphi}\hat{\lambda}$_w	$\hat{\phi}\hat{\varphi}$_a	$\hat{\phi}\hat{\varphi}$_w	$\hat{\phi}$_a	$\hat{\phi}$_w	o_p	o_f
$\hat{\mathbf{a}}$	acc.	w-acc.	acc.	w-acc.	acc.	w-acc.	$\bar{\mathbf{a}}$	-
$\hat{\mathbf{B}}$	acc.	w-acc.	acc.	w-acc.	elicit	elicit	-	$\bar{\mathbf{B}}$
$\hat{\lambda}$	0.5	w-acc.	elicit	elicit	elicit	elicit	0	1

achieved by competing baseline metrics with respect to the ground truth ranking.

We generate 100 random oracle metrics $\overline{\Psi}$. $\overline{\Psi}$'s gives us the ground truth ranking of the above classifiers. We then use our proposed procedure FPME (Algorithm 1) to recover the oracle's metric. For comparison in ranking of real-world classifiers, we choose a few metrics that are routinely employed by practitioners as baselines (see Table 2). The prefixes (i.e. $\hat{\phi}, \hat{\varphi}$, or $\hat{\lambda}$) in name of the baseline metrics denote the components that are set to default metrics, and the suffixes (i.e. 'a' or 'wa') denote whether the assignment is done with *accuracy* (i.e. equal weights) or with *weighted accuracy* (weights are assigned randomly however maintaining the true order of weights as in $\overline{\Psi}$). For example, $\hat{\phi}\hat{\varphi}\hat{\lambda}$_a corresponds to the metric where $\hat{\phi}, \hat{\varphi}, \hat{\lambda}$ are set to standard classification accuracy. Similarly, $\hat{\phi}$_w denote a metric where the misclassification cost $\hat{\phi}$ is set to weighted accuracy but both $\hat{\varphi}$ and $\hat{\lambda}$ are elicited using Part 2 and Part 3 of the FPME procedure (Algorithm 1), respectively. Assigning weighted accuracy versions is a commonplace since sometimes the order of the costs associated with the types of mistakes in misclassification cost $\overline{\phi}$ or fairness violation $\overline{\varphi}$ or preference for fairness violation over misclassification $\overline{\lambda}$ is known but not the actual cost. Another example is $\hat{\phi}\hat{\varphi}$_a which corresponds to the metric where $\hat{\phi}, \hat{\varphi}$ are set to accuracy and only the trade-off $\hat{\lambda}$ is elicited using Part 3 of the FPME procedure (Algorithm 1). This is similar to prior work by Zhang et al. [45] who assumed the classification error and fairness violation known, so only the trade-off has to be elicited – however they also assume direct ratio queries, which can be challenging in practice. Our approach applies much simnpler pairwise preference queries. Lastly, o_p and o_f represent *only predictive performance* with $\lambda = 0$ and *only fairness* with $\lambda = 1$, respectively.

Figure 5 shows average NDCG (with exponential gain) [40] and Kendall-tau coefficient [37] over 100 metrics $\overline{\Psi}$ and their respective estimates by the competing baseline metrics. We see that FPME, wherein we elicit $\hat{\phi}, \hat{\varphi}$, and $\hat{\lambda}$ in sequence, achieves the highest possible NDCG and Kendall-tau coefficient. Even though we make some elicitation error in recovery (Section 8), we achieve almost perfect results while ranking the classifiers. To connect to practice, this implies that when given a set of classifiers, ranking based on elicited metrics will align most closely to ranking based on the true metric, as compared to ranking classifiers based on default metrics. This is a crucial advantage of metric elicitation for practical purposes. In this experiment, baseline metrics achieve inferior ranking of classifiers in comparison to the rankings achieved

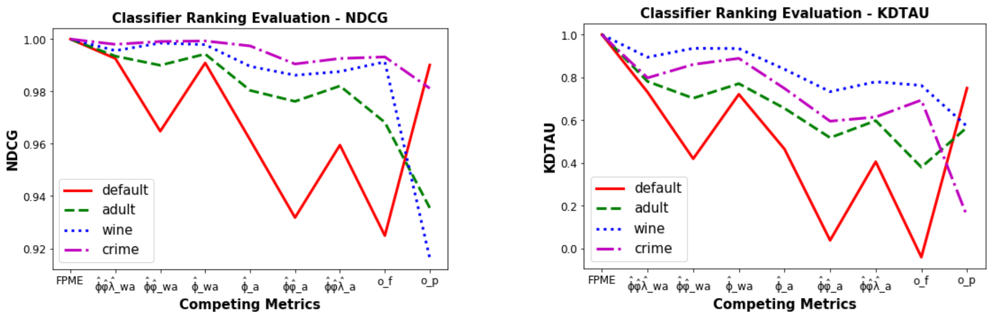

Figure 5: Ranking Performance of competing metrics while ranking real-world classifiers.

by metrics that are elicited using the proposed FPME procedure. Figure 5 also suggests that it is beneficial to elicit all three components $(\bar{\mathbf{a}}, \bar{\mathbf{B}}, \bar{\lambda})$ of the metric in Definition 1, rather than pre-define a component and elicit the rest. For the *crime* dataset, some methods also achieve high NDCG values, so ranking at the top is good; however Kendall-tau coefficient is weak which suggests that overall ranking is poor. With the exception of the *default* dataset, the weighted versions are better than equally weighted versions in ranking. This is expected because in weighted versions, at least order of the preference for the type of costs matches with the oracle's preferences.

7 Related Work

Some early attempts to eliciting individual fairness metrics [18, 28] are distinct from ours – as we are focused on the more prevalent setting of group fairness, yet for which there are no existing approaches to our knowledge. Zhang et al. [45] propose an approach that elicits only the trade-off between accuracy and fairness using complicated ratio queries. We, on the other hand, elicit classification cost, fairness violation, and the trade-off together as a non-linear function, all using much simpler pairwise comparison queries. Prior work for constrained classification focus on learning classifiers under constraints for fairness [13, 14, 44, 29]. We take the regularization view of algorithmic fairness, where a fairness violation is embedded in the metric definition instead of as constraints [21, 6, 9, 2, 27]. From the elicitation perspective, the closest line of work to ours is Hiranandani et al. [16, 17], who propose the problem of ME but solve it only for a simpler setting of classification without fairness. As we move to multiclass, multigroup fair performance ME, we find that the complexity of both the form of the metrics and the query space increases. This results in starkly different elicitation strategy with novel methods required to provide query complexity guarantees. Learning (linear) functions passively using pairwise comparisons is a mature field [20, 15, 35], but these approaches fail to control sample (i.e. query) complexity. Active learning in fairness [32] is a related direction; however the aim there is to learn a fair classifier based on fixed metric instead of eliciting the metric itself.

8 Discussion Points and Future Work

- **Transportability:** Our elicitation procedure is independent of the population $\mathbb{P}$ as long as there exists a sphere of rates which is feasible for all groups. Thus, any metric that is learned using one dataset or model class (i.e. by estimated $\hat{\mathbb{P}}$) can be applied to other applications and datasets, as long as the expert believes the context and tradeoffs are the same.

- **Extensions.** The FPME procedure can be modified to leverage the structure in the metric or groups to decrease the query complexity, e.g., when the fairness violation weights are the same for all pairs of groups, the procedure in Section C.2.2 requires only one partitioning of groups to elicit the metric $\hat{\varphi}$. Such modifications are easy to incorporate. In the future, we plan to extend our approach to more complex metrics such as linear-fractional functions of predictive rates and discrepancies.
- **Practicality.** Like standard ME tasks [16, 17], FPME can be applied by posing classifier comparisons directly via A/B testing [39] or via interpretable learning techniques [36, 10]. We also aim to build novel visualizations similar to [45] for comparing classifiers or rates for real user studies.
- **Optimal bounds.** We conjecture that our query complexity bounds are tight; however, we leave this detail for the future. In conclusion, we elicit a more complex (non-linear) group fair-metric with the same query complexity order as standard classification linear elicitation procedures [17].

9 Conclusion

We study the space of multiclass, multigroup predicitve rates and propose a novel, provably query efficient strategy to elicit group-fair performance metrics. The proposed procedure only requires pairwise preference feedback over classifiers and and is robust to finite sample and feedback noise.

References

[1] N. Abe, B. Zadrozny, and J. Langford. An iterative method for multi-class cost-sensitive learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 3–11. ACM, 2004.

[2] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias risk assessments in criminal sentencing. *ProPublica, May*, 23, 2016.

[4] S. Barocas, M. Hardt, and A. Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[5] S. Barocas and A. D. Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[6] Y. Bechavod and K. Ligett. Learning fair classifiers: A regularization-inspired approach. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2017.

[7] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

[8] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

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

[10] F. Doshi-Velez and B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. *ArXiv e-prints:1702.08608*, 2017.

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

[12] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338, 2019.

[13] G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

[14] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[15] R. Herbrich. Large margin rank boundaries for ordinal regression. In *Advances in large margin classifiers*, pages 115–132. The MIT Press, 2000.

[16] G. Hiranandani, S. Boodaghians, R. Mehta, and O. Koyejo. Performance metric elicitation from pairwise classifier comparisons. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 371–379, 2019.

[17] G. Hiranandani, S. Boodaghians, R. Mehta, and O. O. Koyejo. Multiclass performance metric elicitation. In *Advances in Neural Information Processing Systems*, pages 9351–9360, 2019.

[18] C. Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

[19] T. Joachims. Svmlight: Support vector machine. *SVM-Light Support Vector Machine http://svmlight. joachims. org/, University of Dortmund*, 19(4), 1999.

[20] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.

[21] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2012.

[22] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems*, pages 3146–3154, 2017.

[23] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, and M. Klein. *Logistic regression*. Springer, 2002.

[24] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[25] O. O. Koyejo, N. Natarajan, P. K. Ravikumar, and I. S. Dhillon. Consistent multilabel classification. In *NIPS*, pages 3321–3329, 2015.

[26] P. Lahoti, K. P. Gummadi, and G. Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019.

[27] A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[28] D. Mukherjee, M. Yurochkin, M. Banerjee, and Y. Sun. Two simple ways to learn individual fairness metric from data. In *ICML*, 2020.

[29] H. Narasimhan. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 1646–1654, 2018.

[30] H. Narasimhan, A. Cotter, and M. Gupta. Optimizing generalized rate metrics with three players. In *Advances in Neural Information Processing Systems*, pages 10746–10757, 2019.

[31] H. Narasimhan, H. Ramaswamy, A. Saha, and S. Agarwal. Consistent multiclass algorithms for complex performance measures. In *ICML*, pages 2398–2407, 2015.

[32] A. Noriega-Campero, M. A. Bakker, B. Garcia-Bulle, and A. Pentland. Active fairness in algorithmic decision making. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 77–83, 2019.

[33] S. Opotow. Affirmative action, fairness, and the scope of justice. *Journal of Social Issues*, 52(4):19–24, 1996.

[34] S. K. Pal and S. Mitra. Multilayer perceptron, fuzzy sets, classifiaction. 1992.

[35] M. Peyrard, T. Botschen, and I. Gurevych. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84, 2017.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *ACM SIGKDD*, pages 1135–1144. ACM, 2016.

[37] G. S. Shieh. A weighted kendall's tau statistic. *Statistics & probability letters*, 39(1):17–24, 1998.

[38] A. Singla, E. Horvitz, P. Kohli, and A. Krause. Learning to hire teams. In *Third AAAI Conference on Human Computation and Crowdsourcing*, 2015.

[39] G. Tamburrelli and A. Margara. Towards automated *A/B* testing. In *International Symposium on Search Based Software Engineering*, pages 184–198. Springer, 2014.

[40] H. Valizadegan, R. Jin, R. Zhang, and J. Mao. Learning to rank by optimizing ndcg measure. In *Advances in neural information processing systems*, pages 1883–1891, 2009.

[41] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953, 2017.

[42] S. Yang and D. Q. Naiman. Multiclass cancer classification based on gene expression comparison. *Statistical applications in genetics and molecular biology*, 13(4):477–496, 2014.

[43] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

[44] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017.

[45] Y. Zhang, R. Bellamy, and K. Varshney. Joint optimization of ai fairness and utility: A human-centered approach. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 400–406, 2020.

Appendices

A Standard Linear Performance Metric Elicitation

As explained in Section 2.3, we use the linear metric elicitation procedure [17] as a subroutine in order to elicit a more complicated metric as defined in Definition 1. For completeness, we provide the details here.

The linear metric elicitation procedure proposed in [17] assumes an enclosed sphere $\mathcal{S} \subset \mathcal{Z}$, where $\mathcal{Z}$ is the q-dimensional space of classifier statistics that are feasible, i.e., can be achieved by some classifier. Let the the radius of the sphere $\mathcal{S}$ be ρ. We extend the linear metric elicitation procedure (Algorithm 2 in [17]) to elicit any linear metric (without the monotonicity condition) defined over the space $\mathcal{Z}$. This is because in Section 4.2, we require to elicit slopes that are not necessarily for monotonic metrics (e.g., see Equation (11)). Let the oracle's scale invariant metric be $\xi(\mathbf{z}) \coloneqq \langle \mathbf{a}, \mathbf{z} \rangle$, such that $\|\mathbf{a}\|_2 = 1$. Analogously, the oracle queries are $\Omega\left(\mathbf{z}_1, \mathbf{z}_2\,;\,\xi\right) \coloneqq \mathbb{1}[\xi(\mathbf{z}_1) > \xi(\mathbf{z}_2)]$. We start by outlining a trivial Lemma from [17].

Lemma 2. *[17] Let ξ be a linear metric parametrized by $\mathbf{a}$ such that $\|\mathbf{a}\|_2 = 1$, then the unique optimal classifier statistic $\bar{\mathbf{z}}$ over the sphere $\mathcal{S}$ is a point on the boundary of $\mathcal{S}$ given by $\bar{\mathbf{z}} = \rho\mathbf{a} + \mathbf{o}$, where $\mathbf{o}$ is the center of the sphere $\mathcal{S}$.*

Given a linear performance metric, Lemma 2 provides a unique point in the query space which lies on the boundary of the sphere $\partial\mathcal{S}$. Moreover, the converse also holds true that given a point on the boundary of the sphere $\partial\mathcal{S}$, one may recover the linear metric for which the given point is optimal. Thus, in order to elicit a linear metric, Hiranandani et al. [17] essentially search for the optimal statistic (over the surface of the sphere) using pairwise queries to the oracle which in turn reveals the true metric. The algorithm is summarized in Algorithm 2. The algorithm also uses the following standard paramterization for the surface of the sphere $\partial\mathcal{S}$.

Parameterizing the boundary of the enclosed sphere $\partial\mathcal{S}$. Let $\boldsymbol{\theta}$ be a $(q-1)$-dimensional vector of angles, where all the angles except the primary angle are in $[0, \pi]$, and the primary angle is in $[0, 2\pi]$. A linear performance metric with $\|\mathbf{a}\|_2 = 1$ is constructed by setting $a_i = \Pi_{j=1}^{i-1} \sin\theta_j \cos\theta_i$ for $i \in [q-1]$ and $a_q = \Pi_{j=1}^{q-1} \sin\theta_j$. By using Lemma 2, the metric's optimal classifier statistic over the sphere $\mathcal{S}$ is easy to compute. Thus, varying $\boldsymbol{\theta}$ in this procedure, parametrizes the surface of the sphere $\partial\mathcal{S}$. We denote this parametrization by $\mu(\boldsymbol{\theta})$, where $\mu : [0, \pi]^{q-2} \times [0, 2\pi] \rightarrow \partial\mathcal{S}$.

Description of Algorithm 2:[1] Suppose that the oracle's linear metric is ξ parametrized by $\mathbf{a}$ where $\|\mathbf{a}\|_2 = 1$ (Section 2.3). Using the parametrization $\mu(\boldsymbol{\theta})$ of the surface of the sphere $\partial\mathcal{S}$ as explained above, Algorithm 2 returns an estimate $\hat{\mathbf{a}}$ with $\|\hat{\mathbf{a}}\|_2 = 1$. Line 2-6 in Algorithm 2 recovers the orthant of the optimal statistic over the sphere by posing q trivial queries. Once the search orthant of the optimal statistic is fixed, the procedure is same as Algorithm 2 of [17]. In each iteration of the for loop, the algorithm updates one angle θ_j keeping other angles fixed by a binary-search procedure, where the *ShrinkInterval* subroutine (illustrated in Figure 6) shrinks the interval $[\theta_j^a, \theta_j^b]$ by half based on the responses. Then the algorithm

[1] The superscripts in Algorithm 2 denote iterates. Please do not confuse it with the sensitive group index.

Algorithm 2 Standard Linear Performance Metric Elicitation

Input: Query space $\mathcal{S}$, binary-search tolerance $\epsilon > 0$, oracle $\Omega(\cdot, \cdot\,; \xi)$ with metric ξ

for $i = 1, 2, \cdots q$ **do**
  Set $\mathbf{a} = \mathbf{a}' = (1/\sqrt{q}, \ldots, 1/\sqrt{q})$.
  Set $a'_i = -1/\sqrt{q}$.
  Compute the optimal $\bar{z}^{(\mathbf{a})}$ and $\bar{z}^{(\mathbf{a}')}$ over the sphere $\mathcal{S}$ using Lemma 2
  Query $\Omega(\bar{\mathbf{z}}^{(\mathbf{a})}, \bar{\mathbf{z}}^{(\mathbf{a}')}; \xi)$
{Fix the search orthant based on the above oracle responses}

Initialize: $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}$ {$\boldsymbol{\theta}^{(1)}$ is any point in the search orthant.}
for $t = 1, 2, \cdots, T = 4(q-1)$ **do**
  Set $\boldsymbol{\theta}^{(a)} = \boldsymbol{\theta}^{(c)} = \boldsymbol{\theta}^{(d)} = \boldsymbol{\theta}^{(e)} = \boldsymbol{\theta}^{(b)} = \boldsymbol{\theta}^{(t)}$.
  while $\left|\theta_j^{(b)} - \theta_j^{(a)}\right| > \epsilon$ **do**
    Set $\theta_j^{(c)} = \frac{3\theta_j^{(a)} + \theta_j^{(b)}}{4}$, $\theta_j^{(d)} = \frac{\theta_j^{(a)} + \theta_j^{(b)}}{2}$, and $\theta_j^{(e)} = \frac{\theta_j^{(a)} + 3\theta_j^{(b)}}{4}$.
    Set $\bar{\mathbf{z}}^{(a)} = \mu(\boldsymbol{\theta}^{(a)})$ (i.e. parametrization of $\partial\mathcal{S}$). Similarly, set $\bar{\mathbf{z}}^{(c)}, \bar{\mathbf{z}}^{(d)}, \bar{\mathbf{z}}^{(e)}, \bar{\mathbf{z}}^{(b)}$
    Query $\Omega(\bar{\mathbf{z}}^{(c)}, \bar{\mathbf{z}}^{(a)}\,; \xi), \Omega(\bar{\mathbf{z}}^{(d)}, \bar{\mathbf{z}}^{(c)}\,; \xi)$, $\Omega(\bar{\mathbf{z}}^{(e)}, \bar{\mathbf{z}}^{(d)}\,; \xi), \Omega(\bar{\mathbf{z}}^{(b)}, \bar{\mathbf{z}}^{(e)}\,; \xi)$.
    $[\theta_j^{(a)}, \theta_j^{(b)}] \leftarrow$ *ShrinkInterval* (responses) {see Figure 6}
  Set $\theta_j^{(d)} = \frac{1}{2}(\theta_j^{(a)} + \theta_j^{(b)})$
  Set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(d)}$.
Output: $\hat{a}_i = \Pi_{j=1}^{i-1} \sin\theta_j^{(T)} \cos\theta_i{}^{(T)}\, \forall i \in [q-1],\ \hat{a}_q = \Pi_{j=1}^{q-1} \sin\theta_j^{(T)}$

cyclically updates each angle until it converges to a metric sufficiently close to the true metric. The number of cycles in coordinate-wise search is fixed to four.

B Proofs and Details of Section 3

Proof of Proposition 1. The set of rates $\mathcal{R}^g$ for a group g satisfies the following properties:

- *Convex*: Let us take two classifiers $h_1^g, h_2^g \in \mathcal{H}^g$ which achieve the rates $\mathbf{r}_1^g, \mathbf{r}_2^g \in \mathcal{R}^g$. We need to check whether or not the convex combination $\alpha\mathbf{r}_1^g + (1-\alpha)\mathbf{r}_2^g$ is feasible, i.e., there exists some classifier which achieve this rate. Consider a classifier h^g, which with probability α predicts what classifier h_1^g predicts and with probability $1-\alpha$ predicts what classifier h_2^g predicts. Then the elements of the rate matrix $R_{ij}^g(h)$ is given by:

$$\begin{aligned}
R_{ij}^g(h) &= \mathbb{P}(h^g = j | Y = i) \\
&= \mathbb{P}(h_1^g = j | h^g = h_1^g, Y = i)\mathbb{P}(h^g = h_1^g) + \mathbb{P}(h_2^g = j | h^g = h_2^g, Y = i)\mathbb{P}(h^g = h_2^g) \\
&= \alpha\mathbf{r}_1^g + (1-\alpha)\mathbf{r}_2^g.
\end{aligned}$$

  Therefore, $\mathcal{R}^g\ \forall\, g \in [m]$ is convex.

- *Bounded:* Since $R_{ij}^g(h) = P[h = j | Y = i] = P[h = j, Y = i]/P[Y = i] \leq 1$ for all $i, j \in [k]$, $\mathcal{R}^g \subseteq [0,1]^q$.

Subroutine *ShrinkInterval*

Input: Oracle responses for $\Omega(\bar{\mathbf{z}}^{(c)}, \bar{\mathbf{z}}^{(a)}\,;\,\xi)$, $\Omega(\bar{\mathbf{z}}^{(d)}, \bar{\mathbf{z}}^{(c)}\,;\,\xi)$, $\Omega(\bar{\mathbf{z}}^{(e)}, \bar{\mathbf{z}}^{(d)}\,;\,\xi), \Omega(\bar{\mathbf{z}}^{(b)}, \bar{\mathbf{z}}^{(e)}\,;\,\xi)$

If $(\bar{\mathbf{z}}^{(a)} \succ \bar{\mathbf{z}}^{(c)})$ Set $\theta_j^{(b)} = \theta_j^{(d)}$.

elseif $(\bar{\mathbf{z}}^{(a)} \prec \bar{\mathbf{z}}^{(c)} \succ \bar{\mathbf{z}}^{(d)})$ Set $\theta_j^{(b)} = \theta_j^{(d)}$.

elseif $(\bar{\mathbf{z}}^{(c)} \prec \bar{\mathbf{z}}^{(d)} \succ \bar{\mathbf{z}}^{(e)})$ Set $\theta_j^{(a)} = \theta_j^{(c)}$, $\theta_j^{(b)} = \theta_j^{(e)}$.

elseif $(\bar{\mathbf{z}}^{(d)} \prec \bar{\mathbf{z}}^{(e)} \succ \bar{\mathbf{z}}^{(b)})$ Set $\theta_j^{(a)} = \theta_j^{(d)}$.

else Set $\theta_j^{(a)} = \theta_j^{(d)}$.

Output: $[\theta_j^{(a)}, \theta_j^{(b)}]$.

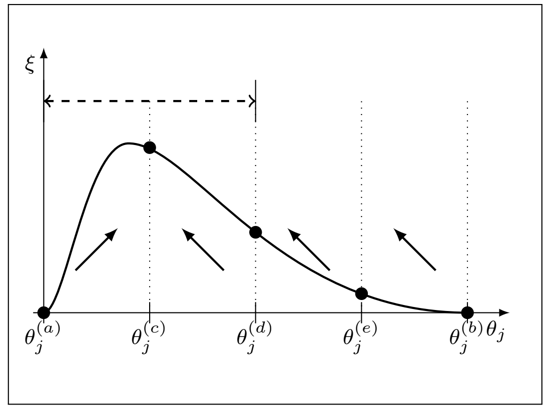

Figure 6: (Left): Subroutine *ShrinkInterval*. (Right): Visual intuition of the subroutine *ShrinkInterval* [17]; the subroutine shrinks the current interval to half based on oracle responses to the four queries.

- $\mathbf{e}_i$*'s and* $\mathbf{o}$ *are always achieved:* The classifier which always predicts class i, will achieve the rate $\mathbf{e}_i$. Thus, $\mathbf{e}_i \in \mathcal{R}^g \,\forall\, i \in [k], g \in [m]$ are feasible. Just like the convexity proof, a classifier which predicts similar to one of the trivial classifiers with probability $1/k$ achieves rate $\mathbf{o}$.

- $\mathbf{e}_i$*'s are vertices:* Any supporting hyperplane with slope $\ell_{1i} < \ell_{1j} < 0$ and $\ell_{1p} = 0$ for $p \in [k], p \neq i, j$ will be supported by $\mathbf{e}_1$ (corresponding to the trivial classifier which predict class 1). Thus, $\mathbf{e}_i$'s are vertices of the convex set. As long as the class-conditional distributions are not identical, i.e., there is some signal for non-trivial classification conditioned on each group [17], one can construct a ball around the trivial rate $\mathbf{o}$ and thus $\mathbf{o}$ lies in the interior.

□

B.1 Finding the Sphere $\mathcal{S}_\rho$

In this section, we discuss how a sufficiently large sphere $\mathcal{S}_\rho$ with radius ρ may be found. The following discussion is extended from [17] to multiple groups setting and provided here for completeness.

The following optimization problem is a special case of OP2 in [29]. The problem corresponds to feasiblity check problem for a given rate $\mathbf{r}_0$ achieved by all groups within small error $\epsilon > 0$.

$$\min_{\mathbf{r}^g \in \mathcal{R}^g \,\forall g \in [m]} 0 \qquad s.t.\;\; \|\mathbf{r}^g - \mathbf{r}_0\|_2 \leq \epsilon \quad \forall\, g \in [m]. \tag{OP1}$$

The above problem checks the feasibility and if a solution to the above problem exists, then Algorithm 1 of [29] returns it. The approach in [29] constructs a classifier whose group-wise rates are ϵ-close to the given rate $\mathbf{r}_0$.

Furthermore, Algorithm 3 computes a value of $\rho \geq \tilde{s}/k$, where $\tilde{s}$ is the radius of the largest ball contained in the set $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$. Notice that the approach in [29] is consistent, thus we should get a good estimate of the sphere, provided we have sufficient samples. The algorithm runs offline and does not impact query complexity.

Algorithm 3 Obtaining the sphere $\mathcal{S}_\rho$ with radius ρ

1: **Input:** The center $\mathbf{o}$ of the feasible region of rates across groups.
2: **for** $j = 1, 2, \cdots, q$ **do**
3: Let $\mathbf{r}_j$ be the standard basis vector for the j-th dimension.
4: Compute the maximum ℓ_j such that $\mathbf{o} + \ell_j \mathbf{r}_j$ is feasible for all groups by solving (OP1).
5: Let $CONV$ be the convex hull of $\{\mathbf{o} \pm \ell_j \mathbf{r}_j\}_{j=1}^q$.
6: Compute the radius s of the largest ball which can fit inside of $CONV$, centered at $\mathbf{o}$.
7: **Output:** Sphere $\mathcal{S}_\rho$ with radius $\rho = s$ centered at $\mathbf{o}$.

Lemma 3. *[17] Let $\tilde{s}$ be the radius of the largest ball centered at $\mathbf{o}$ in $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$. Then Algorithm 3 returns a radius $\rho \geq \tilde{s}/k$.*

Proof. Let ℓ_j be as computed in the algorithm and $\ell := \min_j \ell_j$, then we have $\ell \geq \tilde{s}$. Moreover, the region $CONV$ contains the convex hull of $\{o \pm \ell \mathbf{e}_j\}_{j=1}^q$; however, this region contains a ball of radius $\ell/\sqrt{q} = \ell/\sqrt{k^2 - k} \geq \ell/k \geq \tilde{s}/k$, and thus $\rho \geq \tilde{s}/k$. □

C Derivations of Section 4

C.1 Eliciting the Misclassification Cost $\overline{\phi}(\mathbf{r})$; Part 1 in Figure 3

The key to eliciting $\overline{\phi}$ is to remove the effect of fairness violation $\overline{\varphi}$ in the oracle responses. As explained in Section 4.1, we run the SLPME procedure (Algorithm 2) with the q-dimensional query space $\mathcal{S}_\rho$, binary search tolerance ϵ, the equivalent oracle with metric $\overline{\Psi} \circ \nu$. From Remark 1, this subroutine returns a slope $\mathbf{f}$ with $\|\mathbf{f}\|_2 = 1$ such that:

$$\frac{(1-\overline{\lambda})\overline{a}_i}{(1-\overline{\lambda})\overline{a}_j} = \frac{f_i}{f_j} \implies \frac{\overline{a}_i}{\overline{a}_j} = \frac{f_i}{f_j}. \tag{19}$$

Thus, we set $\hat{\mathbf{a}} := \mathbf{f}$ (line 1, Algorithm 1).

C.2 Eliciting the Fairness Violation $\overline{\varphi}(\mathbf{r}^{1:m})$; Part 2 in Figure 3

C.2.1 Eliciting the Fairness Violation $\overline{\varphi}(\mathbf{r}^{1:m})$ for $m = 2$; lines 2-5 in Algorithm 1

For $m = 2$, we have only one vector of unfairness weights $\mathbf{b}^{12}$, which we now aim to elicit given $\hat{\mathbf{a}}$. As discussed in Section 4.2.1, we fix trivial rates (through trivial classifiers) to one group and allow non-trivial rates from $\mathcal{S}_\rho$ on another group. This essentially makes the metric in Definition 1 linear. The elicitation procedure is as follows.

Fix trivial classifier predicting class 1 for group 2 i.e. fix $h^2(x) = 1 \, \forall \, x \in \mathcal{X}$, and thus $\mathbf{r}^2 = \mathbf{e}_1$. For group 1, we constrain the confusion rates to lie in the sphere $\mathcal{S}_\rho$ i.e. $\mathbf{r}^1 = \mathbf{s}$ for $\mathbf{s} \in \mathcal{S}_\rho$. Such a set up is governed by the parametrization $\nu'(\cdot, 1)$ in equation (10), and the metric in Definition 1 amounts to:

$$\overline{\Psi}(\nu'(\mathbf{s}, 1); \overline{\mathbf{a}}, \overline{\mathbf{b}}^{12}, \overline{\lambda}) = (1-\overline{\lambda})\langle \overline{\mathbf{a}} \odot (1 - \boldsymbol{\tau}^2), \mathbf{s}\rangle + \overline{\lambda}\langle \overline{\mathbf{b}}^{12}, |\mathbf{e}_1 - \mathbf{s}|\rangle + c_1. \tag{20}$$

The above is a function of $\mathbf{s} \in \mathcal{S}_\rho$. Since $\mathbf{e}_i$'s are binary vectors and since $0 \leq \mathbf{s} \leq 1$, the sign of the absolute function with respect to $\mathbf{s}$ can be recovered. Recall that the rates are defined in row major form of the rate matrices, thus $\mathbf{e}_1$ is 1 at every $(k + j * (k-1))$-th coordinate, where $j \in \{0, \ldots, k-2\}$, and 0 otherwise. The coordinates where the confusion rates are 1 in $\mathbf{e}_1$, the absolute function opens with a negative sign (wrt. $\mathbf{s}$) and with a positive sign otherwise. In particular, define a q-dimensional vector $\mathbf{w}_1$ with entries -1 at every $(k + j * (k-1))$-th coordinate, where $j \in \{0, \ldots, k-2\}$, and 1 otherwise. One may then write the metric $\overline{\Psi}$ as:

$$\overline{\Psi}(\nu'(\mathbf{s}, 1)\,;\, \bar{\mathbf{a}}, \bar{\mathbf{b}}^{12}, \bar{\lambda}) = \langle (1-\bar{\lambda})\bar{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^2) + \bar{\lambda}\mathbf{w}_1 \odot \bar{\mathbf{b}}^{12}, \mathbf{s}\rangle + c_1. \tag{21}$$

This is again a linear metric elicitation problem where $\mathbf{s} \in \mathcal{S}$. We may again use the SLPME procedure (Algorithm 2), which outputs a (normalized) slope $\breve{\mathbf{f}}$ with $\|\breve{\mathbf{f}}\|_2 = 1$ in line 3 of Algorithm 1. Using Remark 1, we get $q-1$ independent equations and may represent every element of $\bar{\mathbf{b}}^{12}$ based on one element, say $\bar{b}^{12}_{k-1}$, i.e.:

$$\begin{aligned}
&\frac{\breve{f}_{k-1}}{\breve{f}_i} = \frac{(1-\bar{\lambda})(1-\tau^2_{k-1})\bar{a}_{k-1} + \bar{\lambda}\bar{b}^{12}_{k-1}}{(1-\bar{\lambda})(1-\tau^2_i)\bar{a}_i + \bar{\lambda}w_{1i}\bar{b}^{12}_i} \qquad \forall\, i \in [q]. \\
\implies &\bar{\lambda}\bar{\mathbf{b}}^{12} = \mathbf{w}_1 \odot \left[\left(\frac{(1-\bar{\lambda})(1-\tau^2_{k-1})\bar{a}_{k-1} + \bar{\lambda}\bar{b}^{12}_{k-1}}{\breve{f}_{k-1}}\right)\breve{\mathbf{f}} - (1-\bar{\lambda})((1-\boldsymbol{\tau}^2) \odot \bar{\mathbf{a}})\right].
\end{aligned} \tag{22}$$

In order to elicit entire $\bar{\mathbf{b}}^{12}$, we need one more linear relation such as (22). So, we now fix the trivial classifier predicting class k for group 2 i.e. fix $h^2(x) = k\, \forall\, \mathbf{x} \in \mathcal{X}$, and thus $\mathbf{r}^2 = \mathbf{e}_k$. For group 1, we constrain the rates to again lie in the sphere $\mathcal{S}_\rho$ i.e. $\mathbf{r}^1 = \mathbf{s}$ for $\mathbf{s} \in \mathcal{S}_\rho$. Such a setup is governed by the parametrization $\nu'(\cdot, k)$ (10). Since the rate vectors are in row major form of the rate matrices, notice that $\mathbf{e}_k$ is 1 at every $(k-1+j*(k-1))$-th coordinate, where $j \in \{0, \ldots, k-2\}$, and 0 otherwise. In particular, define a q-dimensional vector $\mathbf{w}_k$ with entries -1 at every $(k-1+j*(k-1))$-th coordinate, where $j \in \{0, \ldots, k-2\}$, and 1 otherwise. One may then write the metric $\overline{\Psi}$ as:

$$\overline{\Psi}(\nu'(\mathbf{s}, k); \bar{\mathbf{a}}, \bar{\mathbf{b}}^{12}, \bar{\lambda}) = (1-\bar{\lambda})\langle \bar{\mathbf{a}} \odot (1 - \boldsymbol{\tau}^2), \mathbf{s}\rangle + \bar{\lambda}\langle \bar{\mathbf{b}}^{12}, |\mathbf{e}_k - \mathbf{s}|\rangle + c_k. \tag{23}$$

This is a linear metric elicitation problem where $\mathbf{s} \in \mathcal{S}$. Thus, line 4 of Algorithm 1 applies SLPME subroutine (Algorithm 2), which outputs a (normalized) slope $\tilde{\mathbf{f}}$ with $\|\tilde{\mathbf{f}}\|_2 = 1$. Using Remark 1, we extract the following relation between two of its coordinates, say the $(k-1)$-th and $((k-1)^2+1)$-th coordinates:

$$\frac{\tilde{f}_{k-1}}{\tilde{f}_{(k-1)^2+1}} = \frac{(1-\bar{\lambda})(1-\tau^2_{k-1})\bar{a}_{k-1} - \bar{\lambda}\bar{b}^{12}_{k-1}}{(1-\bar{\lambda})(1-\tau^2_{(k-1)^2+1})\bar{a}_{(k-1)^2+1} + \bar{\lambda}\bar{b}^{12}_{(k-1)^2+1}}. \tag{24}$$

Combining equations (22) and (24) and replacing the true $\bar{\mathbf{a}}$ with the estimated $\hat{\mathbf{a}}$ from

Section 4.1, we have an estimate of the scaled substitute as:

$$\tilde{\mathbf{b}}^{12} = \mathbf{w}_1 \odot \left[\delta \breve{\mathbf{f}}^{12} - \hat{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^2)\right], \tag{25}$$

$$\text{where } \delta = \frac{2(1-\tau^2_{k-1})\hat{a}_{k-1}}{\breve{f}_{k-1}} \left[\frac{\frac{(1-\tau^2_{(k-1)^2+1})\hat{a}_{(k-1)^2+1}}{(1-\tau^2_{k-1})\hat{a}_{k-1}} - \frac{\tilde{f}_{(k-1)^2+1}}{\tilde{f}_{k-1}}}{\left(\frac{\breve{f}_{(k-1)^2+1}}{\breve{f}_{k-1}} - \frac{\tilde{f}_{(k-1)^2+1}}{\tilde{f}_{k-1}}\right)} \right]$$

and $\tilde{\mathbf{b}}$ is a scaled substitute defined as $\tilde{\mathbf{b}}^{12} := \frac{\bar{\lambda}}{(1-\bar{\lambda})}\overline{\mathbf{b}}^{12}$, which nonetheless is computable from (25). Since we require a solution $\hat{\mathbf{b}}$ such that $\|\hat{\mathbf{b}}\|_2 = 1$ (Definition 1), we normalize $\tilde{\mathbf{b}}$ and get the final solution:

$$\hat{\mathbf{b}}^{12} = \frac{\tilde{\mathbf{b}}^{12}}{\|\tilde{\mathbf{b}}^{12}\|_2}. \tag{26}$$

Notice that, due to the above normalization, the solution is independent of the true trade-off $\bar{\lambda}$.

C.2.2 Eliciting the Fairness Violation $\overline{\varphi}(\mathbf{r}^{1:m})$ for $m > 2$; line 6-11 in Algorithm 1

Consider a non-empty set of sets $\mathcal{M} \subset 2^{[m]} \setminus \{\varnothing, [m]\}$. We will later discuss how to choose $\mathcal{M}$ for efficient elicitation. When $m > 2$, we partition the set of groups $[m]$ into two sets of groups. Let $\sigma \in \mathcal{M}$ and $[m] \setminus \sigma$ be one such partition of the m groups defined by the set σ. We follow exactly similar procedure as in the previous section i.e. fixing trivial rates (through trivial classifiers) on the groups in σ and allowing non-trivial rates from $\mathcal{S}_\rho$ on the groups in $[m] \setminus \sigma$. In particular, consider a paramterization $\nu' : (\mathcal{S}_\rho, \mathcal{M}, [k]) \to \mathcal{R}^{1:m}$ defined as:

$$\nu'(\mathbf{s}, \sigma, i) := \mathbf{r}^{1:m} \quad \text{such that} \quad \mathbf{r}^g = \begin{cases} \mathbf{e}_i & \text{if } g \in \sigma \\ \mathbf{s} & \text{o.w.} \end{cases} \tag{27}$$

i.e., ν' assigns trivial confusion rates $\mathbf{e}_i$ on the groups in σ and assigns $\mathbf{s} \in \mathcal{S}_\rho$ on the rest of the groups. Similar to the previous section, we first fix trivial classifier predicting class 1 for groups in σ and constrain the rates for groups in $[m] \setminus \sigma$ to be on the sphere $\mathcal{S}_\rho$. Such a setup is governed by the parametrization $\nu'(\cdot, \sigma, 1)$ in equation (13). Specifically, fixing $h^g(\mathbf{x}) = 1 \; \forall \; g \in \sigma$ would entail the metric in Definition 1 to be:

$$\overline{\Psi}(\nu'(\mathbf{s}, \sigma, 1); \bar{\mathbf{a}}, \overline{\mathbf{B}}, \bar{\lambda}) = (1-\bar{\lambda})\langle \bar{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^\sigma), \mathbf{s}\rangle + \lambda \langle \overline{\boldsymbol{\eta}}^\sigma, |\mathbf{e}_1 - \mathbf{s}|\rangle + c_1, \tag{28}$$

where $\boldsymbol{\tau}^\sigma = \sum_{g \in \sigma} \boldsymbol{\tau}^g$ and $\overline{\boldsymbol{\eta}}^\sigma = \sum_{u,v \in [m], v > u} \mathbb{1}\left[|\{u, v\} \cap \sigma| = 1\right] \overline{\mathbf{b}}^{uv}$. Similar to the previous section, since $\mathbf{e}_i$'s are binary vectors, the sign of the absolute function wrt. $\mathbf{s}$ can be recovered. In particular, the metric amounts to:

$$\overline{\Psi}(\nu'(\mathbf{s}, \sigma, 1); \bar{\mathbf{a}}, \overline{\mathbf{B}}, \bar{\lambda}) = \langle (1-\bar{\lambda})\bar{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^2) + \bar{\lambda}\mathbf{w}_1 \odot \overline{\boldsymbol{\eta}}^\sigma, \mathbf{s}\rangle + c_1, \tag{29}$$

where $\mathbf{w}_1 := 1 - 2\mathbf{e}_1$ and c_1 is a constant not affecting the responses. Notice that (28) and (29) are analogous to (20) and (21), respectively, except that $\boldsymbol{\tau}^2$ is replaced by $\boldsymbol{\tau}^\sigma$ and $\overline{\mathbf{b}}^{12}$ is replaced by $\overline{\boldsymbol{\eta}}^\sigma$. This is a linear metric in $\mathbf{s}$. We again the use the SLPME procedure in line 8 of Algorithm 1, which outputs a normalized slope $\breve{\mathbf{f}}^\sigma$ such that $\|\breve{\mathbf{f}}^\sigma\|_2 = 1$, and thus we get an analogous solution to (22) as:

$$\bar{\lambda}\bar{\boldsymbol{\eta}}^\sigma = \mathbf{w}_1 \odot \left[\left(\frac{(1-\bar{\lambda})(1-\tau^\sigma_{k-1})\bar{a}_{k-1} + \bar{\lambda}\bar{\eta}^\sigma_{k-1}}{\breve{f}^\sigma_{k-1}} \right) \breve{\mathbf{f}}^\sigma - (1-\bar{\lambda})((\mathbf{1}-\boldsymbol{\tau}^\sigma) \odot \bar{\mathbf{a}} \right]. \tag{30}$$

In order to elicit entire $\bar{\boldsymbol{\eta}}^\sigma$, we need one more linear relation such as (30). So, we now fix the trivial rates through trivial classifier predicting class k for the groups in σ i.e. fix $h^g(x) = k \, \forall \, \mathbf{x} \in \mathcal{X}$ if $g \in \sigma$, and thus $\mathbf{r}^g = \mathbf{e}_k$ for all groups $g \in \sigma$. For the rest of the groups, we constrain the confusion rates to again lie in the sphere $\mathcal{S}_\rho$ i.e. $\mathbf{r}^g = \mathbf{s}$ for $\mathbf{s} \in \mathcal{S}_\rho$ for all groups $g \in [m] \setminus \sigma$. Such a setup is governed by the parametrization $\nu'(\cdot, \sigma, k)$ (13). The metric $\bar{\Psi}$ in Definition 1 amounts to:

$$\bar{\Psi}(\nu'(\mathbf{s}, \sigma, k); \bar{\mathbf{a}}, \bar{\mathbf{B}}, \bar{\lambda}) = (1-\bar{\lambda})\langle \bar{\mathbf{a}} \odot (1-\boldsymbol{\tau}^\sigma), \mathbf{s} \rangle + \bar{\lambda}\langle \bar{\boldsymbol{\eta}}^\sigma, |\mathbf{e}_k - \mathbf{s}| \rangle + c_k. \tag{31}$$

Thus by running SLPME procedure again in line 9 of Algorithm 1 results in $\tilde{\mathbf{f}}^{12}$ with $\|\tilde{\mathbf{f}}^{12}\|_2 = 1$. Using Remark 1, we extract the following relation between the $(k-1)$-th and $((k-1)^2+1)$-th coordinates:

$$\frac{\tilde{f}^\sigma_{k-1}}{\tilde{f}^\sigma_{(k-1)^2+1}} = \frac{(1-\bar{\lambda})(1-\tau^\sigma_{k-1})\bar{a}_{k-1} - \bar{\lambda}\bar{\eta}^\sigma_{k-1}}{(1-\bar{\lambda})(1-\tau^\sigma_{(k-1)^2+1})\bar{a}_{(k-1)^2+1} + \bar{\lambda}\bar{\eta}^\sigma_{(k-1)^2+1}}. \tag{32}$$

Combining equations (30) and (32), we have:

$$\sum\nolimits_{u,v} \mathbb{1}\left[|\{u,v\} \cap \sigma| = 1\right] \tilde{\mathbf{b}}^{uv} = \boldsymbol{\gamma}^\sigma, \quad \text{where} \tag{33}$$

$$\boldsymbol{\gamma}^\sigma = \mathbf{w}_1 \odot \left[\delta^\sigma \mathbf{f}^\sigma - \hat{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau}^\sigma)\right], \; \delta^\sigma = \frac{2(1-\tau^\sigma_{k-1})\hat{a}_{k-1}}{f^\sigma_{k-1}} \left[\frac{\frac{(1-\tau^\sigma_{(k-1)^2+1})\hat{a}_{(k-1)^2+1}}{(1-\tau^\sigma_{k-1})\hat{a}_{k-1}} - \frac{\tilde{f}^\sigma_{(k-1)^2+1}}{\tilde{f}^\sigma_{k-1}}}{\left(\frac{f^\sigma_{(k-1)^2+1}}{f^\sigma_{k-1}} - \frac{\tilde{f}^\sigma_{(k-1)^2+1}}{\tilde{f}^\sigma_{k-1}} \right)} \right],$$

and $\tilde{\mathbf{b}}^{uv} := \bar{\lambda}\bar{\mathbf{b}}^{uv}/(1-\bar{\lambda})$ is a scaled version of the true (unknown) $\bar{\mathbf{b}}$, which nonetheless can be computed from (33).

By two runs of SLPME algorithm, we can get $\boldsymbol{\gamma}^\sigma$ and solve (33). However, the left hand side of (33) does not allow us to recover the $\tilde{\mathbf{b}}$'s separately and provides only one equation. Let us denote the Equation (33) by ℓ^σ corresponding to the set σ. In order to elicit all $\tilde{\mathbf{b}}$'s we need a system of $M := \binom{m}{2}$ independent equations. This is easily achievable by choosing M σ's so that we get M set of unique equations like (33). Let $\mathcal{M}$ be those set of sets. In most cases, pairing two groups to have trivial rates (through trivial classifiers) and rest of the groups to have rates from the sphere $\mathcal{S}$ will work. For example, when $m = 3$, fixing $\mathcal{M} = \{\{1,2\},\{1,3\},\{2,3\}\}$ suffices. Thus, running over all the choices of sets of groups $\sigma \in \mathcal{M}$ provides the system of equations $\mathcal{L} := \cup_{\sigma \in \mathcal{M}} \ell^\sigma$ (line 10 in Algorithm 1), which is formally described as follows:

$$\begin{bmatrix} \Xi & 0 & \dots & 0 \\ 0 & \Xi & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Xi \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}}_{(1)} \\ \tilde{\mathbf{b}}_{(2)} \\ \dots \\ \tilde{\mathbf{b}}_{(q)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_{(1)} \\ \boldsymbol{\gamma}_{(2)} \\ \dots \\ \boldsymbol{\gamma}_{(q)} \end{bmatrix}, \tag{34}$$

where $\tilde{\mathbf{b}}_{(i)} = (\tilde{b}_i^1, \tilde{b}_i^2, \cdots, \tilde{b}_i^M)$ and $\boldsymbol{\gamma}_{(i)} = (\gamma_i^1, \gamma_i^2, \cdots, \gamma_i^M)$ are vectorized versions of the i-th entry across groups for $i \in [q]$, and $\Xi \in \{0,1\}^{M\times M}$ is a binary full-rank matrix denoting membership of groups in the set $\sigma \in \mathcal{M}$. For instance, for the choice of $\mathcal{M} = \{\{1,2\},\{1,3\},\{2,3\}\}$ when $m = 3$ gives:

$$\Xi = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

From technical point of view, one may choose any $\mathcal{M}$ such that the resulting group membership matrix Ξ is non-singular. Hence the solution of the system of equations $\mathcal{L}$ is:

$$\begin{bmatrix} \tilde{\mathbf{b}}_{(1)} \\ \tilde{\mathbf{b}}_{(2)} \\ \dots \\ \tilde{\mathbf{b}}_{(q)} \end{bmatrix} = \begin{bmatrix} \Xi & 0 & \dots & 0 \\ 0 & \Xi & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Xi \end{bmatrix}^{(-1)} \begin{bmatrix} \boldsymbol{\gamma}_{(1)} \\ \boldsymbol{\gamma}_{(2)} \\ \dots \\ \boldsymbol{\gamma}_{(q)} \end{bmatrix}. \tag{35}$$

When we normalize $\tilde{\mathbf{b}}$, we get the final fairness violation weight estimates as:

$$\hat{\mathbf{b}}^{uv} = \frac{\tilde{\mathbf{b}}^{uv}}{\sum_{u,v=1,v>u}^{m} \|\tilde{\mathbf{b}}^{uv}\|_2} \quad \text{for} \quad u, v \in [m], v > u. \tag{36}$$

Notice that, due to the above normalization, the solution is again independent of the true trade-off $\bar{\lambda}$.

C.3 Eliciting Trade-off $\bar{\lambda}$; Part 3 in Figure 3

Using the parametrization ν'' from (16), the metric in Definition 1 reduces to a linear metric in $\mathbf{s}^+$ as discussed in (17), i.e:

$$\overline{\Psi}(\nu''(\mathbf{s}^+)\,;\,\bar{\mathbf{a}}, \overline{\mathbf{B}}, \bar{\lambda}) = \langle (1-\bar{\lambda})\boldsymbol{\tau}^1 \odot \bar{\mathbf{a}} + \bar{\lambda} \sum\nolimits_{v=2}^{m} \bar{\mathbf{b}}^{1v}, \mathbf{s}^+ \rangle + c. \tag{37}$$

We first show the proof of Lemma 1 and then discuss the trade-off elicitation algorithm (Algorithm 4).

Proof of Lemma 1. For simplicity, let us abuse notation for this proof and denote $\boldsymbol{\tau}^1 \odot \bar{\mathbf{a}}$ simply by $\mathbf{a}$, $\sum_{v=2}^{m} \bar{\mathbf{b}}^{1v}$ simply by $\mathbf{b}$, and $\mathcal{S}_{\varrho}^{+}$ simply by $\mathcal{S}$.

$\mathcal{S}$ is a convex set. Let $\mathcal{Z} = \{\mathbf{z} = (z_1, z_2)\,|\, z_1 = <\mathbf{a}, \mathbf{s}>, z_2 = <\mathbf{b}, \mathbf{s}>, \mathbf{s} \in \mathcal{S}\}$.

Claim: $\mathcal{Z}$ is convex.

Let $z, z' \in \mathcal{Z}$.

$$\begin{aligned} \alpha z_1 + (1-\alpha) z_1' &= \alpha <\mathbf{a}, \mathbf{s}> + (1-\alpha) <\mathbf{a}, \mathbf{s}'> &= <\mathbf{a}, \alpha\mathbf{s} + (1-\alpha)\mathbf{s}'> \\ \alpha z_2 + (1-\alpha) z_2' &= \alpha <\mathbf{b}, \mathbf{s}> + (1-\alpha) <\mathbf{b}, \mathbf{s}'> &= <\mathbf{b}, \alpha\mathbf{s} + (1-\alpha)\mathbf{s}'> \end{aligned}$$

Since $\alpha\mathbf{s} + (1-\alpha)\mathbf{s}' \in \mathcal{S}$, $\alpha z + (1-\alpha) z' \in \mathcal{Z}$. Hence $\mathcal{Z}$ is convex.

Claim: The boundary of the set $\mathcal{Z}$ is a strictly convex curve with no vertices for $\mathbf{a} \neq \mathbf{b}$.

Recall that, the required function is given by:

$$\vartheta(\lambda) = \max_{\mathbf{z}\in\mathcal{Z}}(1-\lambda)z_1 + \lambda z_2 + c \tag{38}$$

(i) Since the set $\mathcal{Z}$ is convex, every boundary point is supported by a hyperplane.

(ii) Since $\mathbf{a} \neq \mathbf{b}$, notice that the slope is uniquely defined by λ. Since the sphere $\mathcal{S}$ is strictly convex, the above linear functional defined by λ is maximized by a unique point in $\mathcal{Z}$ (similar to Lemma 2). Thus, the the hyperplane is tangent at a unique point on the boundary of $\mathcal{Z}$.

(iii) It only remains to show that there are no vertices on the boundary of $\mathcal{Z}$. Recall that a vertex exists if (and only if) some point is supported by more than one tangent hyperplane in two dimensional space. This means there are two values of λ that achieve the same maximizer. This is contradictory since there are no two linear functionals that achieve the same maximizer on $\mathcal{S}$.

This implies that the boundary of $\mathcal{Z}$ is strictly convex curve with no vertices. Since we are interested in the maximization of ϑ, let us call this boundary as the upper boundary and denote it by $\partial\mathcal{Z}_+$.

Claim: Let $\upsilon : [0,1] \to \partial\mathcal{Z}_+$ be continuous, bijective, parametrizations of the upper boundary. Let $\vartheta : \mathcal{Z} \to \mathbb{R}$ be a quasiconcave function which is monotone increasing in both z_1 and z_2. Then the composition $\vartheta \circ \upsilon : [0,1] \to \mathbb{R}$ is strictly quasiconcave (and therefore unimodal with no flat regions) on the interval $[0,1]$.

Let S be some superlevel set of the quasiconcave function ϑ. Since υ is a continuous bijection and since the boundary $\partial\mathcal{Z}_+$ is a strictly convex curve with no vertices, wlog., for any $r < s < t$, $z_1(\upsilon(r)) < z_1(\upsilon(s)) < z_1(\upsilon(t))$, and $z_2(\upsilon(r)) > z_2(\upsilon(s)) > z_2(\upsilon(t))$. (otherwise, swap r and t). Since the boundary $\partial\mathcal{Z}_+$ is a strictly convex curve, then $\upsilon(s)$ must be greater (component-wise) a point in the convex combination of $\upsilon(r)$ and $\upsilon(t)$. Let us denote that point by u. Since ϑ is monotone increasing, then $x \in S$ implies that $y \in S$, too, for all $y \geq x$ componentwise. Therefore, $\vartheta(\upsilon(s)) \leq \vartheta(u)$. Since S is convex, $u \in S$ and thus $\upsilon(s) \in S$.

This implies that $\upsilon^{-1}(\partial\mathcal{Z}_+ \cap S)$ is an interval; hence it is convex, which in turn tells us that the superlevel sets of $\vartheta \circ \upsilon$ are convex. So, $\vartheta \circ \upsilon$ is quasiconcave, as desired. This implies unimodaltiy, because a function defined on real line which has more than one local maximum can not be quasiconcave. Moreover, since there are no vertices on the boundary $\partial\mathcal{Z}_+$, the $\vartheta \circ \upsilon : [0,1] \to \mathbb{R}$ is strictly quasiconcave (and thus unimodal with no flat regions) on the interval $[0,1]$. This completes the proof of Lemma 1. $\square$

Description of Algorithm 4:[2] Given the unimodality of $\vartheta(\lambda)$ from Lemma 1, we devise the binary-search procedure Algorithm 4 for eliciting the true trade-off $\overline{\lambda}$. The algorithm takes in input the query space $\mathcal{S}_\varrho^+$, binary-search tolerance ϵ, an equivalent oracle $\Omega(\cdot,\cdot \;\; \Psi\circ\nu'')$ with metric $\Psi\circ\nu''$ as explained in Section 4.3, the elicited $\hat{\mathbf{a}}$ from Section 4.1, and the elicited $\hat{\mathbf{B}}$ from Section 4.2. The algorithm finds the maximizer of the function $\hat{\vartheta}(\lambda)$ defined analogously to (18), where $\bar{\mathbf{a}}, \overline{\mathbf{B}}$ are replaced by $\hat{\mathbf{a}}, \hat{\mathbf{B}}$. The algorithm poses four queries to the oracle and shrink the interval $[\lambda^{(a)}, \lambda^{(b)}]$ into half based on the responses using a subroutine analogous to *ShrinkInterval* shown in Figure 6. The algorithm stops when the length of the search interval $[\lambda^{(a)}, \lambda^{(b)}]$ is less than the tolerance ϵ.

[2] The superscripts in Algorithm 2 denote iterates. Please do not confuse it with the sensitive group index.

Algorithm 4 Eliciting the trade-off $\overline{\lambda}$

1: **Input:** Query space $\mathcal{S}_\varrho^+$, binary-search tolerance $\epsilon > 0$, oracle $\Omega(\cdot, \cdot, \overline{\Psi} \circ \nu'')$ with metric $\overline{\Psi} \circ \nu''$, $\hat{\mathbf{a}}, \hat{\mathbf{B}}$.
2: **Initialize:** $\lambda^{(a)} = 0$, $\lambda^{(b)} = 1$.
3: **while** $\left|\lambda^{(b)} - \lambda^{(a)}\right| > \epsilon$ **do**
4: Set $\lambda^{(c)} = \frac{3\lambda^{(a)}+\lambda^{(b)}}{4}$, $\lambda^{(d)} = \frac{\lambda^{(a)}+\lambda^{(b)}}{2}$, $\lambda^{(e)} = \frac{\lambda^{(a)}+3\lambda^{(b)}}{4}$
5: Set $\mathbf{s}^{(a)} = \underset{\mathbf{s}^+ \in \mathcal{S}_\varrho^+}{\operatorname{argmax}} \langle (1-\lambda_a)\boldsymbol{\tau}^1 \odot \hat{\mathbf{a}} + \lambda_a \sum_{v=2}^{m} \hat{\mathbf{b}}^{1v}, \mathbf{s}^+ \rangle$ using Lemma 2
6: Similarly, set $\mathbf{s}^{(c)}$, $\mathbf{s}^{(d)}$, $\mathbf{s}^{(e)}$, $\mathbf{s}^{(b)}$.
7: Query $\Omega(\mathbf{s}^{(c)}, \mathbf{s}^{(a)}\,;\, \Psi \circ \nu'')$, $\Omega(\mathbf{s}^{(d)}, \mathbf{s}^{(c)}\,;\, \Psi \circ \nu'')$, $\Omega(\mathbf{s}^{(e)}, \mathbf{s}^{(d)}\,;\, \Psi \circ \nu'')$, and $\Omega(\mathbf{s}^{(b)}, \mathbf{s}^{(e)}\,;\, \Psi \circ \nu'')$.
8: $[\lambda^{(a)}, \lambda^{(b)}] \leftarrow$ *ShrinkInterval* (responses) using a subroutine analogous to the routine shown in Figure 6.
9: **Output:** $\hat{\lambda} = \frac{\lambda^{(a)}+\lambda^{(b)}}{2}$.

D Proof of Section 5

Proof of Theorem 1. We break this proof into three parts.

1. *Elicitation guarantees for the misclassification cost $\hat{\phi}$ (i.e., $\hat{\mathbf{a}}$)*

   Since Algorithm 1 elicits a linear metric using the q-dimensional sphere $\mathcal{S}$, the guarantees on $\hat{\mathbf{a}}$ follows from Theorem 2 of [17]. Thus, under Assumption 2, the output $\hat{\mathbf{a}}$ from line 1 of Algorithm 1 satisfies $\|\mathbf{a}^* - \hat{\mathbf{a}}\|_2 \leq O(\sqrt{q}(\epsilon + \sqrt{\epsilon_\Omega/\rho}))$ after $O\left(q \log \frac{\pi}{2\epsilon}\right)$ queries.

2. *Elicitation guarantees for the fairness violation cost $\hat{\varphi}$ (i.e., $\hat{\mathbf{B}}$)*

   We start with the definition of true $\boldsymbol{\gamma}$ (i.e. when all the elicited entities are true) from (33) and let us drop the superscript σ for simplicity. Furthermore, let $\epsilon + \sqrt{\epsilon_\Omega/\rho}$ be denoted by ϵ.

$$\boldsymbol{\gamma} = \mathbf{w}_1 \odot \left[\delta\breve{\mathbf{f}} - \bar{\mathbf{a}} \odot (\mathbf{1} - \boldsymbol{\tau})\right] \quad \text{where } \delta = \frac{2(1-\tau_{k-1})\bar{a}_{k-1}}{\breve{f}_{k-1}} \left[\frac{\frac{(1-\tau_{(k-1)^2+1})\bar{a}_{(k-1)^2+1}}{(1-\tau_{k-1})\bar{a}_{k-1}} - \frac{\tilde{f}_{(k-1)^2+1}}{\tilde{f}_{k-1}}}{\left(\frac{\breve{f}_{(k-1)^2+1}}{\breve{f}_{k-1}} - \frac{\tilde{f}_{(k-1)^2+1}}{\tilde{f}_{k-1}}\right)}\right].$$

   Let us look at the derivative of the i-th coordinate of $\boldsymbol{\gamma}$.

$$\frac{\partial \gamma_i}{\partial a_j} = \begin{cases} 0 & \text{if } j \neq i, j \neq k-1, j \neq (k-1)^2+1 \\ -\tau_i & \text{if } j = i \\ c_{i,1} & \text{if } j = k-1 \\ c_{i,2} & \text{if } j = (k-1)^2+1, \end{cases}$$

   where $c_{i,1}$ and $c_{i,2}$ are some bounded constants due to Assumption 2. Here, we clip the estimates at the lower bounds of the true entities from Assumption 2. Similarly, $\partial\gamma_i/\partial f_j$ is bounded as well due to the regularity Assumption 2. This means that γ_i is Lipschitz in 2-norm wrt. $\mathbf{a}$ and $\mathbf{f}$. Thus,

$$\|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_\infty \leq c_3\|\bar{\mathbf{a}} - \hat{\mathbf{a}}\|_2 + c_4\|\breve{\mathbf{f}} - \hat{\breve{\mathbf{f}}}\|_2,$$

for some Lipschits constants c_3 and c_4. From the bounds of Part 1 of this proof, we have:

$$\|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}\|_\infty \leq O(\sqrt{q}\epsilon).$$

Recall the construction of $\tilde{\mathbf{b}}_{(i)}$ from (34). We then have from the solution of system of equations (35) that:

$$\tilde{\mathbf{b}}_{(i)} = \Xi^{-1}\boldsymbol{\gamma}_{(i)} \quad \forall\, i \in [q],$$

where $\tilde{\mathbf{b}}_{(i)} = (\tilde{b}_i^1, \tilde{b}_i^2, \cdots, \tilde{b}_i^M)$ and $\tilde{\boldsymbol{\gamma}}_{(i)} = (\gamma_i^1, \gamma_i^2, \cdots, \gamma_i^M)$ are vectorized versions of the i-th entry across groups for $i \in [q]$. $\Xi \in \{0,1\}^{M\times M}$ is a full-rank symmetric matrix with bounded infinity norm $\|\Xi^{-1}\|_\infty \leq c$ (here, infinity norm of a matrix is defined as the maximum absolute row sum of the matrix). Thus we have:

$$\|\tilde{\mathbf{b}}_{(i)} - \hat{\tilde{\mathbf{b}}}_{(i)}\|_\infty = \|\Xi^{-1}\boldsymbol{\gamma}_{(i)} - \Xi^{-1}\hat{\boldsymbol{\gamma}}_{(i)}\|_\infty = \|\Xi^{-1}(\boldsymbol{\gamma}_{(i)} - \hat{\boldsymbol{\gamma}}_{(i)})\|_\infty \leq \|\Xi^{-1}\|_\infty\|\boldsymbol{\gamma}_{(i)} - \hat{\boldsymbol{\gamma}}_{(i)}\|_\infty,$$

which gives

$$\|\tilde{\mathbf{b}}_{(i)} - \hat{\tilde{\mathbf{b}}}_{(i)}\|_\infty \leq O(\sqrt{q}\epsilon).$$

Now, our final estimate is the normalized form of $\hat{\tilde{\mathbf{b}}}$ from (36), so the final error in the stacked version $vec(\overline{\mathbf{B}})$ and $vec(\hat{\mathbf{B}})$ is:

$$\|vec(\overline{\mathbf{B}}) - vec(\hat{\mathbf{B}})\|_\infty \leq O(\sqrt{q}\epsilon). \tag{39}$$

Since there are $q \times M$ entities in $vec(\mathbf{B})$, we have:

$$\|vec(\overline{\mathbf{B}}) - vec(\hat{\mathbf{B}})\|_2 \leq O(\sqrt{qM}\sqrt{q}\epsilon) = O(mq\epsilon). \tag{40}$$

Due to elicitation on sphere and the oracle noise ϵ_Ω as defined in Definition 4, we can replace ϵ with $\epsilon + \sqrt{\epsilon_\Omega/\rho}$ back to get the final bound on fairness violation weights as in Theorem 1.

3. *Elicitation guarantees for the trade-off parameter (i.e., $\hat{\lambda}$)*

   The metric for our purpose is a linear metric in $\mathbf{s}^+ \in \mathcal{S}_\rho^+$ with the following slope:

$$\overline{\Psi}(\nu''(\mathbf{s}^+)\,;\,\bar{\mathbf{a}}, \overline{\mathbf{B}}, \bar{\lambda}) = \langle (1-\bar{\lambda})\boldsymbol{\tau}^1 \odot \bar{\mathbf{a}} + \bar{\lambda}\sum_{v=2}^{m} \overline{\mathbf{b}}^{1v}, \mathbf{s}^+\rangle. \tag{41}$$

   Since we elicit λ through queries over a surface of the sphere, we pose this problem as finding the right angle (slope) defined by the true $\bar{\lambda}$. Note that $\bar{\lambda}$ is what we want to elicit;

however, due to oracle noise ϵ_Ω, we can only aim to achieve a target angle λ_t. Moreover, we do not have true $\bar{\mathbf{a}}$ and $\overline{\mathbf{B}}$ but have only estimates $\hat{\mathbf{a}}$ and $\hat{\mathbf{B}}$. Thus we query proxy solutions always and can only aim to achieve an estimated version λ_e of the target angle. Lastly, Algorithm 4 is stopped within an ϵ threhsold, thus the final solution $\hat{\lambda}$ is within ϵ distance from λ_e. In total, we want to find:

$$|\bar{\lambda} - \hat{\lambda}| \leq \underbrace{|\bar{\lambda} - \lambda_t|}_{\text{oracle error}} + \underbrace{|\lambda_t - \lambda_e|}_{\text{estimation error}} + \underbrace{|\lambda_e - \hat{\lambda}|}_{\text{optimization error}} .$$

- optimization error: $|\lambda_e - \hat{\lambda}| \leq \epsilon$.
- oracle error: Notice that the oracle correctly answers as long as $\varrho(1 - \cos(\bar{\lambda} - \lambda_t)) > \epsilon_\Omega$. This is due to the fact that the metric is a 1-Lipschitz linear function, and the optimal value on the sphere of radius ϱ is ϱ. However, as $1 - \cos(x) \geq x^2/3$, so oracle is correct as long as $|\bar{\lambda} - \lambda_e| \geq \sqrt{3\epsilon_\Omega/\varrho}$. Given this condition, the binary search proceeds in the correct direction.
- estimation error: We make this error because we only have access to the estimated $\hat{\mathbf{a}}$ and $\hat{\mathbf{B}}$ not the true $\bar{\mathbf{a}}$ and $\overline{\mathbf{B}}$. However, since the metric in (41) is Lipschitz in $\bar{\mathbf{a}}$ and $\sum_{v=2}^m \overline{\mathbf{b}}^{1v}$, this error can be treated as oracle feedback noise where the oracle responses with the estimated $\hat{\mathbf{a}}$ and $\hat{\mathbf{B}}$. Thus, if we replace ϵ_Ω from the previous point to the error in $\hat{\mathbf{a}}$ and $\sum_{v=2}^m \hat{\mathbf{b}}^{1v}$, the binary search Algorithm 4 moves in the right direction as long as

  $$|\lambda_t - \lambda_e| \geq O\left(\sqrt{\frac{\|\bar{\mathbf{a}} - \hat{\mathbf{a}}\|_2 + \sum_{v=2}^m \|\overline{\mathbf{b}}^{1v} - \hat{\mathbf{b}}^{1v}\|_2}{\varrho}}\right) = O\left(\sqrt{mq(\epsilon + \sqrt{\epsilon_\Omega/\rho})/\varrho}\right),$$

  where we have used (40) to bound the error in $\{\hat{\mathbf{b}}^{1v}\}_{v=2}^m$.

Combining the three error bounds above gives us the desired result for trade-off parameter in Theorem 1.

□