Quadratic Metric Elicitation with Application to Fairness

Gaurush Hiranandani
UIUC
gaurush2@illinois.edu

Jatin Mathur
UIUC
jatinm2@illinois.edu

Harikrishna Narasimhan
Google Research
hnarasimhan@google.com

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

November 4, 2020

Abstract

Metric elicitation is a recent framework for eliciting performance metrics that best reflect implicit user preferences. This framework enables a practitioner to adjust the performance metrics based on the application, context, and population at hand. However, available elicitation strategies have been limited to linear (or fractional-linear) functions of predictive rates. In this paper, we develop an approach to elicit from a wider range of complex multiclass metrics defined by quadratic functions of rates by exploiting their local linear structure. We apply this strategy to elicit quadratic metrics for group-based fairness, and also discuss how it can be generalized to higher-order polynomials. Our elicitation strategies require only relative preference feedback and are robust to both feedback and finite sample noise.

1 Introduction

Given a classification problem, which performance metric should the classifier optimize? This question is often faced by practitioners while developing machine learning solutions. For example, consider cancer diagnosis where the doctor applies a cost-sensitive predictive model to classify patients into cancer categories [53, 56]. Although it is clear that the chosen costs directly determine the model decisions and thus patient outcomes, it is not clear how to quantify expert intuition into precise quantitative cost trade-offs, i.e. the performance metric. Indeed this is also true for a variety of other domains where picking the right metric is a critical challenge [8].

Hiranandani et al. [16, 17] addressed this issue by formalizing the problem of *Metric Elicitation (ME)*, where the goal is to estimate a performance metric using preference feedback from a user. The motivation is that by employing metrics that reflect a user's innate trade-offs, one can learn models that best capture the user preferences [16]. As humans are often inaccurate in

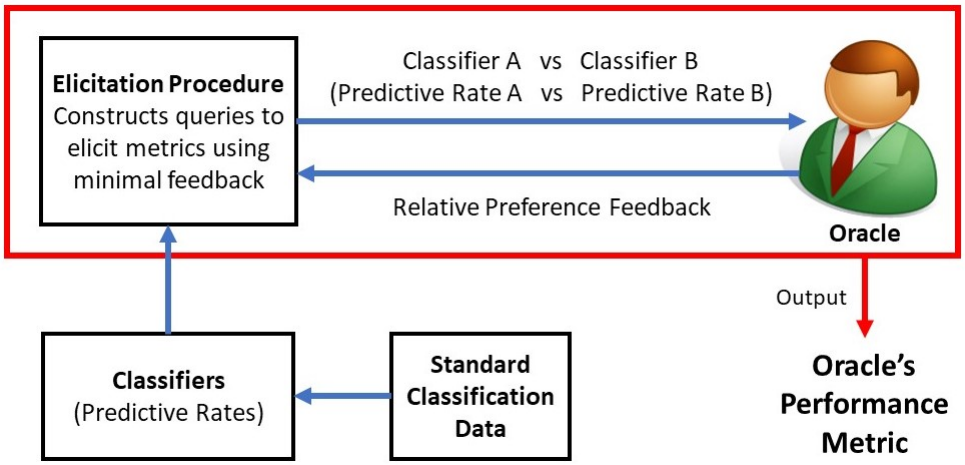

Figure 1: Metric Elicitation Framework [16].

providing absolute preferences [44], Hiranandani et al. [16] propose to use pairwise comparison queries, where the user (oracle) is asked to compare two classifiers and provide a relative preference.

Using such queries, ME aims to recover the oracle's metric. Figure 1 (reproduced from Hiranandani et al. [16]) depicts this framework.

A limitation of existing ME strategies is that they only handle metrics that are linear or quasi-linear functions of predictive rates, which can be restrictive in domains where the metrics are more complex and nuanced, e.g., [35, 38, 14]. In this paper, we propose strategies for eliciting metrics defined by *quadratic* functions of rates, which encompass linear metrics as special cases. We are thus able to handle a more general family of metrics that can better capture a practitioner's innate preferences, and include many new metrics used for class-imbalanced learning [31, 35], distribution matching [13, 11, 38] and group-based fairness [14] (see Section 2.3 for examples).

Our key idea is to approximate the quadratic metric locally by linear functions and apply linear ME as a subroutine to elicit the local linear approximations. The challenge is to choose a small number of center points to approximate the metric and then reconstruct the quadratic metric from the elicited local approximations. We address this by exploiting the geometry of the space of predictive rates and the smoothness properties of the metric and provide an efficient elicitation procedure with a query complexity that is *linear* in the number of unknown entities.

An important application for quadratic metric elicitation is *fairness in machine learning* [23, 10, 14, 28, 52, 36]. While several group-based fairness metrics have been proposed to capture bias in automated decision-making, selecting the right metric for an application remains a crucial challenge [55]. Recently, Hiranandani et al. [18] proposed an approach for eliciting group fairness metrics, where they measure the discrepancy in fairness using the absolute differences in predictive rates across multiple sensitive groups. A limitation of their work is that they only consider fairness metrics that are linear in the group discrepancies.

We extend their setup using our quadratic elicitation approach and allow for more general fairness metrics defined by quadratic functions of (signed) group discrepancies. Like Hiranandani

et al., we provide a procedure to jointly elicit three terms: (i) predictive performance defined by a weighted error metric, (ii) a quadratic fairness violation metric, and (iii) a trade-off between the predictive performance and fairness violation. Lastly, our quadratic elicitation strategy can be further generalized to higher-order polynomial functions. The idea is to approximate a d-th order polynomial locally with $(d-1)$-th order polynomials and recursively apply our procedure to the lower-order polynomials.

In summary, the following are our main contributions:

- We propose a novel quadratic metric elicitation algorithm for multiclass classification (Section 3).
- We adapt our approach for group-fair classification, and show how to jointly elicit the predictive performance and fairness violation metrics, and also the trade-off point between them (Section 4).
- We prove robustness of the proposals under feedback and classifier estimation noise (Section 5).
- We empirically validate the proposed solutions on simulated oracles and show their robustness to multiple classes and groups (Section 6).
- We discuss how our strategy can be generalized to elicit higher-order polynomials of rates (Section 8).

All our procedures require only pairwise preference feedback from the oracle and use binary-search based subroutines. Moreover, they can be applied by querying preferences either over classifiers or over rates.

Notation. For $k \in \mathbb{Z}_+$, we denote $[k] = \{1, 2, \cdots, k\}$ and use Δ_k to denote the $(k-1)$-dimensional simplex. We denote the inner product of vectors by $\langle \cdot, \cdot \rangle$ and the Hadamard product by $\odot$. For a matrix $\mathbf{A}$, *off-diag*$(\mathbf{A})$ returns a vector of off-diagonal elements of $\mathbf{A}$ in row-major form. We denote the 2-norm of a vector by $\| \cdot \|_2$ and the Frobenius norm of a matrix by $\| \cdot \|_F$. We use $\boldsymbol{\alpha}_i \in \mathbb{R}^q$ to denote the i-th standard basis vector, where the i-th coordinate is 1 and the others are 0.

2 Background

We consider the standard k-class classification setting with $X \in \mathcal{X}$ and $Y \in [k]$ representing the input and output random variables, respectively. We assume access to a sample $\{(\mathbf{x}, y)_i\}_{i=1}^n$ of n examples generated *iid* from a distribution $\mathbb{P}(X, Y)$. We work with (randomized) classifiers $h : \mathcal{X} \to \Delta_k$, and use $\mathcal{H} = \{h : \mathcal{X} \to \Delta_k\}$ to denote the set of all classifiers.

Predictive rates: We define the predictive rate matrix for a classifier h by $\mathbf{R}(h, \mathbb{P}) \in \mathbb{R}^{k \times k}$, where the ij-th entry is the fraction of label-i examples for which the classifier h predicts j:

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

Notice that each diagonal element of this matrix can be written in terms of the off-diagonal elements as follows:

$$R_{ii}(h, \mathbb{P}) = 1 - \sum_{j=1, j\neq i}^{k} R_{ij}(h, \mathbb{P}). \tag{2}$$

Using this decomposition, we can uniquely represent a rate matrix with its $q := (k^2 - k)$ off-diagonal elements and can concisely write it as a vector $\mathbf{r}(h, \mathbb{P}) = \textit{off-diag}(\mathbf{R}(h, \mathbb{P}))$. So we will interchangeably refer to the rate matrix as a *'vector of rates'*.

Metrics: We consider performance metrics that are defined by a general function $\phi : [0, 1]^q \to \mathbb{R}$ of rates:

$$\phi(\mathbf{r}(h, \mathbb{P})).$$

This includes the (weighted) error rate $\phi^{\text{err}}(\mathbf{r}(h, \mathbb{P})) = \sum_i a_i r_i(h, \mathbb{P})$, for weights $a_i \in \mathbb{R}_+$, the F-measure and many more metrics [49]. Without loss of generality (wlog), we treat metrics as costs, i.e. lower values are better. Since the scale of the metric does not affect the learning problem [40], we allow $\phi : [0, 1]^q \to [-1, 1]$.

Feasible rates: We will restrict our attention to only those rates that are feasible, i.e., can be achieved by some classifier. The set of all feasible rates is given by:

$$\mathcal{R} = \{\mathbf{r}(h, \mathbb{P}) \,:\, h \in \mathcal{H}\}.$$

For simplicity, we will suppress the dependence on $\mathbb{P}$ and h if it is clear from the context.

2.1 Metric Elicitation: Problem Setup

We now describe the problem of *Metric Elicitation*. Our definitions follow from Hiranandani et al. [17]. There's an *unknown* metric ϕ, and we seek to elicit its form by posing queries to an *oracle* asking which of two classifiers is more preferred by it. The oracle has access to the underlying metric ϕ and provides answers by comparing its value on the two classifiers.

Definition 1 (Oracle Query)**.** *Given two classifiers h_1, h_2 (equiv. to rates $\mathbf{r}_1, \mathbf{r}_2$ respectively), a query to the Oracle (with metric ϕ) is represented by:*

$$\Gamma(h_1, h_2\,;\, \phi) = \Omega(\mathbf{r}_1, \mathbf{r}_2\,;\, \phi) = \mathbb{1}[\phi(\mathbf{r}_1) > \phi(\mathbf{r}_2)], \tag{3}$$

where $\Gamma : \mathcal{H} \times \mathcal{H} \to \{0, 1\}$ and $\Omega : \mathcal{R} \times \mathcal{R} \to \{0, 1\}$. The query asks whether h_1 is preferred to h_2 (equiv. if $\mathbf{r}_1$ is preferred to $\mathbf{r}_2$), as measured by ϕ.

In practice, the oracle can be an expert, a group of experts, or an entire user population. The ME framework can be applied by posing classifier comparisons directly via interpretable learning techniques [46, 9] or via A/B testing [50]. For example, in an internet-based applications one may perform A/B testing by deploying two classifiers A and B with two different sub-populations of users and use their level of engagement to decide which of the two classifiers is preferred. For other applications, we may present to the user, visualizations of the predictive rates for two different classifiers (e.g., [55, 3]), and have the user provide pairwise feedback.

Since the metrics we consider are functions of only the predictive rates, queries comparing classifiers are the same as queries on the associated rates. So for convenience, we will have our

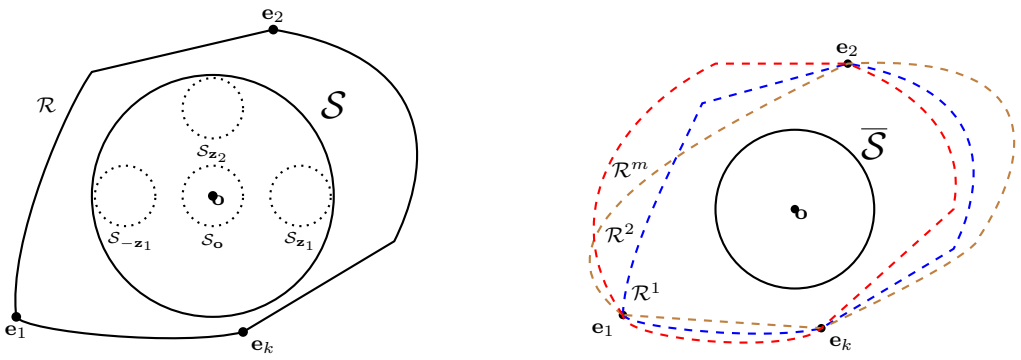

Figure 2: (a) Geometry of set of predictive rates $\mathcal{R}$: A convex set enclosing a sphere $\mathcal{S}$ with trivial rates $\mathbf{e}_i \,\forall\, i \in [k]$ as vertices; (b) Geometry of the product set of group rates $\mathcal{R}^1 \times \cdots \times \mathcal{R}^m$ (best seen in color) [18]; $\mathcal{R}^u \,\forall\, u \in [m]$ are convex sets with common vertices $\mathbf{e}_i \,\forall\, i \in [k]$ and enclose a sphere $\overline{\mathcal{S}} \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$.

algorithms pose queries comparing two (feasible) rates, but they can be equivalently seen as comparing two classifiers. Indeed given a feasible rate, one can efficiently find the associated classifier (see Appendix B.1 for details). We next formally state the ME problem.

Definition 2 (Metric Elicitation with Pairwise Queries (given $\{(\mathbf{x}, y)_i\}_{i=1}^n$) [16, 17])**.** *Suppose that the oracle's (unknown) performance metric is ϕ. Using oracle queries of the form $\Omega(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2\,;\,\phi)$, where $\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2$ are the estimated rates from samples, recover a metric $\hat{\phi}$ such that $\|\phi - \hat{\phi}\| < \kappa$ under a suitable norm $\|\cdot\|$ for sufficiently small error tolerance $\kappa > 0$.*

The performance of ME is evaluated both by the query complexity and the quality of the elicited metric [16, 17]. As is standard in the decision theory literature [29, 16, 17], we present our ME approach by first assuming access to population quantities such as the population rates $\mathbf{r}(h, \mathbb{P})$, then examine estimation error from finite samples, i.e., with empirical rates $\hat{\mathbf{r}}(h, \{(\mathbf{x}, y)_i\}_{i=1}^n)$.

2.2 Linear Performance Metric Elicitation

As a warm up, we give a brief overview of the Linear Performance Metric Elicitation (LPME) procedure of [17], which we will use as a subroutine while eliciting quadratic metrics. Here we assume that the oracle's metric is a linear function of rates $\phi^{\text{lin}}(\mathbf{r}) \coloneqq \langle \mathbf{a}, \mathbf{r} \rangle$, for some unknown costs $\mathbf{a} \in \mathbf{R}^q$ with $\|\mathbf{a}\|_2 = 1$ (wlog. due to scale invariance). In other words, when provided two rate vectors $\mathbf{r}_1$ and $\mathbf{r}_2$, the oracle returns $\mathbb{1}[\langle \mathbf{a}, \mathbf{r}_1 \rangle > \langle \mathbf{a}, \mathbf{r}_2 \rangle]$. The goal is to elicit $\mathbf{a}$ using pairwise queries.

When the number of classes $k = 2$, the coefficients $\mathbf{a}$ can be elicited using a simple one-dimensional binary search. When $k > 2$, one can apply a coordinate-wise procedure, performing a binary search in one coordinate, while keeping the others fixed. The efficacy of this procedure, however, hinges on the geometry of the underlying set of feasible rates $\mathcal{R}$, which we discuss below.

We first make a mild assumption which ensures that there is some signal for non-trivial classification [17].

Assumption 1. *The conditional-class distributions are distinct, i.e.,* $\forall i \neq j$, $P(Y = i|X) \neq P(Y = j|X)$.

Let $\mathbf{e}_i \in \{0, 1\}^q$ denote the rate profile achieved by a trivial classifier that predicts class i for all inputs.

Proposition 1 (Geometry of $\mathcal{R}$; Figure 2(a))**.** *The set of rates* $\mathcal{R} \subseteq [0, 1]^q$ *is convex, has vertices* $\{\mathbf{e}_i\}_{i=1}^k$*, and contains the rate profile* $\mathbf{o} = \frac{1}{k}\sum_{i=1}^k \mathbf{e}_i$ *in the interior. Moreover,* $\mathbf{o}$ *is achieved by a classifier which for any input predicts each class with equal probability.*

Remark 1 (Existence of sphere $\mathcal{S}$)**.** *Since* $\mathcal{R}$ *is convex and contains the point* $\mathbf{o}$ *in the interior, there exists a sphere* $\mathcal{S} \subset \mathcal{R}$ *of non-zero radius* ρ *centered at* $\mathbf{o}$.

By restricting the coordinate-wise binary search procedure to posing queries from within a sphere, LPME can be equivalently seen as minimizing a strongly-convex function and shown to converge to a solution $\hat{\mathbf{a}}$ close to $\mathbf{a}$. Specifically, the LPME procedure takes any sphere $\mathcal{S} \subset \mathcal{R}$, binary-search tolerance ϵ, and the oracle Ω (with metric ϕ^{lin}) as input, and by posing $O(q \log(1/\epsilon))$ queries recovers coefficients $\hat{\mathbf{a}}$ with $\|\mathbf{a} - \hat{\mathbf{a}}\|_2 \leq O(\sqrt{q}\epsilon)$. The details can be found in Algorithm 2 in [17] and are also provided in Appendix A for completeness.

Remark 2 (LPME Guarantee)**.** *Given* any *q-dimensional sphere* $\mathcal{S} \subset \mathcal{R}$ *and an oracle* Ω *for an unknown metric* $\phi^{\text{lin}}(\mathbf{r}) := \langle \mathbf{a}, \mathbf{r} \rangle$*, the LPME 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.,* $a_i/a_j \approx \hat{a}_i/\hat{a}_j \; \forall \; i, j \in [q]$.

The algorithm estimates the direction (slope) of the coefficient vector $\mathbf{a}$, and not its magnitude. Also note the algorithm takes as input an *arbitrary* sphere $\mathcal{S} \subset \mathcal{R}$, and restricts its queries to rate vectors within the sphere. In Appendix B.1, we discuss an efficient procedure [17] for identifying a sphere of suitable radius.

2.3 Quadratic Performance Metrics

Equipped with the LPME subroutine, our aim is to elicit metrics that are quadratic functions of rates.

Definition 3 (Quadratic Metric)**.** *For a vector* $\mathbf{a} \in \mathbb{R}^q$ *and a symmetric matrix* $\mathbf{B} \in \mathbb{R}^{q \times q}$ *with* $\|\mathbf{a}\|_2 + 1/2\|\mathbf{B}\|_F = 1$ *(wlog. due to scale invariance):*

$$\phi^{\text{quad}}(\mathbf{r}\,;\,\mathbf{a}, \mathbf{B}) = \langle \mathbf{a}, \mathbf{r} \rangle + \frac{1}{2}\mathbf{r}^T\mathbf{B}\mathbf{r}. \tag{4}$$

This family trivially includes the linear metrics discussed in the previous section [49, 16, 17] as well as many modern complex metrics outlined below:

Example 1 (Class-imbalanced learning)**.** In problems with imbalanced class proportions, it is common to use evaluation metrics that emphasize equal performance across all classes. One such metric is Q-mean [30, 33, 35], which is the quadratic mean of the normalized class errors: $\phi^{\text{qmean}}(\mathbf{r}) = 1/k \sum_{i=1}^k \left(\sum_{j=1}^{k-1} r_{(i-1)(k-1)+j}\right)^2$.

Example 2 (Distribution matching)**.** In certain applications, one requires the proportion of predictions made by a classifier for each class (i.e., the coverage) to match a target distribution $\boldsymbol{\pi} \in \Delta_k$ [13, 38, 39, 7]. One evaluation measure used for this task is the squared difference between the per-class coverage and the target distribution:

$$\phi^{\text{cov}}(\mathbf{r}) = \sum_{i=1}^{k} \left(\text{cov}_i(\mathbf{r}) - \pi_i\right)^2,$$

where

$$\text{cov}_i(\mathbf{r}) = 1 - \sum_{j=1}^{k-1} r_{(i-1)(k-1)+j} + \sum_{j>i} r_{(j-1)(k-1)+i} + \sum_{j<i} r_{(j-1)(k-1)+i-1}.$$

Similar quadratic metrics can be found in the quantification literature where the target π_i is set to the class prior $\mathbb{P}(Y = i)$ [11, 12, 25]. Our setup also captures more general quadratic distance measures between distributions, e.g. $(\text{cov}(\mathbf{r}) - \boldsymbol{\pi})^{\mathbf{T}}\mathbf{Q}(\text{cov}(\mathbf{r}) - \boldsymbol{\pi})$ for a positive semi-definite matrix $\mathbf{Q} \in PSD_k$ [32].

Example 3. (Fairness violation) A popular criterion for group-based fairness is equalized odds, which requires equal predictive rates across different protected groups [14, 4]. The equalized odds violation can be measured by the squared differences between the rates for each group. With m groups and $\mathbf{r}^g$ denoting the rate vector evaluated on examples from group g, this is given by: $\phi^{\text{EO}}((\mathbf{r}^1, \ldots, \mathbf{r}^m)) = \sum_{v>u} \sum_{i=1}^{q} \left(r_i^u - r_i^v\right)^2$.

Other fairness criteria for two classes that can be written as a quadratic penalty include equal opportunity $\phi^{\text{EOpp}}((\mathbf{r}^1, \ldots, \mathbf{r}^m)) = \sum_{v>u}(r_1^u - r_1^v)^2$ [14], balance for the negative class $\phi^{\text{BN}}((\mathbf{r}^1, \ldots, \mathbf{r}^m)) = (r_2^u - r_2^v)^2$ [28], error-rate balance $\phi^{\text{EB}}((\mathbf{r}^1, \ldots, \mathbf{r}^m)) = 0.5 \sum_{v>u}(r_1^u - r_1^v)^2 + (r_2^u - r_2^v)^2$ [5], etc. and their weighted variants. In Section 4, we consider metrics that trade-off between a weighted error term and a quadratic fairness term.

We will need the following assumption on the metric.

Assumption 2. *The largest singular value* $\sigma_{\max}$ *of* $\mathbf{B}$ *is bounded. In other words, the* ϕ^{quad} *is* $\sigma_{\max}$*-smooth. Further, the gradient of* ϕ *at the trivial rate* $\mathbf{o}$ *is non-zero, i.e.,* $\nabla\phi^{\text{quad}}(\mathbf{r})|_{\mathbf{r}=\mathbf{o}} = \mathbf{a} + \mathbf{B}\mathbf{o} \neq 0$.

The smoothness assumption implies that ϕ^{quad} is locally linear around a given rate. The non-zero gradient assumption is also very reasonable for a convex ϕ^{quad}, where it merely implies that the optimal classifier for the metric is not the uniform random classifier.

3 Quadratic Performance Metric Elicitation

We are now ready to present our approach for Quadratic Performance Metric Elicitation (QPME). Here we assume that the oracle's unknown metric is the quadratic metric (Definition 3) and seek to estimate its parameters $(\mathbf{a}, \mathbf{B})$ by posing queries to the oracle.

The idea is to approximate the metric at a given rate vector by a linear function and use the LPME routine to estimate the local slope. This can be done by restricting LPME to a small sphere $\mathcal{S}$ around the given point. The challenge, of course, is to pick a small number of points to perform this local approximation and to reconstruct the original metric from the estimated local slopes.

3.1 Local Linear Approximation

We will find it convenient to work with a shifted version of the quadratic metric, centered at the point $\mathbf{o}$, the uniform random rate vector (see Proposition 1):

$$\begin{aligned}\phi^{\text{quad}}(\mathbf{r};\, \mathbf{a}, \mathbf{B}) &= \langle \mathbf{d}, \mathbf{r} - \mathbf{o} \rangle + \frac{1}{2}(\mathbf{r} - \mathbf{o})^T \mathbf{B}(\mathbf{r} - \mathbf{o}) + c \\ &= \overline{\phi}(\mathbf{r};\, \mathbf{d}, \mathbf{B}) + c, \end{aligned} \tag{5}$$

where $\mathbf{d} = \mathbf{a} + \mathbf{B}\mathbf{o}$ and c is a constant independent of $\mathbf{r}$, and so the oracle can be equivalently seen as responding with the shifted metric $\overline{\phi}(\mathbf{r};\, \mathbf{d}, \mathbf{B})$.

Let z be a fixed point in $\mathcal{R}$. The smoothness property of the metric (from Assumption 2) gives us that in a small neighborhood around $\mathbf{z}$, the metric can be closely approximated by its first-order Taylor expansion, i.e.,

$$\overline{\phi}(\mathbf{r};\, \mathbf{d}, \mathbf{B}) \approx \langle \mathbf{d} + \mathbf{B}(\mathbf{z} - \mathbf{o}), \mathbf{r} \rangle + c', \tag{6}$$

for a constant c'. So if we apply LPME to the metric $\overline{\phi}$ with the queries $(\mathbf{r}_1, \mathbf{r}_2)$ to the oracle restricted to a small ball around $\mathbf{z}$, the procedure effectively estimates the slope of the vector $\mathbf{d} + \mathbf{B}(\mathbf{z} - \mathbf{o})$ in the above linear function (up to a small approximation error).

We next show how we use this idea of applying LPME to small neighborhoods around selected points to elicit the coefficients $\mathbf{a}$ and $\mathbf{B}$ for the original metric in (4). For simplicity, we will assume that the oracle's feedback is noise-free and later show robustness to noise and the query complexity guarantees in Section 5.

3.2 Eliciting Metric Coefficients

We outline the main steps (see Algorithm 1) below:

Estimate coefficients d (Line 1). We first wish to estimate the linear portion $\mathbf{d}$ of the metric $\overline{\phi}$ in (5). For this, we apply the LPME subroutine to a small ball $\mathcal{S}_{\mathbf{o}} \subset \mathcal{S}$ of radius $\varrho < \rho$ around the point $\mathbf{o}$. See Figure 2(a) for an illustration. Within this ball, the metric $\overline{\phi}$ approximately equals the linear function $\langle \mathbf{d}, \mathbf{r} \rangle + c'$ using (6), and so the subroutine's output gives us an estimate of the slope of $\mathbf{d}$. Specifically, we have from Remark 2 that the returned estimates $\mathbf{f}_0 = (f_{10}, \ldots, f_{q0})$ approximately satisfy the following $(q-1)$ equations:

$$\frac{d_i}{d_1} = \frac{f_{i0}}{f_{10}} \qquad \forall\, i \in \{2, \ldots, q\}. \tag{7}$$

Estimate coefficients B (Lines 2–4). Next, we wish to estimate each column of the matrix $\mathbf{B}$ of the metric $\overline{\phi}$ in (5). For this, we apply LPME to small neighborhoods around points in the direction of standard basis vectors $\boldsymbol{\alpha}_j \in \mathbb{R}^q$, $j = 1, \ldots, q$. Note that within a small ball around $\mathbf{o} + \boldsymbol{\alpha}_j$, the metric $\overline{\phi}$ is approximately the linear function $\langle \mathbf{d} + \mathbf{B}_{:,j}, \mathbf{r} \rangle + c'$, and so the LPME procedure when applied to this region will give us an estimate of the slope of $\mathbf{d} + \mathbf{B}_{:,j}$. However, to ensure that the center point we choose is a feasible rate, we will have to re-scale the standard basis, and apply the subroutine to balls $\mathcal{S}_{\mathbf{z}_j}$ of radius $\varrho < \rho$

Algorithm 1: QPM Elicitation

Input: $\mathcal{S}$, Search tolerance $\epsilon > 0$, Oracle Ω with metric $\bar{\phi}$

1: $\mathbf{f}_0 \leftarrow \text{LPME}(\mathcal{S}_\mathbf{o}, \epsilon, \Omega)$ with $\mathcal{S}_\mathbf{o} \subset \mathcal{S}$ and obtain (7)

2: **For** $j \in \{1, 2, \ldots, q\}$ **do**

3: $\quad \mathbf{f}_j \leftarrow \text{LPME}\left(\mathcal{S}_{\mathbf{z}_j}, \epsilon, \Omega\right)$ with $\mathcal{S}_{\mathbf{z}_j} \subset \mathcal{S}$ and obtain (8)

4: $\mathbf{f}_1^- \leftarrow \text{LPME}(\mathcal{S}_{-\mathbf{z}_1}, \epsilon, \Omega)$ with $\mathcal{S}_{-\mathbf{z}_1} \subset \mathcal{S}$ and obtain (9)

5: $\hat{\mathbf{a}}, \hat{\mathbf{B}} \leftarrow$ normalized solution dervied from (10)

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

centered at $\mathbf{z}_j = \mathbf{o} + (\rho - \varrho)\boldsymbol{\alpha}_j$. See Figure 2(a) for the visual intuition. The returned estimates $\mathbf{f}_j = (f_{1j}, \ldots, f_{qj})$ approximately satisfy:

$$\frac{d_i + (\rho - \varrho)B_{ij}}{d_1 + (\rho - \varrho)B_{1j}} = \frac{f_{ij}}{f_{1j}} \quad \forall\, i \in \{2, \ldots, q\},\ j \leq i. \tag{8}$$

Since the matrix $\mathbf{B}$ is symmetric, this gives us $q(q+1)/2$ equations. There are $q(q+1)/2 + q$ unknown entities in $\mathbf{a}$ and $\mathbf{B}$, and to estimate them we need 1 more equation beside the normalization condition. For this, we apply LPME to a sphere $\mathcal{S}_{-\mathbf{z}_1}$ of radius ϱ around rate $-\mathbf{z}_1$ as shown in Figure 2(a). The returned slopes $\mathbf{f}_1^- = (f_{11}^-, \ldots, f_{q1}^-)$ approximately satisfy:

$$\frac{d_2 - (\rho - \varrho)B_{21}}{d_1 - (\rho - \varrho)B_{11}} = \frac{f_{21}^-}{f_{11}^-}. \tag{9}$$

Putting it together (Line 5). By combining (7), (8) and (9), we express each entry of $\mathbf{B}$ in terms of d_1:

$$B_{ij} = \left(F_{i,1,j}(1 + F_{j,1,1}) - F_{i,1,j}F_{j,1,0}d_1 - F_{i,1,0} + F_{i,1,j}\frac{F_{2,1,1}^- + F_{2,1,1} - 2F_{2,1,0}}{F_{2,1,1}^- - F_{2,1,1}}\right)d_1, \tag{10}$$

where $F_{i,j,l} = \frac{f_{il}}{f_{jl}}$ and $F_{i,j,l}^- = \frac{f_{il}^-}{f_{jl}^-}$. Using $\mathbf{d} = \mathbf{a} + \mathbf{B}\mathbf{o}$ and the fact that the coefficients are normalized, i.e., $\|\mathbf{a}\|_2 + \frac{1}{2}\|\mathbf{B}\|_F = 1$, we can obtain estimates for $\mathbf{B}$ and $\mathbf{a}$ independent of d_1. See Appendix C for the details.

The derivation so far assumes $d_1 \neq 0$. This is based on Assumption 2 which states that at least one coordinate of $\mathbf{d}$ is non-zero, and we've assumed w.l.o.g. that this is d_1. In practice, we can identify a non-zero coordinate using q queries of the form $(\varrho\boldsymbol{\alpha}_i + \mathbf{o}, \mathbf{o}), \forall i \in [q]$.

4 Application to Quadratic Fairness Metrics

We next describe an application of our quadratic elicitation strategy to *fairness in machine learning*. We consider the setup in Hiranandani et al. (2020) [18], where the goal is to elicit a metric that trades-off between a predictive performance and fairness violation [23, 14, 5, 4, 36], and extend their approach to handle general quadratic fairness metrics.

4.1 Fairness Preliminaries

We consider a k-class problem comprising m groups and use $g \in [m]$ to denote the group membership. The groups are assumed to be disjoint, fixed, and known apriori [14, 1, 2]. We

have access to a dataset of size n denoted by $\{(\mathbf{x}, g, y)_i\}_{i=1}^n$, generated *iid* from a distribution $\mathbb{P}(X, G, Y)$. In this case, we will work with a separate (randomized) classifiers $h^g : \mathcal{X} \to \Delta_k$ for each group g, and use $\mathcal{H}^g = \{h^g : \mathcal{X} \to \Delta_k\}$ to denote the set of all classifiers for a group g.

Group predictive rates: Similar to (1), we denote the group-conditional rate matrix for a classifier h^g by $\mathbf{R}^g(h^g, \mathbb{P}) \in \mathbb{R}^{k \times k}$, where the ij-th entry is given by:

$$R^g_{ij}(h^g, \mathbb{P}) \coloneqq \mathbb{P}(h^g = j | Y = i, G = g) \, \forall \, i, j \in [k]. \tag{11}$$

Similar to (2), we denote the group rates by vectors $\mathbf{r}^g(h^g, \mathbb{P}) = \textit{off-diag}(\mathbf{R}^g(h^g, \mathbb{P}))$, and the set of feasible rates for group g by $\mathcal{R}^g = \{\mathbf{r}^g(h^g, \mathbb{P}) \, : \, h^g \in \mathcal{H}^g\}$.

Rates for overall classifier: We construct the overall classifier $h : (\mathcal{X}, [m]) \to \Delta_k$ by predicting with classifier h^g for group g, i.e. $h(\mathbf{x}, g) \coloneqq h^g(\mathbf{x})$. We will be interested in both the predictive performance of the overall classifier and its fairness violation. For the former, we will measure the overall rate matrix for h (1):

$$R_{ij} \coloneqq \mathbb{P}(h = j | Y = i) = \textstyle\sum_{g=1}^m t^g_i R^g_{ij}, \tag{12}$$

where $t^g_i \coloneqq \mathbb{P}(G = g | Y = i)$ is the prevalence of group g within class i. For the latter, we will need the m group-specific rates, denoted together as a tuple:

$$\mathbf{r}^{1:m} \coloneqq (\mathbf{r}^1, \ldots, \mathbf{r}^m) \in \mathcal{R}^1 \times \cdots \times \mathcal{R}^m =: \mathcal{R}^{1:m}.$$

Lastly, the overall rates in (12) can be succinctly written as a flattened vector $\mathbf{r} \in [0, 1]^q$, and can be expressed in terms of the group-specific rates as $\mathbf{r} = \sum_{g=1}^m \boldsymbol{\tau}^g \odot \mathbf{r}^g$, where $\boldsymbol{\tau}^g \coloneqq \textit{off-diag}([\mathbf{t}^g \; \mathbf{t}^g \; \ldots \; \mathbf{t}^g])$.

4.2 Fair (Quadratic) Metric Elicitation

We seek to elicit a performance metric which trades-off between predictive performance defined by a linear function of the overall rates $\mathbf{r}$, and fairness violation defined by a quadratic function of the group rates $\mathbf{r}^{1:m}$.

Definition 4. (Fair (Quadratic) Performance Metric) *For misclassification costs* $\mathbf{a} \in \mathbb{R}^q$, $\mathbf{a} \geq 0$, *fairness violation costs* $\mathbb{B} = \{\mathbf{B}^{uv} \in PSD_q\}_{u,v=1,v>u}^m$, *and a trade-off parameter* $\lambda \in [0, 1]$, *we define:*

$$\phi^{\text{fair}}(\mathbf{r}^{1:m}; \mathbf{a}, \mathbb{B}, \lambda) \; \coloneqq \; (1 - \lambda)\langle \mathbf{a}, \mathbf{r} \rangle + \lambda \frac{1}{2} \left(\textstyle\sum_{v>u} (\mathbf{r}^u - \mathbf{r}^v)^T \mathbf{B}^{uv} (\mathbf{r}^u - \mathbf{r}^v) \right), \tag{13}$$

where the parameters $\mathbf{a}$ *and* $\mathbf{B}^{uv}$*'s are normalized:* $\|\mathbf{a}\|_2 = 1$, $\frac{1}{2} \sum_{v>u}^m \|\mathbf{B}^{uv}\|_F = 1$.

Hiranandani et al. [18] consider a similar setup, but only handle fairness terms that are *linear* in the (absolute) group discrepancies. We allow for more general quadratic violation terms, and enable a practitioner to specify more nuanced fairness criteria. See Section 2.3 for examples of quadratic fairness metrics.

The coefficients $\mathbf{a}, \mathbf{B}^{uv}$'s are separately normalized so that the predictive performance and fairness violation are in the same scale, and we can additionally elicit the trade-off parameter λ. Analogous to Definition 1–2, we present the problem of fair quadratic metric elicitation.

Definition 5 (Fair Quadratic Metric Elicitation with Pairwise Queries (given $\{(\mathbf{x}, g, y)_i\}_{i=1}^n$) [18])**.** *Let Ω be an oracle for the (unknown) metric ϕ^{fair}, which for any given $\mathbf{r}_1^{1:m}, \mathbf{r}_2^{1:m}$, outputs $\Omega(\mathbf{r}_1^{1:m}, \mathbf{r}_2^{1:m}) = \mathbb{1}[\phi^{\text{fair}}(\mathbf{r}_1^{1:m}) > \phi^{\text{fair}}(\mathbf{r}_2^{1:m})]$. Using oracle queries of the form $\Omega(\hat{\mathbf{r}}_1^{1:m}, \hat{\mathbf{r}}_2^{1:m})$, where $\hat{\mathbf{r}}_1^{1:m}, \hat{\mathbf{r}}_2^{1:m}$ are the estimated rates from samples, recover a metric $\hat{\phi}^{\text{fair}} = (\hat{\mathbf{a}}, \hat{\mathbb{B}}, \hat{\lambda})$ such that $\|\phi^{\text{fair}} - \hat{\phi}^{\text{fair}}\| < \kappa$ under a suitable norm $\|\cdot\|$ for sufficiently small error tolerance $\kappa > 0$.*

Similar to Section 2.2, we make observations about the space of feasible rates $\mathcal{R}^{1:m} = \mathcal{R}^1 \times \cdots \times \mathcal{R}^m$. We begin with a mild distributional assumption.

Assumption 3. *For each group $g \in [m]$, the conditional-class distributions $P(Y = j|X, G = g)$, $j = 1, \ldots, q$, are not identical, i.e. there is some signal for non-trivial classification for each group [18].*

Proposition 2 (Geometry of $\mathcal{R}^{1:m}$; Figure 2(b))**.** *For each group g, a trivial classifier which predicts class i on all inputs results in the same rate vector $\mathbf{e}_i$. The rate space $\mathcal{R}^g$ for each group g is convex and so is the intersection $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$. Moreover, the intersection contains the rate profile $\mathbf{o} = \frac{1}{k}\sum_{i=1}^k \mathbf{e}_i$ (defined by the uniform random classifier) in the interior.*

Remark 3 (Existence of sphere $\overline{\mathcal{S}}$ in $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$)**.** *There exists a sphere $\overline{\mathcal{S}} \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ of non-zero radius ρ centered at $\mathbf{o}$. Thus, a rate $\mathbf{s} \in \overline{\mathcal{S}}$ is feasible for each of the m groups, i.e. $\mathbf{s}$ is achievable by some classifier h^g for each group $g \in [m]$.*

Because we allow separate classifiers for the m groups, the above remark implies that a rate profile of the form $\mathbf{r}^{1:m} = (\mathbf{s}^1, \ldots, \mathbf{s}^m)$ for arbitrary points $\mathbf{s}^1, \ldots, \mathbf{s}^m \in \overline{\mathcal{S}}$ is achievable for some choice of group-specific classifiers $h^1, \ldots, h^m$. We will find this observation useful in the elicitation algorithm we describe next.

4.3 Eliciting Metric Parameters $(\mathbf{a}, \mathbb{B}, \lambda)$

We present an elicitation strategy for the fair quadratic metric in Definition 4 by adapting the QPME algorithm from Section 3. For easy exposition, we focus on a setting with two groups, i.e. $m = 2$, and extend our approach to multiple groups in Appendix D.

Observe that for a rate profile $\mathbf{r}^{1:2} = (\mathbf{s}, \mathbf{o})$, where the first group is assigned an arbitrary point in $\overline{\mathcal{S}}$ and the second group is assigned the uniform random classifier's rate $\mathbf{o}$, the fair quadratic metric (13) becomes:

$$\begin{aligned}
&\phi^{\text{fair}}((\mathbf{s}, \mathbf{o}); \mathbf{a}, \mathbf{B}^{12}, \lambda) \\
&\coloneqq (1-\lambda)\langle \mathbf{a}, \boldsymbol{\tau}^1 \odot \mathbf{s} + \boldsymbol{\tau}^2 \odot \mathbf{o}\rangle + \frac{\lambda}{2}(\mathbf{s}-\mathbf{o})^T\mathbf{B}^{12}(\mathbf{s}-\mathbf{o}) \\
&\coloneqq \langle \mathbf{d}, \mathbf{s}-\mathbf{o}\rangle + \frac{1}{2}(\mathbf{s}-\mathbf{o})^T\mathbf{B}(\mathbf{s}-\mathbf{o}) \\
&\coloneqq \overline{\phi}(\mathbf{s}; \mathbf{d}, \mathbf{B}),
\end{aligned} \tag{14}$$

where $\mathbf{d} = (1-\lambda)\boldsymbol{\tau}^1 \odot \mathbf{a}$ and $\mathbf{B} = \lambda\mathbf{B}^{12}$, and we use $\boldsymbol{\tau}^1 + \boldsymbol{\tau}^2 = \mathbf{1}$ (the vector of ones) for the second step.

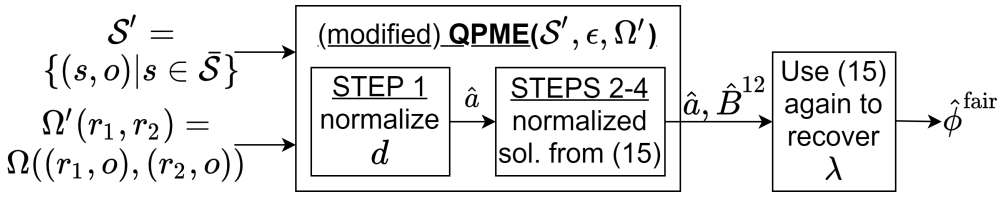

Figure 3: Eliciting Fair Quadratic Metrics (Definition 5) for two groups. We formulate a q-dimensional elicitation problem and use a variant of QPME (Algorithm 1).

The metric $\overline{\phi}$ above is a particular instance of the quadratic metric in (5) in q dimensions. We can thus apply a slight variant of the QPME procedure in Algorithm 1 to solve the q-dimensional quadratic metric elicitation problem over the sphere $\mathcal{S}' = \{(\mathbf{s}, \mathbf{o}) \,|\, \mathbf{s} \in \overline{\mathcal{S}}\}$ with the modified oracle $\Omega'(\mathbf{r}_1, \mathbf{r}_2) = \Omega((\mathbf{r}_1, \mathbf{o}), (\mathbf{r}_2, \mathbf{o}))$.

The only change needed to be made to the algorithm is in line 5, where we need to take into account the changed relationship between $\mathbf{d}$ and $\mathbf{a}$, and need to separately (not jointly) normalize the linear and quadratic coefficients. With this change, the output of the algorithm directly gives us the required estimates. Specifically, we have from step 1 of Algorithm 1 and (7) an estimate $\hat{d}_i = (1-\lambda)\tau_i^1 \hat{a}_i$. By normalizing $\mathbf{d}$, we directly get an estimate $\hat{\mathbf{a}} = \frac{\mathbf{d}}{\|\mathbf{d}\|}$ for the linear coefficients. Similarly, steps 2-4 of Algorithm 1 and (10) gives us:

$$\begin{aligned}\hat{B}_{ij} &= \lambda \hat{B}_{ij}^{12} \\ &= \left(F_{i,1,j}(1 + F_{j,1,1}) - F_{i,1,j}F_{j,1,0}d_1 - F_{i,1,0} + F_{i,1,j}\frac{F^-_{2,1,1} + F_{2,1,1} - 2F_{2,1,0}}{F^-_{2,1,1} - F_{2,1,1}}\right)(1-\lambda)\tau_1^1 \hat{a}_1. \qquad (15)\end{aligned}$$

Again by normalizing we directly get estimates $\hat{\mathbf{B}}^{12} = \hat{\mathbf{B}}/\|\hat{\mathbf{B}}\|_F$ for the quadratic coefficients.

Finally, because the linear and quadratic coefficients are separately normalized, the estimates $\hat{\mathbf{a}}$, $\hat{\mathbf{B}}^{12}$ are independent of the trade-off parameter λ. Given estimates $\hat{B}_{ij}^{12}$ and $\hat{a}_1$, we can now additionally estimate the trade-off parameter $\hat{\lambda}$ from equation (15). See Figure 3 for an illustration of the procedure.

The proposed approach for the fair (quadratic) performance metric elicitation easily extends to multiple groups by applying the QPME procedure described above multiple times after fixing one cluster of groups to the point $\mathbf{o}$ and the remaining to the same point $\mathbf{s}$ in the intersection sphere $\overline{\mathcal{S}}$. See Appendix D for details.

In Appendix D.1, we also provide an alternate binary search based method similar to Hiranandani et al. [18] for eliciting the trade-off parameter λ when the linear predictive and quadratic fairness coefficients are already known. This is along similar lines to the application considered by Zhang et al. [55], but unlike them, instead of ratio queries, we require simpler pairwise queries.

5 Guarantees

We discuss elicitation guarantees for the QPME procedure (Algorithm 1) under the following feedback model, which is useful in practice. The guarantees for the fair metric elicitation follow directly as a consequence.

Definition 6 (Oracle Feedback Noise: $\epsilon_\Omega \geq 0$)**.** *Given rates* $\mathbf{r}_1, \mathbf{r}_2$*, the oracle responds correctly iff* $|\phi^{\text{quad}}(\mathbf{r}_1) - \phi^{\text{quad}}(\mathbf{r}_2)| > \epsilon_\Omega$ *and may be incorrect otherwise.*

In words, the oracle may respond incorrectly if the rates are very close as measured by the quadratic metric ϕ^{quad}. Also, since deriving the final metric involves offline computations including certain ratios, we discuss guarantees under the following regularity assumption that ensures all components are well defined.

Assumption 4. *For the shifted quadratic metric* $\bar{\phi}$ *in* (5)*, the gradients at the rate profiles* $\mathbf{o}$*,* $-\mathbf{z}_1$*, and* $\{\mathbf{z}_1, \ldots, \mathbf{z}_q\}$*, are element-wise non-zero vectors. More formally,* $\exists$ *constants* $c_0, c_{-1}, c_1, \ldots c_q$ *s.t.* $\min_i |d_i| > c_0$*,* $\min_i |(d - B(z_1 - o))_i| > c_{-1}$*, and* $\min_i |(d + B(z_j - o))_i| > c_j \; \forall \; j \in [q]$*. Additionally,* $\rho > \varrho \gg \epsilon_\Omega$*.*

Theorem 1. *Given* $\epsilon, \epsilon_\Omega \geq 0$*, and a 1-Lipschitz metric* ϕ^{quad} *(Definition 3) parametrized by* $\mathbf{a}, \mathbf{B}$*, under Assumptions 1, 2 and 4, after* $O\left(q^2 \log \frac{1}{\epsilon}\right)$ *queries Algorithm 1 returns a metric* $\hat{\phi}^{\text{quad}} = (\hat{\mathbf{a}}, \hat{\mathbf{B}})$ *such that:*

- $\|\mathbf{a} - \hat{\mathbf{a}}\|_2 \leq O\left(q(\epsilon + \sqrt{\sigma_{max} + \epsilon_\Omega/\varrho})\right)$.
- $\|\mathbf{B} - \hat{\mathbf{B}}\|_F \leq O\left(q\sqrt{q}(\epsilon + \sqrt{\sigma_{max} + \epsilon_\Omega/\varrho})\right)$.

The theorem shows that the QPME procedure is robust to noise, and its query complexity depends only *linearly* in the number of unknown entities.

We stress that despite eliciting a more complex (nonlinear) metric, the query complexity is of the same order as prior methods for linear elicitation [16, 17]. Moreover, since sample estimates of rates are consistent estimators, and the metrics discussed are 1-Lipschitz w.r.t. rates, with high probability, we gather correct oracle feedback from querying with finite sample estimates $\Omega(\hat{\mathbf{r}}_1, \hat{\mathbf{r}}_2)$ instead of querying with population statistics $\Omega(\mathbf{r}_1, \mathbf{r}_2)$, as long as we have sufficient samples. Other than this, Algorithm 1 is agnostic to finite sample errors as long as the sphere $\mathcal{S}$ is in the feasible space $\mathcal{R}$.

6 Experiments

We empirically evaluate our approach on simulated oracles. We first present results on a synthetically generated query space in Section 6.1 and then include results on real-world datasets in Section 6.2 . We run our elicitation procedures with tolerance $\epsilon = 10^{-2}$.

6.1 Recovery Quality

Eliciting quadratic metrics. We first apply QPME (Algorithm 1) to elicit quadratic metrics in Definition 3. We assume access to a q-dimensional sphere $\mathcal{S}$ centered at rate $\mathbf{o}$ with radius $\rho = 0.2$, from which we query rate vectors $\mathbf{r}$. Recall that in practice, Remark 1 guarantees the existence of such a sphere within the feasible region $\mathcal{R}$. We randomly generate quadratic metrics ϕ^{quad} parametrized by $(\mathbf{a}, \mathbf{B})$ and repeat the experiment over 100 trials for varying

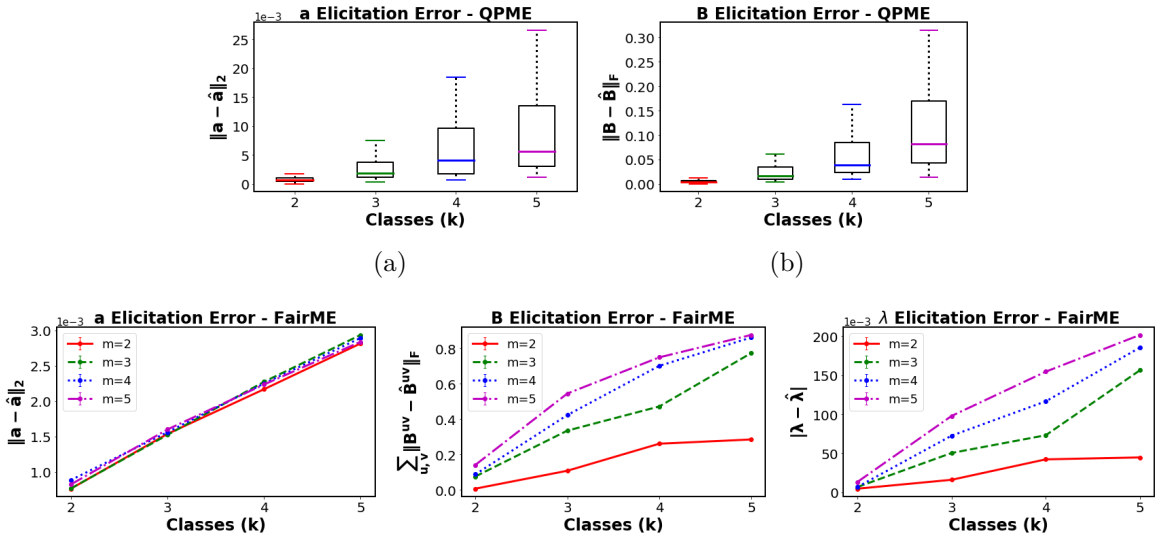

Figure 4: Elicitation error as a function of number of classes k and groups m for quadratic metrics in Definition 3 (a–b) and fairness metrics in Definition 4 (c–e). The results are averaged over 100 random metrics.

numbers of classes $k \in \{2, 3, 4, 5\}$ (equiv. $q \in \{2, 6, 12, 20\}$). In Figures 4(a)–4(b), we show box plots [34] of the ℓ_2 (Frobenius) norm between the true and elicited linear (quadratic) coefficients. We generally find that QPME is able to elicit metrics close to the true ones. This holds for varying k, showing the effectiveness of our approach in handling multiclass metrics. The larger standard deviation for $k = 5$ is due to Assumption 4 failing to hold in a small number of trials and the resulting coefficient estimates not being as accurate. We discuss this in detail in Appendix F.1.

Eliciting fairness metrics. We next apply the elicitation procedure in Figure 3 to elicit the fairness metrics in Definition 4. We randomly generate oracle metrics ϕ^{fair} parametrized by $(\mathbf{a}, \mathbb{B}, \lambda)$ and repeat the experiment over 100 trials and with different number of classes and groups $k, m \in \{2, 3, 4, 5\}$. Figures 4(c)–4(e) show the mean elicitation errors for the the three elicited parameters. For the linear predictive performance term, the error $\|\mathbf{a} - \hat{\mathbf{a}}\|_2$ increases only with the number of classes k and not groups m, as this term is independent of the number of groups. For the quadratic violation term, the error $\sum_{u,v} \|\mathbf{B}^{uv} - \hat{\mathbf{B}}^{uv}\|_F$ increases with both the number of classes k and groups m. This is because the QPME procedure is run $\binom{m}{2}$ times for eliciting $\binom{m}{2}$ matrices $\{\mathbf{B}^{uv}\}_{v>u}$, and so the elicitation error accumulates with increasing m. Lastly, the elicited trade-off $\hat{\lambda}$ is seen to be close to the true λ as well.

6.2 Ranking of Real-World Classifiers

Performance metrics provide quantifiable scores to classifiers. This score is then often used to rank classifiers and select the best set of classifiers in practice. In this section, we discuss the benefits of elicited metrics in comparison to some default metrics while ranking real-world classifiers.

Table 1: Dataset statistics

Dataset	k	#samples	#features
default	2	30000	33
adult	2	43156	74
sensIT Vehicle	3	98528	50
covtype	7	581012	54

For this experiment, we work with four real world datasets with varying number of classes $k \in \{2, 3, 7\}$. See Table 1 for details of the datasets. We use 60% of each dataset to train classifiers. The rest of the data is used to compute (testing) predictive rates. For each dataset, we create a pool of 80 classifiers by tweaking hyper-parameters in some famous machine learning models that are routinely used in practice. Specifically, we create 20 classifiers each from logistic regression models [27], multi-layer perceptron models [42], LightGBM models [26], and support vector machines [21]. We compare ranking of these 80 classifiers provided by competing baseline metrics with respect to the ground truth ranking, which is provided by the oracle's true metric.

We generate a random quadratic metric ϕ^{quad} following Definition 3. We treat the true ϕ^{quad} as oracle's metric. It provides us the ground truth ranking of the classifiers in the pool. We then use our proposed procedure QPME (Algorithm 1) to recover the oracle's metric. For comparison in ranking of real-world classifiers, we choose two linear metrics that are routinely employed by practitioners as baselines. The first is accuracy $\phi^{acc} = 1/\sqrt{q}\langle \mathbf{1}, \mathbf{r} \rangle$, and the second is weighted accuracy, where we just use the linear part $\langle \mathbf{a}, \mathbf{r} \rangle$ of the oracle's true quadratic metric $\langle \mathbf{a}, \mathbf{r} \rangle + \mathbf{r}^T \mathbf{B} \mathbf{r}$. We repeat this experiment over 100 trials.

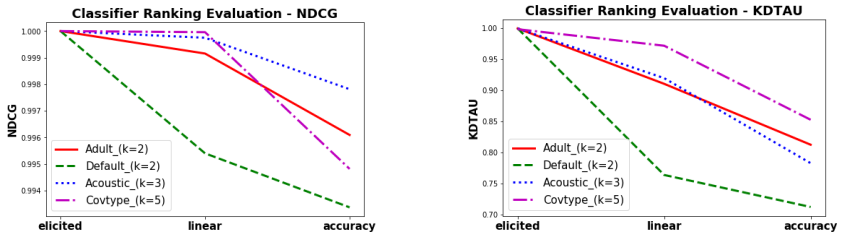

Figure 5: Performance of competing metrics while ranking real-world classifiers. 'Elicited' is the metric elicited by QPME, 'linear' is the metric that comprises only the linear part of the oracle's true quadratic metric, and 'accuracy' is the linear metric which weigh all classification errors equally (often used in practice).

We report NDCG (with exponential gain) [51] and Kendall-tau coefficient [48] averaged over the 100 trials in Figure 5. We observe consistently for all the datasets that the elicited metrics using the QPME procedure achieve the highest possible NDCG and Kendall-tau coefficient of 1. As we saw in Section 5, QPME may incur elicitation error, and thus the elicited metrics may not be very accurate; however, Figure 5 shows that the elicited metrics may still achieve near-optimal ranking results. This implies that when given a set of classifiers, ranking based on elicited metric scores align most closely to true ranking in comparison to ranking based on default metric scores. Consequentially, the elicited metrics may allow us to select or discard classifiers for a given task. This is advantageous in practice. For the *covtype* dataset, we see

that the *linear* metric also achieves high NDCG values, so perhaps ranking at the top is quite accurate; however Kendall-tau coefficient is low suggesting that the overall ranking of classifiers is poor. We also observe that, in general, the weighted version (*linear* metric) is better than *accuracy* while ranking classifiers.

7 Related Work

Hiranandani et al. [16] formalize the problem of ME for binary classification and then later extend it to the multiclass setting [17]. Their focus, however, is on eliciting linear and fractional-linear metrics; whereas, we are interested in more complex quadratic metrics. Learning linear functions passively using pairwise comparisons is a mature field [22, 15, 43], but unlike their active learning counter-parts [47, 20, 24], these methods are not query efficient. Other related work include active classification [47, 24, 41], which learn classifiers for a fixed (known) metric. In contrast, we seek to elicit an unknown metric by posing queries to an oracle. There is also some work on active linear elicitation, e.g. Qian et al. [45], but they do not provide theoretical bounds and work with a different query space. We are unaware of prior work on eliciting a quadratic metric, either passively or actively through pairwise comparisons.

The use of metric elicitation for fairness is relatively new, with some work on eliciting *individual* fairness metrics [19, 37]. Hiranandani et al. [18] is the only work we are aware of that elicits *group* fairness metrics, which we extend to handle a more general family of metrics. Zhang et al. [55] propose an approach to elicit the trade-off between accuracy and fairness using complex ratio queries. In contrast, we jointly elicit the predictive performance, fairness violation, and trade-off using simpler pairwise comparison queries. Lastly, prior work has also focused on learning classifiers under constraints for fairness [13, 14, 54, 38]. We take the regularization view of algorithmic fairness, where the fairness violation is included in the objective [23, 4, 6, 1, 36].

8 Discussion

We have provided an efficient elicitation strategy for quadratic metrics and shown application to fairness. A notable advantage of our proposal is that it is independent of the population $\mathbb{P}$. Thus any metric that is learned using one dataset or model class can be applied to other applications and datasets, as long as the expert believes the context and tradeoffs are the same.

Our approach can be generalized to *higher-order polynomials* of rates. Consider e.g. a cubic polynomial:

$$\phi^{\text{cubic}}(\mathbf{r}) \coloneqq \sum_i a_i r_i + \sum_{i,j} B_{ij} r_i r_j + \sum_{i,j,l} C_{ijl} r_i r_j r_l,$$

where $\mathbf{B}$ and $\mathbf{C}$ are symmetric. A quadratic approximation to this metric around a point $\mathbf{z}$ is given by: $\sum_i a_i r_i + \sum_{i,j} B_{ij} r_i r_j + 6 \sum_{i,j,l} C_{ijl}(r_i - z_i)(r_j - z_j) z_l + c$. We can estimate the parameters of this approximation by applying QPME with the metric centered at an appropriate point, and its queries restricted to a small neighborhood around $\mathbf{z}$. Repeating this over multiple points, we can recover the metric $\hat{\phi}^{\text{cubic}} = (\hat{\mathbf{a}}, \hat{\mathbf{B}}, \hat{\mathbf{C}})$ with as many queries as the number of unknowns. For a d-th order polynomial, one can recursively apply this procedure to estimate $(d-1)$-th order approximations at multiple points, and similarly derive the polynomial coefficients from the estimated local approximations.

Interestingly, the query complexity for these complex non-linear metrics has the same dependence on the number of unknowns as that for linear metrics [17]. We look forward to future work showing optimality of our query complexity bounds, improving the bounds under structural assumptions on the parameters, and conducting user studies presenting intuitive visualizations of rates [55, 3] to receive human preference feedback.

References

[1] 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.

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

[3] E. Beauxis-Aussalet and L. Hardman. Visualization of confusion matrix for non-expert users. In *IEEE Conference on Visual Analytics Science and Technology (VAST)-Poster Proceedings*, 2014.

[4] 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.

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

[6] 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.

[7] A. Cotter, H. Narasimhan, and M. Gupta. On making stochastic classifiers deterministic. In *NeurIPS*, 2019.

[8] P. Dmitriev and X. Wu. Measuring metrics. In *CIKM*, 2016.

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

[10] 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.

[11] A. Esuli and F. Sebastiani. Optimizing text quantifiers for multivariate loss functions. *ACM Transactions on Knowledge Discovery and Data*, 9(4):Article 27, 2015.

[12] W. Gao and F. Sebastiani. Tweet sentiment: From classification to quantification. In *ASONAM*, 2015.

[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] G. Hiranandani, H. Narasimhan, and O. Koyejo. Fair performance metric elicitation. *arXiv preprint arXiv:2006.12732*, 2020.

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

[20] K. G. Jamieson and R. Nowak. Active ranking using pairwise comparisons. In *NIPS*, pages 2240–2248, 2011.

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

[22] 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.

[23] 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.

[24] D. M. Kane, S. Lovett, S. Moran, and J. Zhang. Active classification with comparison queries. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 355–366. IEEE, 2017.

[25] P. Kar, S. Li, H. Narasimhan, S. Chawla, and F. Sebastiani. Online optimization methods for the quantification problem. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1625–1634, 2016.

[26] 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.

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

[28] 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.

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

[30] S. Lawrence, I. Burns, A. Back, A.-C. Tsoi, and C. Giles. Neural network classification and prior class probabilities. In *Neural Networks: Tricks of the Trade*, LNCS, pages 1524:299–313. Springer, 1998.

[31] S. Lawrence, I. Burns, A. Back, A. C. Tsoi, and C. L. Giles. Neural network classification and prior class probabilities. In *Neural networks: tricks of the trade*, pages 299–313. Springer, 1998.

[32] B. G. Lindsay, M. Markatou, S. Ray, K. Yang, S.-C. Chen, et al. Quadratic distances on probabilities: A unified foundation. *The Annals of Statistics*, 36(2):983–1006, 2008.

[33] W. Liu and S. Chawla. A quadratic mean based supervised learning model for managing data skewness. In *SDM*, 2011.

[34] R. McGill, J. W. Tukey, and W. A. Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[35] A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611, 2013.

[36] 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.

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

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

[39] 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.

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

[41] 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.

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

[43] 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.

[44] B. Qian, X. Wang, F. Wang, H. Li, J. Ye, and I. Davidson. Active learning from relative queries. In *IJCAI*, pages 1614–1620, 2013.

[45] L. Qian, J. Gao, and H. Jagadish. Learning user preferences by adaptive pairwise comparison. *Proceedings of the VLDB Endowment*, 8(11):1322–1333, 2015.

[46] 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.

[47] B. Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

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

[49] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

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

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] Z.-H. Zhou and X.-Y. Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010.

Appendices

A Linear Performance Metric Elicitation (LPME)

In this section, we shed more light on the procedure from [17] that elicits a multiclass linear metric. We call it the Linear Performance Metric Elicitation (LPME) procedure. As discussed in Algorithm 1, we use this as a subroutine to elicit metrics in the quadratic family.

LPME exploits the enclosed sphere $\mathcal{S} \subset \mathcal{R}$ for eliciting linear multiclass metrics. Let the sphere $\mathcal{S}$'s radius be $\rho > 0$, and the oracle's scale invariant metric be $\phi^{\text{lin}}(\mathbf{r}) \coloneqq \langle \mathbf{a}, \mathbf{r} \rangle$ such that $\|\mathbf{a}\|_2 = 1$. The oracle queries are $\Omega\left(\mathbf{r}_1, \mathbf{r}_2\,;\, \phi^{\text{lin}}\right) \coloneqq \mathbb{1}[\phi^{\text{lin}}(\mathbf{r}_1) > \phi^{\text{lin}}(\mathbf{r}_2)]$. We first outline a trivial Lemma from [17].

Lemma 1. *[17] Let a normalized vector* $\mathbf{a}$ *with* $\|\mathbf{a}\|_2 = 1$ *parametrize a linear metric* $\phi^{lin} \coloneqq \langle \mathbf{a}, \mathbf{r} \rangle$*, then the unique optimal rate* $\bar{\mathbf{r}}$ *over* $\mathcal{S}$ *is a rate on the boundary of* $\mathcal{S}$ *given by* $\bar{\mathbf{r}} = \rho\mathbf{a} + \mathbf{o}$*, where* $\mathbf{o}$ *is the center of* $\mathcal{S}$*.*

Lemma 1 provides a way to define a one-to-one correspondence between a linear performance metric and its optimal rate. That is, given a linear performance metric, using Lemma 1, we may get a unique point in the query space lying on the boundary of the sphere $\partial\mathcal{S}$. Moreover, the converse is also true; i.e., given a feasible rate on the boundary of the sphere $\partial\mathcal{S}$, one may recover the linear metric for which the given rate is optimal. Thus, for eliciting a linear metric, Hiranandani et al. [17] essentially search for the optimal rate (over the sphere $\mathcal{S}$) using pairwise queries to the oracle. The optimal rate by virtue of Lemma 1 reveals the true metric. The LPME subroutine is summarized in Algorithm 2. Intuitively, Algorithm 2 minimizes a strongly convex function denoting distance of query points from a supporting hyperplane whose slope is the true metric (see Figure 2(c) in [17]). The procedure 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. In $\boldsymbol{\theta}$, all the angles except the primary angle are in $[0, \pi]$, and the primary angle is in $[0, 2\pi]$. A scale invariant linear performance metric with $\|\mathbf{a}\|_2 = 1$ can be constructed by assigning $a_i = \Pi_{j=1}^{i-1} \sin\theta_j \cos\theta_i$ for $i \in [q-1]$ and $a_q = \Pi_{j=1}^{q-1} \sin\theta_j$. Since we can easily compute the metric's optimal rate over $\mathcal{S}$ using Lemma 1, by varying $\boldsymbol{\theta}$ in this procedure, we parametrize the surface of the sphere $\partial\mathcal{S}$. We denote this parametrization by $\mu(\boldsymbol{\theta})$, where $\mu : [0, \pi]^{q-2} \times [0, 2\pi] \to \partial\mathcal{S}$.

Description of Algorithm 2: Let the oracle's metric be $\phi^{\text{lin}} = \langle \mathbf{a}, \mathbf{r} \rangle$ such that $\|\mathbf{a}\|_2 = 1$ (Section 2.2). Using the parametrization $\mu(\boldsymbol{\theta})$ for the boundary of the sphere $\partial\mathcal{S}$, Algorithm 2 returns an estimate $\hat{\mathbf{a}}$ with $\|\hat{\mathbf{a}}\|_2 = 1$. Line 2-6 recover the search orthant of the optimal rate over the sphere by posing q trivial queries. Once the search orthant of the optimal rate is known, the algorithm in each iteration of the for loop (line 9-19) updates one angle θ_j keeping other angles fixed by the *ShrinkInterval* subroutine. The *ShrinkInterval* subroutine (illustrated in Figure 6) is binary-search based routine that shrinks the interval $[\theta_j^a, \theta_j^b]$ by half based on the oracle responses to four queries. Then the algorithm cyclically updates each angle until it converges to a metric sufficiently close to the true metric. We fix the number of cycles in coordinate-wise binary search to four.

Algorithm 2 Linear Performance Metric Elicitation

Input: Query space $\mathcal{S} \subset \mathcal{R}$, binary-search tolerance $\epsilon > 0$, oracle $\Omega(\cdot, \cdot\,; \phi^{\text{lin}})$ with metric ϕ^{lin}

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 $\bar{s}^{(\mathbf{a})}$ and $\bar{s}^{(\mathbf{a}')}$ over the sphere $\mathcal{S}$ using Lemma 1
  Query $\Omega(\bar{\mathbf{s}}^{(\mathbf{a})}, \bar{\mathbf{s}}^{(\mathbf{a}')}; \phi^{\text{lin}})$
  {These queries reveal the search orthant}

Start with coordinate $j = 1$.
Initialize: $\boldsymbol{\theta} = \boldsymbol{\theta}^{(1)}$ {$\boldsymbol{\theta}^{(1)}$ is a point in the search orthant.}
for $t = 1, 2, \cdots, T = 4(q-1)$ **do**
  Set $\boldsymbol{\theta}^{(a)} = \boldsymbol{\theta}^{(c)} = \boldsymbol{\theta}^{(d)} = \boldsymbol{\theta}^{(e)} = \boldsymbol{\theta}^{(b)} = \boldsymbol{\theta}^{(t)}$.
  Set $\theta_j^{(a)}$ and $\theta_j^{(b)}$ to be the min and max angle, respectively, based on the search orthant
  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 $\bar{\mathbf{r}}^{(a)} = \mu(\boldsymbol{\theta}^{(a)})$ (i.e. parametrization of $\partial\mathcal{S}$). Similarly, set $\bar{\mathbf{r}}^{(c)}, \bar{\mathbf{r}}^{(d)}, \bar{\mathbf{r}}^{(e)}, \bar{\mathbf{r}}^{(b)}$
    Query $\Omega(\bar{\mathbf{r}}^{(c)}, \bar{\mathbf{r}}^{(a)}\,; \phi^{\text{lin}}), \Omega(\bar{\mathbf{r}}^{(d)}, \bar{\mathbf{r}}^{(c)}\,; \phi^{\text{lin}}), \Omega(\bar{\mathbf{r}}^{(e)}, \bar{\mathbf{r}}^{(d)}\,; \phi^{\text{lin}}), \Omega(\bar{\mathbf{r}}^{(b)}, \bar{\mathbf{r}}^{(e)}\,; \phi^{\text{lin}}))$.
    $[\theta_j^{(a)}, \theta_j^{(b)}] \leftarrow$ *ShrinkInterval* (responses) {see Figure 6}
  Set $\theta_j^{(d)} = \frac{1}{2}(\theta_j^{(a)} + \theta_j^{(b)})$
  Set $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(d)}$.
  Update coordinate $j \leftarrow j + 1$ cyclically.
Output: $\hat{a}_i = \Pi_{j=1}^{i-1} \sin\theta_j^{(T)} \cos\theta_i^{(T)}\ \forall i \in [q-1]$, $\hat{a}_q = \Pi_{j=1}^{q-1} \sin\theta_j^{(T)}$

B Geometry of the Feasible Space (Proofs of Section 2.2 and Section 4.2)

Proof of Proposition 1 and Proposition 2. We prove Proposition 2. The proof of Proposition 1 is analogous where the probability measures (corresponding to classifiers and their rates) are not conditioned on any group.

The group-specific set of rates $\mathcal{R}^g$ for a group g has the following properties [18]:

- *Convex*: Suppose there are two classifiers $h_1^g, h_2^g \in \mathcal{H}^g$ that achieve the rates $\mathbf{r}_1^g, \mathbf{r}_2^g \in \mathcal{R}^g$, respectively. Consider a classifier h^g that predicts what classifier h_1^g predicts with probability γ and predicts what classifier h_2^g predicts with probability $1-\gamma$. Then the rate matrix of the classifier h^g is given by:

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

Subroutine *ShrinkInterval*

Input: Oracle responses for $\Omega(\bar{\mathbf{r}}^{(c)}, \bar{\mathbf{r}}^{(a)}\,;\,\phi^{\text{lin}})$, $\Omega(\bar{\mathbf{r}}^{(d)}, \bar{\mathbf{r}}^{(c)}\,;\,\phi^{\text{lin}})$, $\Omega(\bar{\mathbf{r}}^{(e)}, \bar{\mathbf{r}}^{(d)}\,;\,\phi^{\text{lin}})$, $\Omega(\bar{\mathbf{r}}^{(b)}, \bar{\mathbf{r}}^{(e)}\,;\,\phi^{\text{lin}})$

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

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

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

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

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

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

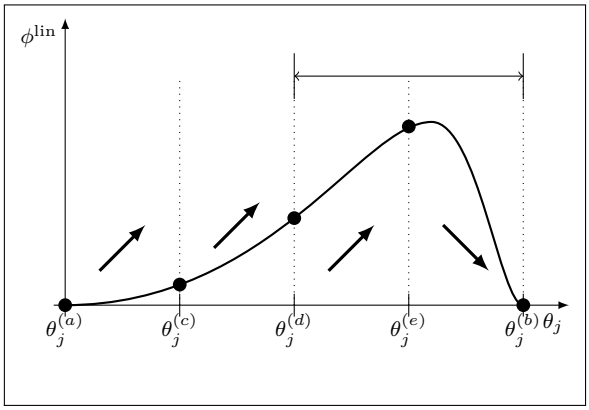

Figure 6: (Left): The *ShrinkInterval* subroutine used in line 16 of Algorithm 2 (Right): Visual illustration of the subroutine *ShrinkInterval* [17]; *ShrinkInterval* shrinks the current search interval to half based on oracle responses to four queries.

The above equations shows that the convex combination of any two rates is feasible as well, i.e., one can construct a randomized classifier which will achieve the convex combination of rates. Hence, $\mathcal{R}^g \;\forall\; g \in [m]$ is convex. Since intersection of convex sets is convex, the intersection set $\mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ is convex as well.

- *Bounded:* Since $R_{ij}^g(h) = P[h = j | Y = i] \leq 1$ for all $i, j \in [k]$, $\mathcal{R}^g \subseteq [0,1]^q$.
- *The rate profiles* $\mathbf{o}$ *and* $\mathbf{e}_i$*'s are always achievable:* A random uniform classifier, i.e, the classifier, which given any input, predicts all classes with probability $1/k$ achieves the rate profile $\mathbf{o}$. A classifier that always predicts class i achieves the rate $\mathbf{e}_i$. Thus, $\mathbf{e}_i \in \mathcal{R}^g \,\forall\, i \in [k], g \in [m]$ are always feasible.
- $\mathbf{e}_i$*'s are vertices:* All the supporting hyperplanes with positive slope, i.e., $\ell_{1i} < \ell_{1j} < 0$ and $\ell_{1l} = 0$ for $p \in [k], l \neq i, j$ will be supported by $\mathbf{e}_1$. Thus, $\mathbf{e}_i$'s are vertices of the convex set $\mathcal{R}^g$. Due to Assumption 1, one can construct a ball around the trivial rate $\mathbf{o}$ and thus $\mathbf{o}$ lies in the interior.

All the points above apply to space of overall rates $\mathcal{R}$ as well using which Proposition 1 is also proved. □

B.1 Finding the Sphere $\mathcal{S} \subset \mathcal{R}$

In this section, we provide details regarding how a sphere $\mathcal{S}$ with sufficiently large radius ρ inside the feasible region $\mathcal{R}$ may be found (see Figure 2(a)). The following discussion is borrowed from [17] and provided here for completeness.

The following optimization problem is a special case of OP2 in [38]. The problem is associated with a feasibility check problem. Given a rate profile $\mathbf{r}_0$, the optimization routine tries to construct a classifier that achieves the rate $\mathbf{r}_0$ within small error $\epsilon > 0$.

$$\min_{\mathbf{r} \in \mathcal{R}} \; 0 \qquad s.t. \;\; \|\mathbf{r} - \mathbf{r}_0\|_2 \leq \epsilon. \tag{OP1}$$

Algorithm 3 Obtaining the sphere $\mathcal{S} \subset \mathcal{R}$ (Figure 2(a)) of radius ρ centered at $\mathbf{o}$

for $j = 1, 2, \cdots, q$ **do**
  Let $\boldsymbol{\alpha}_j$ be the standard basis vector.
  Compute the maximum constant c_j such that $\mathbf{o} + c_j\boldsymbol{\alpha}_j$ is feasible by solving (OP1).
Let $CONV$ denote the convex hull of $\{\mathbf{o} \pm c_j\boldsymbol{\alpha}_j\}_{j=1}^{q}$. It will be centered at $\mathbf{o}$.
Compute the radius ρ of the largest ball that fits in $CONV$.
Output: Sphere $\mathcal{S}$ with radius ρ centered at $\mathbf{o}$.

The above optimization problem checks the feasibility, and if there exists a solution to the above problem, then Algorithm 1 of [38] returns it. Furthermore, Algorithm 3 computes a value of $\rho \geq \tilde{p}/k$, where $\tilde{p}$ is the radius of the largest ball contained in the set $\mathcal{R}$. Also, the approach in [38] is consistent, thus we should get a good estimate of the sphere, provided we sufficiently large number of samples. The algorithm is completely offline and does not impact oracle query complexity.

Lemma 2. *[17] Let $\tilde{p}$ denote the radius of the largest ball in $\mathcal{R}$ centered at $\mathbf{o}$. Then Algorithm 3 returns a sphere with radius $\rho \geq \tilde{p}/k$, where k is the number of classes.*

The idea in Algorithm 3 can be trivially extended to finding a sphere $\overline{S} \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ corresponding to Remark 3.

C Quadratic Performance Metric Elicitation Procedure

In this section, we describe how the subroutine calls to LPME in Algorithm 1 elicit a quadratic metric in Definition 3. We start with the shifted metric of Equation (6). Also, as explained in the main paper, we may assume $d_1 \neq 0$ due to Assumption 2. We can derive the following solution using any non-zero coordinate of $\mathbf{d}$, instead of d_1. We can identify a non-zero coordinate using q trivial queries of the form $(\varrho\boldsymbol{\alpha}_i + \mathbf{o}, \mathbf{o}), \forall i \in [q]$.

1. From line 1 of Algorithm 1, we get local linear approximation at $\mathbf{o}$. Using Remark 2, we have (7) which is

$$d_i = \frac{f_{i0}}{f_{10}} d_1 \qquad \forall\, i \in \{2, \ldots, q\}. \tag{16}$$

2. Similarly, if we apply LPME on small balls around rate profiles $\mathbf{z}_j$, Remark 2 gives us:

$$\frac{d_i + (\rho - \varrho)B_{ij}}{d_1 + (\rho - \varrho)B_{1j}} = \frac{f_{ij}}{f_{1j}} \quad \forall\, i \in \{2, \ldots, q\},\ j \leq i. \tag{17}$$

$$
\begin{aligned}
&\implies d_i + (\rho-\varrho)B_{ij} = \frac{f_{ij}}{f_{1j}}(d_1 + (\rho-\varrho)B_{1j}) \\
&\implies (\rho-\varrho)B_{ij} = \frac{f_{ij}}{f_{1j}}(d_1 + (\rho-\varrho)B_{j1}) - d_i \\
&\implies (\rho-\varrho)B_{ij} = \frac{f_{ij}}{f_{1j}}(d_1 + \frac{f_{j1}}{f_{11}}(d_1 + (\rho-\varrho)B_{11}) - d_j) - \frac{f_{i0}}{f_{10}}d_1 \\
&\implies (\rho-\varrho)B_{ij} = \left(\frac{f_{ij}}{f_{1j}} - \frac{f_{i0}}{f_{10}} + \frac{f_{ij}}{f_{1j}}\left(\frac{f_{j1}}{f_{11}} - \frac{f_{j0}}{f_{10}}\right)\right) d_1 + (\rho-\varrho)\frac{f_{j1}}{f_{11}}B_{11}, \qquad (18)
\end{aligned}
$$

where we have used that the matrix $\mathbf{B}$ is symmetric in the second step, and (16) in the last two steps. We can represent each element in terms of B_{11} and d_1. So, a relation between B_{11} and d_1 may allow us to represent each element of $\mathbf{a}$ and $\mathbf{B}$ in terms of d_1.

3. Therefore, by applying LPME on small balls around rate profiles $-\mathbf{z}_1$, Remark 2 gives us (9):

$$
\frac{d_2 - (\rho-\varrho)B_{21}}{d_1 - (\rho-\varrho)B_{11}} = \frac{f^-_{21}}{f^-_{11}}. \qquad (19)
$$

4. Using (17) and (19), we have:

$$
(\rho-\varrho)B_{11} = \frac{\frac{f^-_{21}}{f^-_{11}} + \frac{f_{21}}{f_{11}} - 2\frac{f_{20}}{f_{10}}}{\frac{f^-_{21}}{f^-_{11}} - \frac{f_{21}}{f_{11}}} d_1. \qquad (20)
$$

Putting (20) in (18), we get:

$$
\begin{aligned}
B_{ij} &= \left[\frac{f_{ij}}{f_{1j}}\left(1 + \frac{f_{j1}}{f_{11}}\right) - \frac{f_{ij}}{f_{1j}}\frac{f_{j0}}{f_{10}} - \frac{f_{i0}}{f_{10}} + \frac{f_{ij}}{f_{1j}}\frac{f_{j1}}{f_{11}}\frac{\frac{f^-_{21}}{f^-_{11}} + \frac{f_{21}}{f_{11}} - 2\frac{f_{20}}{f_{10}}}{\frac{f^-_{21}}{f^-_{11}} - \frac{f_{21}}{f_{11}}}\right] d_1 \\
&= \left(F_{i,1,j}(1 + F_{j,1,1}) - F_{i,1,j}F_{j,1,0} - F_{i,1,0} + F_{i,1,j}\frac{F^-_{2,1,1} + F_{2,1,1} - 2F_{2,1,0}}{F^-_{2,1,1} - F_{2,1,1}}\right) d_1, \qquad (21)
\end{aligned}
$$

where $F_{i,j,l} = \frac{f_{il}}{f_{jl}}$ and $F^-_{i,j,l} = \frac{f^-_{il}}{f^-_{jl}}$. As $\mathbf{a} = \mathbf{d} + \mathbf{Bo}$, we can represent each element of $\mathbf{a}$ and $\mathbf{B}$ using using (16) and (21) in terms of d_1. We can then use the normalization condition $\|\mathbf{a}\|_2 + \frac{1}{2}\|\mathbf{B}\|_F = 1$ to get estimates of $\mathbf{a}, \mathbf{B}$ which are independent of d_1.

This completes the derivation of solution from QPME (section 3).

D Fair (Quadratic) Performance Metric Elicitation Procedure

Algorithm 4: FPM Elicitation

Input: Query set $\mathcal{S}'$, search tolerance $\epsilon > 0$, oracle Ω'

1. Let $\mathcal{L} \leftarrow \varnothing$
2: **For** $\sigma \in \mathcal{M}$ **do**
3: $\quad \boldsymbol{\beta}^{\sigma} \leftarrow \text{QPME}(\mathcal{S}', \epsilon, \Omega')$
4: $\quad$ Let ℓ^{σ} be Eq. (25), extend $\mathcal{L} \leftarrow \mathcal{L} \cup \{\ell^{\sigma}\}$
5: $\hat{\mathbb{B}} \leftarrow$ normalized solution from (28) using $\mathcal{L}$
6: $\hat{\lambda} \leftarrow$ trace back normalized solution from (28) for any σ

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

We first discuss eliciting the fair (quadratic) metric in Definition 4, where all the parameters are unknown. We then provide an alternate procedure for eliciting just the trade-off parameter λ when the predictive performance and fairness violation coefficients are known. The latter is a separate application as discussed in [55]. However, unlike Zhang et al. [55], instead of ratio queries, we use simpler pairwise comparison queries.

In this section, we work with any number of groups $m \geq 2$. The idea, however, remains the same as described in the main paper for number of groups $m = 2$. We specifically select queries from the sphere $\overline{\mathcal{S}} \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$, which is common to all the group-specific feasible region of rates, so to reduce the problem into multiple instances of the proposed QPME procedure of Section 3.

Suppose that the oracle's fair performance metric is ϕ^{fair} parametrized by $(\mathbf{a}, \mathbb{B}, \lambda)$ as in Definition 4. The overall fair metric elicitation procedure framework is summarized in Algorithm 4. The framework exploits the sphere $\overline{\mathcal{S}} \subset \mathcal{R}^1 \cap \cdots \cap \mathcal{R}^m$ and uses the QPME procedure (Algorithm 1) as a subroutine multiple times.

Let us consider a non-empty set of sets $\mathcal{M} \subset 2^{[m]} \setminus \{\varnothing, [m]\}$. We will later discuss how to choose such a set $\mathcal{M}$. 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 of groups σ. For example, when $m = 3$, one may choose the set of groups $\sigma = \{1, 2\}$.

Now, consider a sphere $\mathcal{S}'$ whose elements $\mathbf{r}^{1:m} \in \mathcal{S}'$ are given by:

$$\mathbf{r}^{g} = \begin{cases} \mathbf{s} & \text{if } g \in \sigma \\ \mathbf{o} & \text{o.w.} \end{cases} \tag{22}$$

This is an extension of the sphere $\mathcal{S}'$ defined in the main paper for the $m > 2$ case. Elements in $\mathcal{S}'$ have rate profiles $\mathbf{s} \in \overline{\mathcal{S}}$ to the groups in σ and trivial rate profile $\mathbf{o}$ to the remaining groups in $[m] \setminus \sigma$. Analogously, the modified oracle $\Omega'(\mathbf{r}_1, \mathbf{r}_2) = \Omega((\mathbf{r}_1^{1:m}), (\mathbf{r}_2^{1:m}))$, where $\mathbf{r}_1^{1:m}, \mathbf{r}_2^{1:m}$ are the elements of the spheres $\mathcal{S}'$ above. Thus, for elements in $\mathcal{S}'$, the metric in Definition 4 reduces to:

$$\phi^{\text{fair}}(\mathbf{r}^{1:m} \in \mathcal{S}'; \mathbf{a}, \mathbb{B}, \lambda) = (1 - \lambda)\langle \mathbf{a} \odot \boldsymbol{\tau}^{\sigma}, \mathbf{s} - \mathbf{o} \rangle + \lambda \frac{1}{2} (\mathbf{s} - \mathbf{o})^T \mathbf{W}^{\sigma} (\mathbf{s} - \mathbf{o}) + c^{\sigma} \tag{23}$$

where $\boldsymbol{\tau}^\sigma = \sum_{g\in\sigma} \boldsymbol{\tau}^g$, $\mathbf{W}^\sigma = \sum_{u\in\sigma, v\in[m]\setminus\sigma} B^{uv}$, and c^σ is a constant not affecting the oracle responses.

The above metric is a particular instance of $\bar{\phi}(\mathbf{s}; \mathbf{d}, \mathbf{B})$ in (5) with $\mathbf{d} \coloneqq (1-\lambda)\mathbf{a} \odot \boldsymbol{\tau}^\sigma$ and $\mathbf{B} \coloneqq \lambda \mathbf{W}^\sigma$; thus, we apply QPME procedure as a subroutine in Algorithm 2 to elicit the metric in (23).

The only change needed to be made to the algorithm is in line 5, where we need to take into account the changed relationship between $\mathbf{d}$ and $\mathbf{a}$, and need to separately (not jointly) normalize the linear and quadratic coefficients. With this change, the output of the algorithm directly gives us the required estimates. Specifically, we have from step 1 of Algorithm 1 and (7) an estimate

$$\frac{d_i}{d_1} = \frac{\tau_i^\sigma a_i}{\tau_1^\sigma a_1} = \frac{f_{i0}}{f_{10}} \implies a_i = \frac{f_{i0}}{f_{10}} \frac{\tau_1^\sigma}{\tau_i^\sigma} a_1. \tag{24}$$

Using the normalization condition (i.e., $\|\mathbf{a}\|_2 = 1$), we directly get an estimate $\hat{\mathbf{a}}$ for the linear coefficients. Similarly, steps 2-4 of Algorithm 1 and (10) gives us:

$$\begin{aligned}
\hat{B}_{ij} &= \sum_{u\in\sigma, v\in[m]\setminus\sigma} \tilde{B}_{ij}^{uv} \\
&= \left(F_{i,1,j}^\sigma (1 + F_{j,1,1}^\sigma) - F_{i,1,j}^\sigma F_{j,1,0}^\sigma d_1 - F_{i,1,0}^\sigma + F_{i,1,j}^\sigma \frac{F_{2,1,1}^{-,\sigma} + F_{2,1,1}^\sigma - 2F_{2,1,0}^\sigma}{F_{2,1,1}^{-,\sigma} - F_{2,1,1}^\sigma} \right) \tau_1^1 \hat{a}_1 \\
&= \beta^\sigma,
\end{aligned} \tag{25}$$

where the above solution is similar to the two group case in (15), but here it is corresponding to a partition of groups defined by σ, and $\tilde{\mathbf{B}}^{uv} \coloneqq \lambda \mathbf{B}^{uv}/(1-\lambda)$ is a scaled version of the true (unknown) $\mathbf{B}^{uv}$. Let equation (25) be denoted by ℓ^σ. Also, let the right hand side term of (25) be denoted by β^σ.

Since we want to elicit $\binom{m}{2}$ fairness violation weight matrices in $\mathbb{B}$, we require $\binom{m}{2}$ ways of partitioning the groups into two sets so that we construct $\binom{m}{2}$ independent matrix equations similar to (25). Let $\mathcal{M}$ be those set of sets. Thus, running over all the choices of sets of groups $\sigma \in \mathcal{M}$ provides the system of equations $\mathcal{L} \coloneqq \cup_{\sigma\in\mathcal{M}} \ell^\sigma$ (line 4 in Algorithm 2), which is:

$$\begin{bmatrix} \Xi & 0 & \dots & 0 \\ 0 & \Xi & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Xi \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{b}}_{(11)} \\ \tilde{\mathbf{b}}_{(12)} \\ \dots \\ \tilde{\mathbf{b}}_{(qq)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{(11)} \\ \boldsymbol{\beta}_{(12)} \\ \dots \\ \boldsymbol{\beta}_{(qq)} \end{bmatrix}, \tag{26}$$

where $\tilde{\mathbf{b}}_{(ij)} = (\tilde{b}_{ij}^1, \tilde{b}_{ij}^2, \cdots, \tilde{b}_{ij}^{\binom{m}{2}})$ and $\boldsymbol{\gamma}_{(ij)} = (\beta_{ij}^1, \beta_{ij}^2, \cdots, \beta_{ij}^{\binom{m}{2}})$ are vectorized versions of the ij-th entry across groups for $i, j \in [q]$, and $\Xi \in \{0,1\}^{\binom{m}{2}\times\binom{m}{2}}$ is a binary full-rank matrix denoting membership of groups in the set σ. For example, when one chooses $\mathcal{M} = \{\{1,2\}, \{1,3\}, \{2,3\}\}$ for $m = 3$, Ξ is given by:

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

One may choose any set of sets $\mathcal{M}$ that allows the resulting group membership matrix Ξ to be non-singular. The solution of the system of equations $\mathcal{L}$ is:

$$\begin{bmatrix} \tilde{\mathbf{b}}_{(11)} \\ \tilde{\mathbf{b}}_{(12)} \\ \dots \\ \tilde{\mathbf{b}}_{(qq)} \end{bmatrix} = \begin{bmatrix} \Xi & 0 & \dots & 0 \\ 0 & \Xi & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \Xi \end{bmatrix}^{(-1)} \begin{bmatrix} \boldsymbol{\beta}_{(11)} \\ \boldsymbol{\beta}_{(12)} \\ \dots \\ \boldsymbol{\beta}_{(qq)} \end{bmatrix}. \tag{27}$$

When all $\tilde{\mathbf{B}}^{uv}$'s are normalized, we have the estimated fairness violation weight matrices as:

$$\hat{\mathbf{B}}^{uv} = \frac{\tilde{\mathbf{B}}^{uv}}{\sum_{u,v=1,v>u}^{m} \|\tilde{\mathbf{B}}^{uv}\|_F} \quad \text{for} \quad u, v \in [m], v > u. \tag{28}$$

Due to the above normalization, the solution is again independent of the true trade-off λ.

Given estimates $\hat{B}_{ij}^{uv}$ and $\hat{a}_1$, we can now additionally estimate the trade-off parameter $\hat{\lambda}$ from ℓ^σ (25) for any $\sigma \in \mathcal{M}$. This completes the fair (quadratic) metric elicitation procedure.

D.1 Eliciting Trade-off λ when (linear) predictive performance and (quadratic) fairness violation coefficients are known

We now provide an alternate binary search based method similar to Hiranandani et al. [18] for eliciting the trade-off parameter λ when the linear predictive and quadratic fairness coefficients are already known. This is along similar lines to the application considered by Zhang et al. [55], but unlike them, instead of ratio queries, we require simpler pairwise queries.

Here, the key insight is to approximate the non-linearity posed by the fairness violation in Definition 4, which then reduces the problem to a one-dimensional binary search. We have:

$$\phi^{\text{fair}}(\mathbf{r}^{1:m}\,;\,\mathbf{a},\mathbb{B},\lambda) \coloneqq (1-\lambda)\langle \mathbf{a}, \mathbf{r}\rangle + \lambda \frac{1}{2}\left(\sum_{u,v=1,v>u}^{m} (\mathbf{r}^u - \mathbf{r}^v)^T \mathbf{B}^{uv} (\mathbf{r}^u - \mathbf{r}^v)\right). \tag{29}$$

To this end, we define a new sphere $\mathcal{S}' = \{(\mathbf{s}, \mathbf{o}, \dots, \mathbf{o}) | \mathbf{s} \in \overline{\mathcal{S}}\}$. The elements in $\mathcal{S}'$ is the set of rate profiles whose first group achieves rates $\mathbf{s} \in \overline{\mathcal{S}}$ and rest of the groups achieve trivial rate $\mathbf{o}$ (corresponding to uniform random classifier). For any element in $\mathcal{S}'$, the associated discrepancy terms $(\mathbf{r}^u - \mathbf{r}^v) = 0$ for $u, v \neq 1$. Thus for elements in $\mathcal{S}'$, the metric in Definition 4 reduces to:

$$\phi^{\text{fair}}((\mathbf{s}, \mathbf{o}, \dots, \mathbf{o})\,;\,\mathbf{a},\mathbb{B},\lambda) = (1-\lambda)\langle \boldsymbol{\tau}^1 \odot \mathbf{a}, \mathbf{s} - \mathbf{o}\rangle + \lambda \frac{1}{2}(\mathbf{s}-\mathbf{o})^T \sum_{v=2}^{m} \mathbf{B}^{1v}(\mathbf{s}-\mathbf{o}) + c. \tag{30}$$

Additionally, we consider a small sphere $\overline{\mathcal{S}}'_{\mathbf{z}_1}$, where $\mathbf{z}_1 \coloneqq (\rho - \varrho)\boldsymbol{\alpha}_1 + \mathbf{o}$, similar to what is shown in Figure 2(a). We may approximate the quadratic term on the right hand side above by its first order Taylor approximation as follows:

$$\begin{aligned} \phi^{\text{fair}}((\mathbf{s}, \mathbf{o}, \dots, \mathbf{o}); \mathbf{a}, \mathbb{B}, \lambda) &\approx \phi^{\text{fair, apx}}((\mathbf{s}, \mathbf{o}, \dots, \mathbf{o}); \mathbf{a}, \mathbb{B}, \lambda) \\ &= \langle (1-\lambda)\boldsymbol{\tau}^1 \odot \mathbf{a} + \lambda \sum_{v=2}^{m} \mathbf{B}^{1v}(\mathbf{z}_1 - \mathbf{o}), \mathbf{s}\rangle \end{aligned} \tag{31}$$

for $\mathbf{s}$ in a small neighbourhood around the rate profile $\mathbf{z}_1$. Since the metric is essentially linear in $\mathbf{s}$, the following lemma from [18] shows that the metric in (31) is quasiconcave in λ.

Algorithm 5 Elicit trade-off λ when predictive performance and fairness violation are known

Input: Query space $\overline{\mathcal{S}}'_{\mathbf{z}_1}$, binary-search tolerance $\epsilon > 0$, oracle Ω
Initialize: $\lambda^{(a)} = 0$, $\lambda^{(b)} = 1$.
while $\left|\lambda^{(b)} - \lambda^{(a)}\right| > \epsilon$ **do**
  Set $\lambda^{(c)} = \frac{3\lambda^{(a)}+\lambda^{(b)}}{4}$, $\lambda^{(d)} = \frac{\lambda^{(a)}+\lambda^{(b)}}{2}$, $\lambda^{(e)} = \frac{\lambda^{(a)}+3\lambda^{(b)}}{4}$
  Set $\mathbf{s}^{(a)} = \underset{\mathbf{s}\in\overline{\mathcal{S}}'_{\mathbf{z}_1}}{\operatorname{argmax}} \langle (1-\lambda^{(a)})\boldsymbol{\tau}^1 \odot \hat{\mathbf{a}} + \lambda^{(a)} \sum_{v=2}^{m} \hat{\mathbf{B}}^{1v}(\mathbf{z}_1 - \mathbf{o}), \mathbf{s}\rangle$ using Lemma 1
  Similarly, set $\mathbf{s}^{(c)}$, $\mathbf{s}^{(d)}$, $\mathbf{s}^{(e)}$, $\mathbf{s}^{(b)}$.
  Query $\Omega(\mathbf{s}^{(c)}, \mathbf{s}^{(a)})$, $\Omega(\mathbf{s}^{(d)}, \mathbf{s}^{(c)})$, $\Omega(\mathbf{s}^{(e)}, \mathbf{s}^{(d)})$, and $\Omega(\mathbf{s}^{(b)}, \mathbf{s}^{(e)})$.
  $[\lambda^{(a)}, \lambda^{(b)}] \leftarrow$ *ShrinkInterval* (responses) using a subroutine analogous to the routine in Figure 6.
Output: $\hat{\lambda} = \frac{\lambda^{(a)}+\lambda^{(b)}}{2}$.

Lemma 3. *Under the regularity assumption that* $\langle \boldsymbol{\tau}^1 \odot \mathbf{a}, \sum_{v=2}^{m} \mathbf{B}^{1v}(\mathbf{z}_1 - \mathbf{o})\rangle \neq 1$*, the function*

$$\vartheta(\lambda) := \max_{\mathbf{s}\in\overline{\mathcal{S}}'_{\mathbf{z}_1}} \phi^{fair,\ apx}((\mathbf{s}, \mathbf{o}, \ldots, \mathbf{o}); \mathbf{a}, \mathbb{B}, \lambda) \tag{32}$$

is strictly quasiconcave (and therefore unimodal) in λ*.*

The unimodality of $\vartheta(\lambda)$ allows us to perform the one-dimensional binary search in Algorithm 5 using the query space $\overline{\mathcal{S}}'_{\mathbf{z}_1}$, tolerance ϵ, and the oracle Ω. The binary search algorithm is same as Algorithm 4 in [18] and provided here for completeness.

E Proof of Section 5

Proof of Theorem 1. We first find the smoothness coefficient of the metric in Definition 3.

A function ϕ is said to be L-smooth if for some bounded constant L, we have:

$$\|\nabla\phi(x) - \nabla\phi(y)\|_2 \leq L\|x - y\|_2.$$

For the metric in Definition 3, we have:

$$\begin{aligned}\|\nabla\phi^{\text{quad}}(x) - \nabla\phi^{\text{quad}}(y)\|_2 &= \|\mathbf{a} + \mathbf{B}\mathbf{x} - (\mathbf{a} + \mathbf{B}\mathbf{y})\|_2 \\ &\leq \|\mathbf{B}\|_2\|x - y\|_2 \\ &= \sigma_{\max}\|x - y\|_2,\end{aligned}$$

where $\sigma_{\max}$ is the maximum singular value of the matrix $\mathbf{B}$. By Assumption 2, $\sigma_{\max}$ is bounded; hence, the metrics in Definition 3 are $\sigma_{\max}$-smooth.

Now, we look at the error in Taylor series approximation when we approximate the metric in ϕ^{quad} in Definition 4 with a linear approximation. Our metric is

$$\phi^{\text{quad}}(\mathbf{r}) = \langle \mathbf{a}, \mathbf{r}\rangle + \frac{1}{2}\mathbf{r}^T\mathbf{B}\mathbf{r}.$$

We approximate it with the first order Taylor polynomial around a point $\mathbf{z}$, which we define as follows:

$$T_1(\mathbf{r}) = \langle \mathbf{a}, \mathbf{z} \rangle + \frac{1}{2}\mathbf{z}^T\mathbf{B}\mathbf{z} + \langle \mathbf{a} + \mathbf{B}\mathbf{z}, \mathbf{r} \rangle$$

The bound on the error in this approximation is:

$$\begin{aligned}
|E(\mathbf{r})| &= |\phi^{\text{quad}}(\mathbf{r}) - T_1(\mathbf{r})| \\
&= \frac{1}{2}|(\mathbf{r}-\mathbf{z})^T \Delta\phi^{\text{quad}}|_{\mathbf{c}}(\mathbf{r}-\mathbf{z})| \\
&= \frac{1}{2}|(\mathbf{r}-\mathbf{z})^T \mathbf{B}(\mathbf{r}-\mathbf{z})| \\
&\leq \frac{1}{2}\sigma_{\max}\|\mathbf{r}-\mathbf{z}\|_2 \\
&= \frac{1}{2}\sigma_{\max}\varrho,
\end{aligned}$$

where Hessian at any point $\mathbf{c}$, $\Delta\phi^{\text{quad}}|_{\mathbf{c}}$ is the matrix $\mathbf{B}$, and the singular values are same as absolute of the eigenvalues of the matrix $\mathbf{B}$, as $\mathbf{B}$ is a symmetric matrix. So when the oracle is asked $\Omega(\mathbf{r}_1, \mathbf{r}_2) = \mathbb{1}[\phi^{\text{quad}}(\mathbf{r}_1) > \phi^{\text{quad}}(\mathbf{r}_2)]$, the approximation error can be treated as feedback error from the oracle with feedback noise $2 \times \frac{1}{2}\sigma_{\max}\rho$. Thus, the overall feedback noise by the oracle is $\epsilon_\Omega + \sigma_{\max}\rho$.

For the purpose of this proof, let us denote $\epsilon + \sqrt{\epsilon_\Omega/\rho + \sigma_{\max}}$ by ϵ. We first prove guarantees for the matrix $\mathbf{B}$ and then for the vector $\mathbf{a}$. We write Equation (10) in the following form assuming $d_1 = 1$ (since we normalize the coefficients at the end due to scale invariance):

$$B_{ij} = F_{ij} = \left[\frac{f_{ij}}{f_{1j}}\left(1 + \frac{f_{j1}}{f_{11}}\right) - \frac{f_{ij}}{f_{1j}}\frac{f_{j0}}{f_{10}} - \frac{f_{i0}}{f_{10}} + \frac{f_{ij}}{f_{1j}}\frac{f_{j1}}{f_{11}}\frac{\frac{f^-_{21}}{f^-_{11}} + \frac{f_{21}}{f_{11}} - 2\frac{f_{20}}{f_{10}}}{\frac{f^-_{21}}{f^-_{11}} - \frac{f_{21}}{f_{11}}}\right].$$

$$\begin{aligned}
\implies \mathbf{B}[:, j] &= \mathbf{f}_j\left(\frac{1}{f_{1j}} + \frac{f_{j1}}{f_{1j}f_{11}} + \frac{f_{j0}}{f_{1j}f_{10}} + \frac{f_{j1}}{f_{1j}f_{11}}\left(\frac{\frac{f^-_{21}}{f^-_{11}} + \frac{f_{21}}{f_{11}} - 2\frac{f_{20}}{f_{10}}}{\frac{f^-_{21}}{f^-_{11}} - \frac{f_{21}}{f_{11}}}\right)\right) + \mathbf{f}_0\frac{1}{f_{10}} \\
&= c_j\mathbf{f}_j + c_0\mathbf{f}_0, \qquad (33)
\end{aligned}$$

where $\mathbf{B}[:, j]$ is the j-th column of the matrix $\mathbf{B}$, and the constants c_j and c_0 are well-defined due to the regularity Assumption 4. Notice that,

$$\frac{\partial \mathbf{B}[:, j]}{\partial \mathbf{f}_j} = diag(\mathbf{c}'_j) \odot \mathbf{I} \quad \text{, and} \quad \frac{\partial \mathbf{B}[:, j]}{\partial \mathbf{f}_0} = diag(\mathbf{c}'_0) \odot \mathbf{I},$$

where $\mathbf{c}'_j, \mathbf{c}'_0$ are vector of Lipschitz constants (bounded due to Assumption 4). This implies

$$\begin{aligned}
\|\overline{\mathbf{B}}[:, j] - \hat{\mathbf{B}}[:, j]\|_2 &\leq c'_j\|\bar{\mathbf{f}}_j - \hat{\mathbf{f}}_j\|_2 + c'_0\|\bar{\mathbf{f}}_0 - \hat{\mathbf{f}}_0\|_2 \\
&\leq c'_j\sqrt{q}\epsilon + c'_0\sqrt{q}\epsilon = O(\sqrt{q}\epsilon),
\end{aligned}$$

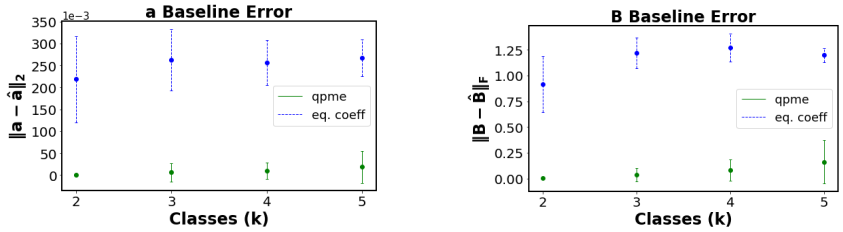

Figure 7: Elicitation error in comparison to a baseline which assigns equal coefficients.

where we have used LPME guarantees for eliciting slopes (linear performance metrics) from Section 2.2.

The above inequality provides bounds on each column of $\mathbf{B}$. Since $\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2$, we have $\max_{ij} |B_{ij} - \hat{B}_{ij}| \leq O(\sqrt{q}\epsilon)$, and consequentially, $\|\mathbf{B} - \hat{\mathbf{B}}\|_F \leq O(q\sqrt{q}\epsilon)$. Due to elicitation on sphere and the oracle noise ϵ_Ω as defined in Definition 6, we can replace ϵ with $\epsilon + \sqrt{\epsilon_\Omega/\rho + \sigma_{\max}}$ back to get the final bound on fairness violation weights as in Theorem 1.

Now let us look at guarantees for $\mathbf{a}$. Since $\mathbf{a} = \mathbf{d} - \mathbf{B}\mathbf{o}$ from (5), we can write

$$\mathbf{a} = c_0 \mathbf{f}_0 - \sum_{j=1}^{q} o_j \mathbf{B}[:, j],$$

where $c_0 = 1/f_{10}$. Since $\mathbf{o}$ is the rate achieved by random classifier, $o_j = 1/k \ \forall j \in [k]$, and thus we have

$$\frac{\partial \mathbf{a}}{\partial \mathbf{f}_0} = c_0 \mathbf{I} \quad \text{and} \quad \frac{\partial \mathbf{a}}{\partial \mathbf{B}[:, j]} = \frac{1}{k} \mathbf{I}.$$

Thus,

$$\|\mathbf{a} - \hat{\mathbf{a}}\|_2 \leq c_0' \sqrt{q}\epsilon + \frac{1}{k} \sum_{j=1}^{q} \sqrt{q}\epsilon = c_0' \sqrt{q}\epsilon + \frac{1}{\sqrt{q}} \sum_{j=1}^{q} c_j' \sqrt{q}\epsilon = O(q\epsilon),$$

where c_0', c_j''s are some Lipschitz constants (bounded due to Assumption 4), and we have used the fact that $q = k^2 - k$ in the second step. □

F Extended Experiments

F.1 More Details on Simulated Experiments on Quadratic Metric Elicitation (Section 6)

In Figures 4(a)–4(b), we show box plots [34] of the ℓ_2 (Frobenius) norm between the true and elicited linear (quadratic) coefficients. We generally find that QPME is able to elicit metrics close to the true ones.

To reinforce this point, we also compare the elicitation error of the QPME procedure and the elicitation error of a baseline which assigns equal coefficients to $\mathbf{a}$ and $\mathbf{B}$ in Figure 7. We see

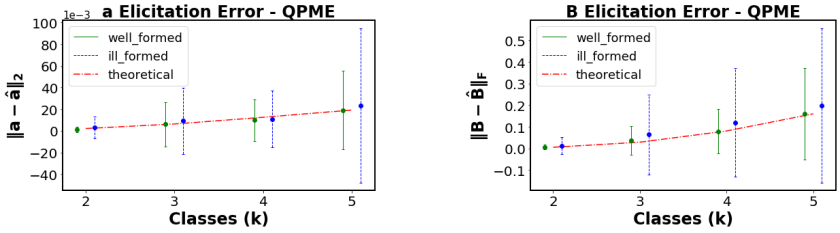

Figure 8: Elicitation error for metrics following Assumption 4 vs elicitation error for completely random metrics.

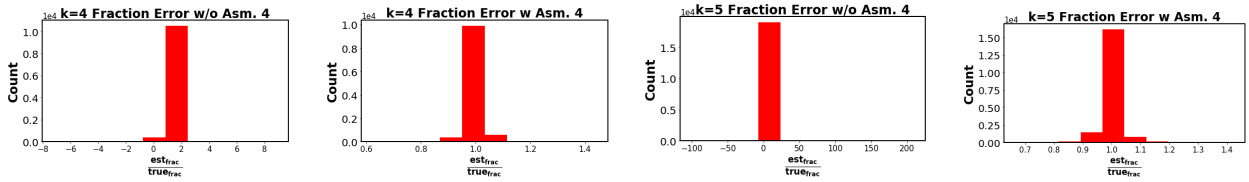

Figure 9: Ratio of estimated fractions to true fractions over 1000 simulated runs with and without Assumption 4.

that the elicitation error of the baseline is order of magnitude higher than the elicitation error of the QPME procedure. This holds for varying k showing that the QPME procedure is able to elicit oracle's multiclass quadratic metrics very well.

Effect of Assumption 4. We mentioned in Section 6 that in a small number of trials, Assumption 4 failed to hold with sufficiently large constants $c_0, c_{-1}, c_1 \ldots, c_q$. We now analyze in greater detail the effect of this regularity assumption in eliciting quadratic metrics and understand how the lower bounding constants impact the elicitation error. Assumption 4 effectively ensures that the ratios computed in (10) are well-defined. To this end, we generate two sets of 100 quadratic metrics. One set is generated following Assumption 4 with all the (entry-wise) lower bounds in the gradient being greater than 10^{-2}, and the other is generated randomly without any regularity condition. For both sets, we run QPME and elicit the corresponding metrics.

In Figure 8, we see that the elicitation error is much higher when the regularity Assumption 4 is not followed, owing to the fact that the ratio computation in (10) is more susceptible to errors when gradient coordinates approach zero in some cases of randomly generated metrics. The dash-dotted curve (in red color) shows the trajectory of the theoretical bounds with increasing q (within a constant factor). In Figure 8, we see that the mean of ℓ_2 (analogously, Frobenius) norm better follow the theoretical bound trajectory in the case when regularity Assumption 4 is followed by the metrics.

We next analyze the ratio of estimated fractions to the true fractions used in (10) over 1000 simulated runs. Ideally, this ratio should be 1, but as we see in Figure 9, these estimated ratios can be off by a significant amount for a few trials when the metrics are generated randomly. The estimated ratios, however, are more stable under Assumption 4. Since we multiply fractions in (10), even then we may observe the compounding effect of fraction estimation errors in the final estimates. Hence, we see for $k = 5$ in Figure 4(a)-4(b), the standard deviation is high due to few trials where the lower bound of 10^{-2} on the constants in Assumption 4 may not be enough. However, majority of the trials as shown in Figure 4(a)-4(b) and Figure 7 incur low elicitation error.