Performance Metric Elicitation from Pairwise Classifier Comparisons

Gaurush Hiranandani
UIUC

Shant Boodaghians
UIUC

Ruta Mehta
UIUC

Oluwasanmi Koyejo
UIUC

Abstract

Given a binary prediction problem, which performance metric should the classifier optimize? We address this question by formalizing the problem of *Metric Elicitation*. The goal of metric elicitation is to discover the performance metric of a practitioner, which reflects her innate rewards (costs) for correct (incorrect) classification. In particular, we focus on eliciting binary classification performance metrics from pairwise feedback, where a practitioner is queried to provide relative preference between two classifiers. By exploiting key geometric properties of the space of confusion matrices, we obtain provably query efficient algorithms for eliciting linear and linear-fractional performance metrics. We further show that our method is robust to feedback and finite sample noise.

1 INTRODUCTION

Selecting an appropriate performance metric is crucial to the real-world utility of predictive machine learning. Specialized teams of statisticians and economists are routinely hired in the industry to monitor many metrics – since optimizing the wrong metric directly translates into lost revenue [6]. Medical predictions are another important application, where ignoring cost sensitive trade-offs can directly impact lives [23]. Unfortunately, there is scant formal guidance within the literature for how a practitioner/user might choose a metric, beyond a few common default choices [4, 10, 22], and even less guidance on selecting a metric which reflects the preferences of the practitioners/users.

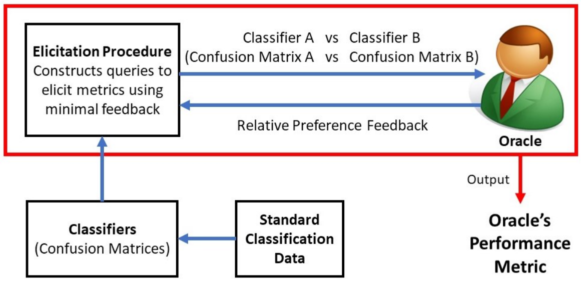

Elicitation Procedure
Constructs queries to elicit metrics using minimal feedback
Classifier A vs Classifier B
(Confusion Matrix A vs Confusion Matrix B)
Relative Preference Feedback
Oracle
Output
Oracle's Performance Metric
Classifiers
(Confusion Matrices)
Standard Classification Data

Fig. 1: Metric Elicitation framework.

Metric Elicitation: Motivated by the principle that the performance metric which best reflects implicit user tradeoffs results in learning models that best resonate with user preferences [9, 22], we introduce a framework, *metric elicitation (ME)*, for determining the binary classification performance metric from user feedback. Since human feedback is costly, the goal is to use as little feedback as possible. On its face, ME simply requires querying a user (oracle) to determine the quality she assigns to classifiers that are learned from standard classification data; however, humans are often inaccurate in providing absolute preferences [19]. Therefore, we propose to employ pairwise comparison queries, where the user (oracle) is asked to compare two classifiers and provide an indicator of relative preference. Based on that relative preference feedback, we elicit the innate performance metric of the user (oracle). See Figure 1 for visual intuition of the framework.

Our approach is inspired by a large literature in economics and psychology on *preference elicitation* [21, 16, 27, 3]. Here, the goal is to learn user preferences from purchases at posted prices. Since there is no notion of prices or purchases in *ME* for machine learning, standard approaches from these studies do not apply. In addition, we emphasize that the notion of pairwise classifier comparison is not new and is already prevalent in the industry. An example is A/B testing [26], where the whole population of users acts as an oracle.[1] Similarly, classifier comparison by a single

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

[1] In A/B testing, sub-populations of users are shown classifier A vs. classifier B, and their responses determine the overall preference. Interestingly, while each person is shown a sample output from one of the classifiers, the entire user population acts as the oracle for comparing classifiers.

expert is becoming commonplace due to advances in the field of interpretable machine learning [20, 7].

In this first edition of ME, we focus on the most common performance metrics which are functions of the confusion matrix [14, 17, 22], particularly, linear and ratio-of-linear functions.[2] This includes almost all modern metrics such as accuracy, F_β-Measure, Jaccard Similarity Coefficient [22], etc. By construction, pairwise classifier comparisons may be conceptually represented by their associated pairwise confusion matrix comparisons. Despite this apparent simplification, the problem remains challenging because one can only query feasible confusion matrices, i.e. confusion matrices for which there exists a classifier. As we show, our characterization of the space of confusion matrices enables the design of efficient binary-search type procedures that identify the innate performance metric of the oracle. While classifier (confusion matrix) comparisons may introduce additional noise, our approach remains robust, both to noise from classifier (confusion matrix) estimation, and to noise in the comparison itself. Thus, our work directly results in a practical algorithm.

Example: Consider the case of cancer diagnosis, where a doctor's unknown, innate performance metric is a linear function of the confusion matrix, i.e., she has some innate reward values for True Positives and True Negatives – equivalently (equiv.), costs for False Positives and False Negatives – based on known consequences of misdiagnosis. Here, the doctor takes the role of the oracle. Our proposed approach exploit the space of confusion matrices associated with all possible classifiers that can be learned from standard classification data and determine the underlying rewards (equiv., costs) provably using the least possible number of pairwise comparison queries posed to the doctor.

Our contributions are summarized as follows:

- We propose the technical problem of *Metric Elicitation*, a framework for determining supervised learning metrics from user feedback. For the case of pairwise feedback, we show that under certain conditions ME is equivalent to learning preferences between pairs of confusion matrices.
- When the underlying metric is linear, we propose a binary search algorithm that can recover the metric with query complexity that decays logarithmically with the desired resolution. We further show that our query-complexity rates match the lower bound.
- We extend the elicitation algorithm to more complex linear-fractional performance metrics.
- We prove robustness of the proposed approach under feedback and classifier estimation noise.

2 BACKGROUND

Let $X \in \mathcal{X}$ and $Y \in \{0,1\}$ represent the input and output random variables respectively (0 = negative class, 1 = positive class). We assume a dataset of size n, $\{(x_i, y_i)\}_{i=1}^n$, generated *iid* from a data generating distribution $\mathbb{P} \overset{\text{iid}}{\sim} (X,Y)$. Let f_X be the marginal distribution for $\mathcal{X}$. Let $\eta(x) = \mathbb{P}(Y=1|X=x)$ and $\zeta = \mathbb{P}(Y=1)$ represent the conditional and the unconditional probability of the positive class, respectively. Note that the earlier term is a function of the input x; whereas, the latter is a constant. We denote a classifier by h, and let $\mathcal{H} = \{h : \mathcal{X} \to [0,1]\}$ be the set of all classifiers. A confusion matrix for a classifier h is denoted by $C(h,\mathbb{P}) \in \mathbb{R}^{2\times 2}$, comprising true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) and is given by:

$$\begin{aligned} C_{11} &= TP(h,\mathbb{P}) = \mathbb{P}(Y=1, h=1), \\ C_{01} &= FP(h,\mathbb{P}) = \mathbb{P}(Y=0, h=1), \\ C_{10} &= FN(h,\mathbb{P}) = \mathbb{P}(Y=1, h=0), \\ C_{00} &= TN(h,\mathbb{P}) = \mathbb{P}(Y=0, h=0). \end{aligned} \tag{1}$$

Clearly, $\sum_{i,j} C_{ij} = 1$. We denote the set of all confusion matrices by $\mathcal{C} = \{C(h,\mathbb{P}) : h \in \mathcal{H}\}$. Under the population law $\mathbb{P}$, the components of the confusion matrix can be further decomposed as: $FN(h,\mathbb{P}) = \zeta - TP(h,\mathbb{P})$ and $FP(h,\mathbb{P}) = 1 - \zeta - TN(h,\mathbb{P})$. This decomposition reduces the four dimensional space to two dimensional space. Therefore, the set of confusion matrices can be defined as $\mathcal{C} = \{(TP(h,\mathbb{P}), TN(h,\mathbb{P})) : h \in \mathcal{H}\}$. For clarity, we will suppress the dependence on $\mathbb{P}$ in our notation. In addition, we will subsume the notation h if it is implicit from the context and denote the confusion matrix by $C = (TP, TN)$. We represent the boundary of the set $\mathcal{C}$ by $\partial\mathcal{C}$. Any hyperplane (line) ℓ in the (tp, tn) coordinate system is given by:

$$\ell := a \cdot tp + b \cdot tn = c, \qquad \text{where } a, b, c \in \mathbb{R}.$$

Let $\phi : [0,1]^{2\times 2} \to \mathbb{R}$ be the performance metric for a classifier h determined by its confusion matrix $C(h)$. Without loss of generality (WLOG), we assume that ϕ is a utility, so that larger values are better.

2.1 Types of Performance Metrics

We consider two of the most common families of binary classification metrics, namely linear and linear-fractional functions of the confusion matrix (1).

Definition 1. *Linear Performance Metric (LPM): We denote this family by φ_{LPM}. Given constants (representing costs or weights) $\{a_{11}, a_{01}, a_{10}, a_{00}\} \in \mathbb{R}^4$, we define the metric as:*

$$\begin{aligned} \phi(C) &= a_{11}TP + a_{01}FP + a_{10}FN + a_{00}TN \\ &= m_{11}TP + m_{00}TN + m_0, \end{aligned} \tag{2}$$

[2] Metrics depending on factors such as model complexity and interpretability are beyond the scope of this manuscript.

where $m_{11} = (a_{11} - a_{10})$, $m_{00} = (a_{00} - a_{01})$, *and* $m_0 = a_{10}\zeta + a_{01}(1-\zeta)$.

Example 1. *Weighted Accuracy (WA) [24]:*

$$WA = w_1 TP + w_2 TN,$$

where $w_1, w_2 \in [0,1]$ *(*w_1, w_2 *can be shifted and scaled to* $[0,1]$ *without changing the learning problem [17]).*

Definition 2. *Linear-Fractional Performance Metric (LFPM): We denote this family by* φ_{LFPM}*. Given constants (representing costs or weights)* $\{a_{11}, a_{01}, a_{10}, a_{00}, b_{11}, b_{01}, b_{10}, b_{00}\} \in \mathbb{R}^8$*, we define the metric as:*

$$\begin{aligned}\phi(C) &= \frac{a_{11}TP + a_{01}FP + a_{10}FN + a_{00}TN}{b_{11}TP + b_{01}FP + b_{10}FN + b_{00}TN} \\ &= \frac{p_{11}TP + p_{00}TN + p_0}{q_{11}TP + q_{00}TN + q_0},\end{aligned} \tag{3}$$

where $p_{11} = (a_{11} - a_{10})$, $p_{00} = (a_{00} - a_{01})$, $q_{11} = (b_{11} - b_{10})$, $q_{00} = (b_{00} - b_{01})$, $p_0 = a_{10}\zeta + a_{01}(1-\zeta)$, $q_0 = b_{10}\zeta + b_{01}(1-\zeta)$.

Example 2. *The* F_β *measure and the Jaccard similarity coefficient (JAC) [22]:*

$$F_\beta = \frac{TP}{\frac{TP}{1+\beta^2} - \frac{TN}{1+\beta^2} + \frac{\beta^2\zeta + 1 - \zeta}{1+\beta^2}}, \; JAC = \frac{TP}{1 - TN} \tag{4}$$

2.2 Bayes Optimal and Inverse Bayes Optimal Classifiers

Given a performance metric ϕ, the Bayes utility $\overline{\tau}$ is the optimal value of the performance metric over all classifiers, i.e., $\overline{\tau} = \sup_{h\in\mathcal{H}} \phi(C(h)) = \sup_{C\in\mathcal{C}} \phi(C)$. The Bayes classifier $\overline{h}$ (when it exists) is the classifier that optimizes the performance metric, so $\overline{h} = \arg\max_{h\in\mathcal{H}} \phi(C(h))$. Similarly, the Bayes confusion matrix is given by $\overline{C} = \arg\max_{C\in\mathcal{C}} \phi(C)$. We further define the inverse Bayes utility $\underline{\tau} = \inf_{h\in\mathcal{H}} \phi(C(h)) = \inf_{C\in\mathcal{C}} \phi(C)$. The inverse Bayes classifier is given by $\underline{h} = \arg\min_{h\in\mathcal{H}} \phi(C(h))$. Similarly, the inverse Bayes confusion matrix is given by $\underline{C} = \arg\min_{C\in\mathcal{C}} \phi(C)$. Notice that for $\phi \in \varphi_{LPM}$ (2), the Bayes classifier predicts the label which maximizes the expected utility conditioned on the instance, as discussed below.

Proposition 1. *Let* $\phi \in \varphi_{LPM}$*, then*

$$\overline{h}(x) = \left\{ \begin{array}{ll} \mathbb{1}[\eta(x) \geq \frac{m_{00}}{m_{11}+m_{00}}], & m_{11} + m_{00} \geq 0 \\ \mathbb{1}[\frac{m_{00}}{m_{11}+m_{00}} \geq \eta(x)], & o.w. \end{array} \right\}$$

is a Bayes optimal classifier w.r.t ϕ*. Further, the inverse Bayes classifier is given by* $\underline{h} = 1 - \overline{h}$*.*

2.3 Problem Setup

We first formalize *oracle query.* Recall that by the definition of confusion matrices (1), there exists a surjective mapping from $\mathcal{H} \to \mathcal{C}$. An oracle is queried to determine relative preference between two classifiers. However, since we only consider metrics which are functions of the confusion matrix, a comparison query over classifiers becomes equivalent to a comparison query over confusion matrices in our setting.

Definition 3. *Oracle Query: Given two classifiers* h, h' *(equiv. to confusion matrices* C, C' *respectively), a query to the Oracle (with metric* ϕ*) is represented by:*

$$\Gamma(h, h') = \Omega(C, C') = \mathbb{1}[\phi(C) > \phi(C')] =: \mathbb{1}[C \succ C'], \tag{5}$$

where $\Gamma : \mathcal{H} \times \mathcal{H} \to \{0,1\}$ *and* $\Omega : \mathcal{C} \times \mathcal{C} \to \{0,1\}$*. The query denotes whether* h *is preferred to* h' *(equiv. to* C *is preferred to* C'*) as measured according to* ϕ*.*

We emphasize that depending on practical convenience, the oracle may be asked to compare either confusion matrices or classifiers achieving the corresponding confusion matrices, via approaches discussed in Section 1. Henceforth, for simplicity of notation, we will treat any comparison query as confusion matrix comparison query. Next, we state the metric elicitation problem.

Definition 4. *Metric Elicitation (given* $\mathbb{P}$*): Suppose that the oracle's true, unknown performance metric is* ϕ*. Recover a metric* $\hat{\phi}$ *by querying the oracle for as few pairwise comparisons of the form* $\Omega(C, C')$*, such that* $\|\phi - \hat{\phi}\|_{--} < \kappa$ *for sufficiently small* $\mathbb{R} \ni \kappa > 0$ *and for any suitable norm* $\|\cdot\|_{--}$*.*

Notice that Definition 4 involves true population quantities C, C' (See (1)). However, in practice, we are given only finite samples. This leads to a more practical definition of metric elicitation problem.

Definition 5. *Metric Elicitation (given* $\{(x_i, y_i)\}_{i=1}^n$*): The same problem as stated in Definition 4, except that the queries are of the form* $\Omega(\hat{C}, \hat{C}')$*, where* $\hat{C}, \hat{C}'$ *are the estimated confusion matrices from the samples.*

Ultimately, we want to perform ME as described in Definition 5. A good approach to do so is to first solve ME as defined in Definition 4, i.e, ME assuming access to the appropriate population quantities, and then consider practical implementation using finite data. This is a standard approach in decision theory (see e.g. [15]), where estimation error from finite samples is adjudged as a noise source and handled accordingly.

3 CONFUSION MATRICES

ME will require confusion matrices that are achieved by all possible classifiers, thus it is necessary to characterize the set $\mathcal{C}$ in a way which is useful for the task.

Assumption 1. *We assume* $g(t) = \mathbb{P}[\eta(X) \geq t]$ *is continuous and strictly decreasing for* $t \in [0,1]$.

This is equivalent to standard assumptions [14] that the event $\eta(X) = t$ has positive density but zero probability. Note that this requires X to have no point mass.

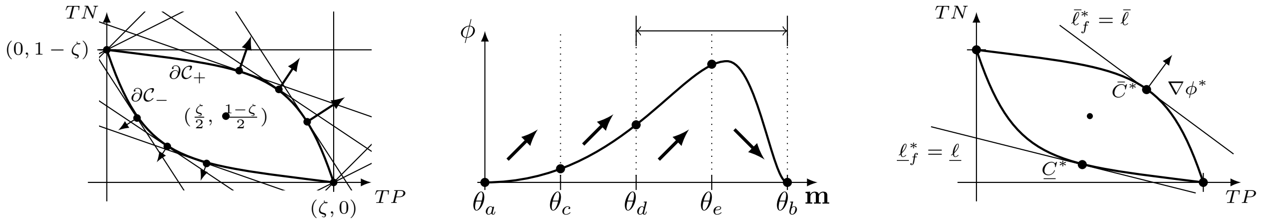


Fig. 2: (a) Supporting hyperplanes (with normal vectors) and resulting geometry of $\mathcal{C}$; (b) Sketch of Algorithm 1; (c) Maximizer $\overline{C}^*$ and minimizer $\underline{C}^*$ along with the supporting hyperplanes for LFPMs.

Proposition 2. *(Properties of $\mathcal{C}$ — Figure 2(a).) The set of confusion matrices $\mathcal{C}$ is convex, closed, contained in the rectangle $[0,\zeta]\times[0,1-\zeta]$ (bounded), and 180° rotationally symmetric around the center-point $(\frac{\zeta}{2},\frac{1-\zeta}{2})$. Under Assumption 1, $(0,1-\zeta)$ and $(\zeta,0)$ are the only vertices of $\mathcal{C}$, and $\mathcal{C}$ is strictly convex. Thus, any supporting hyperplane of $\mathcal{C}$ is tangent at only one point.*[3]

3.1 LPM Parametrization and Connection with Supporting Hyperplanes of $\mathcal{C}$

For an LPM ϕ (2), Proposition 2 guarantees the existence of a unique Bayes confusion matrix on the boundary $\partial\mathcal{C}$. This is because optimum for a linear function over a strictly convex set is unique and lies on the boundary [2]. Note that any linear function with the same trade-offs for TP and TN, i.e. same (m_{11},m_{00}), is maximized at the same boundary point regardless of the bias term m_0. Thus, different LPMs can be generated by varying trade-offs $\mathbf{m}=(m_{11},m_{00})$ such that $\|\mathbf{m}\|=1$ and $m_0=0$. The condition $\|\mathbf{m}\|=1$ does not affect the learning problem as discussed in Example 1. In other words, the performance metric is scale invariant. This allows us to represent the family of linear metrics φ_{LPM} by a single parameter $\theta\in[0,2\pi]$:

$$\varphi_{LPM}=\{\mathbf{m}=(\cos\theta,\sin\theta):\theta\in[0,2\pi]\}. \quad (6)$$

Given $\mathbf{m}$ (equiv. to θ), we can recover the Bayes classifier using Proposition 1, and then the Bayes confusion matrix $\overline{C}_\theta=\overline{C}_\mathbf{m}=(\overline{TP}_\mathbf{m},\overline{TN}_\mathbf{m})$ using (1). Under Assumption 1, due to strict convexity of $\mathcal{C}$, the Bayes confusion matrix $\overline{C}_\mathbf{m}$ is unique; therefore, we have that

$$\langle\mathbf{m},C\rangle<\langle\mathbf{m},\overline{C}_\mathbf{m}\rangle \qquad \forall\, C\in\mathcal{C},C\neq\overline{C}_\mathbf{m}. \quad (7)$$

Notice the connection between the linear performance metrics and the supporting hyperplanes of the set $\mathcal{C}$ (see Figure 2(a)). Given $\mathbf{m}$, there exists a supporting hyperplane tangent to $\mathcal{C}$ at only $\overline{C}_\mathbf{m}$ defined as follows:

$$\overline{\ell}_\mathbf{m}:=m_{11}\cdot tp+m_{00}\cdot tn=m_{11}\overline{TP}_\mathbf{m}+m_{00}\overline{TN}_\mathbf{m}. \quad (8)$$

Clearly, if m_{11} and m_{00} are of opposite sign (i.e., $\theta\in(\pi/2,\pi)\cup(3\pi/2,2\pi)$), then $\overline{h}_\mathbf{m}$ is the trivial classifier predicting either 1 or 0 everywhere. In other words, if the slope of the hyperplane is positive, then it touches the set $\mathcal{C}$ either at $(\zeta,0)$ or $(0,1-\zeta)$. When $m_{11},m_{00}\neq 0$ with the same sign (i.e., $\theta\in(0,\pi/2)\cup(\pi,3\pi/2)$), then the Bayes confusion matrix is away from the two vertices. Now, we may split the boundary $\partial\mathcal{C}$ as follows:

Definition 6. *The Bayes confusion matrices for LPMs with $m_{11},m_{00}\geq 0$ ($\theta\in[0,\pi/2]$) form the upper boundary, denoted by $\partial\mathcal{C}_+$. The Bayes confusion matrices for LPMs with $m_{11},m_{00}<0$ ($\theta\in(\pi,3\pi/2)$) form the lower boundary, denoted by $\partial\mathcal{C}_-$. From Proposition 1, it follows that the confusion matrices in $\partial\mathcal{C}_+$ and $\partial\mathcal{C}_-$ correspond to the classifiers of the form $\mathbb{1}[\eta(x)\geq\delta]$ and $\mathbb{1}[\delta\geq\eta(x)]$, respectively, for some $\delta\in[0,1]$.*

4 ALGORITHMS

In this section, we propose binary-search type algorithms, which exploit the geometry of the set $\mathcal{C}$ (Section 3) to find the maximizer / minimizer and the associated supporting hyperplanes for any quasiconcave / quasiconvex metrics. These algorithms are then used to elicit LPMs and LFPMs, both of which belong to both quasiconcave and quasiconvex function families.

We allow *noisy* oracles; however, for simplicity, we will first discuss algorithms and elicitation with no-noise, and then show that they are robust to the noisy feedback (Section 6). Moreover, as one typically prefers metrics which reward correct classification, we first discuss metrics that are monotonically increasing in both TP and TN. The monotonically decreasing case is discussed in Appendix D as a natural extension.

The following lemma for any quasiconcave and quasiconvex metrics forms the basis of our proposed algorithms.

Lemma 1. *Let $\rho^+:[0,1]\to\partial\mathcal{C}_+$, $\rho^-:[0,1]\to\partial\mathcal{C}_-$ be continuous, bijective, parametrizations of the upper and lower boundary, respectively. Let $\phi:\mathcal{C}\to\mathbb{R}$ be a quasiconcave function, and $\psi:\mathcal{C}\to\mathbb{R}$ be a quasiconvex function, which are monotone increasing in both TP and TN. Then the composition $\phi\circ\rho^+:[0,1]\to\mathbb{R}$ is quasiconcave (and therefore unimodal) on the interval $[0,1]$, and $\psi\circ\rho^-:[0,1]\to\mathbb{R}$ is quasiconvex (and therefore unimodal) on the interval $[0,1]$.*

[3] Additional visual intuition about the geometry of C (via an example) is given in Appendix A.

Algorithm 1 Quasiconcave Metric Maximization

```
1: Input: $\epsilon > 0$ and oracle $\Omega$.
2: Initialize: $\theta_a = 0$, $\theta_b = \frac{\pi}{2}$.
3: while $|\theta_b - \theta_a| > \epsilon$ do
4:     Set $\theta_c = \frac{3\theta_a+\theta_b}{4}$, $\theta_d = \frac{\theta_a+\theta_b}{2}$, and $\theta_e = \frac{\theta_a+3\theta_b}{4}$.
       Set corresponding slopes ($\mathbf{m}$'s) using (6).
5:     Obtain $\overline{h}_{\theta_a}$,$\overline{h}_{\theta_c}$,$\overline{h}_{\theta_d}$, $\overline{h}_{\theta_e}$, $\overline{h}_{\theta_b}$ using Proposition 1.
       Compute $\overline{C}_{\theta_a}$,$\overline{C}_{\theta_c}$,$\overline{C}_{\theta_d}$,$\overline{C}_{\theta_e}$, $\overline{C}_{\theta_b}$ using (1).
6:     Query $\Omega(\overline{C}_{\theta_c}, \overline{C}_{\theta_a}), \Omega(\overline{C}_{\theta_d}, \overline{C}_{\theta_c}), \Omega(\overline{C}_{\theta_e}, \overline{C}_{\theta_d})$,
       and $\Omega(\overline{C}_{\theta_b}, \overline{C}_{\theta_e})$.
7:     If $\overline{C}_{\theta} \succ \overline{C}_{\theta'} \prec \overline{C}_{\theta''}$ for consecutive $\theta < \theta' < \theta''$,
       assume the default order $\overline{C}_{\theta} \prec \overline{C}_{\theta'} \prec \overline{C}_{\theta''}$.
8:     if $(\overline{C}_{\theta_a} \succ \overline{C}_{\theta_c})$ Set $\theta_b = \theta_d$.
9:     elseif $(\overline{C}_{\theta_a} \prec \overline{C}_{\theta_c} \succ \overline{C}_{\theta_d})$ Set $\theta_b = \theta_d$.
10:    elseif $(\overline{C}_{\theta_c} \prec \overline{C}_{\theta_d} \succ \overline{C}_{\theta_e})$ Set $\theta_a = \theta_c$, $\theta_b = \theta_e$.
11:    elseif $(\overline{C}_{\theta_d} \prec \overline{C}_{\theta_e} \succ \overline{C}_{\theta_b})$ Set $\theta_a = \theta_d$.
12:    else Set $\theta_a = \theta_d$.
13: Output: $\overline{\mathbf{m}}, \overline{C}$, and $\overline{\ell}$, where $\overline{\mathbf{m}} = \mathbf{m}_d\ (\theta_d)$, $\overline{C} = \overline{C}_{\theta_d}$, and $\overline{\ell} := \langle \overline{\mathbf{m}}, (tp, tn) \rangle = \langle \overline{\mathbf{m}}, \overline{C} \rangle$.
```

The unimodality of quasiconcave (quasiconvex) metrics on the upper (lower) boundary of the set $\mathcal{C}$ along with the one-dimensional parametrization of $\mathbf{m}$ using $\theta \in [0, 2\pi]$ (Section 3) allows us to devise binary-search-type methods to find the maximizer $\overline{C}$, the minimizer $\underline{C}$, and the first order approximation of ϕ at these points, i.e., the supporting hyperplanes at $\overline{C}$ and $\underline{C}$.

Algorithm 1. *Maximizing quasiconcave metrics and finding supporting hyperplanes at the optimum:* Since ϕ is monotonically increasing in both *TP* and *TN*, and $\mathcal{C}$ is convex, the maximizer must be on the upper boundary. Hence, we start with the interval $[\theta_a = 0, \theta_b = \frac{\pi}{2}]$ (Definition 6). We divide it into four equal parts and set slopes using (6) in line 4 (see Figure 2(b) for visual intuition). Then, we compute the Bayes classifiers using Proposition 1 and the associated Bayes confusion matrices in line 5. We pose four pairwise queries to the oracle in line 6. Line 7 gives the default direction to binary search in case of out-of-order responses.[4] In lines 8-12, we shrink the search interval by half based on oracle responses. We stop when the search interval becomes smaller than a given $\epsilon > 0$ (tolerance). Lastly, we output the slope $\overline{\mathbf{m}}$, the Bayes confusion matrix $\overline{C}$, and the supporting hyperplane $\overline{\ell}$ at that point.

Algorithm 2. *Minimizing quasiconvex metrics and finding supporting hyperplane at the optimum:* The same algorithm can be used for quasiconvex minimization with only two changes. First, we start with $\theta \in [\pi, \frac{3}{2}\pi]$, because the optimum will lie on the lower boundary $\partial\mathcal{C}_-$. Second, we check for $C \prec C'$ whenever Algorithm 1 checks for $C \succ C'$, and vice versa. Here, we output the counterparts, i.e., slope $\underline{\mathbf{m}}$, inverse Bayes Confusion matrix $\underline{C}$, and supporting hyperplane $\underline{\ell}$.

LPM Elicitation (True metric $\phi^* = \mathbf{m}^*$)

1. Run Algorithm 1 to get $\overline{C}^*$ and a hyperplane $\overline{\ell}$.
2. Set the elicited metric to be the slope of $\overline{\ell}$.

LFPM Elicitation (True metric ϕ^*)

1. Run Algorithm 1 to get $\overline{C}^*$, a hyperplane $\overline{\ell}$, and SoE (9).
2. Run Algorithm 2 to get $\underline{C}^*$, a hyperplane $\underline{\ell}$, and SoE (10).
3. Run the oracle-query independent Algorithm 3 to get the elicited metric, which satisfies both the SoEs.

Fig. 3: LPM and LFPM elicitation procedures.

5 METRIC ELICITATION

In this section, we discuss how Algorithms 1, 2, and 3 (described later) are used as subroutines to elicit LPMs and LFPMs. See Figure 3 for a brief summary.

5.1 Eliciting LPMs

Suppose that the oracle's metric is $\varphi_{LPM} \ni \phi^* = \mathbf{m}^*$, where, WLOG, $\|\mathbf{m}^*\| = 1$ and $m_0^* = 0$ (Section 3). Application of Algorithm 1 to the oracle, who responds according to $\mathbf{m}^*$, returns the maximizer and supporting hyperplane at that point. Since the true performance metric is linear, we take the elicited metric, $\hat{\mathbf{m}}$, to be the slope of the resulting supporting hyperplane.

5.2 Eliciting LFPMs

An LFPM is given by (3), where p_{11}, p_{00}, q_{11}, and q_{00} are not simultaneously zero. Also, it is bounded over $\mathcal{C}$. As scaling and shifting does not change the linear-fractional form, *WLOG*, we may take $\phi(C) \in [0, 1]\, \forall C \in \mathcal{C}$ with positive numerator and denominator.

Assumption 2. *Let $\phi \in \varphi_{LFPM}$ (3). We assume that $p_{11}, p_{00} \geq 0$, $p_{11} \geq q_{11}$, $p_{00} \geq q_{00}$, $p_0 = 0$, $q_0 = (p_{11} - q_{11})\zeta + (p_{00} - q_{00})(1 - \zeta)$, and $p_{11} + p_{00} = 1$.*

Proposition 3. *The conditions in Assumption 2 are sufficient for $\phi \in \varphi_{LFPM}$ to be bounded in $[0, 1]$ and simultaneously monotonically increasing in TP and TN.*

The conditions in Assumption 2 are reasonable as we want to elicit any unknown bounded, monotonically increasing LFPM. To no surprise, examples outlined in (4) and Koyejo et al. [14] satisfy these conditions. We first provide intuition for eliciting LFPMs (Figure 3). We obtain two hyperplanes: one at the maximizer on the upper boundary, and other at the minimizer on the lower boundary. This results in two nonlinear systems of equations (SoEs) having only one degree of freedom, but they are satisfied by the true unknown metric. Thus, the elicited metric is one where solutions to the two systems match pointwise on the confusion matrices. Formally, suppose that the oracle's metric is:

[4] Due to finite samples, $\mathcal{C}$'s boundary may have staircase-type bumps in practice. This may lead to out-of-order responses, even when the metric is unimodal *w.r.t.* θ.

$$\phi^*(C) = \frac{p^*_{11}TP + p^*_{00}TN}{q^*_{11}TP + q^*_{00}TN + q^*_0}.$$

Let $\overline{\tau}^*$ and $\underline{\tau}^*$ be the maximum and minimum value of ϕ^* over $\mathcal{C}$, respectively, i.e., $\underline{\tau}^* \leq \phi^*(C) \leq \overline{\tau}^* \; \forall \, C \in \mathcal{C}$. Under Assumption 1, we have a hyperplane

$$\overline{\ell}^*_f := (p^*_{11} - \overline{\tau}^* q^*_{11})tp + (p^*_{11} - \overline{\tau}^* q^*_{11})tn = \overline{\tau}^* q^*_0$$

touching the set $\mathcal{C}$ only at $(\overline{TP}^*, \overline{TN}^*)$ on the upper boundary $\partial\mathcal{C}_+$. Similarly, we have a hyperplane

$$\underline{\ell}^*_f := (p^*_{11} - \underline{\tau}^* q^*_{11})tp + (p^*_{00} - \underline{\tau}^* q^*_{00})tn = \underline{\tau}^* q^*_0,$$

which touches the set $\mathcal{C}$ only at $(\underline{TP}^*, \underline{TN}^*)$ on the lower boundary $\partial\mathcal{C}_-$. To help with intuition, see Figure 2(c). Since LFPM is quasiconcave, Algorithm 1 returns a hyperplane $\overline{\ell} := \overline{m}_{11}tp + \overline{m}_{00}tn = \overline{C}_0$, where $\overline{C}_0 = \overline{m}_{11}\overline{TP}^* + \overline{m}_{00}\overline{TN}^*$. This is equivalent to $\overline{\ell}^*_f$ up to a constant multiple; therefore, the true metric is the solution to the following non-linear SoE:

$$p^*_{11} - \overline{\tau}^* q^*_{11} = \alpha \overline{m}_{11}, p^*_{00} - \overline{\tau}^* q^*_{00} = \alpha \overline{m}_{00}, \overline{\tau}^* q^*_0 = \alpha \overline{C}_0,$$

where $\alpha \geq 0$, because LHS and $\overline{m}$'s are non-negative. Additionally, we ignore the case when $\alpha = 0$, since this would imply a constant ϕ. Next, we may divide the above equations by $\alpha > 0$ on both sides so that all the coefficients $\overline{p}^*$'s and $\overline{q}^*$'s are factored by α. This does not change ϕ^*; thus, the SoE becomes:

$$p'_{11} - \overline{\tau}^* q'_{11} = \overline{m}_{11}, p'_{00} - \overline{\tau}^* q'_{00} = \overline{m}_{00}, \overline{\tau}^* q'_0 = \overline{C}_0. \tag{9}$$

Notice that none of the conditions in Assumption 2 are changed except $p'_{11} + p'_{00} = 1$. However, we may still use this condition to learn a constant α times the true metric, which does not harm the elicitation problem.

As LFPM is also quasiconvex, Algorithm 2 outputs a hyperplane $\underline{\ell} := \underline{m}_{11}tp + \underline{m}_{00}tn = \underline{C}_0$, where $\underline{C}_0 = \underline{m}_{11}\underline{TP}^* + \underline{m}_{00}\underline{TN}^*$. This is equivalent to $\underline{\ell}^*_f$ up to a constant multiple; thus, the true metric is also the solution of the following SoE:

$$p^*_{11} - \underline{\tau}^* q^*_{11} = \gamma \underline{m}_{11}, p^*_{00} - \underline{\tau}^* q^*_{00} = \gamma \underline{m}_{00}, \underline{\tau}^* q^*_0 = \gamma \underline{C}_0,$$

where $\gamma \leq 0$ since LHS is positive, but $\underline{m}$'s are negative. Again, we may assume $\gamma < 0$. By dividing the above equations by $-\gamma$ on both sides, all the coefficients p^*'s and q^*'s are factored by $-\gamma$. This does not change ϕ^*; thus, the system of equations becomes the following:

$$p''_{11} - \underline{\tau}^* q''_{11} = \underline{m}_{11}, p''_{00} - \underline{\tau}^* q''_{00} = \underline{m}_{00}, \underline{\tau}^* q''_0 = \underline{C}_0. \tag{10}$$

Proposition 4. *Under Assumption 2, knowing p'_{11} solves the system of equations (9) as follows:*

$$p'_{00} = 1 - p'_{11}, \; q'_0 = \overline{C}_0 \frac{P'}{Q'},$$

$$q'_{11} = (p'_{11} - \overline{m}_{11})\frac{P'}{Q'}, \; q'_{00} = (p'_{00} - \overline{m}_{00})\frac{P'}{Q'}, \tag{11}$$

where $P' = p'_{11}\zeta + p'_{00}(1-\zeta)$ and $Q' = P' + \overline{C}_0 - \overline{m}_{11}\zeta - \overline{m}_{00}(1-\zeta)$. Thus, it elicits the LFPM.

Algorithm 3 Grid Search for Best Ratio

```
Input: $k, \Delta$.
Initialize: $\sigma_{opt} = \infty, p'_{11,opt} = 0$.
Generate $C_1, ..., C_k$ on $\partial\mathcal{C}_+$ and $\partial\mathcal{C}_-$ (Section 3).
for ($p'_{11} = 0$; $p'_{11} \leq 1$; $p'_{11} = p'_{11} + \Delta$) do
    Compute $\phi'$, $\phi''$ using Proposition 4. Compute array $r = [\frac{\phi'(C_1)}{\phi''(C_1)}, ..., \frac{\phi'(C_k)}{\phi''(C_k)}]$. Set $\sigma = \text{std}(r)$.
    if ($\sigma < \sigma_{opt}$) Set $\sigma_{opt} = \sigma$ and $p'_{11,opt} = p'_{11}$.
Output: $p'_{11,opt}$.
```

Now assume we know p'_{11}. Using Proposition 4, we may solve the system (9) and obtain a metric, say ϕ'. System (10) can be solved analogously, provided we know p''_{11}, to get a metric, say ϕ''. Notice that when $p^*_{11}/p^*_{00} = p'_{11}/p'_{00} = p''_{11}/p''_{00}$, then $\phi^*(C) = \phi'(C)/\alpha = -\phi''(C)/\gamma$. This means that when the true ratios of p's are known, then ϕ', ϕ'' are constant multiples of each other. So, to know the true p'_{11} (or, p''_{11}) is to search the grid $[0, 1]$ and select the one where the ratios of ϕ' and ϕ'' are constant on a number of confusion matrices. Since we can generate many confusion matrices on $\partial\mathcal{C}_+$ and $\partial\mathcal{C}_-$ (vary δ in Definition 6), we can estimate the ratio p'_{11} to p'_{00} using grid search based Algorithm 3. We may then use Proposition 4 for the output of Algorithm 3 and set the elicited metric $\hat{\phi} = \phi'$. Note that Algorithm 3 is independent of oracle queries and easy to implement, thus it is suitable for the purpose.

6 GUARANTEES

In this section, we discuss guarantees for the elicitation procedures (Section 5) in the presence of (a) confusion matrices' estimation noise from finite samples and (b) oracle feedback noise with the following notion.

Definition 7. *Oracle Feedback Noise ($\epsilon_\Omega \geq 0$): The oracle may provide wrong answers whenever $|\phi(C) - \phi(C')| < \epsilon_\Omega$. Otherwise, it provides correct answers.*

Simply put, if the confusion matrices are close as measured by ϕ, then the oracle responses can be wrong. Moving forward to the guarantees, we make two assumptions which hold in most common settings.

Assumption 3. *Let $\{\hat{\eta}_i(x)\}_{i=1}^n$ be a sequence of estimates of $\eta(x)$ depending on the sample size. We assume that $\|\eta - \hat{\eta}_i\|_\infty \xrightarrow{P} 0$.*

Assumption 4. *For quasiconcave ϕ, recall that the Bayes classifier is of the form $h = \mathbb{1}[\eta(x) \geq \delta]$. Let $\overline{\delta}$ be the threshold that maximizes ϕ. We assume that the probability that $\eta(X)$ lies near $\overline{\delta}$ is bounded from below and above. Formally, $k_0\nu \leq$*

$\mathbb{P}\left[(\bar{\delta} - \eta(X)) \in [0, \nu]\right], \mathbb{P}\left[(\eta(X) - \bar{\delta}) \in [0, \nu]\right] \le k_1 \nu$ *for any* $0 < \nu \le \frac{2}{k_0}\sqrt{k_1 \epsilon_\Omega}$ *and some* $k_1 \ge k_0 > 0$.

Assumption 3 is arguably natural, as most estimation is parametric, where the function classes are sufficiently well behaved. Assumption 4 ensures that near the optimal threshold $\bar{\delta}$, the values of $\eta(X)$ have bounded density. In other words, when X has no point mass, the slope of $\eta(X)$ where it attains the optimal threshold $\bar{\delta}$ is neither vertical nor horizontal. We start with guarantees for the algorithms in their respective tasks.

Theorem 1. *Given* $\epsilon, \epsilon_\Omega \ge 0$ *and a 1-Lipschitz metric* ϕ *that is monotonically increasing in TP, TN. If it is quasiconcave (quasiconvex) then Algorithm 1 (Algorithm 2) finds an approximate maximizer* $\overline{C}$ *(minimizer* $\underline{C}$*). Furthemore, (i) the algorithm returns the supporting hyperplane at that point, (ii) the value of* ϕ *at that point is within* $O(\sqrt{\epsilon_\Omega} + \epsilon)$ *of the optimum, and (iii) the number of queries is* $O(\log \frac{1}{\epsilon})$.

Lemma 2. *Under our model, no algorithm can find the maximizer (minimizer) in fewer than* $O(\log \frac{1}{\epsilon})$ *queries.*

Theorem 1 and Lemma 2, guarantee that Algorithm 1 (Algorithm 2), for a quasiconcave (quasiconvex) metric, finds a confusion matrix and a hypeplane which is close to the true maximizer (minimizer) and its associated supporting hyperplane, using just the optimal number of queries. Further, since binary search always tends towards the optimal whenever responses are correct, the algorithms necessarily terminate within a confidence interval of the true maximizer. Thus, we can take ϵ sufficiently small so that the only error that arises is due to the feedback noise ϵ_Ω. Now, we present our main result which guarantees effective LPM elicitation. Guarantees in LFPM elicitation follow naturally as discussed in the proof of Theorem 2 (Appendix B).

Theorem 2. *Let* $\varphi_{LPM} \ni \phi^* = \mathbf{m}^*$ *be the true performance metric. Under Assumption 4, given* $\epsilon > 0$*, LPM elicitation (Section 5.1) outputs a performance metric* $\hat{\phi} = \hat{\mathbf{m}}$*, such that* $\|\mathbf{m}^* - \hat{\mathbf{m}}\|_\infty \le \sqrt{2}\epsilon + \frac{2}{k_0}\sqrt{2k_1\epsilon_\Omega}$.

So far, we assumed access to the confusion matrices. However, in practice, we need to estimate them using samples $\{(x_i, y_i)\}_{i=1}^n$. We now discuss robustness of the algorithms working with samples. Recall that, as a standard consequence of Chernoff-type bounds [1], sample estimates of true-positive and true-negative are consistent estimators. Therefore, with high probability, we can estimate the confusion matrix within any desired tolerance, provided we have sufficient samples. This implies that we can also estimate the ϕ values within any tolerance since LPM and LFPM are 1-Lipschitz due to (6) and Assumption 2, respectively. Thus, with high probability, the elicitation procedures gather correct oracle's preferences within feedback noise ϵ_Ω. Further, we may prove the following lemma which allow us to control the error in optimal classifiers from using the estimated $\hat{\eta}(x)$ rather than the true $\eta(x)$.

Table 1: LPM elicitation at tolerance $\epsilon = 0.02$ radians.

$\phi^* = \mathbf{m}^*$	$\hat{\phi} = \hat{\mathbf{m}}$	$\phi^* = \mathbf{m}^*$	$\hat{\phi} = \hat{\mathbf{m}}$
(0.98,0.17)	(0.99,0.17)	(-0.94,-0.34)	(-0.94,-0.34)
(0.64,0.77)	(0.64,0.77)	(-0.50,-0.87)	(-0.50,-0.87)

Lemma 3. *Let* h_θ *and* $\hat{h}_\theta$ *be two classifiers estimated using* η *and* $\hat{\eta}$*, respectively. Further, let* $\bar{\theta}$ *be such that* $h_{\bar{\theta}} = \arg\max_\theta \phi(h_\theta)$*. Then* $\|C(\hat{h}_{\bar{\theta}}) - C(h_{\bar{\theta}})\|_\infty = O(\|\hat{\eta}_n - \eta\|_\infty)$.

The errors due to using $\hat{\eta}$, instead of true η may propel in the results discussed earlier, however, only in the bounded sense. This shows that our elicitation approach is robust to feedback and finite sample noise.

7 EXPERIMENTS

In this section, we empirically validate the theory and investigate the sensitivity due to sample estimates.[5]

7.1 Synthetic Data Experiments

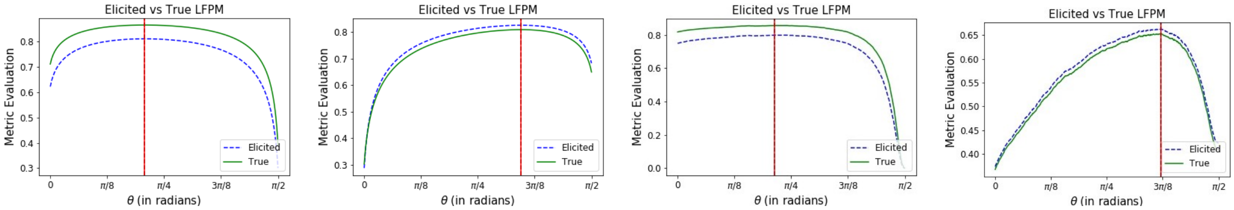
We assume a joint probability for $\mathcal{X} = [-1, 1]$ and $\mathcal{Y} = \{0, 1\}$ given by $f_X = \mathbb{U}[-1, 1]$ and $\eta(x) = \frac{1}{1+e^{ax}}$, where $\mathbb{U}[-1, 1]$ is the uniform distribution on $[-1, 1]$, and a is a parameter controlling the degree of noise in the labels. We fix $a = 5$ in our experiments. To verify LPM elicitation, we first define a true metric ϕ^*. This specifies the query outputs in line 6 of Algorithm 1 (Algorithm 2). Then we run LPM elicitation procedure (Section 5.1) to check whether or not we compute the same metric. Some results are shown in Table 1. We elicit the true metrics even for $\epsilon = 0.02$ radians.

Next, we elicit LFPM. We define a true metric ϕ^* by $\{(p_{11}^*, p_{00}^*), (q_{11}^*, q_{00}^*, q_0^*)\}$. Then we follow the LFPM elicitation procedure (Section 5.2), where Algorithms 1 and 2 are run with $\epsilon = 0.05$ and Algorithm 3 is run with $k = 2000$ and $\Delta = 0.01$. The elicited metric $\hat{\phi}$ is denoted by $\{(\hat{p}_{11}, \hat{p}_{00}), (\hat{q}_{11}, \hat{q}_{00}, \hat{q}_0)\}$ and presented in Table 2 (Column 2). We also present mean (α) and standard deviation (σ) of the ratio of the elicited metric $\hat{\phi}$ to the true metric ϕ^* over a subset of confusion matrices (columns 3 and 4). For improved comparisons, Figure 4 shows the true and elicited metrics evaluated on selected pairs of $(TP, TN) \in \partial\mathcal{C}_+$. The metrics are plotted together after sorting the slope parameter θ. Clearly, the elicited metric is a constant multiple of the true metric. We also see that the *argmax* of the true and elicited metric coincide, thus validating Theorem 1.

[5] A subset of results is shown here. Please refer Appendix C for extended set of results.

Table 2: LFPM Elicitation for synthetic distribution (Section 7.1) and Magic (M) dataset (Section 7.2). α and σ are the mean and standard deviation of $\hat{\phi}/\phi^*$ evaluated over a subset of confusion matrices used in Algorithm 3.

True Metric	Results on Synthetic Distribution (Section 7.1)			Results on Real World Dataset M (Section 7.2)		
$(p^*_{11}, p^*_{00}), (q^*_{11}, q^*_{00}, q^*_0)$	$(\hat{p}_{11}, \hat{p}_{00}), (\hat{q}_{11}, \hat{q}_{00}, \hat{q}_0)$	α	σ	$(\hat{p}_{11}, \hat{p}_{00}), (\hat{q}_{11}, \hat{q}_{00}, \hat{q}_0)$	α	σ
(1.00,0.00),(0.50,-0.50,0.50)	(1.00,0.00),(0.25,-0.75,0.75)	0.92	0.03	(1.00,0.00),(0.25,-0.75,0.75)	0.90	0.06
(0.20,0.80),(-0.40,-0.20,0.80)	(0.12, 0.88),(-0.43, 0.002, 0.71)	1.02	0.006	(0.19,0.81),(-0.38,-0.13,0.70)	1.02	0.004



(a) Table 2, line 1, column 2 (b) Table 2, line 2, column 2 (c) Table 2, line 1, column 5 (d) Table 2, line 2, column 5

Fig. 4: True (solid green) and elicited (dashed blue) LFPMs for synthetic distribution and dataset M from Table 2. The solid red and coinciding dashed black vertical lines are *argmax* of the true and elicited metric, respectively.

7.2 Real-World Data Experiments

Now, we validate the elicitation procedures with two real-world datasets. The datasets are: (a) Breast Cancer (BC) Wisconsin Diagnostic dataset [25] containing 569 instances, and (b) Magic (M) dataset [8] containing 19020 instances. For both the datasets, we standardize the features and split the data into two parts $\mathcal{S}_1$ and $\mathcal{S}_2$. On $\mathcal{S}_1$, we learn the estimator $\hat{\eta}$ using regularized logistic regression model. We use $\mathcal{S}_2$ for making predictions and computing sample confusion matrices.

We randomly selected twenty-eight LPMs by choosing θ^* ($\mathbf{m}^*$). We then used Algorithm 1 (Algortihm 2) with different tolerance ϵ and for different datasets and recovered the estimate $\hat{\mathbf{m}}$ using LPM elicitation. In Table 5 of Appendix C, we report the proportion of the number of times when our procedure failed to recover the true $\mathbf{m}^*$. We see improved elicitation for dataset M, suggesting that ME improves with larger datasets. In particular, for dataset M, we elicit all the metrics within threshold $\epsilon = 0.11$ radians. We also observe that $\epsilon = 0.02$ is an overly tight tolerance for both the datasets leading to many failures. This is because the elicitation routine gets stuck at the closest achievable confusion matrix from finite samples, which need not be optimal within the given (small) tolerance.

Next, we evaluate LFPM elicitation using dataset M. We define the same true metrics and follow the same LFPM elicitation process as defined in Section 7.1. In Table 2 (columns 5, 6, and 7), we present the elicitation results along with mean α and standard deviation σ of the ratio of the elicited metric and the true metric. We also show the true and elicited metrics evaluated on the selected pairs of $(TP, TN) \in \partial\mathcal{C}_+$ in Figure 4, ordered by the parameter θ. We see that the elicited metrics are equivalent to the true metrics up to a constant.

8 RELATED WORK

Our work may be compared to ranking from pairwise comparisons [28]. However, we note that our results depend on novel geometric ideas on the space of confusion matrices. Thus, instead of a ranking problem, we show that ME in standard models can be reduced to just finding the maximizer (and minimizer) of an unknown function which in turn yields the true metric – resulting in low query complexity. A direct ranking approach adds unnecessary complexity to achieve the same task. Further, in contrast to our approach, most large margin ordinal regression based ranking [11] fail to control which samples are queried. There is another line of work, which actively controls the query samples for ranking, e.g., [12]. However, to our knowledge, this requires that the number of objects is finite and finite dimensional – thus cannot be directly applied to ME without significant modifications, e.g. exploiting confusion matrix properties, as we have. Learning a performance metric which correlates with human preferences has been studied before [13, 18]; however, these studies learn a regression function over some predefined features which is fundamentally different from our problem. Lastly, while [4, 10] address how one might qualitatively choose between metrics, none addresses our central contribution – a principled approach for eliciting the ideal metric from user feedback.

9 CONCLUSION

We conceptualize *metric elicitation* and elicit linear and linear-fractional metrics using preference feedback over pairs of classifiers. We propose provably query efficient and robust algorithms which exploit key properties of the set of confusion matrices. In future, we plan to explore metric elicitation beyond binary classification.

References

[1] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[2] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] D. Braziunas and C. Boutilier. Minimax regret based elicitation of generalized additive utilities. In *UAI*, pages 25–32, 2007.

[4] R. Caruana and A. Niculescu-Mizil. Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *ACM SIGKDD*, pages 69–78. ACM, 2004.

[5] T. H. Cormen. *Introduction to algorithms*. MIT press, 2009.

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

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

[8] J. Dvořák and P. Savickỳ. Softening splits in decision trees using simulated annealing. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 721–729. Springer, 2007.

[9] C. Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[10] C. Ferri, J. Hernández-Orallo, and R. Modroiu. An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1):27–38, 2009.

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

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

[13] F. Janssen and J. Furnkranz. On meta-learning rule learning heuristics. In *ICDM*, pages 529–534. IEEE, 2007.

[14] O. O. Koyejo, N. Natarajan, P. K. Ravikumar, and I. S. Dhillon. Consistent binary classification with generalized performance metrics. In *NIPS*, pages 2744–2752, 2014.

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

[16] A. Mas-Colell. The recoverability of consumers' preferences from market demand behavior. *Econometrica: Journal of the Econometric Society*, pages 1409–1430, 1977.

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

[18] 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.

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

[20] 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.

[21] P. A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938.

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

[23] H. C. Sox. *Medical decision making*. ACP Press, 1988.

[24] I. Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2):225–287, 2007.

[25] W. N. Street, W. H. Wolberg, and O. L. Mangasarian. Nuclear feature extraction for breast tumor diagnosis. In *Biomedical Image Processing and Biomedical Visualization*, volume 1905, pages 861–871. International Society for Optics and Photonics, 1993.

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

[27] H. R. Varian. *Revealed preference. In Samuelsonian Economics and the 21st Century by M. Szenberg and L. Ramrattand and A. A. Gottesman (editors)*. Oxford University Press, 2005.

[28] F. Wauthier, M. Jordan, and N. Jojic. Efficient ranking from pairwise comparisons. In *ICML*, pages 109–117, 2013.