Electronic Journal of Statistics
Vol. 11 (2017) 5198–5231
ISSN: 1935-7524
https://doi.org/10.1214/17-EJS1376SI

On the interpretability of conditional probability estimates in the agnostic setting

Yihan Gao, Aditya Parameswaran and Jian Peng

University of Illinois at Urbana-Champaign
201 N Goodwin Ave
Urbana, IL 61801
e-mail: ygao34@illinois.edu; adityagp@illinois.edu; jianpeng@illinois.edu

Abstract: We study the interpretability of conditional probability estimates for binary classification under the agnostic setting or scenario. Under the agnostic setting, conditional probability estimates do not necessarily reflect the true conditional probabilities. Instead, they have a certain calibration property: among all data points that the classifier has predicted $\mathcal{P}(Y = 1|X) = p$, p portion of them actually have label $Y = 1$. For cost-sensitive decision problems, this calibration property provides adequate support for us to use Bayes Decision Rule. In this paper, we define a novel measure for the calibration property together with its empirical counterpart, and prove a uniform convergence result between them. This new measure enables us to formally justify the calibration property of conditional probability estimations. It also provides new insights on the problem of estimating and calibrating conditional probabilities, and allows us to reliably estimate the expected cost of decision rules when applied to an unlabeled dataset.

Received June 2017.

1. Introduction

Many binary classification algorithms, such as naive Bayes and logistic regression, naturally produce confidence measures in the form of conditional probability of labels. These confidence measures are usually interpreted as the conditional probability of the label $y = 1$ given the feature x. An important research question is how to justify these conditional probabilities, i.e., how to prove the trustworthiness of such results.

In classical statistics, this question is usually studied under the realizable assumption, which assumes that the true underlying probability distribution has the same parametric form as the model assumption. More explicitly, statisticians usually construct a parametric conditional distribution $\mathcal{P}(Y|X, \theta)$, and assume that the true conditional distribution is also of this form (with unknown θ). The justification of conditional probabilities can then be achieved by using either hypothesis testing or confidence interval estimation on θ.

However, in modern data analysis workflows, the realizable assumption is often violated, e.g., data analysts usually try out several off-the-shelf classification algorithms to identify those that work the best. This setting is often called

agnostic — essentially implying that we do not have any knowledge about the underlying distribution. Under the agnostic setting, conditional probability estimates can no longer be justified by standard statistical tools, as most hypothesis testing methods are designed to distinguish two parameter areas in the hypothesis space (e.g., $\theta < \theta_0$ v.s. $\theta \geq \theta_0$), and confidence intervals require realizable assumption to be interpretable.

In this paper, we study the interpretability of conditional probability estimates in the agnostic binary classification setting: what kind of guarantees can we have without making any assumption on the underlying distribution? Justifying these conditional probability estimates is important for applications that explicitly utilize them, including medical diagnostic systems [6] and fraud detection [8]. In such applications, the misclassification loss function is often asymmetric (i.e., false positive and false negative incur different loss), and accurate conditional probability estimates are crucial empirically. In particular, in medical diagnostic systems, a false positive means additional tests are needed, while a false negative could potentially be fatal.

Summary of notation

We focus on the binary classification problem in this paper. Let us first define some notation here that will be used throughout the paper:

- $\mathcal{X}$ denotes the feature space and $\mathcal{Y} = \{\pm 1\}$ denotes the label space.
- $\mathcal{P}$ denotes the underlying probability distribution over $\mathcal{X} \times \mathcal{Y}$ that governs the generation of datasets.
- $D = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ denotes a set of i.i.d. data points from $\mathcal{P}$.
- A *fuzzy classifier* is a function from $\mathcal{X}$ to $[0, 1]$ where the output denotes the estimated conditional probability of $\mathcal{P}(Y = 1|X)$.

Interpretations of conditional probability estimates

Ideally, we hope that our conditional probability estimates can be interpreted as the true conditional probabilities. This interpretation is justified if we can prove that the conditional probability estimates are close to the true values. Let $l_1(f, \mathcal{P})$ be the l_1 distance between the true distribution and the estimated distribution as a measure of the "correctness" of conditional probability estimates:

$$l_1(f, \mathcal{P}) = \mathbb{E}_{X \sim \mathcal{P}} |f(X) - \mathcal{P}(Y = 1|X)|$$

Here X is a random variable representing the feature vector of a sample data point, Y is the label of X and $f(X)$ is a fuzzy classifier that estimates $\mathcal{P}(Y = 1|X)$. If we can prove that $l_1(f, \mathcal{P}) \leq \epsilon$ for some small ϵ, then the output of f can be approximately interpreted as the true conditional probability.

Unfortunately, as we will show in this paper, it is impossible to guarantee any reasonably small upper bound for $l_1(f, \mathcal{P})$ under the agnostic setting. In fact,

as we will demonstrate in this paper, for many such situations, the estimated conditional probabilities are usually no longer close to the true values in practice.

Therefore, instead of trying to bound the l_1 distance, we develop an alternative interpretation for these conditional probability estimates. We introduce the following calibration definition for fuzzy classifiers:

Definition 1. *Let $\mathcal{X}$ be the feature space, $\mathcal{Y} = \{\pm 1\}$ be the label space and $\mathcal{P}$ be the distribution over $\mathcal{X} \times \mathcal{Y}$. Let $f : \mathcal{X} \to [0, 1]$ be a fuzzy classifier, then we say f is calibrated if for any $p_1 < p_2$, we have:*

$$\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)] = \mathcal{P}(Y = 1, p_1 < f(X) \leq p_2)$$

Intuitively, a fuzzy classifier is calibrated if its output correctly reflects the relative frequency of labels among instances they believe to be similar. For instance, suppose the classifier output $f(X) = p$ for n data points, then roughly there are np data points with label $Y = 1$. We also define a measure of how close f is to be calibrated:

Definition 2. *A fuzzy classifier f is ϵ-calibrated if*

$$c(f) = \sup_{p_1<p_2} |\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)] - \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)| \leq \epsilon$$

f is ϵ-empirically calibrated with respect to D if

$$c_{emp}(f, D) = \frac{1}{n} \sup_{p_1<p_2} |\sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2} f(X_i) - \sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2, Y_i=1}| \leq \epsilon$$

where $D = \{(X_i, Y_i), \ldots, (X_n, Y_n)\}$ is a size n dataset consisting of i.i.d. examples from $\mathcal{P}$.

Note that the empirical calibration measure $c_{emp}(f, D)$ can be efficiently computed on a finite dataset. We further prove that, with bounds on a certain complexity measure related to the hypothesis class, $c_{emp}(f, D)$ converges uniformly to $c(f)$ over all functions f in that hypothesis class. Therefore, the calibration property of these classifiers can be demonstrated by showing that they are empirically calibrated on the training data.

The calibration definition is motivated by analyzing the properties of commonly used conditional probability estimation algorithms: many such algorithms will generate classifiers that are naturally calibrated. Our calibration definition justifies the common practice of using calibrated conditional probability estimates as true conditional probabilities: we show that if the fuzzy classifier is (almost) calibrated and the output of the classifier is the only source of information, then applying Bayes Decision Rule on the conditional probability estimates would result in a (near) optimal strategy.

The uniform convergence result of $c_{emp}(f, D)$ and $c(f)$ has several applications. First, it can be directly used to prove a fuzzy classifier is (almost) calibrated, which makes the conditional probability estimates interpretable to users. Second, it suggests that we need to minimize the empirical calibration

measure to obtain calibrated classifiers, which is a new direction for designing conditional probability estimation algorithms. Third, taking uncalibrated conditional probability estimates as input, we can calibrate them by minimizing the calibration measure[1]. Finally, by using a calibrated fuzzy classifier f, one can reliably estimate the label frequency in any decision region $p_1 < f(X) \leq p_2$ for an unlabeled dataset, and thereby estimate the expected costs of decision rules with confidence bounds.

Paper outline

The rest of this paper is organized as following. In Section 2, we argue that the l_1 distance cannot be provably bounded under the agnostic setting (Theorem 1) and then motivate our calibration definition. In Section 3 we present the uniform convergence result (Theorem 2) and discuss the potential applications. In Section 4, we report experiments that illustrate the behavior of our calibration measure on several common classification algorithms. In Section 5, we discuss some potential extensions of the calibration measure to multi-class settings.

Related work

Our definition of calibration is similar to the definition of calibration in prediction theory [9], where the goal is also to make predicted probability values match the relative frequency of correct predictions. In prediction theory, the problem is formulated from a game-theoretic point of view: the sequence generator is assumed to be malevolent, and the goal is to design algorithms to achieve this calibration guarantee no matter what strategy the sequence generator uses.

There is another calibration measure in literature, which is derived from the Brier score decomposition [14]. It is defined as the weighted average of the squared distance between the actual relatively frequency of Y label and the predicted value for each unique prediction $f(X)$. Compared to our calibration measure, this calibration measure requires *binning* (i.e., $f(X)$ can only take value from a small predefined set), and does not enjoy the properties established in this paper.

To the best of our knowledge, there is no other work addressing the interpretability of conditional probability estimates in the agnostic setting. Our definition of calibration is also connected to the problem of calibrating conditional probability estimates, which has been studied in many papers [21, 17].

This paper is an extended version of our earlier conference paper [10], with the following new contents:

- The claim in Section 2.4 has been revised, which now shows that the optimality gap of Bayes Decision Rule is proportional to the calibration measure $c(f)$.

[1] In fact, one of the most well-known calibration algorithm, the isotonic regression algorithm, can be interpreted this way.

- A new section (Section 2.5) is added to discuss the intuition behind our calibration measure.
- An improved bound (Claim 4) between $c(f)$ and $c_{emp}(f, D)$ has been derived for the important case where we use an independent validation dataset to directly estimate $c(f)$ through $c_{emp}(f, D)$.
- A new application of our calibration measure is discussed in Section 3.3.4: we show that it is possible to use (almost) calibrated fuzzy classifiers to estimate the expected cost of decision rules (with confidence bound) when applied to an unlabeled dataset. The confidence interval length for such an estimation depends linearly on the calibration measure of the fuzzy classifier.
- A new experiment is added in Section 4: we demonstrate the application of estimating Receiver Operating Characteristics (ROC) curve using calibrated conditional probability estimates. We show that the expected deviation between the estimated ROC curve and the true ROC curve can be roughly estimated via the calibration measure.
- Finally, a new discussion section (Section 5) is added to discuss the potential extensions of the calibration measure to multi-class settings.

2. The calibration definition: Motivation & impossibility result

2.1. Impossibility result for l_1 distance

Recall that the l_1 distance between f and $\mathcal{P}$ is defined as:

$$l_1(f, \mathcal{P}) = \mathbb{E}_{X\sim\mathcal{P}}|f(X) - \mathcal{P}(Y = 1|X)|$$

Suppose f is our conditional probability estimator that we learned from the training dataset. We attempt to prove that the l_1 distance between f and $\mathcal{P}$ is small. In the agnostic setting, we do not know anything about $\mathcal{P}$, and the only tool we can utilize is a validation dataset D_{val} that consists of i.i.d. samples from $\mathcal{P}$. Therefore, our best hope would be a prover $A_f(D)$ that:

- Returns 1 with high probability if $l_1(f, \mathcal{P})$ is small.
- Returns 0 with high probability if $l_1(f, \mathcal{P})$ is large.

The following theorem states that no such prover exists, and the proof can be found in the appendix.

Theorem 1. *Let $\mathcal{Q}$ be a probability distribution over a discrete feature space $\mathcal{X}$, and $f : \mathcal{X} \to [0, 1]$ be a fuzzy classifier. Define B_f as:*

$$B_f = \mathbb{E}_{X\sim\mathcal{Q}} \min(f(X), 1 - f(X))$$

If there exists $\epsilon > 0$ such that $\forall x \in \mathcal{X}, \mathcal{Q}(x) < \frac{\epsilon}{n^2}$, then there is no prover $A_f : \{\mathcal{X} \times \mathcal{Y}\}^n \to \{0, 1\}$ for f satisfying the following two conditions:

For any probability distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$ such that $\mathcal{P}_X = \mathcal{Q}$ (i.e., $\forall x \in \mathcal{X}, \sum_{y\in\mathcal{Y}} \mathcal{P}(x, y) = \mathcal{Q}(x)$), suppose $D_{val} \in \{\mathcal{X} \times \mathcal{Y}\}^n$ is a validation dataset consisting of n i.i.d. samples from $\mathcal{P}$:

1. *If* $l_1(f, \mathcal{P}) = 0$, *then* $\mathbf{P}_{D_{val}}(A_f(D_{val}) = 1) > \frac{1+\epsilon}{2}$.
2. *If* $l_1(f, \mathcal{P}) \geq B_f$, *then* $\mathbf{P}_{D_{val}}(A_f(D_{val}) = 1) < \frac{1-\epsilon}{2}$.

The assumption we made in Theorem 1 (i.e., $\forall x \in \mathcal{X}, \mathcal{Q}(x) < \frac{\epsilon}{n^2}$) is to exclude the scenario where a significant amount of probability mass concentrates on a few data points so that their corresponding conditional probability can be estimated via repeated sampling. Note that the statement is not true in the extreme case where all probability mass concentrates on one single data point (i.e., $\exists x \in X, Q(x) = 1$). The assumption is true when the feature space $\mathcal{X}$ is large enough that it is almost impossible for any data point to have significant enough probability mass to be sampled more than once in the training dataset.

Note that Theorem 1 implies that it would be impossible to guarantee a small upper bound of $l_1(f, \mathcal{P})$ with high probability under the agnostic setting, if the dataset contains no (or very few) duplicate data points. Thus, we cannot interpret the conditional probability estimates as the true conditional probabilities under such setting. This result motivates us to develop a new interpretation of the conditional probability estimates, together with a new measure of "correctness" to justify the conditional probability estimates.

2.2. $l_1(f, \mathcal{P})$ *in practice*

The fact that we cannot guarantee an upper bound of the l_1 distance is not merely a theoretical artifact. In fact, when we are under the agnostic setting, the value of $l_1(f, \mathcal{P})$ is often very large in practice. Here we use the following document categorization example to demonstrate this point.

Example 1. *Denote Z to be the collection of all English words. In this problem the feature space $\mathcal{X} = Z^*$ is the collection of all possible word sequences, and $\mathcal{Y}$ denotes whether this document belongs to a certain topic (say, football). Let $\mathcal{P}$ be the following data generation process: X is generate from the Latent Dirichlet Allocation model [5], and Y is chosen randomly according to the topic mixture.*

We use logistic regression, which is parameterized by a weight function $w : Z \to \mathbb{R}$, and two additional parameters a and b. For each document $X = z_1 z_2 \dots z_k$, the output of the classifier is:

$$f(X) = \frac{1}{1 + \exp(-a \sum_{i=1}^{k} w(z_i) - b)}$$

The reason we are using automatically generated documents instead of true documents here is that the conditional probabilities $P(Y|X)$ are directly computable (otherwise we cannot evaluate $l_1(f, \mathcal{P})$). We conducted an experimental simulation for this example, and the experimental details can be found in the appendix. Here we summarize the major findings: the logistic regression classifier has very large l_1 error, which is probably due to the discrepancy between the logistic regression model and the underlying model. However, the logistic regression classifier is almost naturally calibrated in this example. This is not a coincidence, and we will discuss the corresponding intuition in Section 2.3.

2.3. The motivation of the calibration property

Let us revisit Example 1. This time, we fix the word weight function w. In this case, every document X can be represented using a single parameter $w(X) = \sum_i w(z_i)$, and we search for the optimal a and b such that the log-likelihood is maximized. This is illustrated in Figure 1.

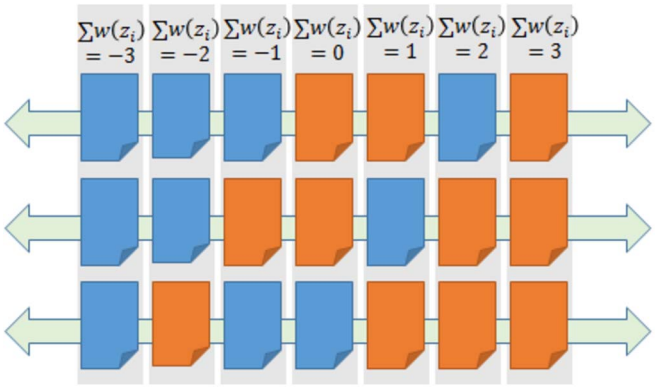


FIG 1. *Example 1 with Fixed Word Weight*

Now, intuitively, to maximize the log-likelihood, we need the sigmoid function $(1+\exp(-aw(X)-b))^{-1}$ to match the conditional probability of Y conditioned on $w(X)$: $\mathcal{P}(Y=1|w(X))$. Therefore, for the optimal a and b, we could say that the following property is roughly correct:

$$\mathcal{P}(Y=1|w(X)) \approx \frac{1}{1+\exp(-aw(X)-b)}$$

In other words,

$$\forall 0 \leq p \leq 1, \mathbb{E}[\mathcal{P}(Y=1|X)|f(X)=p] \approx p$$

Let us examine this example more closely. The reason why the logistic regression classifier tells us that $f(X) \approx p$ is because of the following: among all the documents with similar weight $w(X)$, about p portion of them actually belong to the topic in the training dataset. This leads to an important observation: logistic regression classifiers estimate the conditional probabilities by computing the relative frequency of labels among documents it believes to be similar.

This behavior is not unique to logistic regression. Many other algorithms, including decision tree classifiers, nearest neighbor (NN) classifiers, and neural networks, exhibit similar behavior:

- In decision trees, all data points reaching the same decision leaf are considered similar.
- In NN classifiers, all data points with the same nearest neighbors are considered similar.

- In neural networks, all data points reaching the same output layer values are considered similar.

We can abstract the above conditional probability estimators as the following two-step process:

1. Partition the feature space $\mathcal{X}$ into several regions.
2. Estimate the relative frequency of labels among all data points inside each region.

The definition of the calibration property follows easily from the above two-step process. We can argue that the classifier is approximately calibrated, if for each region S in the feature space $\mathcal{X}$, the output conditional probability of data points in S is close to the actual relative frequency of labels in S. The definition for the calibration property then follows from the fact that all data points inside each region have the same output conditional probabilities:

$$\forall p_1 < p_2, \quad \mathcal{P}(Y = 1 | p_1 < f(X) \leq p_2) = \mathbb{E}_{X \sim \mathcal{P}}[f(X) | p_1 < f(X) \leq p_2]$$

2.4. Using calibrated conditional probabilities in decision making

The calibration property justifies the common practice of using estimated conditional probabilities in decision making. Consider the following scenario: we have a set of actions $\mathcal{A} = \{A_1, A_2, \ldots, A_k\}$, and each action $A_i = (a_i, b_i)$ would incur a_i/b_i cost for each positive/negative instance respectively. In this case, the optimal strategy is to choose the action that minimizes the expected cost:

$$A^* = \arg\min_{A_i} [\mathcal{P}(Y = 1|X) a_i + \mathcal{P}(Y = 0|X) b_i]$$

Now, consider the more practical setting where we no longer know the actual value of $\mathcal{P}(Y = 1|X)$, but instead only have access to a calibrated fuzzy classifier f that estimates the conditional probabilities. If we can only use $f(X)$ to make decision, then the best strategy is to use $f(X)$ in the same way as $\mathcal{P}(Y = 1|X)$:

Claim 1. *Suppose we are given an ϵ-calibrated fuzzy classifier $f : \mathcal{X} \to [0, 1]$ (i.e., $c(f) \leq \epsilon$), and we need to make decisions solely based on the output of f. Denote our decision D as a collection of mutually disjoint intervals:*

$$D = \{(l_1, r_1], (l_2, r_2], \ldots, (l_k, r_k]\}$$

indicating that our decision for X is A_i iff $f(X) \in (l_i, r_i]$. Then we have,

$$|\mathbb{E}_X \mathcal{L}_{\mathcal{P},D}(X) - \mathbb{E}_X \mathcal{L}_{f,D}(X)| \leq \epsilon \sum_{i=1}^{k} |a_i - b_i|$$

where a_i, b_i are the cost of each positive/negative instance for action A_i (as defined in the beginning of Section 2.4), and $\mathcal{L}_{\mathcal{P},D}(X)$ and $\mathcal{L}_{f,D}(X)$ are the

true/estimated expected cost of D for data point X respectively:

$$\mathcal{L}_{\mathcal{P},D}(X) = \sum_{i=1}^{k} \mathbb{1}_{l_i < f(X) \leq r_i} [a_i \mathcal{P}(Y = 1|X) + b_i \mathcal{P}(Y = 0|X)]$$

$$\mathcal{L}_{f,D}(X) = \sum_{i=1}^{k} \mathbb{1}_{l_i < f(X) \leq r_i} [a_i f(X) + b_i (1 - f(X))]$$

Proof. By rearranging terms, we have:

$$\begin{aligned}
&\mathbb{E}_X \mathcal{L}_{\mathcal{P},D}(X) - \mathbb{E}_X \mathcal{L}_{f,D}(X) \\
=&\sum_{i=1}^{k} (a_i - b_i) \mathbb{E}_X \mathbb{1}_{l_i < f(X) \leq r_i} [\mathcal{P}(Y = 1|X) - f(X)] \\
=&\sum_{i=1}^{k} (a_i - b_i) [\mathcal{P}(Y = 1, l_i < f(X) \leq r_i) - \mathbb{E}_X \mathbb{1}_{l_i < f(X) \leq r_i} f(X)]
\end{aligned}$$

Since f is ϵ-calibrated, we have:

$$\forall p_1 < p_2, |\mathcal{P}(Y = 1, p_1 < f(X) \leq p_2) - \mathbb{E}_X \mathbb{1}_{p_1 < f(X) \leq p_2} f(X)| \leq \epsilon$$

Substituting into the previous equation, we get the desired result. □

Intuitively, the term $\mathcal{L}_{\mathcal{P},D}(X)$ is the expected cost of X if we take actions according to D, and the term $\mathcal{L}_{f,D}(X)$ is basically the same as $\mathcal{L}_{\mathcal{P},D}(X)$ except $\mathcal{P}(Y = 1|X)$ is replaced by $f(X)$. Claim 1 states that if f is almost calibrated, then the true expected cost will be very close to the cost estimated through f. Therefore, by applying Bayes Decision Rule on $f(X)$, we are also minimizing the true expected cost.

2.5. Intuition of the calibration measure

Recall that the calibration measure $c(f)$ is defined as:

$$c(f) = \sup_{p_1 < p_2} |\mathbb{E}_{X \sim \mathcal{P}} [\mathbb{1}_{p_1 < f(X) \leq p_2} f(X)] - \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)|$$

This expression has a very intuitive interpretation: the second term $\mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)$ is the fraction of data points satisfying both $p_1 < f(X) \leq p_2$ and $Y = 1$, while the first term $\mathbb{E}_{X \sim \mathcal{P}} [\mathbb{1}_{p_1 < f(X_i) \leq p_2} f(X_i)]$ means the estimated[2] fraction of data points satisfying both $p_1 < f(X) \leq p_2$ and $Y = 1$. Finally, the calibration measure $c(f)$ is simply defined to be the supreme absolute difference between these two terms over all possible regions $(p_1, p_2]$.

Therefore, if $c(f) < \epsilon$, then the difference between terms $\mathbb{E}_{X \sim \mathcal{P}} [\mathbb{1}_{p_1 < f(X) \leq p_2} f(X)]$ and $\mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)$ would be at most

[2]This estimation is made by assuming $\forall X, \mathcal{P}(Y = 1|X) = f(X)$

ϵ for any p_1, p_2. When ϵ is small enough, one can estimate the value of $\mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)$ through $\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)]$. It is worth noting that the latter term does not depend on the true label Y, and therefore can be estimated even on an unlabeled dataset. This observation has some interesting implications, which will be discussed in Section 3.3.4.

3. Uniform convergence of the calibration measure

3.1. The uniform convergence result

Let $\mathcal{G}$ be a collection of functions from $\mathcal{X} \times \mathcal{Y}$ to $[0, 1]$, the Rademacher Complexity [1][3] of $\mathcal{G}$ with respect to D is defined as [18]:

$$R_D(\mathcal{G}) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{g\in\mathcal{G}} \sum_{i=1}^{n} \sigma_i g(X_i, Y_i)]$$

Then we have the following result (the proof can be found in the appendix):

Theorem 2. *Let $\mathcal{F}$ be a set of fuzzy classifiers, i.e., functions from $\mathcal{X}$ to $[0, 1]$. Let $\mathcal{H}$ be the set of binary classifiers obtained by thresholding the output of fuzzy classifiers in $\mathcal{F}$:*

$$\mathcal{H} = \{\mathbb{1}_{p_1<f(X)\leq p_2} : p_1, p_2 \in \mathbb{R}, f \in \mathcal{F}\}$$

Suppose the Rademacher Complexity of $\mathcal{H}$ satisfies:

$$2\mathbb{E}_D R_D(\mathcal{H}) + \sqrt{\frac{2\ln(8/\delta)}{n}} < \frac{\epsilon}{2}$$

Then,

$$\mathbf{Pr}_D(\sup_{f\in\mathcal{F}} |c(f) - c_{emp}(f, D)| > \epsilon) < \delta$$

3.2. The hypothesis class $\mathcal{H}$

In Theorem 2, $\mathcal{H}$ is the collection of binary classifiers obtained by thresholding the output of a fuzzy classifier in $\mathcal{F}$. For many hypothesis classes $\mathcal{F}$, the Rademacher Complexity of $\mathcal{H}$ can be naturally bounded. For instance, if $\mathcal{F}$ is the d-dimensional generalized linear classifiers with monotone link function, then $\mathbb{E}_D R_D(\mathcal{H})$ can be bounded by $O(\sqrt{d \log n/n})$. We remark that $\mathcal{H}$ is different from the hypothesis class $\mathcal{H}_{p_1,p_2}$, where the thresholds are fixed in advance:

$$\mathcal{H}_{p_1,p_2} = \{\mathbb{1}_{p_1<f(X)\leq p_2} : f \in \mathcal{F}\}$$

In general, the gap between the Rademacher Complexities of $\mathcal{H}$ and $\mathcal{H}_{p_1,p_2}$ can be arbitrarily large. The following example illustrates this point.

[3]Our definition of Rademacher Complexity comes from Shalev-Shwartz and Ben-David's textbook [18], which is slightly different from the original definition in Bartlett and Mendelson's paper [1].

Example 2. *Let $\mathcal{X} = \{1, \ldots, n\}$, and $A_1, A_2, \ldots, A_{2^n}$ be a sequence of sets containing all subsets of $\mathcal{X}$. Let $\mathcal{F}$ be the following hypothesis space:*

$$\mathcal{F} = \{f_i(x) = \frac{i}{2^n} - \frac{1}{2^{n+1}}\mathbb{1}_{x \in A_i} : i \in \{1, 2, \ldots, 2^n\}\}$$

Intuitively, $\mathcal{F}$ contains 2^n fuzzy classifiers, the ith classifier produces a output of either $\frac{i}{2^n}$ or $\frac{i}{2^n} - \frac{1}{2^{n+1}}$ depending on whether $x \in A_i$. One can easily verify that for any p_1, p_2, the VC-dimension [19] of $\mathcal{H}_{p_1,p_2}$ is at most 2, but the VC-dimension of $\mathcal{H}$ is n.

However, if for any $x \in \mathcal{X}, f \in \mathcal{F}$, we have $f(x) \in P^*$ with $|P^*| < \infty$, then $R_D(\mathcal{H})$ can be bounded using the maximum VC-dimension of $\mathcal{H}_{p_1,p_2}$ and $\log |P^*|$ (the proof can be found in the appendix):

Claim 2. *Suppose that for any $f \in \mathcal{F}, x \in \mathcal{X}$, we have $f(x) \in P^*$ where P^* is a finite set, and for all $p_1, p_2 \in \mathbb{R}$, the VC-dimension of hypothesis space $\mathcal{H}_{p_1,p_2}$ is at most d. If the size n of the dataset D satisfies $n > d + 1$, then we have:*

$$R_D(\mathcal{H}) \leq \sqrt{\frac{2d(\ln \frac{n}{d} + 1) + 4\ln(|P^*| + 1)}{n}}$$

Therefore, as long as $|P^*|$ is $O(1)$, we still have the $O(n^{-0.5})$ convergence rate for $R_D(\mathcal{H})$, and thus the resulting conditional probability estimates would still be nearly calibrated for large enough n.

3.3. Applications of Theorem 2

3.3.1. Verifying the calibration of classifier

The first application of Theorem 2 is that we can verify whether the learned classifier f is calibrated. For simple hypothesis classes $\mathcal{F}$ (e.g., logistic regression), the corresponding hypothesis space $\mathcal{H}$ has low Rademacher Complexity. In this case, Theorem 2 naturally guarantees the generalization of calibration measure.

There are also cases where the Rademacher Complexity of $\mathcal{H}$ is not small. One notable example is SVM classifiers with Platt Scaling [17]:

Claim 3. *Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\forall x \in \mathcal{X}, ||x||_2 \leq 1$. Let $\mathcal{F}$ be the following hypothesis class:*

$$\mathcal{F} = \{x \to \frac{1}{1 + \exp(aw^T x + b)} : w \in \mathbb{R}^d, ||w||_2 \leq B, a, b \in \mathbb{R}\}$$

If the training data size $n < d$ and x_i are linearly independent, then $R_D(\mathcal{H}) \geq \frac{1}{2}$.

Proof. For any $\sigma \in \{\pm 1\}^n$, we can find a vector w such that for every x_i, we have $w^T x_i = \sigma_i$ (this is always possible since the number of equations n is

less than the dimensionality d). Let $w^* = \frac{Bw}{||w||_2}$ so that $||w^*||_2 = B$, and let $a = \lambda||w||_2/B$ and $b = 0$. Then we have:

$$f(x_i) = \frac{1}{1 + \exp(a(w^*)^T x_i + b)} = \frac{1}{1 + e^{\lambda\sigma_i}}$$

Let $\lambda \to -\infty$, then $\sum_{i=1}^n \sigma_i f(X_i) \to \sum_{i=1}^n \mathbb{1}_{\sigma_i=1}$, and the conclusion of the claim follows easily. □

In the case of SVM, the dimensionality of the feature space is usually much larger than the training dataset size (this is especially true for kernel SVM). In such situation, we can no longer verify the calibration property using only the training data, and a separate validation dataset is needed to calibrate the classifier (as suggested by Platt [17]). When verifying the calibration of a classifier on a validation dataset, we have the following result (the proof can be found in the appendix):

Claim 4. *Let f be any given fuzzy classifier, and D be a* ***validation*** *dataset consisting of i.i.d. samples from $\mathcal{P}$ (i.e., D is not used when training f), then:*

$$\mathbf{Pr}(c(f) \leq c_{emp}(f, D) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}) \geq 1 - \delta$$

$$\mathbf{Pr}(c(f) \geq c_{emp}(f, D) - [16\sqrt{2\pi} + 2\sqrt{2\ln\frac{8}{\delta}}]\sqrt{\frac{1}{n}}) \geq 1 - \delta$$

3.3.2. Implications on learning algorithm design

Standard conditional probability estimation algorithms usually maximize the likelihood of the training data to find the best fuzzy classifier. However, since we can only guarantee the calibration property of conditional probability estimates under the agnostic setting, any calibrated classifier is as good as the maximum likelihood estimation in terms of interpretability. Therefore, likelihood maximization is not necessarily the only method for estimating conditional probabilities.

There are other loss functions that are already widely used for binary classification. For example, hinge loss is at the foundation of large margin classifiers. Based on our discussion in this paper, we believe that these loss functions can also be used for conditional probability estimation. For example, Theorem 2 suggests the following constrained optimization problem:

$$\min \mathcal{L}(f, D) \quad s.t. \quad c_{emp}(f, D) = 0$$

where $\mathcal{L}(f, D)$ is the loss function we want to minimize. By optimizing over the space of empirically calibrated classifiers, we can ensure that the resulting classifier is also calibrated with respect to $\mathcal{P}$.

3.3.3. Connection to the calibration problem

Suppose that we are given an uncalibrated fuzzy classifier $f_0 : \mathcal{X} \to [0,1]$, and we want to find a function $g : [0,1] \to [0,1]$, so that $g \circ f_0$ presents a better conditional probability estimation. This is the problem of classifier calibration, which has been studied in many papers [21, 17].

Traditionally, calibration algorithms find the best link function g by maximizing likelihood or minimizing squared loss. In this paper, we suggest a different approach to the calibration problem. We can find the best g by minimizing the empirical calibration measure $c_{\text{emp}}(g \circ f_0)$. Let us assume w.l.o.g. that the training dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ satisfies

$$f_0(x_1) \leq \ldots \leq f_0(x_n)$$

and that g is monotonically nondecreasing. Then we have,

$$\begin{aligned} c_{\text{emp}}(g \circ f_0, D) =& \frac{1}{n} \sup_{p_1,p_2} \left| \sum_{i=1}^{n} \mathbb{1}_{p_1 < g(f_0(x_i)) \leq p_2} (\mathbb{1}_{y_i=1} - g(f_0(x_i))) \right| \\ \leq& \frac{1}{n} \max_{a,b} \left| \sum_{a<i\leq b} (\mathbb{1}_{y_i=1} - g(f_0(x_i))) \right| \end{aligned}$$

This expression can be used as the objective function for calibration: we search over the space of hypothesis $\mathcal{G}$ to find a function g that minimizes this objective function. Compared to other loss functions, the benefits of minimizing this objective function is that the resulting classifier is more likely to be calibrated, and therefore provides more interpretable conditional probability estimates.

In fact, one of the most well-known calibration algorithms, the isotonic regression algorithm, can be viewed as minimizing this objective function (the proof can be found in the appendix):

Claim 5. *Let $\mathcal{G}$ be the set of all continuous nondecreasing functions from $[0,1]$ to $[0,1]$. Then the optimal solution found by the isotonic regression algorithm (Algorithm 1) not only minimizes the squared loss*

$$\mathcal{L}_2(g) = \sum_{i=1}^{n} (\mathbb{1}_{y_i=1} - g(f_0(x_i)))^2$$

as shown in [15], but also minimizes

$$\mathcal{L}_c(g) = \max_{a,b} \left| \sum_{a<i\leq b} (\mathbb{1}_{y_i=1} - g(f_0(x_i))) \right|$$

Using this connection we proved several interesting properties of the isotonic regression algorithm [15] (pseudo-code can be found in Algorithm 1 for reference):

Algorithm 1: Isotonic Regression Calibration (PAV Algorithm) [15]

1. Order the data points so that $f_0(x_1) \leq f_0(x_2) \leq \ldots \leq f_0(x_n)$
2. For $i = 0, \ldots, n$, Compute $P_i = (i, S_i = \sum_{j \leq i} \mathbb{1}_{y_j=1})$
3. Let $cv(P)$ be the lower boundary of the convex hull of the set of points P_i
   Remark: Implementing this step using the Graham's algorithm [11] would result in the exact same algorithmic procedure as in [15].
4. For $i = 0, \ldots, n$, Let $Z_i =$ intersection of $cv(P)$ and the line $x = i$
5. Compute $z_i = Z_i - Z_{i-1}$
6. Let $g(f_0(x_i)) = z_i$, extrapolate these points to get continuous nondecreasing function g.

Claim 6. *Let g^* be the output calibrating function of Algorithm 1, then:*

1. *The empirical calibration measure $c_{emp}(g^* \circ f_0, D)$ of the calibrated classifier is always 0.*
2. *For any asymmetric loss $(1-p, p)$ (i.e., each false negative incurs $1-p$ cost and each false positive incurs p cost), the empirical loss of the calibrated classifier is always no greater that of the original classifier (both using the optimal decision threshold p):*

$$\sum_{i=1}^{n} [(1-p)\mathbb{1}_{g^*(f_0(x_i)) \leq p, y_i=1} + p\mathbb{1}_{g^*(f_0(x_i)) > p, y_i=0}]$$
$$\leq \sum_{i=1}^{n} [(1-p)\mathbb{1}_{f_0(x_i) \leq p, y_i=1} + p\mathbb{1}_{f_0(x_i) > p, y_i=0}]$$

In particular, when $p = 0.5$, the empirical accuracy of the calibrated classifier is always greater than or equal to the empirical accuracy of the original classifier.

We also used Theorem 2 to prove some non-asymptotic convergence results for the PAV Algorithm, which can be found in the appendix.

3.3.4. Estimating & optimizing expected loss over unlabeled data

Recall the discussion in Section 2.5, the calibration measure $c(f)$ has the following interpretation: for every pair p_1 and p_2, the difference between $\mathbb{E}_{X \sim \mathcal{P}}[\mathbb{1}_{p_1 < f(X) \leq p_2} f(X)]$ and $\mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)$ is at most $c(f)$:

$$\forall p_1, p_2, |\mathbb{E}_{X \sim \mathcal{P}}[\mathbb{1}_{p_1 < f(X) \leq p_2} f(X)] - \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)| \leq c(f)$$

Similarly, for the empirical calibration measure $c_{emp}(f, D)$, we have the following interpretation: for any dataset D, the difference between $\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2} f(x_i)$

and $\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2, y_i = 1}$ is at most $c_{emp}(f, D)n$:

$$\forall p_1, p_2, |\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2} f(x_i) - \sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2, y_i = 1}| \leq c_{emp}(f, D)n \quad (1)$$

Now consider an unlabeled dataset D, the first quantity $\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2} f(x_i)$ can be estimated on D since it does not involve the label y_i. Suppose that $c_{emp}(f, D)$ is sufficiently small, then Equation 1 tells that $\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2} f(x_i)$ is approximately equal to $\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2, y_i = 1}$ for every $p_1 < p_2$.

The latter term $\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2, y_i = 1}$ is of practical interest since it is directly related to the costs of decision rules. As in Section 2.4, if we denote the action sets as $\mathcal{A} = \{(a_1, b_1), \ldots, (a_k, b_k)\}$, and the decision as $\mathcal{D} = \{(l_1, r_1], (l_2, r_2], \ldots, (l_k, r_k]\}$, then the total cost for the dataset on decision $\mathcal{D}$ is:

$$\begin{aligned}\mathcal{L}(\mathcal{D}) &= \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}_{l_j < f(x_i) \leq r_j} [\mathbb{1}_{y_i = 1} a_j + (1 - \mathbb{1}_{y_i = 1}) b_j] \\ &= \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}_{l_j < f(x_i) \leq r_j} b_j + \sum_{j=1}^{k} (a_j - b_j) \sum_{i=1}^{n} \mathbb{1}_{l_j < f(x_i) \leq r_j, y_i = 1}\end{aligned}$$

Substituting in the approximation in Equation (1), we have

$$\begin{aligned}|\mathcal{L}(\mathcal{D}) - \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}_{l_j < f(x_i) \leq r_j} b_j - \sum_{j=1}^{k} (a_j - b_j) \sum_{i=1}^{n} \mathbb{1}_{l_j < f(x_i) \leq r_j} f(x_i)| \\ \leq c_{emp}(f, D)n \sum_{j=1}^{k} |a_j - b_j| \quad (2)\end{aligned}$$

Intuitively, the above result implies that the total cost of every decision $\mathcal{D}$ can be estimated by the following quantity,

$$\mathcal{L}(\mathcal{D}) \approx \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}_{l_j < f(x_i) \leq r_j} b_j + \sum_{j=1}^{k} (a_j - b_j) \sum_{i=1}^{n} \mathbb{1}_{l_j < f(x_i) \leq r_j} f(x_i)$$

and the error for such an estimation is at most $c_{emp}(f, D)n \sum_{j=1}^{k} |a_j - b_j|$. This result allows us to explicitly search for the decision that optimizes the right hand side of the above equation, which can be useful when Bayes Decision Rule is not directly applicable (e.g., when each action has a certain start-up cost).

Equation 2 also gives us a confidence interval on the total cost, and its length depends linearly on the empirical calibration measure $c_{emp}(f, D)$. As we have discussed in Section 3.3.1 and 3.3.3, the calibration measure scales as $O(n^{-1/2})$ after the calibration procedure. Therefore, one can obtain tighter confidence bound by using larger validation dataset if desired.

4. Empirical behavior of the calibration measure

In this section, we conduct some preliminary experiments to demonstrate the behavior of the calibration measure on some common algorithms. We use two binary classification datasets from the UCI Repository[4]: ADULT[5] and COVTYPE[6]. COVTYPE has been converted to a binary classification problem by treating the largest class as positive and the rest as negative. Five algorithms have been used in these experiments: naive Bayes(NB), boosted decision trees, SVM[7], logistic regression(LR), random forest(RF).

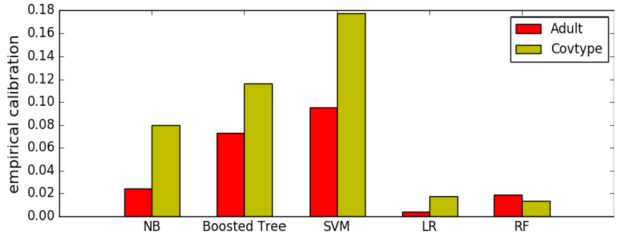


FIG 2. *The empirical calibration error*

Figure 2 shows the empirical calibration error c_{emp} on test datasets for all methods. From the experimental results, it appears that Logistic Regression and Random Forest naturally produce calibrated classifiers, which is intuitive as we discussed in the paper. The calibration measure of Naive Bayes seems to depend on the dataset. For large margin methods (SVM and boosted trees), the calibration measures are high, meaning that they are not calibrated on these two datasets.

There is also an interesting connection between the calibration error and the benefit of applying a calibration algorithm, which is illustrated in Figure 3. In this experiment, we used a loss parameter p to control the asymmetric loss: each false negative incurs $1-p$ cost and each false positive incurs p cost. All the algorithms are first trained on the training dataset, then calibrated on a separate calibration dataset of size 2000 using isotonic regression. For each algorithm, we compute the prior-calibration and post-calibration average losses on the testing dataset using the following decision rule: For each data point X, we predict $Y = 1$ if and only if we predict that $\mathbf{Pr}(Y = 1|X) \geq p$. Finally, we report the ratio between two losses:

$$\text{loss ratio} = \frac{\text{the average loss after calibration}}{\text{the average loss before calibration}}$$

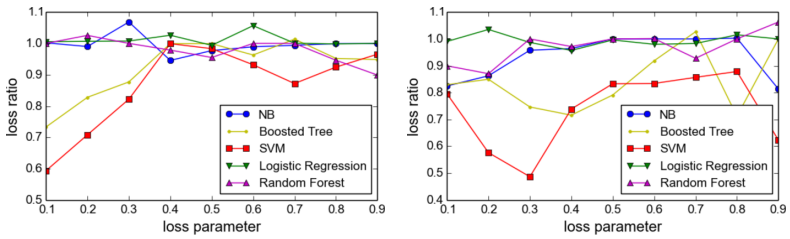
[4]These datasets are chosen from the datasets used in Niculescu-Mizil and Caruana's work [15]. We only used two datasets because the experiments are only explorative (i.e., identifying potential properties of the calibration measure). More rigorous experiments are needed to formally verify these properties.

[5]https://archive.ics.uci.edu/ml/datasets/Adult.

[6]https://archive.ics.uci.edu/ml/datasets/Covertype.

[7]For SVM and boosting, we rescale the output score to $[0, 1]$ by $(x - \min)/(\max - \min)$ as in Niculescu-Mizil and Caruana's paper [15]

As we can see in the Figure 3, the calibration procedure on average reduces



(a) Adult (b) Covtype

FIG 3. *The loss ratio on two datasets*

the cost by 3%-5% for naive Bayes and random forest, 20% for SVM, 12% for boosted trees, and close to 0% for logistic regression. Comparing with the results in Figure 2, the two algorithms that benefit most from calibration (i.e., SVM and boosted trees) also has high empirical calibration error. This result suggests that if an algorithm already has a low calibration error to begin with, then it is not likely to benefit much from the calibration process. This finding could potentially help us decide whether we need to calibrate the current classifier using isotonic regression [15].

It is also possible to use calibrated conditional probability estimates to compute the Receiver Operating Characteristic (ROC) curve of classifiers. More specifically, we estimate both the true positive rates and false positive rates via conditional probability estimates:

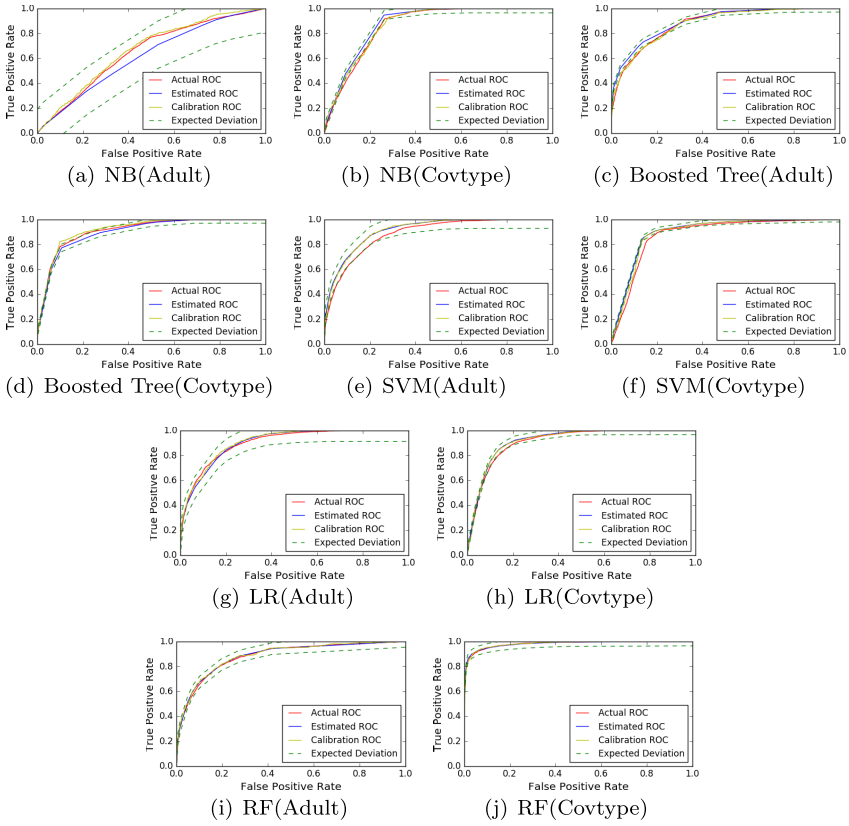
$$TPR = \frac{\sum_i f(x_i)\mathbb{1}_{f(x_i)<p}}{\sum_i f(x_i)} \qquad FPR = \frac{\sum_i (1-f(x_i))\mathbb{1}_{f(x_i)<p}}{n - \sum_i f(x_i)}$$

Figure 4 shows the estimated ROC curve for all 10 pairs of dataset-algorithm combination. In this experiment, all algorithms are first trained on the training dataset, then calibrated on the 2000-size calibration dataset. The calibrated classifiers are then used to estimate the ROC curve on the test dataset using the formulas above. Figure 4 also shows the true ROC curve of the test dataset (red curve) and the ROC curve of the calibration dataset (yellow curve) for comparison. As we can see, the estimated ROC curve matches reasonably well with the actual ROC curve, and roughly resembles the calibration ROC curve.

Additionally, we evaluated the calibration measure $c(f)$ of the calibrated classifiers using a separate validation dataset, and used this value to plot the expected deviation region of the ROC curve, via the following formulas:

$$TPR_u = \frac{\sum_i f(x_i)\mathbb{1}_{f(x_i)<p} + c_{emp}(f, D_{val})}{\sum_i f(x_i)}$$

$$TPR_l = \frac{\sum_i f(x_i)\mathbb{1}_{f(x_i)<p} - c_{emp}(f, D_{val})}{\sum_i f(x_i)}$$



(a) NB(Adult) (b) NB(Covtype) (c) Boosted Tree(Adult)

(d) Boosted Tree(Covtype) (e) SVM(Adult) (f) SVM(Covtype)

(g) LR(Adult) (h) LR(Covtype)

(i) RF(Adult) (j) RF(Covtype)

FIG 4. *The estimated and actual ROC on two datasets*

where $c_{emp}(f, D_{val})$ is empirical calibration measure of f on the validation dataset D_{val}.

As we can see from Figure 4, the expected deviation region matches pretty accurately with empirical data. This result suggests that we can use the calibration measure $c(f)$ to evaluate the reliability of ROC curves, and the plotted deviation region can be useful for decision makers.

5. Discussion: Potential extensions to multi-class setting

In this section, we discuss the potential extensions of the calibration measures to multi-class setting. The simplest such extension is to take supreme over all the binary calibration measures for each possible label:

$$c(f) = \sup_{y^* \in \mathcal{Y}} \sup_{p_1 < p_2} |\mathbb{E}_{X \sim \mathcal{P}}[\mathbb{1}_{p_1 < f(X, y^*) \leq p_2} f(X, y^*)] - \mathcal{P}(p_1 < f(X, y^*) \leq p_2, Y = y^*)|$$

where $f(X, y*)$ denotes the estimated conditional probability of $\mathcal{P}(Y = y^*|X)$. Intuitively, $c(f)$ is small if the classifier is calibrated with respect to every possible label $y^* \in \mathcal{Y}$.

However, this calibration property is not very useful in practice: even if the calibration error is 0 according to the above definition, we can still only guarantees the conditional probability estimates to be calibrated with respect to all the simple regions of the following form:

$$\{X : p_1 < f(X, y^*) \leq p_2, y^* \in \mathcal{Y}, p_1, p_2 \in \mathbb{R}\}$$

which is usually not good enough in practice. To see this, consider the generalization of Claim 1 to multi-class setting: assume for simplicity that there are 3 possible class labels: $\mathcal{Y} = \{1, 2, 3\}$[8]. Denote $\mathcal{A} = \{A_1, \ldots, A_k\}$ to be the collection of actions available to us, with each action $A_i = (a_i, b_i, c_i)$ incurring cost $a_i/b_i/c_i$ for each data point with label 1/2/3 respectively. Under this notation, the Bayes-optimal decision region for each action is the following:

$$\begin{aligned} D_i = \{X : \forall j \neq i, \quad & a_i f(X, 1) + b_i f(X, 2) + c_i f(X, 3) \\ & \leq a_j f(X, 1) + b_j f(X, 1) + c_j f(X, 1)\} \end{aligned}$$

Note that unlike the binary setting where all decision regions are intervals, the shape of decision regions here depend on the action set cardinality k. In order to generalize Claim 1 to multi-class setting, we need to make sure that conditional probability estimates are calibrated with respect to all possible decision regions (i.e., all possible intersections of $k - 1$ half-spaces). Denote $\mathcal{S}_k$ to be collection of all possible sets obtained by intersecting k half-planes, then we have the following definition for the multi-class calibration measure:

$$c_k(f) = \sup_{y^* \in \mathcal{Y}} \sup_{S \in \mathcal{S}_{k-1}} |\mathbb{E}_{X \sim \mathcal{P}}[\mathbb{1}_{X \in S} f(X, y^*)] - \mathcal{P}(X \in S, Y = y^*)|$$

and we can denote the limit of $c_k(f)$ with $k \to \infty$ as $c(f)$:

$$c(f) = \lim_{k \to \infty} c_k(f) = \sup_{y^* \in \mathcal{Y}} \sup_{S \in \mathcal{C}} |\mathbb{E}_{X \sim \mathcal{P}}[\mathbb{1}_{X \in S} f(X, y^*)] - \mathcal{P}(X \in S, Y = y^*)|$$

where $\mathcal{C}$ is the collection of all convex sets.

The above definition appears to be a more "natural" extension of the calibration measures to multi-class setting. However, there are still many other questions that remain unanswered, for instance:

- How to compute $c_{emp,k}(f)$ when k is large, or $c_{emp}(f)$ as $k \to \infty$?
- Is it still possible to calibrate the classifiers in multi-class setting?
- Is $|c(f) - c_{emp}(f)|$ still converging to 0 as $n \to \infty$?

We haven't been able to find definitive answers to these questions. Hopefully, future works along this line would bring us answers and improve our understanding regarding the interpretability of multi-class conditional probability estimates.

[8]This is only to simplify notations, the argument generalizes to more possible label scenarios

6. Conclusion

In this paper, we discussed the interpretability of conditional probability estimates under the agnostic assumption. We proved that it is impossible to upper bound the l_1 error of conditional probability estimates under such scenario. Instead, we defined a novel measure of calibration to provide interpretability for conditional probability estimates. The uniform convergence result between the measure and its empirical counterpart allows us to empirically verify the calibration property without making any assumption on the underlying distribution: the classifier is (almost) calibrated if and only if the empirical calibration measure is low. Our result provides new insights on conditional probability estimation: ensuring empirical calibration is already sufficient for providing interpretable conditional probability estimates, and thus many other loss functions (e.g., hinge loss) can also be utilized for estimating conditional probabilities. Finally, our calibration measure allows us to estimate and optimize the total cost for decision making: if the fuzzy classifier is (almost) calibrated, then the total cost of decision rules can be reliably estimated using the conditional probability estimates, which is available even for unlabeled datasets.

Appendix

Experimental simulation of Example 1

Here we experimentally simulate Example 1 to illustrate that the logistic regression classifier has large l_1 error. We use Latent Dirichlet Allocation (LDA) [5], the state of the art generative model for documents, to generate datasets. The detailed experiment settings are listed below:

- The dataset consists of 20000 documents, the number of topics is 20, the dictionary size is 1000, and the average number of words in each document is 200.
- We use the non-informative Dirichlet prior $\alpha = (1, 1, \ldots, 1)$ over topics. The word distribution in each topic follows power law with a random order among words.
- For each document, we randomly sample with replacement 10 topic labels from the topic distribution.

Average l_1 Error	Empirical Calibration
0.1270 ± 0.0008	0.0083 ± 0.0003
Trivial l_1 Error	Frequency of Labels
0.2022 ± 0.0001	0.3448 ± 0.0001

TABLE 1
L_1 error and empirical calibration

Table 1 reports the mean experiment results and the standard deviation across five runs. For reference we also include the relative frequency of labels,

and the l_1 error achieved by the trivial classifier that always output the global relative frequency of labels as conditional probability.

As we can see from Table 1, the logistic regression only achieves 0.13 average l_1 error, while even the trivial classifier can achieve 0.2. This implies that logistic regression performed very badly with respect to the l_1 error in this example. However, the empirical calibration measure of logistic regression classifier is relatively low (0.01), indicating that the classifier is almost calibrated.

Proof of Theorem 1

Proof. The proof relies on the following lemma:

Lemma 1. *Let $\mathcal{P}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is discrete. Let D be a size n i.i.d. sample set from $\mathcal{P}$. Let V be a verifier of $\mathcal{P}$ given D (i.e., V is a function from $\{\mathcal{X} \times \mathcal{Y}\}^n$ to $\{0,1\}$), such that*

1. *With probability at least $1-\delta_1$, a dataset D with n i.i.d. samples from $\mathcal{P}$ will pass V:*
$$\mathbf{Pr}_D(V(D)=1) \geq 1-\delta_1$$
2. *With probability at least $1-\delta_2$, a dataset D with n i.i.d. samples from $\mathcal{P}$ satisfies:*
$$\mathbf{Pr}(\forall i \neq j, X_i \neq X_j) \geq 1-\delta_2$$

Then there exists another probability distribution $\mathcal{P}'$ such that:

1. *With probability at least $1-\delta_1-\delta_2$, a data D' with n i.i.d. samples from $\mathcal{P}'$ will also pass V.*
$$\mathbf{Pr}_{D'}(V(D')=1) \geq 1-\delta_1-\delta_2$$
2. $$\forall X \in \mathcal{X}, \sum_{Y\in\mathcal{Y}} \mathcal{P}(X,Y) = \sum_{Y\in\mathcal{Y}} \mathcal{P}'(X,Y)$$
3. $$\forall X \in \mathcal{X}, \mathcal{P}'(Y=1|X) = 0 \text{ or } 1$$

Proof. First we construct the following distribution over all possible $\mathcal{P}'$ satisfying the last two conditions:

$$\mathbf{Pr}(\mathcal{P}') = \prod_{X\in\mathcal{X}} Q(P'(Y=1|X), P(Y=1|X))$$

where $Q(p',p)$ is defined as:

$$Q(p',p) = \begin{cases} p & p'=1 \\ 1-p & p'=0 \end{cases}$$

Now it suffices to show that if we sample $\mathcal{P}'$ according to the above distribution and then sample D' from $\mathcal{P}'$, then with probability at least $1-\delta_1-\delta_2$,

D' will pass V. Assuming this is true, then at least one distribution $\mathcal{P}'$ has to satisfy the first condition, and thereby we have proved the existence of $\mathcal{P}'$.

To compute the probability that D' would pass V, denote $D_X = \{X_1, X_2, \ldots, X_n\}$ and $D_Y = \{Y_1, Y_2, \ldots, Y_n\}$. Note that all $\mathcal{P}'$ have the same marginal distribution over $\mathcal{X}$, therefore:

$$\begin{aligned}&\mathbf{Pr}_{\mathcal{P}',D'}(V(D')=1) = \sum_{\mathcal{P}'}\mathbf{Pr}(\mathcal{P}')\sum_{D'}\mathbf{Pr}(D'|\mathcal{P}')V(D')\\ =&\sum_{D'_X}\mathbf{Pr}(D'_X)\sum_{\mathcal{P}'}\mathbf{Pr}(\mathcal{P}')\sum_{D'_Y}\mathbf{Pr}(D'_Y|\mathcal{P}',D'_X)V(D')\end{aligned}$$

We only consider all those D'_X with distinct X_i values. Based on the assumption, such D'_X accounts for at least $1-\delta_2$ of the probability mass. Now the important observation is that for every fixed D'_X with distinct X values, the marginal distribution of D'_Y given D'_X (i.e. marginalize over $\mathcal{P}'$) is exactly $\mathcal{P}(D'_Y|D'_X)$, the distribution that we sample labels independently from $\mathcal{P}(Y|X)$ for each X'_i in D'_X:

$$\begin{aligned}&\sum_{D'_X}\mathbf{Pr}(D'_X)\sum_{\mathcal{P}'}\mathbf{Pr}(\mathcal{P}')\sum_{D'_Y}\mathbf{Pr}(D'_Y|\mathcal{P}',D'_X)V(D')\\ \geq&\sum_{D'_X}\mathbf{Pr}(D'_X)\mathbb{1}_{\forall i\neq j, X'_i\neq X'_j}\sum_{D'_Y}\mathbf{Pr}(D'_Y|\mathcal{P},D'_X)V(D')\end{aligned}$$

The latter probability is actually the probability that D' will pass V and have distinct X values at the same time. Based on the assumptions in the lemma, it occurs with probability at least $1-\delta_1-\delta_2$. □

Now given this lemma, the proof of Theorem 1 is easy: We show that if any prover A_f satisfies the two conditions in the theorem, it can be used as the verifier V in the lemma such that no $\mathcal{P}'$ can satisfy all three conditions.

Let $\delta_1 = \frac{1-\epsilon}{2}$, then the first assumption in the lemma is satisfied, also since $\forall x \in \mathcal{X}, \mathcal{Q}(x) < \frac{\epsilon}{n^2}$, we have:

$$\forall i \neq j, \mathbf{Pr}(X_i = X_j) = \sum_x \mathcal{Q}(x)^2 \leq \frac{\epsilon}{n^2}$$

By a union bound, we have:

$$\mathbf{Pr}(\forall i \neq j, X_i \neq X_j) \geq 1-\epsilon$$

Therefore we can set $\delta_2 = \epsilon$. By the above lemma, there exists another $\mathcal{P}'$ such that

$$\mathbf{Pr}_{D'\sim\mathcal{P}'}(A_f(D')) \geq \frac{1-\epsilon}{2}$$

and

$$\forall X \in \mathcal{X}, Y \in \mathcal{Y}, \mathcal{P}'(Y|X) = 0 \text{ or } 1$$

On the other hand, note that the l_1 distance between $\mathcal{P}'$ and $\mathcal{P}$ is at least B_f, then by the properties of A_f, D' cannot pass A_f with probability greater than or equal to $\frac{1-\epsilon}{2}$. This contradicts our earlier result. Therefore no such A_f can exist. □

Proof of Theorem 2

Proof. We will use the following uniform convergence result [18]:

Theorem 3 (Uniform Convergence of Functions [18]). *Let D be i.i.d. samples of $(\mathcal{X} \times \mathcal{Y}, \mathcal{P})$, then with probability at least $1-\delta$,*

$$\sup_{g \in \mathcal{G}} |\frac{1}{n} \sum_{i=1}^{n} g(X_i, Y_i) - \mathbb{E}g(X, Y)| \leq 2\mathbb{E}_D R_D(\mathcal{G}) + \sqrt{\frac{2\ln(4/\delta)}{n}} \tag{3}$$

In the following we sometimes allow $\mathcal{G}$ to be a collection of functions from $\mathcal{X}$ to $[0,1]$ in the above results. When used in this sense, we assume that the function will not use y label: $g(x,y) = g(x)$.

Define $\mathcal{F}_{D,p_1,p_2}(f)$ to be the relative frequency of event $\{p_1 < f(X) \leq p_2, Y = 1\}$:

$$\mathcal{F}_{D,p_1,p_2}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{p_1 < f(X_i) \leq p_2, Y_i = 1}$$

Define $\mathcal{F}_{\mathcal{P},p_1,p_2}(f)$ to be the probability of the same event:

$$\mathcal{F}_{\mathcal{P},p_1,p_2}(f) = \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)$$

Define $\mathcal{E}_{D,p_1,p_2}(f)$ as the empirical expectation of $f(X)\mathbb{1}_{p_1 < f(X) \leq p_2}$:

$$\mathcal{E}_{D,p_1,p_2}(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i)\mathbb{1}_{p_1 < f(X_i) \leq p_2}$$

Define $\mathcal{E}_{\mathcal{P},p_1,p_2}(f)$ as the expectation of the same function:

$$\mathcal{E}_{\mathcal{P},p_1,p_2}(f) = \mathbb{E}[f(X)\mathbb{1}_{p_1 < f(X) \leq p_2}]$$

When the context is clear, subscripts p_1 and p_2 can be dropped. Using this notation, we can rewrite $c(f)$ and $c_{\text{emp}}(f, D)$ as follows:

$$c(f) = \sup_{p_1,p_2} |\mathcal{F}_{\mathcal{P}}(f) - \mathcal{E}_{\mathcal{P}}(f)|$$

$$c_{\text{emp}}(f) = \sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{E}_D(f)|$$

Note that:

$$\begin{aligned}
&|\sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{E}_D(f)| - \sup_{p_1,p_2} |\mathcal{F}_S(f) - \mathcal{E}_S(f)|| \\
&\leq \sup_{p_1,p_2} ||\mathcal{F}_D(f) - \mathcal{E}_D(f)| - |\mathcal{F}_S(f) - \mathcal{E}_S(f)|| \\
&\leq \sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{E}_D(f) - \mathcal{F}_S(f) + \mathcal{E}_S(f)| \\
&\leq \sup_{p_1,p_2} (|\mathcal{F}_D(f) - \mathcal{F}_S(f)| + |\mathcal{E}_D(f) - \mathcal{E}_S(f)|) \\
&\leq \sup_{p_1,p_2} |\mathcal{F}_D(f) - \mathcal{F}_S(f)| + \sup_{p_1,p_2} |\mathcal{E}_D(f) - \mathcal{E}_S(f)|
\end{aligned}$$

Therefore it suffices to show that

$$\mathbf{P}(\sup_{f,p_1,p_2} |\mathcal{F}_D(f) - \mathcal{F}_S(f)| + \sup_{f,p_1,p_2} |\mathcal{E}_D(f) - \mathcal{E}_S(f)| > \epsilon) < \delta$$

Define

$$\mathcal{H}_1 = \{\mathbb{1}_{p_1 < f(X) \leq p_2, Y=1} : p_1, p_2 \in \mathbb{R}, f \in \mathcal{F}\}$$
$$\mathcal{H}_2 = \{f(X)\mathbb{1}_{p_1 < f(X) \leq p_2} : p_1, p_2 \in \mathbb{R}, f \in \mathcal{F}\}$$

Then we have the following lemma:

Lemma 2. *Let $\mathcal{H}_1, \mathcal{H}_2$ as defined above, then:*

$$R_D(\mathcal{H}_1) \leq R_D(\mathcal{H}) \quad R_D(\mathcal{H}_2) \leq R_D(\mathcal{H})$$

Proof. For $R_D(\mathcal{H}_1)$, we have:

$$R_D(\mathcal{H}_1) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^{n} \sigma_i \mathbb{1}_{p_1 < f(X_i) \leq p_2, Y_i = 1}]$$

We can replace $\mathbb{1}_{Y_i=1}$ with $\mathbb{E}_{Z_i\in\{\pm1\}} \max(Z_i, Y_i)$ (since Y_i is either -1 or 1):

$$R_D(\mathcal{H}_1) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^{n} \sigma_i \mathbb{1}_{p_1 < f(X_i) \leq p_2} \mathbb{E}_{Z_i\in\{\pm1\}} \max(Z_i, Y_i)]$$

Move the expectation over Z out of the supremum operator, we have:

$$R_D(\mathcal{H}_1) \leq \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n, Z\in\{\pm1\}}[\sup_{p_1,p_2,f} \sum_{i=1}^{n} \mathbb{1}_{p_1 < f(X_i) \leq p_2} \sigma_i \max(Z_i, Y_i)]$$

Now define $T_i = \sigma_i \max(Z_i, Y_i)$, then

$$R_D(\mathcal{H}_1) \leq \frac{1}{n}\mathbb{E}_{Z\in\{\pm1\}}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^{n} \mathbb{1}_{p_1 < f(X_i) \leq p_2} T_i]$$

Note that T_i is always uniformly distributed over $\{\pm1\}$, which is independent of Z_i and Y_i. Therefore,

$$R_D(\mathcal{H}_1) \leq \frac{1}{n}\mathbb{E}_{T\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^{n} \mathbb{1}_{p_1 < f(X_i) \leq p_2} T_i] = R_D(\mathcal{H})$$

For $R_D(\mathcal{H}_2)$, we have:

$$R_D(\mathcal{H}_2) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \sum_{i=1}^{n} \sigma_i f(X_i) \mathbb{1}_{p_1 < f(X_i) \leq p_2}]$$

Replace $f(X_i)$ with $\int_0^1 \mathbb{1}_{t<f(X_i)} dt$, we have

$$R_D(\mathcal{H}_2) = \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1,p_2,f} \int_0^1 \sum_{i=1}^{n} \sigma_i \mathbb{1}_{t<f(X_i)} \mathbb{1}_{p_1 < f(X_i) \leq p_2} dt]$$

Move the integral out of the supremum operator, we have:

$$R_D(\mathcal{H}_2) \leq \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}\int_0^1 [\sup_{p_1,p_2,f}\sum_{i=1}^n \sigma_i \mathbb{1}_{\max(p_1,t)<f(X_i)\leq p_2}]dt$$

Define $p_1' = \max(t, p_1)$, then we have:

$$R_D(\mathcal{H}_2) \leq \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}\int_0^1 [\sup_{p_1'\geq t,p_2,f}\sum_{i=1}^n \sigma_i \mathbb{1}_{p_1'<f(X_i)\leq p_2}]dt$$

Remove the restriction over p_1':

$$R_D(\mathcal{H}_2) \leq \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}\int_0^1 [\sup_{p_1',p_2,f}\sum_{i=1}^n \sigma_i \mathbb{1}_{p_1'<f(X_i)\leq p_2}]dt$$

Now the expression inside the bracket no longer depends on the value of t, therefore we conclude:

$$R_D(\mathcal{H}_2) \leq \frac{1}{n}\mathbb{E}_{\sigma\sim\{\pm1\}^n}[\sup_{p_1',p_2,f}\sum_{i=1}^n \sigma_i \mathbb{1}_{p_1'<f(X_i)\leq p_2}] = R_D(\mathcal{H}) \qquad \square$$

Combining this lemma with the assumptions in the theorem:

$$\mathbb{E}_D R_D(\mathcal{H}_1) + \sqrt{\frac{2\ln(8/\delta)}{n}} < \frac{\epsilon}{2}$$
$$\mathbb{E}_D R_D(\mathcal{H}_2) + \sqrt{\frac{2\ln(8/\delta)}{n}} < \frac{\epsilon}{2}$$

By Equation (3):

$$\mathbf{P}(\sup_{f,p_1,p_2}|\mathcal{F}_D(f) - \mathcal{F}_S(f)| > \frac{\epsilon}{2}) < \frac{\delta}{2}$$
$$\mathbf{P}(\sup_{f,p_1,p_2}|\mathcal{E}_D(f) - \mathcal{E}_S(f)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \qquad \square$$

Proof of Claim 2

Proof. By Massart Lemma [18], we have:

$$R_D(\mathcal{H}) \leq \sqrt{\frac{2\ln|\mathcal{H}(D)|}{n}}$$

where $\mathcal{H}(D)$ is the restriction of $\mathcal{H}$ to D. It suffices to show that

$$|\mathcal{H}(D)| \leq (|P^*|+1)^2(en/d)^d$$

Note that

$$\mathcal{H}(D) = \cup_{p_1,p_2}\mathcal{H}_{p_1,p_2}(D)$$

Since $f(x)$ only takes finite possible values, we only need to consider values of p_1, p_2 in $P^* \cup \{-\infty\}$. Therefore,

$$|\mathcal{H}(D)| \leq \sum_{p_1,p_2 \in P^* \cup \{-\infty\}} |\mathcal{H}_{p_1,p_2}(D)|$$

Since each $\mathcal{H}_{p_1,p_2}$ has VC-dimension at most d, by Sauer's Lemma [18]:

$$\forall p_1, p_2, |\mathcal{H}_{p_1,p_2}(D)| \leq (en/d)^d$$

Combining the last two inequalities, we get the desired result. □

Proof of Claim 4

Proof. We first prove the first inequality. For any p_1, p_2, we have:

$$\begin{aligned}&\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)] - \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)\\ =&\mathbb{E}_D \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2}[f(X_i) - \mathbb{1}_{Y_i=1}]\end{aligned}$$

Therefore, by Hoeffdding's inequality, with probability $1-\delta$:

$$\begin{aligned}&|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2}[f(X_i) - \mathbb{1}_{Y_i=1}]\\ &-\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)] + \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)| \leq \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\end{aligned}$$

Therefore with probability $1-\delta$,

$$\begin{aligned}&|\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2}[f(X_i) - \mathbb{1}_{Y_i=1}]|\\ \geq&|\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)] - \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)| - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\end{aligned}$$

For any $\epsilon > 0$, we can choose p_1 and p_2 such that

$$|\mathbb{E}_{X\sim\mathcal{P}}[\mathbb{1}_{p_1<f(X)\leq p_2} f(X)] - \mathcal{P}(p_1 < f(X) \leq p_2, Y = 1)| > c(f) - \epsilon$$

Then with probability $1-\delta$,

$$c_{emp}(f, D) \geq |\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{p_1<f(X_i)\leq p_2}[f(X_i) - \mathbb{1}_{Y_i=1}]| > c(f) - \epsilon - \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}$$

Since ϵ can be any positive real number, the desired result follows immediately.

To prove the second inequality, by Theorem 2, it suffices to show that

$$\forall D, R_D(\mathcal{H}) \leq \sqrt{\frac{32\pi}{n}}$$

for $\mathcal{F} = \{f\}$. Assume w.l.o.g. that $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ satisfies

$$f(x_1) \leq f(x_2) \leq \ldots \leq f(x_n)$$

Then,

$$R_D(\mathcal{H}) = \frac{1}{n}\mathbb{E}_\sigma \sup_{p_1,p_2} |\sum_{i=1}^{n} \mathbb{1}_{p_1 < f(x_i) \leq p_2} \sigma_i| \leq \frac{1}{n}\mathbb{E}_\sigma \max_{a,b} |\sum_{a<i\leq b} \sigma_i|$$

Denote $S_i = \sum_{j\leq i} \sigma_j$, then S_i is a simple one-dimensional random walk, by the reflection principle of symmetric random walk [7], we have:

$$\forall C \geq 0, \mathbf{Pr}(\sup_i |S_i| > C) \leq 2\mathbf{Pr}(|S_n| > C)$$

Therefore,

$$\mathbb{E}_\sigma[\sup_{i,j} |S_i - S_j|] \leq 2\mathbb{E}_\sigma[\sup_i |S_i|] \leq 4\mathbb{E}_\sigma|S_n|$$

By Hoeffding's inequality,

$$\forall C \geq 0, \mathbf{Pr}(|S_n| \geq C\sqrt{n}) \leq 2\exp(-\frac{1}{2}C^2)$$

Therefore,

$$\begin{aligned} R_D(\mathcal{H}) \leq \frac{4}{n}\mathbb{E}_\sigma|S_n| =& \sqrt{\frac{16}{n}} \int_0^\infty \mathbf{Pr}(|S_n| \geq x\sqrt{n})dx \\ \leq& \sqrt{\frac{64}{n}} \int_0^\infty e^{-\frac{1}{2}x^2} dx = \sqrt{\frac{32\pi}{n}} \end{aligned}$$

□

Proof of Claim 5

Proof. Let $z_i = g(f_0(x_i))$, then we can rewrite the objective function as:

$$\max_{a,b} |\sum_{a<i\leq b} (\mathbb{1}_{y_i=1} - z_i)|$$

To prove Algorithm 1 also minimizes this objective function, we first state the minimization problem as a linear programming:

$$\begin{aligned} \min \xi_1 + \xi_2 \quad \textbf{s.t.} \quad & \xi_1, \xi_2 \geq 0 \\ & 0 \leq z_1 \leq z_2 \leq \ldots \leq z_n \leq 1 \end{aligned}$$

$$\forall 1 \le k \le n, \sum_{i \le k} z_i \ge \sum_{i \le k} \mathbb{1}_{y_i=1} - n\xi_1$$
$$\forall 1 \le k \le n, \sum_{i \le k} z_i \le \sum_{i \le k} \mathbb{1}_{y_i=1} + n\xi_2$$

Define $S_k = \sum_{i \le k} \mathbb{1}_{y_i=1}$ and $Z_k = \sum_{i \le k} z_i$. Then we have the following constraints:

$$\forall 1 \le k \le n-1, Z_k - Z_{k-1} \le Z_{k+1} - Z_k$$
$$\forall 1 \le k \le n, S_k - n\xi_1 \le Z_k \le S_k + n\xi_2$$

Let Z_i^* be the solution produced by Algorithm 1, it should be obvious that $Z_i^* \le S_i$ for all i. Therefore,

$$\xi_2^* = \frac{1}{n} \min_i (S_i - Z_i^*) = 0 \quad \xi_1^* = \frac{1}{n} \max_i (S_i - Z_i^*)$$

We need to prove that $\xi_1^* \le \xi_1 + \xi_2$ for every feasible solution (Z_i, ξ_i). Suppose $\xi_1^* = \frac{1}{n}(S_i - Z_i^*)$, and Z_i^* lies on the line segment $\{(j, S_j), (k, S_k)\}$. Then we have:

$$S_i - n\xi_1^* = Z_i^* = \frac{i-j}{k-j} S_k + \frac{k-i}{k-j} S_j$$

Because of the convexity constraint of Z, it must satisfy the following inequality:

$$Z_i \le \frac{i-j}{k-j} Z_k + \frac{k-i}{k-j} Z_j$$

Computing the difference between these two, we get

$$Z_i - S_i + n\xi_1^* \le \frac{i-j}{k-j}(Z_k - S_k) + \frac{k-i}{k-j}(Z_j - S_j)$$

Substituting in

$$Z_i - S_i \ge -n\xi_1 \quad Z_k - S_k \le n\xi_2 \quad Z_j - S_j \le n\xi_2$$

We get

$$n\xi_1^* \le n\xi_1 + n\xi_2$$

which proves the optimality of Z^*. □

Proof of Claim 6

Proof. Throughout the proof, let C be the convex hull computed in Algorithm 1:

$$C = \{(i_0 = 0, 0), (i_1, S_{i_1}), \ldots, (i_{m-1}, S_{i_{m-1}}), (i_m = n, S_n)\}$$

We will use the following notation:

$$z_i = g^*(f_0(x_i)) \quad Z_k = \sum_{i=1}^{k} z_i \quad S_k = \sum_{i=1}^{k} \mathbb{1}_{y_i=1}$$

1. For any p_1, p_2, let l, r be such that:

$$l = \max_{k \le n, z_k \le p_1} k \qquad r = \max_{k \le n, z_k \le p_2} k$$

If no such k exists for p_1 or p_2 (i.e., $\forall k, z_k > p_i$ for either $i = 1$ or $i = 2$), we simply set l or r to be 0 respectively. By Algorithm 1, we have

$$\forall i_j < k \le i_{j+1}, z_k = \frac{S_{i_{j+1}} - S_{i_j}}{i_{j+1} - i_j}$$

Thus we have $(l, S_l), (r, S_r) \in C$, $Z_l = S_l, Z_r = S_r$, and therefore

$$\begin{aligned} &\sum_{i=1}^{n} \mathbb{1}_{p_1 < z_i \le p_2, y_i = 1} - \sum_{i=1}^{n} \mathbb{1}_{p_1 < z_i \le p_2} z_i \\ =&(Z_r - Z_l) - (S_r - S_l) = 0 \end{aligned}$$

which implies that $c_{emp}(g^* \circ f_0) = 0$

2. Let $a = \max\{i : f_0(x_i) \le p\}, b = \max\{i : z_i \le p\}$, then we need to show that

$$\begin{aligned} &(1-p) \sum_{i=1}^{b} \mathbb{1}_{y_i=1} + p \sum_{i=b+1}^{n} \mathbb{1}_{y_i=0} \\ \le &(1-p) \sum_{i=1}^{a} \mathbb{1}_{y_i=1} + p \sum_{i=a+1}^{n} \mathbb{1}_{y_i=0} \end{aligned}$$

We consider two separate cases:

(a) $a \le b$, in this case we only need to show that

$$\sum_{i=a+1}^{b} [p\mathbb{1}_{y_i=0} - (1-p)\mathbb{1}_{y_i=1}] \ge 0$$

or equivalently,

$$p[(b-a) - (S_b - S_a)] - (1-p)(S_b - S_a) \ge 0$$

Rearrange terms, it suffices to show

$$p(b-a) - (S_b - S_a) \ge 0$$

Since $S_b = Z_b, S_a \ge Z_a$

$$S_b - S_a \le Z_b - Z_a \le z_b(b-a) \le p(b-a)$$

(b) $a > b$, in this case we only need to show

$$\sum_{i=b+1}^{a} [p\mathbb{1}_{y_i=0} - (1-p)\mathbb{1}_{y_i=1}] \le 0$$

or equivalently,

$$p[(a-b) - (S_a - S_b)] - (1-p)(S_a - S_b) \le 0$$

Rearrange terms, it suffices to show

$$p(a-b) - (S_a - S_b) \leq 0$$

Since $S_b = Z_b, S_a \geq Z_a$

$$S_a - S_b \geq Z_a - Z_b \geq z_{b+1}(a-b) \geq p(a-b)$$ □

Convergence of the PAV algorithm

We can also use Theorem 2 to derive the following non-asymptotic convergence result of Algorithm 1.

Claim 7. *Let $F(t) = \mathcal{P}(f_0(X) \leq t)$ be the distribution function of $f_0(X)$, and define $G(t)$ as:*

$$G(t) = \mathcal{P}(f_0(X) \leq t, Y = 1)$$

Let $cv : [0,1] \to [0,1]$ be the convex hull of all points $(F(t), G(t))$ for all $t \in [0,1]$. Define G_e as:

$$G_e(t) = \mathbb{E}[\mathbb{1}_{f_0(X) \leq t} g^*(f_0(X))]$$

Then under the same condition in Theorem 2,

$$\mathbf{P}(\sup_t |G_e(t) - cv(F(t))| > 2\epsilon) < 5\delta$$

In particular, if $\mathcal{P}(Y = 1|f_0(X))$ is monotonically increasing, then

$$\mathbf{P}(\sup_t |G_e(t) - G(t)| > 2\epsilon) < 5\delta$$

The intuition behind this claim can be explained as follows: $F(t)$ is the percentage of data points satisfying $f_0(X) \leq t$, and $G(t)$ is $F(t)$ times the conditional probability of $Y = 1$ in the region $\{f_0(X) \leq t\}$. Now consider points $P_i = (i, S_i)$ in Algorithm 1, it is not hard to show that as $n \to \infty$, the limit of points P_i are the curve $(F(t), G(t)), t \in [0,1]$ (after proper scaling). Similarly, $G_e(t)$ is $F(t)$ times the expected value of $g^*(f_0(X))$ in the region $\{f_0(X) \leq t\}$, and it is not hard to show that $(F(t), G_e(t))$ is the limit of (i, Z_i) (after proper scaling). Now the claim states that in the PAV algorithm, $(F(t), G_e(t))$ converge uniformly to the convex hull of $(F(t), G(t))$, which should not be surprising, since we explicitly computed the convex hull of $\{P_i\}$ in Algorithm 1.

When $\mathcal{P}(Y = 1|f_0(X))$ is monotonically increasing w.r.t. $f_0(X)$, $(F(t), G(t))$ is convex, and Claim 7 immediately implies that $G_e(t)$ will converge uniformly to $G(t)$. In this case, the PAV algorithm will eventually recover the "true" link function $g^*(f_0(X)) = \mathcal{P}(Y = 1|f_0(X))$ given sufficient training samples, and Claim 7 provides a rough estimate of the number of samples required to achieve the desired precision.

Proof. Throughout the proof, let C be the convex hull computed in Algorithm 1:

$$C = \{(i_0 = 0, 0), (i_1, S_{i_1}), \ldots, (i_{m-1}, S_{i_{m-1}}), (i_m = n, S_n)\}$$

We will use the following notation:

$$z_i = g^*(f_0(x_i)) \quad Z_k = \sum_{i=1}^{k} z_i \quad S_k = \sum_{i=1}^{k} \mathbb{1}_{y_i=1}$$

We will use the following facts in the proof of Theorem 2:

$$\mathbf{P}(\sup_{g,p_1,p_2} |\mathcal{F}_D(g \circ f_0) - \mathcal{F}_{\mathcal{P}}(g \circ f_0)| > \frac{\epsilon}{2}) < \frac{\delta}{2}$$
$$\mathbf{P}(\sup_{g,p_1,p_2} |\mathcal{E}_D(g \circ f_0) - \mathcal{E}_{\mathcal{P}}(g \circ f_0)| > \frac{\epsilon}{2}) < \frac{\delta}{2}$$

For any $t \in [0,1]$, let u be any continuous increasing function from $[0,1]$ to $[0,1]$. Let $k = \max\{i : f_0(x_i) \leq t\}, p_1 = -\infty, p_2 = u(t)$ in the above inequalities, then we have:

$$\mathbf{P}(|\frac{1}{n} S_k - G(t)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \tag{4}$$

$$\mathbf{P}(|\frac{1}{n} \sum_{i=1}^{k} u(f_0(x_i)) - \mathbb{E}[\mathbb{1}_{f_0(X) \leq t} u(f_0(X))]| > \frac{\epsilon}{2}) < \frac{\delta}{2}$$

Now we set u to be such that $||u - g^*||_\infty < \lambda$, where $\lambda > 0$ can be arbitrarily small[9]. Let $\lambda \downarrow 0$, then the second inequality implies

$$\mathbf{P}(|\frac{1}{n} Z_k - G_e(t)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \tag{5}$$

Similarly, we can set u to be such that $|u(x) - 1| < \lambda$ for any x. Let $\lambda \downarrow 0$, then the second inequality implies

$$\mathbf{P}(|\frac{1}{n} k - F(t)| > \frac{\epsilon}{2}) < \frac{\delta}{2} \tag{6}$$

For any $t \in [0,1]$, let $k = \max\{i : f_0(x_i) \leq t\}$. Let $[i_{j-1} = l, i_j = r]$ be the segment of C with $l < k \leq r$. Then we have

$$z_{l+1} = \ldots = z_k = \ldots = z_r$$
$$S_l = Z_l = Z_k - (k - l) z_k$$
$$S_r = Z_r = Z_k + (r - k) z_k$$

On the other hand, by (4), with probability at least $1 - \delta$:

$$\frac{1}{n} S_l \geq G(f_0(x_l)) - \frac{\epsilon}{2} \quad \frac{1}{n} S_r \geq G(f_0(x_r)) - \frac{\epsilon}{2}$$

[9]Note that we cannot simply have $u = g^*$ here because u need to be strictly monotonically increasing so that the inverse function is well-defined. The same goes for the next part where we cannot simply set $u \equiv 1$.

Since cv is the convex hull of $(F(t), G(t))$, we have

$$qG(f_0(x_l)) + (1-q)G(f_0(x_r)) \geq cv(F(t))$$

where $q = \frac{F(f_0(x_r))-F(t)}{F(f_0(x_r))-F(f_0(x_l))}$. Combining all, with probability at least $1-\delta$:

$$\frac{1}{n}Z_k + \frac{1}{n}[ql + (1-q)r - k]z_k + \frac{\epsilon}{2} \geq cv(F(t))$$

By (6), with probability at least $1 - \frac{3}{2}\delta$:

$$\frac{1}{n}l \leq F(f_0(x_l)) + \frac{\epsilon}{2} \quad \frac{1}{n}r \leq F(f_0(x_r)) + \frac{\epsilon}{2}$$

$$\frac{1}{n}k \geq F(t) - \frac{\epsilon}{2}$$

Therefore, we have with probability at least $1 - \frac{5}{2}\delta$,

$$\frac{1}{n}Z_k + \frac{3\epsilon}{2} \geq cv(F(t))$$

Then by (5), with probability at least $1 - 3\delta$,

$$G_e(t) + 2\epsilon \geq cv(F(t))$$

Conversely, suppose $(F(t), cv(F(t)))$ is on the line segment between $(F(a), G(a))$ and $(F(b), G(b))$, then

$$G(a) = cv(F(t)) - w(F(t) - F(a))$$

$$G(b) = cv(F(t)) + w(F(b) - F(t))$$

where $w = \frac{G(b)-G(a)}{F(b)-F(a)}$ (if $F(a) = F(b)$ then just let $w = 1$).

By (4) and (5) and the fact that $S_k \geq Z_k$, with probability at least $1 - 2\delta$:

$$G(a) + \epsilon \geq G_e(a) \quad G(b) + \epsilon \geq G_e(b)$$

Also since $(F(t), G_e(t))$ is convex, we have:

$$qG_e(a) + (1-q)G_e(b) \geq G_e(t)$$

where $q = \frac{F(b)-F(t)}{F(b)-F(a)}$. Combining all above, with probability at least $1 - 2\delta$:

$$cv(F(t)) + \epsilon \geq G_e(t)$$

Combining two directions, the proof is complete. □

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003. MR1984026

[2] P. L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *The Journal of Machine Learning Research*, 8:775–790, 2007. MR2307020

[3] P. N. Bennett. Assessing the calibration of naive bayes posterior estimates. Technical report, DTIC Document, 2000.

[4] C. M. Bishop et al. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006. MR2247587

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of machine Learning research*, 3:993–1022, 2003.

[6] G. F. Cooper. Nestor: A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. Technical report, DTIC Document, 1984.

[7] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010. MR2722836

[8] T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997.

[9] D. P. Foster and R. V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998. MR1649119

[10] Y. Gao, A. Parameswaran, and J. Peng. On the interpretability of conditional probability estimates in the agnostic setting. In *Artificial Intelligence and Statistics*, pages 1367–1374, 2017.

[11] R. L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, 1(4):132–133, 1972.

[12] S. M. Kakade, A. Kalai, V. Kanade, and O. Shamir. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

[13] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32. Association for Computational Linguistics, 1999.

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

[15] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632. ACM, 2005.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. MR2854348

[17] J. Platt et al. Probabilistic outputs for support vector machines and com-

parisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

[18] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. MR3277164

[19] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971. MR0288823

[20] V. N. Vapnik and V. Vapnik. *Statistical Learning Theory*, volume 1. Wiley New York, 1998. MR1641250

[21] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, volume 1, pages 609–616. Citeseer, 2001.

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