A UNIFIED FRAMEWORK FOR QUANTILE ELICITATION WITH APPLICATIONS

Arnab Kumar Sarker

A THESIS

in

Data Science

Presented to the Faculties of the University of Pennsylvania in Partial
Fulfillment of the Requirements for the Degree of Master of Science in Engineering

2019

Shivani Agarwal
Supervisor of Thesis

Susan Davidson
Graduate Group Chair

ABSTRACT

We provide a general framework for the problem of quantile elicitation, and discuss applications of the framework in conditional quantile elicitation and multiclass classification. Our framework provides a rich theory for relating quantile elicitation to the problem of binary class probability estimation, and results in the development of new scoring rules for quantile elicitation. In the context of conditional quantile elicitation, we are able to provide regret bounds which quantify the extent to which our predicted quantiles differ from the actual quantiles of a distribution. We then turn our attention to multiclass classification, where we use quantile elicitation techniques to develop what we refer to as approximately consistent surrogate losses.

Contents

1 Introduction

In the field of property elicitation, the goal is to provide an incentive for an agent to truthfully report some statistic of a probability distribution that they believe. This problem has been studied for decades in various forms [21, 29, 32] and is typically solved by using a proper scoring rule based on the agent's report and the actual outcome of the chance experiment; in this thesis we look specifically at the problem of quantile elicitation, in which we wish to obtain an agent's report of a specific quantile of their distribution. Recently, there have been several advances in characterizing solutions to the problem of quantile elicitation [30, 6, 31, 19], and in this thesis, we provide a unifying framework for the elicitation of quantiles.

In constructing our unified framework for the elicitation of quantiles, we draw from the field of machine learning, and define a proper scoring rule inspired by the loss functions used by Pedregosa et al [23] and Langford et al [19]. As a result, we are able to create a general scoring rule for quantile elicitation, which we refer to as the *infinite threshold* scoring rule.

We then discuss two major applications of quantile elicitation. One application in conditional quantile elicitation, and the other in multiclass classification. Conditional quantile elicitation, in which given a set of features one is asked to predict the quantile of the distribution of labels, is a well studied problem in several fields including econometrics, sociology, and ecology [28, 16, 15, 13]. For such applications, the use of conditional quantile elicitation (for example, estimating the conditional median) is preferred to estimation of the conditional mean because quantiles are much more robust to noise than mean estimates. A common example of where this is useful is in wallet estimation, in which one wishes to estimate the amount of money a consumer is willing to spend on a product; because income distributions often have outliers, knowing the quantiles of the conditional distribution of disposable income is a much more valuable statistic [28].

Our second application of quantile elicitation is in machine learning, and specifically in the supervised learning problem of multiclass classification. In recent years, the machine learning community has used surrogate loss minimization as a framework with which to perform supervised learning. Much of this work has been focused on creating calibrated convex surrogates for multiclass learning problems, such as 0-1 classification, subset ranking, and structured prediction [1, 14, 25, 26, 33, 34]. However, creating algorithms based on convex surrogate losses that can be implemented efficiently is not always practical, especially in the case where the number of classes is very large [25]. This setting appears, for example, in image classification. Several prior works have addressed the problem of extreme multiclass classification [5, 14], but none have used quantile estimation as a key component to multiclass classification. We find that we can effectively use conditional quantiles to help perform multiclass classification.

2 Background and Review of Literature

2.1 Property Elicitation and Proper Scoring Rules

We begin with a discussion of property elicitation, as initially described in [21, 29], and further studied in [11, 1, 18, 17, 32]. Suppose we have some random variable Y, drawn from a set of values $\mathcal{Y} \subseteq \mathbb{R}$, where Y is drawn according to $\mathbf{p}$ with cumulative distribution function $F_{\mathbf{p}}(y) = \mathbf{P}(Y \leq y)$. Denoting $\mathcal{P}$ to be the set of all possible probability distributions over $\mathcal{Y}$, we may then define a *property* of a distribution:

Definition 1 (Property of a Distribution)**.** A property $\Gamma : \mathcal{P} \to \mathcal{C}$ of a distribution is a function that maps a distribution function $\mathbf{p} \in \mathcal{P}$ to an element $c \in \mathcal{C}$ for some set $\mathcal{C}$.

Example 1. If $\mathcal{C} = \mathbb{R}$, then $\Gamma(\mathbf{p}) = \int_{-\infty}^{\infty} y dF_{\mathbf{p}}(y)$ is a property of the distribution $\mathbf{p}$, commonly known as the mean.

Example 2. If $\mathcal{Y} = \{0, 1\}$, then the distribution $\mathbf{p}$ places mass on the value 1 with some probability η, and mass on the value 0 with probability $1 - \eta$. Hence, if $\mathcal{C} = [0, 1]$, the function $\Gamma(\mathbf{p}) = \eta$ is a property that fully describes the randomness of Y.

In essence, properties are 'statistics' of a distribution. The problem setting that we consider is that of obtaining such a statistic from an agent who knows the distribution of Y – Property elicitation is also commonly used in other settings, but considering this setting helps us to motivate our key results. To solve this problem, we make use of *scoring rules.*

Consider a setting in which an agent has the probability distribution $\mathbf{p}$ of a random variable Y drawn from a set of values $\mathcal{Y} \subseteq \mathbb{R}$, and we wish to incentivize the agent to truthfully report $\Gamma(\mathbf{p})$ for some property Γ as defined above. In this setting, we receive two values: the agent's reported statistic τ, and the outcome of the chance experiment y, i.e. the value that the random variable Y takes.

Once we receive these two values, we "charge" the agent some amount, so that the agent's goal then becomes to minimize the expected amount that they are charged.[1] Formally, we have the following definition:

Definition 2 (Scoring Rule)**.** A scoring rule is a function $\psi : \mathcal{Y} \times \mathcal{C} \rightarrow \mathbb{R}$ which assigns a score given both the agent's reported value and the outcome that the random variable takes on.

In general, to incentivize the agent, we wish for the scoring rule to be proper:

Definition 3 (Proper Scoring Rule)**.** A scoring rule $\psi : \mathcal{Y} \times \mathcal{C} \rightarrow \mathbb{R}$ is proper if

$$\mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \tau)] \geq \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \Gamma(\mathbf{p}))],$$

for all $\tau \in \mathcal{C}$.

ψ is said to be a *strictly proper scoring rule* if it is uniquely minimized at $\Gamma(\mathbf{p})$, i.e.

$$\mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \tau)] > \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \Gamma(\mathbf{p}))],$$

for all $\tau \neq \Gamma(\mathbf{p})$.

In general, we wish to provide a strictly proper scoring rule to incentivize an agent to be truthful. For example, a constant scoring rule trivially satisfies being proper, but does not explicitly incentivize an agent to be truthful.

2.1.1 A Detailed Example: Bernoulli Random Variables

In this section, we go into detail about a particular setting for proper scoring rules that will prove to be useful for understanding the generalized proper scoring rule of Section 3. In particular, we consider a setting in which $\mathcal{Y} = \{0, 1\}$, so that $\mathbf{p}$ places mass on 1 with probability η and mass on 0 with probability $1 - \eta$, as in Example 2.

[1] It is also common to define the scoring rule as giving the agent some "payout," where the agent's goal is to maximize their payout. Here, we take a "loss" perspective as this will relate to our applications more naturally.

First, consider the case where we wish to determine the value of the property $\Gamma(\mathbf{p}) = \eta$. We refer to this elicitation problem as the *binary class probability estimate problem.* Before we proceed with a sufficient condition for a scoring rule which elicits the binary class probability estimate, we review some definitions of convexity [3]:

Definition 4 (Convex Set)**.** A set $\mathcal{C} \subseteq \mathbb{R}^d$ is convex if, for all $a, b \in \mathcal{C}$, and for any $\theta \in [0, 1]$, we have:

$$\theta a + (1 - \theta) b \in \mathcal{C}$$

Definition 5 (Convex Function)**.** A function $\phi : \mathcal{C} \times \mathbb{R}$ is convex if $\mathcal{C}$ is convex, and for any $a, b \in \mathcal{C}$, and $\theta \in (0, 1)$, ϕ satisfies:

$$\phi\left(\theta a + (1 - \theta) b\right) \leq \theta \phi(a) + (1 - \theta) \phi(b)$$

ϕ is *strictly convex* if the inequality is strict for all $\theta \in (0, 1)$.

If ϕ is differentiable, convexity can equivalently be written:

$$\phi(a) \geq \phi(b) + \nabla \phi(a)^\top (b - a),$$

for all $a, b \in \mathcal{C}$, where ϕ is again strictly convex if the inequality is strict for all a, b.

McCarthy [21] showed the following regarding scoring rules for binary class probability estimation using convex functions:

Lemma 1 (Sufficient Condition for Binary Class Probability Scoring Rule)**.** Let $\phi : [0, 1] \times \mathbb{R}$ be a convex, differentiable function. Then, the function $\psi : \{0, 1\} \times [0, 1] \times \mathbb{R}$, defined as:

$$\psi(y, \widehat{\eta}) = -\Big(\mathbf{1}(y = 0)\left(\phi(\widehat{\eta}) - \widehat{\eta}\phi'(\widehat{\eta})\right) + \mathbf{1}(y = 1)\left(\phi(\widehat{\eta}) + (1 - \widehat{\eta})\phi'(\widehat{\eta})\right)\Big),$$

is a proper scoring rule for $\Gamma(\mathbf{p}) = \eta$. If ϕ is strictly convex, then ψ is a strictly proper scoring rule for $\Gamma(\mathbf{p}) = \eta$.

Proof. We show the proof by noting considering the two terms

$$\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\widehat{\eta})] \text{ and } \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\eta)].$$

In particular, using the above definition for ψ, we have:

$$\begin{aligned}
\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\widehat{\eta})] &= -\Big((1-\eta)\left(\phi(\widehat{\eta}) - \widehat{\eta}\phi'(\widehat{\eta})\right) + \eta\left(\phi(\widehat{\eta}) + (1-\widehat{\eta})\phi'(\widehat{\eta})\right)\Big) \\
&= -\left(\phi(\widehat{\eta}) - (1-\eta)\widehat{\eta}\phi'(\widehat{\eta}) + \eta(1-\widehat{\eta})\phi'(\widehat{\eta})\right) \\
&= -\left(\phi(\widehat{\eta}) + (\eta-\widehat{\eta})\phi'(\widehat{\eta})\right) \\
\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\eta)] &= -\Big((1-\eta)\left(\phi(\eta) - \widehat{\eta}\phi'(\eta)\right) + \eta\left(\phi(\eta) + (1-\eta)\phi'(\eta)\right)\Big) \\
&= -\phi(\eta)
\end{aligned}$$

Since ϕ is convex and differentiable, we have $\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\widehat{\eta})] \geq \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\eta)]$, showing the claim. If ϕ is strictly convex, the inequality becomes strict for $\widehat{\eta} \neq \eta$, showing strict properness as desired. □

Example 3 (Shannon Entropy)**.** Consider the case where $\phi(\widehat{\eta}) = \widehat{\eta}\log\widehat{\eta} + (1-\widehat{\eta})\log(1-\widehat{\eta})$, in other words we consider the negative of the binary Shannon entropy of $\widehat{\eta}$. Then,

$$\psi(y,\widehat{\eta}) = -\mathbf{1}(y=0)\log(1-\widehat{\eta}) - \mathbf{1}(y=1)\log(\widehat{\eta}),$$

which then implies

$$\begin{aligned}
\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\widehat{\eta})] - \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\eta)] &= -(1-\eta)\log(1-\widehat{\eta}) - \eta\log(\widehat{\eta}) \\
&\quad - \left(-(1-\eta)\log(1-\eta) - \eta\log(\eta)\right) \\
&= (1-\eta)\log\left(\frac{1-\eta}{1-\widehat{\eta}}\right) + \eta\log\left(\frac{\eta}{\widehat{\eta}}\right).
\end{aligned}$$

This is the binary KL divergence, known to be non-negative and only be equal to 0 if $\eta = \widehat{\eta}$. Hence, the function $\psi(y,\widehat{\eta}) = -\mathbf{1}(y=0)\log(1-\widehat{\eta}) - \mathbf{1}(y=1)\log(\widehat{\eta})$ is a strictly proper scoring rule.

In fact, the result above can be generalized to a broader class of scoring rules known as *proper composite scoring rules.* To understand this generalization, we first define indirect elicitation, which we will see again in Section 3.1.1.

Definition 6 (Indirect Elicitation)**.** Consider a function $\psi : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ and another function $\text{pred} : \mathcal{H} \to \mathcal{C}$. We say the pair (ψ, pred) indirectly elicits a property $\Gamma : \mathcal{P} \to \mathcal{C}$ if, for arbitrary $\mathbf{p} \in \mathcal{P}$:

$$\forall u^* \in \text{argmin}_{u \in \mathcal{H}} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, u)], \quad \text{pred}(u^*) = \Gamma(\mathbf{p})$$

In the setting of indirect elicitation, one could imagine charging the agent depending on their response $u \in \mathcal{H}$, and then using pred as a transformation to obtain $\Gamma(\mathbf{p})$ from the agent's response u.

Reid and Williamson [27] showed the following regarding indirect elicitation of a binary class probability in the context of binary classification. Here, we show their results in the context of property elicitation, with a significantly altered proof:

Proposition 1. Let $\phi : \mathbb{R} \to \mathbb{R}$ be a convex, differentiable function. Define $\gamma : [0,1] \to \mathbb{R}$ by its inverse:[2]

$$\gamma^{-1}(v) = \frac{\phi'(-v)}{\phi'(-v) + \phi'(v)},$$

and define the scoring rule:

$$\psi(y, v) = \mathbf{1}(y = 0)\phi(-v) + \mathbf{1}(y = 1)\phi(v).$$

Then, (ψ, γ^{-1}) indirectly elicits the property $\Gamma(\mathbf{p}) = \eta$.

[2] The use of the inverse here is a matter of convention, as the authors call such a function an *inverse link function.*

Proof. Consider the set:

$$\begin{aligned}
\text{argmin}_{u\in\mathbb{R}}\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,u)] &= \text{argmin}_{u\in\mathbb{R}}\left((1-\eta)\phi(-u)+\eta\phi(u)\right) \\
&= \left\{u \;\middle|\; -(1-\eta)\phi'(-u)+\eta\phi'(u)=0\right\} \\
&= \left\{u \;\middle|\; \frac{\phi'(u)}{\phi'(-u)} = \frac{1-\eta}{\eta}\right\}.
\end{aligned}$$

Then, for all $u^* \in \text{argmin}_{u\in\mathbb{R}}\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,u)]$, it must be the case:

$$\begin{aligned}
\gamma^{-1}(u^*) &= \frac{\phi'(-u^*)}{\phi'(-u^*)+\phi'(u^*)} \\
&= \frac{1}{1+\frac{\phi'(u^*)}{\phi'(-u^*)}} \\
&= \frac{1}{1+\frac{1-\eta}{\eta}} \\
&= \eta.
\end{aligned}$$

Hence, (ψ,γ^{-1}) indirectly elicits $\Gamma(\mathbf{p})=\eta$. □

Example 4. Let $\phi(v)=e^{-v}$, so that:

$$\gamma^{-1}(v) = \frac{1}{1+e^{-2v}}.$$

Then, we have:

$$\begin{aligned}
\text{argmin}_{u\in\mathbb{R}}\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,u)] &= \text{argmin}_{u\in\mathbb{R}}\left((1-\eta)e^{u}+\eta e^{-u}\right) \\
&= \left\{u \;\middle|\; (1-\eta)e^{u}-\eta e^{-u}=0\right\} \\
&= \left\{\frac{1}{2}\ln\left(\frac{\eta}{1-\eta}\right)\right\}.
\end{aligned}$$

Hence, the only u^* is $\frac{1}{2}\ln\left(\frac{\eta}{1-\eta}\right)$, and:

$$\gamma^{-1}(u^*) = \frac{1}{1+e^{-\ln\left(\frac{\eta}{1-\eta}\right)}} = \frac{1}{\frac{1}{\eta}} = \eta,$$

implying that the pair (ψ, γ^{-1}) indirectly elicits $\Gamma(\mathbf{p}) = \eta$ for $\phi = e^{-u}$.

Now, we turn our attention to the property $\Gamma(\mathbf{p}) = \operatorname{sign}\left(\eta - \frac{1}{2}\right)$, which takes value 1 if $\eta > \frac{1}{2}$, and -1 otherwise. The reader with a background in statistical learning theory will associate this with the Bayes optimal classifier in a binary classification setting.

Again in the setting of binary classification, Bartlett et al [2] showed a result of the following form for the indirect elicitation of $\Gamma(\mathbf{p}) = \operatorname{sign}\left(\eta - \frac{1}{2}\right)$:

Theorem 1 (Bartlett et al [2])**.** Suppose $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function. Then,

$$\psi(y, v) = \mathbf{1}(y = 0)\phi(-v) + \mathbf{1}(y = 1)\phi(v)$$

indirectly elicits $\Gamma(\mathbf{p})$ if and only if ϕ is differentiable at 0 and $\phi'(0) < 0$.

2.2 Quantile Elicitation

In this section, we study specifically proper scoring rules for the elicitation of quantiles of the distribution of a random variable.

2.2.1 Formal Definition of Quantiles

Let us again consider a setting in which we have a random variable Y, drawn from $\mathbf{p}$, which has distribution function $F_{\mathbf{p}} = \mathbf{P}(Y \leq y)$. Further, define the complementary cumulative distribution function $G_{\mathbf{p}} = \mathbf{P}(Y \geq y)$. For $\alpha \in (0, 1)$, the α-quantile of $\mathbf{p}$, denoted $Q_\alpha : \mathcal{P} \to \mathcal{I}$, where $\mathcal{I}$ is the set of closed intervals, is then defined as:

$$Q_\alpha(\mathbf{p}) = \{t \mid F_{\mathbf{p}}(t) \geq \alpha \text{ and } G_{\mathbf{p}}(t) \geq 1 - \alpha\}.$$

We note the following property about Q_α, as mentioned by Schervish et al [30]:

Lemma 2 (Quantiles are Closed Intervals)**.** For all $\mathbf{p}$, $Q_\alpha(\mathbf{p})$ is a closed interval.

Proof. Consider the set $A = \{t \mid F_{\mathbf{p}}(t) \geq \alpha\}$, and the set $B = \{t \mid G_{\mathbf{p}}(t) \geq 1 - \alpha\}$. Consider $s = \inf A$, which we know is finite because $\alpha > 0$. We must have $F_{\mathbf{p}}(t) \leq F_{\mathbf{p}}(s)$ for all $t \in A$, by monotonicity. Suppose towards contradiction that $F_{\mathbf{p}}(s) < \alpha$, so that $F_{\mathbf{p}}(s) = \alpha - \epsilon$ for some $\epsilon > 0$. By right continuity of $F_{\mathbf{p}}$, we see that there exists $s + \delta$ such that $F_{\mathbf{p}}(s + \delta) = \alpha - \epsilon/2 < \alpha$, contradicting the definition of s. Hence, $A = [s, \infty)$.

Similarly, because $G_{\mathbf{p}}(t)$ is left-continuous in t, we see that $B = (-\infty, v]$ for $v = \sup B$. Moreover, $v > s$, because all $x < s$ is in B, and all $x > v$ is in A, again by monotonicity of probability measure. Hence, $Q_\alpha(\mathbf{p}) = A \cap B = [s, v]$, a closed interval. □

From the definition above, we see that Q_α is a property of $\mathbf{p}$, and in the following sections we wish to develop proper scoring rules for Q_α.

2.2.2 Order Sensitive Scoring Rules

One subtype of scoring rules that has been studied in the context of quantile elicitation is that of *order sensitive* scoring rules [4, 32, 31]. Mathematically, we define an order sensitive scoring rule as follows:

Definition 7. Suppose $\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ elicits a property $\Gamma(\mathbf{p}) \in \mathbb{R}$. A scoring rule $\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is order sensitive with respect to $\mathbf{p}$ for Γ if, for $t_1, t_2 \in \mathbb{R}$ such that $t_2 < t_1 \leq \Gamma(\mathbf{p})$, or $\Gamma(\mathbf{p}) \leq t_1 < t_2$, we have

$$\mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, t_2)] > \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, t_1)].$$

This definition applies to quantile elicitation if we assume that the α-quantile of $\mathbf{p}$ is unique, i.e. $Q_\alpha(\mathbf{p}) = \{q\}$ for a single point $q \in \mathbb{R}$. Much of the literature in quantile elicitation, makes this assumption for their theoretical results, but it is often easy to extend the results to interval valued quantiles.

In the next section, we will assume, as done in the literature, that quantiles are singletons, such that the property is defined as a function $Q_\alpha : \mathcal{P} \to \mathbb{R}$. In Section 3, we will extend our results to have two notions of quantile elicitation, which we differentiate as *weak* and *strong* quantile elicitation.

2.2.3 Characterizations of Proper Scoring Rules for Quantiles

In this section, we describe two known characterizations of proper scoring rules for quantiles [30, 32]. In both characterizations, we assume that the quantile $Q_\alpha(\mathbf{p})$ is a unique value as opposed to an interval with non-zero Lebesgue measure, so we discuss the characterizations of scoring rules which elicit the property $q_\alpha \in \mathbb{R}$, where q_α is the only point which satisfies both $F_{\mathbf{p}}(q_\alpha) \geq \alpha$ and $G_{\mathbf{p}}(q_\alpha) \geq 1 - \alpha$.

We discuss two characterizations of proper scoring rules for quantiles: That of order-sensitive proper scoring rules, and of all proper scoring rules for quantiles. Steinwart et al [32] presented the following characterization of proper scoring rules for quantiles which are order sensitive:

Theorem 2 (Steinwart et al [32])**.** Let $\mathcal{P}_0$ be the set of distributions on $\mathcal{Y}$ such that Y has finite mean, and consider $\mathbf{p} \in \mathcal{P}_0$. Then, $\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is a proper scoring rule for the α-quantile if and only if ψ satisfies:

$$\psi(y, q) = \mathbf{1}(y < q)(1 - \alpha)[g(q) - g(y)] + \mathbf{1}(y > q)\alpha[g(y) - g(q)] + \kappa(y),$$

where g is a monotonically increasing function that is finite for all almost all $\mathcal{Y}$ under p, and κ is an arbitrary function in y which is also finite for almost all $\mathcal{Y}$ under $\mathbf{p}$.

Schervish et al [30] showed the following more general characterization of proper scoring rules for quantiles:

Theorem 3 (Schervish et al [30])**.** Let $\mathcal{P}_0$ be the set of distributions on $\mathcal{Y}$ such that Y has finite mean, and consider $\mathbf{p} \in \mathcal{P}_0$. Then, $\psi : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ is a (strictly) proper scoring rule for the α-quantile if and only if ψ satisfies

$$\psi(y,q)-\psi(y,t)=\begin{cases}(1-\alpha)[g(q)-g(t)] & (y\le\min\{t,q\})\\ g(y)-\alpha g(q)-(1-\alpha)g(t) & (q<y<t)\\ -g(y)+\alpha g(q)+(1-\alpha)g(t) & (t<y<q)\\ \alpha[g(t)-g(q)] & (y\ge\max\{t,q\})\end{cases},$$

where g is a (strictly) monotonically increasing real valued function that is finite for all almost all $\mathcal{Y}$ under $\mathbf{p}$.

We see that the characterization provided by Steinwart et al [32] is a special case of that of Schervish et al [30], by the following. Let ψ satisfy Theorem 2. Then,

$$\begin{aligned}\psi(y,q)-\psi(y,t) &= \mathbf{1}(y<q)(1-\alpha)[g(q)-g(y)]+\mathbf{1}(y>q)\alpha[g(y)-g(q)]\\ &\quad -\mathbf{1}(y<t)(1-\alpha)[g(t)-g(y)]-\mathbf{1}(y>t)\alpha[g(y)-g(t)]\\ &= \begin{cases}(1-\alpha)[g(q)-g(t)] & (y\le\min\{t,q\})\\ g(y)-\alpha g(q)-(1-\alpha)g(t) & (q<y<t)\\ -g(y)+\alpha g(q)+(1-\alpha)g(t) & (t<y<q)\\ \alpha[g(t)-g(q)] & (y\ge\max\{t,q\})\end{cases}\end{aligned}$$

As desired.

We now show that both of these definitions apply to the α-pinball loss, also known as the α-check loss, defined as follows:

Definition 8 (α-pinball loss)**.** The α-pinball loss, which we denote $\psi_\alpha:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$, is defined:

$$\psi_\alpha(y,q)=\alpha(y-q)_+ +(1-\alpha)(q-y)_+$$

Proposition 2. The α-pinball loss is a proper scoring rule for the α-quantile.

Proof. This is clear, as it satisfies Theorem 2 with $g(x) = x$ as a monotonically increasing function. □

3 Generalized Proper Scoring Rules for Quantile Elicitation

In this section, we define notions of elicitation of quantiles that account for the fact that quantiles are interval valued, and then we go on to describe another characterization of proper scoring rules for quantile elicitation.

3.1 Weak and Strong Elicitation of Quantiles

3.1.1 Indirect Property Elicitation

Recall the definition of *indirect elicitation*:

Definition 9 (Indirect Elicitation)**.** Consider a function $\psi : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ and another function $\mathrm{pred} : \mathcal{H} \to \mathcal{C}$. We say the pair (ψ, pred) indirectly elicits a property $\Gamma : \mathcal{P} \to \mathcal{C}$ if:

$$\forall u^* \in \mathrm{argmin}_{u \in \mathcal{H}} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, u)], \quad \mathrm{pred}(u^*) = \Gamma(\mathbf{p}).$$

In the context of quantile elicitation, we wish to elicit a property in $\mathcal{C} = \mathcal{I}$, where $\mathcal{I}$ is the set of closed intervals. In the approach of the general proper scoring rule of Section 3.2, we define $\mathcal{H}$ to be the set of all monotonically increasing functions, and use a transformation to obtain an interval from a particular increasing function.

Unfortunately, as predicting a closed interval often provides computational difficulties, and the interval of a quantile occurs with probability 0, it is often preferable to only elicit one value within a quantile. Hence, we differentiate between *weak* and *strong* elicitation of quantiles.

3.1.2 Weak Quantile Elicitation

In cases in which it suffices to only predict one quantile in the interval, we define the following notion of a proper scoring rule for a quantile:

Definition 10 (Weak Indirect Quantile Elicitation)**.** Consider a function $\psi : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ and another function $\text{pred} : \mathcal{H} \to \mathbb{R}$. We say the pair (ψ, pred) weakly indirectly elicits the α-quantile $Q_\alpha : \mathcal{P} \to \mathcal{I}$ evaluated at the distribution$\mathbf{p}$ if:

$$\forall u^* \in \text{argmin}_{u \in \mathcal{H}} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, u)], \quad \text{pred}(u^*) \in Q_\alpha(\mathbf{p}).$$

We say that (ψ, pred) strictly weakly indirectly elicits the α-quantile if it is also the case that:

$$\forall u' \notin \text{argmin}_{u \in \mathcal{H}} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, u)], \quad \text{pred}(u') \notin Q_\alpha(\mathbf{p}).$$

This definition of quantile elicitation coincides with the definition of a proper scoring rule for a quantile when the quantile is a singleton, as discussed in the previous two characterizations of proper scoring rules. We can show that the α-pinball loss weakly indirectly elicits the α-quantile, if we set pred to be the identity[3]

Proposition 3. Let $\text{pred}(x) = x$, and $\mathbf{p}$ be defined over $\mathcal{Y}$ such that Y has finite mean. Then, the pair $(\psi_\alpha, \text{pred})$, where ψ_α is defined as in Definition 8, strictly weakly indirectly elicits the α-quantile of $\mathbf{p}$.

Proof. Consider the definition of $\mathbf{E}_{Y \sim \mathbf{p}}[\psi_\alpha(Y, u)]$:

$$\begin{aligned}
\mathbf{E}_{Y \sim \mathbf{p}}[\psi_\alpha(Y, u)] &= \int_{-\infty}^{\infty} (\alpha(y-u)_+ + (1-\alpha)(u-y)_+)\, dF_{\mathbf{p}}(y) \\
&= \alpha \int_{-\infty}^{\infty} \int_{u}^{y} \mathbf{1}(u < y) d\kappa dF_{\mathbf{p}}(y) + (1-\alpha) \int_{-\infty}^{\infty} \int_{y}^{u} \mathbf{1}(u > y) d\kappa dF_{\mathbf{p}}(y) \\
&= \alpha \int_{u}^{\infty} \int_{u}^{y} d\kappa dF_{\mathbf{p}}(y) + (1-\alpha) \int_{-\infty}^{u} \int_{y}^{u} d\kappa dF_{\mathbf{p}}(y) \\
&= \alpha \int_{u}^{\infty} \int_{u}^{\infty} dF_{\mathbf{p}}(y) d\kappa + (1-\alpha) \int_{-\infty}^{u} \int_{-\infty}^{u} dF_{\mathbf{p}}(y) d\kappa \\
&= \alpha \int_{u}^{\infty} G_{\mathbf{p}}(\kappa) d\kappa + (1-\alpha) \int_{-\infty}^{u} F_{\mathbf{p}}(\kappa) d\kappa.
\end{aligned}$$

[3] By setting pred to be the identity function, we are also saying that the α-pinball loss in some sense weakly *directly* elicits the α-quantile.

Because we assume Y has finite mean, the integral above must exist as the original expectation must exist. Moreover, we see that if u does not satisfy the definition of a quantile, then either $F_{\mathbf{p}}(u) < \alpha$ or $G_{\mathbf{p}}(u) < 1-\alpha$. Consider the case $F_{\mathbf{p}}(u) < \alpha$. Then, if we consider q such that $q \in Q_\alpha(\mathbf{p})$, we have:

$$\begin{aligned}
\mathbf{E}_{Y\sim\mathbf{p}}[\psi_\alpha(Y,u)] - \mathbf{E}_{Y\sim\mathbf{p}}[\psi_\alpha(Y,q)] &= \alpha \int_u^\infty G_{\mathbf{p}}(\kappa)d\kappa + (1-\alpha)\int_{-\infty}^u F_{\mathbf{p}}(\kappa)d\kappa \\
&\quad - \left(\alpha \int_q^\infty G_{\mathbf{p}}(\kappa)d\kappa + (1-\alpha)\int_{-\infty}^q F_{\mathbf{p}}(\kappa)d\kappa\right) \\
&= \alpha \int_u^q G_{\mathbf{p}}(\kappa)d\kappa - (1-\alpha)\int_u^q F_{\mathbf{p}}(\kappa)d\kappa \\
&> 0.
\end{aligned}$$

The final inequality holds because, when we assume $F_{\mathbf{p}}(\kappa) < \alpha$, this implies $G(\kappa) \geq 1-\alpha$, and this must be true for all $u \leq \kappa \leq q_\alpha \leq q$, where q_α is the smallest element of $Q_\alpha(\mathbf{p})$. Because $u < q_\alpha$, as we assumed u does not satisfy the definition of a quantile, it must be the case that the difference $\mathbf{E}_{Y\sim\mathbf{p}}[\psi_\alpha(Y,u)] - \mathbf{E}_{Y\sim\mathbf{p}}[\psi_\alpha(Y,q)]$ is strictly positive as desired. □

3.1.3 Strong Quantile Elicitation

We now present a stronger notion of quantile elicitation, in which we ask that the scoring rule and prediction function elicit the entire interval of the quantile. Formally, the definition is as follows:

Definition 11 (Strong Indirect Quantile Elicitation)**.** Consider a function $\psi : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}$ and another function $\text{pred} : \mathcal{H} \to \mathcal{I}$. We say the pair (ψ, pred) strongly indirectly elicits the α-quantile $Q_\alpha : \mathcal{P} \to \mathcal{I}$ evaluated at the distribution $\mathbf{p}$ if:

$$\forall u^* \in \text{argmin}_{u\in\mathcal{H}} \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,u)], \quad \text{pred}(u^*) = Q_\alpha(\mathbf{p}),$$

We say that (ψ, pred) strictly strongly indirectly elicits the α-quantile if it is also the case that:

$$\forall u' \notin \text{argmin}_{u \in \mathcal{H}} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, u)], \quad \text{pred}(u') \neq Q_\alpha(\mathbf{p}).$$

Remark. Strong indirect quantile elicitation is the same as indirect elicitation of quantiles if we consider quantiles as interval-valued, as they are formally defined.

3.2 General Framework for Quantile Elicitation

We now present a general scoring rule that, depending on the choice of pred function, can either weakly or strongly elicit the α-quantile of F. We call this scoring rule the *infinite threshold* scoring rule, and denote it $\psi_{\phi,\alpha} : \mathcal{Y} \times \mathcal{G} \to \mathbb{R}$, where $\mathcal{G}$ is the set of non-decreasing real-valued functions.

Definition 12 (Infinite Threshold Scoring Rule)**.** For a function $\phi : \mathbb{R} \to \mathbb{R}$, the infinite threshold scoring rule is defined:

$$\psi_{\phi,\alpha}(y, g) = \alpha \int_{t=-\infty}^{y} \phi(-g(t))dt + (1 - \alpha) \int_{t=y}^{\infty} \phi(g(t))dt. \tag{1}$$

Our scoring rule elicits a function g from the agent, and because g is monotonically increasing, it must have a countable number of discontinuities. Hence, we see that $\psi_{\phi,\alpha}$ will be well defined for any choice of $g \in \mathcal{G}$ if the integral is taken as a Lebesgue integral.

This scoring rule is interesting in that its expectation under F can be written as a pointwise function of t, which allows us to easily solve for the optimal function g in terms of the cumulative distribution function. We refer to this as the *decomposability* property. First, note:

$$\begin{aligned}\mathbf{E}_{Y\sim\mathbf{p}}[\psi_{\phi,\alpha}(Y,g)] &= \int_{y=-\infty}^{\infty}\psi_{\phi,\alpha}(y,g)dF_{\mathbf{p}}(y)\\ &= \int_{y=-\infty}^{\infty}\left(\alpha\int_{t=-\infty}^{y}\phi(-g(t))dt + (1-\alpha)\int_{t=y}^{\infty}\phi(g(t))dt\right)dF_{\mathbf{p}}(y)\\ &= \int_{t=-\infty}^{\infty}\alpha(1-F_{\mathbf{p}}(t))\phi(-g(t)) + (1-\alpha)F_{\mathbf{p}}(t)\phi(g(t))dt.\end{aligned}$$

Then, by defining

$$C_\alpha(\beta,q) = \alpha(1-q)\phi(-\beta) + (1-\alpha)q\phi(\beta)\,.$$

we may write

$$\mathbf{E}_{Y\sim\mathbf{p}}[\psi_{\phi,\alpha}(Y,g)] = \int_{t=-\infty}^{\infty}C_\alpha(g(t),F_{\mathbf{p}}(t))dt\,.$$

Now, note that C_α also corresponds to the expected score of a binary margin scoring rule. Hence, by point wise minimization of C_α, we are able to solve for $g(t)$ in terms of $F_{\mathbf{p}}$. We note this idea in the following lemma, which shows that there exists a g^* which is in fact monotonically increasing.

Lemma 3. Suppose ϕ is continuous, and $(1-\alpha)\phi(\beta) - \alpha\phi(-\beta)$ is non-increasing. Let $C^*_\alpha(q) = \inf_\beta C_\alpha(\beta,q)$. Then,

$$\min_g \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,g)] = \int_{t=-\infty}^{\infty}C^*_\alpha(F_{\mathbf{p}}(t))dt\,.$$

Proof. Consider the function $g^*(t) = \sup \text{argmin}_\beta C_\alpha(\beta, F_{\mathbf{p}}(t))$. By construction, we have $g^*(t) \in \text{argmin}_h \mathbf{E}_{Y\sim\mathbf{p}}[\psi_{\phi,\alpha}(Y,h)]$ where h is considered over *all* possible functions of the real line.

We claim g^* is in fact monotonically increasing. Consider two points $t_1 < t_2$. Then, we know $F_{\mathbf{p}}(t_1) \leq F_{\mathbf{p}}(t_2)$. Consider the function $C_\alpha(\beta,q) = \alpha(1-q)\phi(-\beta) + (1-\alpha)q\phi(\beta)$.

Since ϕ is continuous, we know $\text{argmin}_\beta C_\alpha(\beta,q)$ is a closed interval for all q. There are then two cases:

Case 1: $F_{\mathbf{p}}(t_1) = F_{\mathbf{p}}(t_2)$

In this case, $g(t_1) = g(t_2)$, by definition.

Case 2: $F_{\mathbf{p}}(t_1) < F_{\mathbf{p}}(t_2)$

In this case, consider $i = \sup \operatorname{argmin}_\beta C_\alpha(\beta, F_{\mathbf{p}}(t_1))$ and $j = \sup \operatorname{argmin}_\beta C_\alpha(\beta, F_{\mathbf{p}}(t_2))$. Suppose towards contradiction that $i > j$. By definition of argmin, we have the following two inequalities:

- $C_\alpha(i, F_{\mathbf{p}}(t_1)) \leq C_\alpha(j, F_{\mathbf{p}}(t_1))$
- $C_\alpha(j, F_{\mathbf{p}}(t_2)) \leq C_\alpha(i, F_{\mathbf{p}}(t_2))$

Let us write $F_{\mathbf{p}}(t_2) = F_{\mathbf{p}}(t_1) + \epsilon$ for some $\epsilon > 0$. We then see:

$$\begin{aligned}
C_\alpha(i, F_{\mathbf{p}}(t_2)) &= \alpha(1 - F_{\mathbf{p}}(t_2))\phi(-i) + (1-\alpha)F_{\mathbf{p}}(t_2)\phi(i) \\
&= \alpha(1 - F_{\mathbf{p}}(t_1))\phi(-i) + (1-\alpha)F_{\mathbf{p}}(t_1)\phi(i) + \epsilon(\alpha\phi(-i) - (1-\alpha)\phi(i)) \\
&= C_\alpha(i, F_{\mathbf{p}}(t_1)) + \epsilon(\alpha\phi(-i) - (1-\alpha)\phi(i)) \\
&\leq C_\alpha(j, F_{\mathbf{p}}(t_1)) + \epsilon(\alpha\phi(-j) - (1-\alpha)\phi(j)) \\
&= C_\alpha(j, F_{\mathbf{p}}(t_2))\,.
\end{aligned}$$

If the inequality is strict, this contradicts the optimality of $C_\alpha(j, F_{\mathbf{p}}(t_2))$. Otherwise, we see that $C_\alpha(i, F_{\mathbf{p}}(t_2)) = C_\alpha(j, F_{\mathbf{p}}(t_2))$, implying $i = j$, contradicting the assumption $i > j$.

Now, we wish to show that the integral

$$\int_{t=-\infty}^{\infty} C_\alpha(g^*(t), F_{\mathbf{p}}(t))dt$$

exists. First, note that $g^*(t)$ has a countable number of discontinuities; each discontinuity is over an interval, and because rational numbers are dense on the real line, each interval can be mapped to a rational number which can in turn be mapped to a natural number. Similarly, $F_{\mathbf{p}}(t)$ may have a countable number of discontinuities.

Moreover, note that ϕ is convex, and thus unimodal, meaning that it also has a countable number of discontinuities. Thus, $C_\alpha(g^*(t), F_{\mathbf{p}}(t))$ may be discontinuous in a countable number of places, implying that the Lebesgue integral is well defined. □

For concreteness, we provide some examples of the infinite threshold scoring rule:

Example 5. If $\phi(u) = (1-u)_+$, commonly known as the hinge loss, then the surrogate loss becomes:

$$\psi_{\phi,\alpha}(y, g) = \alpha \int_{t=-\infty}^{y} (1 + g(t))_+ dt + (1-\alpha) \int_{t=y}^{\infty} (1 - g(t))_+ dt \,.$$

In order to minimize the expectation, we find that we get:

$$g^*(t) = \mathrm{sign}(F(t) - \alpha) \,.$$

In particular, if we do the minimization over functions of the form $g_v(t) = \mathrm{sign}(t - v)$, where the objective is then to simply learn the parameter v, then $\psi_{\phi,\alpha}(y, g)$ can be rewritten:

$$\psi_{\phi,\alpha}(y, g_v) = 2\alpha(y - v)_+ + 2(1-\alpha)(v - y)_+ = 2\psi_\alpha(y, v) \,.$$

That is, our framework has the α-pinball loss as a special case.

Example 6. If $\phi(u) = e^{-u}$, commonly known as the exponential loss, then the surrogate loss becomes:

$$\psi_{\phi,\alpha}(y, g) = \alpha \int_{t=-\infty}^{y} e^{g(t)} dt + (1-\alpha) \int_{t=y}^{\infty} e^{-g(t)} dt \,.$$

$\phi(u)$	Minimizer $g^*(t)$
$(1-u)_+$	$g^*(t) = \mathrm{sign}(F_{\mathbf{p}}(t) - \alpha)$
e^{-u}	$g^*(t) = \frac{1}{2}\log\left(\frac{F_{\mathbf{p}}(t)(1-\alpha)}{\alpha(1-F_{\mathbf{p}}(t))}\right)$
$\ln(1+e^{-u})$	$g^*(t) = \log\left(\frac{F_{\mathbf{p}}(t)(1-\alpha)}{\alpha(1-F_{\mathbf{p}}(t))}\right)$
$(1-u)^2$	$g^*(t) = \frac{F_{\mathbf{p}}(t)-\alpha}{\alpha(1-F_{\mathbf{p}}(t))+(1-\alpha)F_{\mathbf{p}}(t)}$

Table 1: Examples of $g^*(t)$ for different choices of ϕ when the minimization is over all increasing functions g.

In order to minimize the expectation, we find that we get:

$$g^*(t) = \frac{1}{2}\log\left(\frac{F_{\mathbf{p}}(t)(1-\alpha)}{\alpha(1-F_{\mathbf{p}}(t))}\right) .$$

As one example, we could do the minimization over functions of the form $g_v(t) = t - v$, where the objective is again to simply learn the parameter v. It is important to note here that the assumption that the optimal g^* is of a linear form is not reasonable, except for very specific forms of the CDF, but one can show that for symmetric distributions, this form of the surrogate can effectively elicit the median of a distribution. In this case, $\psi_{\phi,\alpha}(y,g)$ can be rewritten:

$$\psi_{\phi,\alpha}(y, g_v) = \alpha e^{y-v} + (1-\alpha)e^{v-y}$$

Table 1 provides several examples of optimal g^* functions for different forms of ϕ. From this table, we see that, in all presented examples other than that of the hinge loss, the optimal $g^*(t)$ is defined by an invertible transform of $F(t)$.

We then have the following two theorems regarding the infinite threshold surrogate, when combined with different pred functions:

Theorem 4. Let $\phi : \mathbb{R}{\rightarrow}\mathbb{R}_+$ be convex, continuous such that $(1-\alpha)\phi(\beta) - \alpha\phi(-\beta)$ is non-increasing in β. Define $\mathrm{pred}_w : \mathcal{G} \rightarrow \mathbb{R}$ as:

$$\mathrm{pred}_w(g) = \inf\{t \mid g(t) > 0\}\,,$$

and let $\psi_{\phi,\alpha}$ be defined as in Equation 1. Then, $(\psi_{\phi,\alpha}, \text{pred}_w)$ strictly weakly indirectly elicits the α-quantile if ϕ is differentiable at 0, $\phi'(0) < 0$.

Theorem 5. Let $\phi : \mathbb{R} \rightarrow \mathbb{R}_+$ be strictly convex and such that $(1-\alpha)\phi(\beta) - \alpha\phi(-\beta)$ is non-increasing in β, with ϕ is differentiable everywhere and $\phi'(0) < 0$. Define $\text{pred}_s : \mathcal{G} \to \mathbb{R}$ as:

$$\text{pred}_s(g) = \{t \mid g(t) = 0\} \cup \left\{ \inf\{t \mid g(t) > 0\} \right\} \cup \left\{ \sup\{t \mid g(t) < 0\} \right\},$$

and let $\psi_{\phi,\alpha}$ be defined as in Equation 1. Then, $(\psi_{\phi,\alpha}, \text{pred}_s)$ strictly strongly indirectly elicits the α-quantile.

3.2.1 Proof of Theorem 4: Weak Elicitation of Quantiles

Consider an arbitrary non-decreasing function $g^* \in \text{argmin}_g \mathbf{E}_{Y\sim\mathbf{p}}[\psi_{\phi,\alpha}(Y, g)]$. We know by Lemma 3 that g^* must satisfy:

$$\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y, g^*)] = \int_{t=-\infty}^{\infty} C^*_\alpha(F_\mathbf{p}(t))dt$$

We also have:

$$\begin{aligned}
\text{sign}\left(\text{argmin}_{\beta\in\mathbb{R}} C_\alpha(\beta, F_\mathbf{p}(t))\right) &= \text{sign}\left(\text{argmin}_\beta \alpha(1-F(t))\phi(-\beta) + (1-\alpha)F(t)\phi(\beta)\right) \\
&= \text{sign}\Bigg(\text{argmin}_\beta \frac{\alpha(1-F_\mathbf{p}(t))}{\alpha(1-F_\mathbf{p}(t)) + F_\mathbf{p}(t)(1-\alpha)}\phi(-\beta) \\
&\qquad + \frac{F_\mathbf{p}(t)(1-\alpha)}{\alpha(1-F_\mathbf{p}(t)) + F_\mathbf{p}(t)(1-\alpha)}\phi(\beta)\Bigg)
\end{aligned}$$

By Proposition 1, we see that because the scoring rule derived from ϕ can elicit sign$(\eta - 1/2)$, the above has $\beta > 0$ if and only if $\frac{F_\mathbf{p}(t)(1-\alpha)}{\alpha(1-F_\mathbf{p}(t))+F_\mathbf{p}(t)(1-\alpha)} > \frac{1}{2}$ or equivalently $F_\mathbf{p}(t) > \alpha$. Hence,

$$\text{sign}\left(\text{argmin}_{\beta\in\mathbb{R}} C_\alpha(\beta, F_\mathbf{p}(t))\right) = \text{sign}\left(F_\mathbf{p}(t) - \alpha\right) \tag{2}$$

Now, suppose towards contradiction that $\text{pred}_Q(g^*) \notin Q_\alpha(\mathbf{p})$. We know that $Q_\alpha(\mathbf{p})$ is a closed interval, i.e. $Q_\alpha(\mathbf{p}) = [q_\alpha, q^\alpha]$ for $q_\alpha \le q^\alpha$. There are then two cases:

Case 1: $\text{pred}_w(g^*) < q_\alpha$

> In this case, we have $\inf\{t \mid g^*(t) > 0\} < q_\alpha$. Let $t^* = \inf\{t \mid g^*(t) > 0\}$, so that we may write $t^* + \epsilon = q_\alpha$ for some $\epsilon > 0$. Then, by definition of infimum, we know that $g\left(t^* + \frac{\epsilon}{2}\right) > 0$. However, by Equation 2, we know this implies $F_{\mathbf{p}}\left(t^* + \frac{\epsilon}{2}\right) > \alpha$.
>
> However, we have $G_{\mathbf{p}}(q_\alpha) \ge 1 - \alpha$, and that $G_{\mathbf{p}}$ is non-decreasing, so that $G_{\mathbf{p}}\left(t^* + \frac{\epsilon}{2}\right) \ge G_{\mathbf{p}}(q_\alpha) \ge 1 - \alpha$, implying q_α is not the true lower bound of $Q_\alpha(\mathbf{p})$, a contradiction.

Case 2: $\text{pred}_w(g^*) > q^\alpha$

> In this case, we must have: $\inf\{t \mid g^*(t) > 0\} > q^\alpha$. Let $t^* = \inf\{t \mid g^*(t) > 0\}$, so that we may write $t^* = q^\alpha + \epsilon$ for some $\epsilon > 0$. Again, from [30], we know that if $t > q^\alpha$, then $F(t) > \alpha$. Hence, we must have $F_{\mathbf{p}}\left(q_\alpha + \frac{\epsilon}{2}\right) > \alpha$, implying $g\left(q_\alpha + \frac{\epsilon}{2}\right) > 0$ by 2.
>
> However, this contradicts the definition of t^*, as there exists a point $q_\alpha + \frac{\epsilon}{2} < t^*$ such that $g\left(q_\alpha + \frac{\epsilon}{2}\right) > 0$.

Thus, we must have $\text{pred}_w(g^*) \in Q_\alpha(\mathbf{p})$, implying that $(\psi_{\phi,\alpha}, \text{pred}_w)$ indirectly elicits the α-quantile.

3.2.2 Proof of Theorem 5: Strong Elicitation of Quantiles

Consider an arbitrary non-decreasing function $g^* \in \text{argmin}_g \mathbf{E}_{Y\sim\mathbf{p}}[\psi_{\phi,\alpha}(Y, g)]$. We know by Lemma 3 that g^* must satisfy:

$$\mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y, g^*)] = \int_{t=-\infty}^{\infty} C^*_\alpha(F_{\mathbf{p}}(t))dt\,.$$

We now note that, due to the structure of ϕ, we can in fact explicitly solve for $g^*(t)$. Consider the following:

$$\begin{aligned}
\text{argmin}_\beta C_\alpha(\beta, F_{\mathbf{p}}(t)) &= \text{argmin}_\beta \alpha(1 - F_{\mathbf{p}}(t))\phi(-\beta) + (1-\alpha)F_{\mathbf{p}}(t)\phi(\beta) \\
&= \text{argmin}_\beta \frac{\alpha(1 - F_{\mathbf{p}}(t))}{\alpha(1 - F_{\mathbf{p}}(t)) + (1-\alpha)F_{\mathbf{p}}(t)}\phi(-\beta) \\
&\qquad + \frac{(1-\alpha)F_{\mathbf{p}}(t)}{\alpha(1 - F_{\mathbf{p}}(t)) + (1-\alpha)F_{\mathbf{p}}(t)}\phi(\beta) \\
&= \gamma^{-1}\left(\frac{(1-\alpha)F_{\mathbf{p}}(t)}{\alpha(1 - F_{\mathbf{p}}(t)) + (1-\alpha)F_{\mathbf{p}}(t)}\right),
\end{aligned}$$

where the final step and definition of γ^{-1} comes from Proposition 1. Moreover, due to the symmetry of ϕ, we see that $\gamma^{-1}(1/2) = 0$, and since ϕ is strictly convex, we have that γ^{-1} is strictly increasing.

Hence, we have:

$$g^*(t) = \gamma^{-1}\left(\frac{(1-\alpha)F_{\mathbf{p}}(t)}{\alpha(1 - F_{\mathbf{p}}(t)) + (1-\alpha)F_{\mathbf{p}}(t)}\right). \tag{3}$$

Suppose towards contradiction that $\text{pred}_s(g^*) \neq Q_\alpha(\mathbf{p})$. Then, there are two cases:

Case 1: There is $t' \in \text{pred}_s(g^*)$ such that $t' \notin Q_\alpha(\mathbf{p})$.

> There are 3 cases for t': Either $g^*(t') = 0$, $t' = \sup\{t \mid g^*(t) < 0\}$, or $t' = \inf\{t \mid g^*(t) > 0\}$. The first case results in an immediate contradiction, by 3, as if $F(t') = \alpha$, then $t' \in Q_\alpha(\mathbf{p})$ by definition.
>
> Otherwise, suppose $t' = \sup\{t \mid g^*(t) < 0\}$. If $g^*(t') > 0$, then $t = \inf\{t \mid g(t) > 0\}$, and the proof of weak elicitation of quantiles immediately shows that $t' \in Q_\alpha(\mathbf{p})$, a contradiction. Otherwise, if $g^*(t') < 0$, we may repeat the proof of weak elicitation of quantiles, but reverse the sign, again showing that $t' \in Q_\alpha(\mathbf{p})$, as desired.

Case 2: There is $t' \in Q_\alpha(\mathbf{p})$ such that $t' \notin \text{pred}_s(g^*)$.

Consider the value $g(t')$. It must either be the case $g(t') > 0$ or $g(t') < 0$. Without loss of generality, suppose $g(t') > 0$. By 3, it must be the case that $F_{\mathbf{p}}(t') > \alpha$. We must also have $G(t') \geq 1 - \alpha$, since $t' \in Q_\alpha(\mathbf{p})$. This then implies that F places strictly positive probability on t', say $\mathbf{P}(Y = t') = p$. However, this then implies $\sup Q_\alpha(\mathbf{p}) = t'$, so it must be the case that $\inf\{t \mid g^*(t) > 0\}$ contains t', a contradiction.

Therefore, we must have $\mathrm{pred}_s(g^*) = Q_\alpha(\mathbf{p})$, which implies that the pair $(\psi_{\phi,\alpha}, \mathrm{pred}_s)$ strongly indirectly elicits the α-quantile of $\mathbf{p}$.

4 Application to Conditional Quantile Elicitation

In this section, we discuss an application of quantile elicitation to the problem of *conditional* quantile elicitation, also known as *quantile regression*, in which, given some instance features of a particular sample, one wants to know the α-quantile of the conditional distribution of the label of that sample.

Conditional quantile elicitation has several applications, including in Econometrics, Sociology, and Ecology. For example, in the econometric problem of wallet estimation, where one is asked to determine how much a customer is willing to spend on an item, conditional quantiles are quite useful as they are more robust to heavy-tailed income distributions.

In this section, we reuse much notation to reflect similarities to Section 3. For example ψ is used as a surrogate loss here, because we will often use proper scoring rules as our surrogate loss.

4.1 Setting

Formally, we have an instance space $\mathcal{X}$, a label space $\mathcal{Y} = \mathbb{R}$, a prediction space $\widehat{\mathcal{Y}}$, and a set of training examples $S = ((X_1, Y_1), \ldots, (X_m, Y_m))$ drawn from some underlying distribution $\mathcal{D}$. The task is to learn the conditional quantiles of the distribution $\mathbf{P}(\cdot|X = x) = \mathbf{p}_x$ over $\mathcal{Y}$.

In this setting, the goal is to learn a function $h_S : \mathcal{X} \to \widehat{\mathcal{Y}}$ which predicts the α-quantile. Similar to the setting of property elicitation, in learning a function to estimate quantiles, the prediction space varies in the literature. Often, it is assumed that $Q_\alpha(\mathbf{p}_x)$ consists of a single point for each $x \in \mathcal{X}$, so the function learned is of the form $h_S : \mathcal{X} \to \mathbb{R}$. If $Q_\alpha(\mathbf{p}_x)$ is allowed to have non-zero Lebesgue measure, then the function learned is of the form $h_S : \mathcal{X} \to \mathcal{I}$. In this sense, we are again able to differentiate between weak and strong quantile elicitation, as in Section 3.

Indirect Conditional Quantile Elicitation. Often, and again related to the above section, we are able to indirectly elicit a conditional property. Rather than directly learning the function $h_S : \mathcal{X} \to \widehat{\mathcal{Y}}$, we learn a function $f_S : \mathcal{X} \to \mathcal{H}$ in some surrogate space $\mathcal{H}$, and a mapping $\text{pred} : \mathcal{H} \to \widehat{\mathcal{Y}}$. In our context, we will learn f_S using a loss minimization approach.

4.2 Review of Literature

In this section, we describe several results relevant to our framework. The first series of results provides key insights into how the problem of property elicitation can be used for conditional property elicitation. This then allows us to apply many of the results in Section 3 to conditional quantile elicitation.

The next series of results shows the existence of *regret bounds* for conditional quantile elicitation. We are able to contrast such results with our own regret bounds regarding the infinite threshold surrogate.

In the final section, we discuss an approach to conditional quantile elicitation that comes about as a special case of our framework, in which we approximate integrals by sums.

4.2.1 Conditional Property Elicitation

In this section, we provide some results that show that the use of a proper scoring rule can result in conditional property elicitation. We begin with the following definitions and results, from [1].

Definition 13 ($(\ell, \mathcal{P}')$-Calibrated Surrogate Loss)**.** Consider a loss function $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, a surrogate loss $\psi : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}_+$, and a mapping $\text{pred} : \mathcal{H} \to \widehat{\mathcal{Y}}$. Let $Opt(\ell, \mathbf{p}) = \text{argmin}_{\widehat{y} \in \widehat{\mathcal{Y}}} \mathbf{E}_{Y \sim \mathbf{p}}[\ell(Y, \widehat{y})]$. Then, given a set of probability distributions $\mathcal{P}' \subseteq \mathcal{P}$, we say

(ψ, pred) is $(\ell, \mathcal{P}')$-calibrated with respect to ℓ if:

$$\forall \mathbf{p} \in \mathcal{P}', \qquad \inf_{\mathbf{u} \in \mathcal{H} : \mathrm{pred}(\mathbf{u}) \notin Opt(\ell, \mathbf{p})} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] > \min_{\mathbf{u} \in \mathcal{H}} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] \, .$$

Moreover, define the following quantities to denote *generalization error* and *Bayes error*:

Definition 14 (Generalization Error)**.** For a loss $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, the generalization error of a function $h : \mathcal{X} \to Y$ with respect to a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ is defined:

$$\mathrm{er}^{\ell}_{\mathcal{D}}[h] = \mathbf{E}_{(X,Y) \sim \mathcal{D}}[\ell(Y, h(X))] \, .$$

Definition 15 (Bayes Error)**.** For a loss $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, the Bayes error of with respect to a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ is defined:

$$\mathrm{er}^{\ell,*}_{\mathcal{D}} = \inf_{h : \mathcal{X} \to \mathcal{Y}} \mathbf{E}_{(X,Y) \sim \mathcal{D}}[\ell(Y, h(X))] \, .$$

It is typically assumed that the Bayes error is achievable by some function h^*.

We also have the following theorem by Zhang [34] shows the usefulness of algorithms that use $(\ell, \mathcal{P}')$ calibrated surrogate losses:

Theorem 6 (Surrogate Loss Minimization, Zhang [34])**.** Consider a loss function $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, a surrogate loss $\psi : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}_+$, and a mapping $\mathrm{pred} : \mathcal{H} \to \widehat{\mathcal{Y}}$. Then (ψ, pred) is $(\ell, \mathcal{P}')$-calibrated if and only if, for distributions $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ such that, given any $x \in \mathcal{X}$ the conditional distribution of $\mathcal{Y}$ is in $\mathcal{P}'$, and all sequences f_m from training sets $((X_1, Y_1), \ldots, (X_m, Y_m))$ drawn i.i.d. from such distributions $\mathcal{D}$, we have:

$$\mathrm{er}^{\psi}_{\mathcal{D}}[f_m] \to^{P} \mathrm{er}^{\psi,*}_{\mathcal{D}} \qquad \Longrightarrow \qquad \mathrm{er}^{\ell}_{\mathcal{D}}[\mathrm{pred} \circ f_m] \to^{P} \mathrm{er}^{\ell,*}_{\mathcal{D}} \, .$$

Now, we define a similar notion of calibration for a property with respect to a loss ℓ:

Definition 16 (($\ell, \mathcal{P}'$)-Calibrated Property)**.** Consider a loss function $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, a property $\Gamma : \mathcal{P} \to \mathcal{H}$, and a mapping $\text{pred} : \mathcal{H} \to \widehat{\mathcal{Y}}$. Then, given a set of probability distributions $\mathcal{P}' \subseteq \mathcal{P}$, we say (Γ, pred) is $(\ell, \mathcal{P}')$-calibrated, if, for a sequence $\{\mathbf{v}_n\}$:

$$\forall \mathbf{p} \in \mathcal{P}', \quad \mathbf{v}_n \to \Gamma(\mathbf{p}) \quad \implies \quad \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y, \text{pred}(\mathbf{v}_n))] \to \min_{\mathbf{u}\in\mathcal{H}} \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y, \mathbf{u})] \,.$$

Agarwal and Agarwal [1] then showed the following result, relating calibrated properties to calibrated surrogates. In their proof, they considered a discrete prediction space, so here we have adjusted the proof to allow for a continuous prediction space, so long as the prediction with lowest expected loss has strictly lower loss than any other prediction.

Theorem 7 (Calibrated Surrogates from Calibrated Properties, Agarwal and Agarwal [1])**.** Consider discrete spaces $\mathcal{Y}$ and $\widehat{\mathcal{Y}}$. Let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$ and $\mathcal{P}' \subseteq \mathcal{P}$. Let $\Gamma : \mathcal{P} \to \mathcal{H}$, and $\text{pred} : \mathcal{H} \to \widehat{\mathcal{Y}}$, such that (Γ, pred) is $(\ell, \mathcal{P}')$-calibrated. Further, let $\psi : \mathcal{Y} \times \mathcal{H} \to \mathbb{R}_+$ be a strictly proper scoring rule for Γ. Then, the pair (ψ, pred) is $(\ell, \mathcal{P}')$-calibrated.

4.2.2 Minimization of the α-Pinball Loss

The α-pinball loss has been studied widely in quantile regression, for example in [15, 32]. Here, we describe in detail the results of [31], in which surrogate regret bounds are provided for the α-pinball loss in quantile regression.

To understand the results of [31], we must go into detail about specific *types* of quantiles:

Definition 17 (Quantiles of type q)**.** Consider a distribution $\mathbf{p}$ with support on $[-1, 1]$, and label the α-quantile $Q_\alpha(\mathbf{p}) = [q_\alpha, q^\alpha]$. $\mathbf{p}$ is said to have an α-quantile of type $q \in [1, \infty)$ if there exist constants $a_\mathbf{p} \in (0, 2]$ and $b_\mathbf{p} > 0$ such that:

$$\mathbf{P}_{Y\sim\mathbf{p}}(q^\alpha \le Y \le q^\alpha + s) \ge b_\mathbf{p} s^{q-1} \,, \text{ and}$$
$$\mathbf{P}_{Y\sim\mathbf{p}}(q_\alpha - s \le Y \le q_\alpha) \ge b_\mathbf{p} s^{q-1} \,,$$

for all $s \in [0, a_{\mathbf{p}})$.

In essence, this definition quantifies how the probability distribution is spread around the α-quantile.

Moreover, since we are looking at conditional quantile elicitation, to account for the joint distribution we may define α-quantiles of p-average type, as follows.

Definition 18 (Quantiles of p-average type q)**.** Let $p \in (0, \infty)$, and $q \in [1, \infty)$, and let $\mathcal{D}$ be a distribution on $\mathcal{X} \times \mathbb{R}$. We say $\mathcal{D}$ has an α-quantile of p-average type q if $\mathbf{p}_x = \mathbf{P}(\cdot|X = x)$ has an α-quantile of type q for almost all X under D_X, the marginal density of X under $\mathcal{D}$, and the function:

$$\tau(x) = b_{\mathbf{p}_x} a_{\mathbf{p}_x}^{q-1}$$

With $b_{\mathbf{p}_x}$ and $a_{\mathbf{p}_x}$ defined as in Definition 17, is such that τ^{-1} is in the set $L_p(D_X)$.

In this context, Steinwart [31] studied what we would call weak conditional quantile elicitation, in which the goal was to elicit any point in the α-quantile of the conditional distribution. Hence, they defined a notion of distance to a set:

Definition 19 (Distance to a Set)**.** The distance between a point $t \in \mathbb{R}$ and a set $A \subseteq \mathbb{R}$ is defined:

$$\operatorname{dist}(t, A) = \inf_{s \in A} |t - s|$$

Remark. The distance between a quantile and a point can be defined using a minimum rather than an infimum, because quantiles are closed intervals.

We may now state the regret bound shown in [31], which relates the value of the pinball loss to the distance between an algorithms output and the true quantiles:

Theorem 8 (Regret Bound for Pinball Loss, Steinwart [31])**.** Let ψ_α be the α-pinball loss, and $p \in (0, \infty)$ and $q \in [1, \infty)$. Further, let $r = \frac{pq}{p+1}$. Let $\mathcal{D}$ have an α-quantile of p-average type q. Then, for all $f : \mathcal{X} \to [-1, 1]$, it is verified that:

$$||\text{dist}(f, Q_\alpha)||_{L_r(D_X)} \le 2^{1-1/q} q^{1/q} ||\tau^{-1}||_{L_p(D_X)} \left(\text{er}_{\mathcal{D}}^{\psi_\alpha}[f] - \text{er}_{\mathcal{D}}^{\psi_\alpha,*}\right)^{1/q} .$$

4.2.3 Reduction Approaches to Conditional Quantile Elicitation

In terms of algorithmic approaches to quantile elicitation, Langford et al [19] showed that conditional quantile elicitation could be approached by using a reduction to binary classification. In particular, they provided the following algorithm, Quanting-Train to learn a model to predict quantiles:

Algorithm 1 Quanting-Train (Weighted binary classification algorithm A, training sample S, quantile α)

```
for $t \in [0,1]$ do
    $S_t = \{\}$
    for each $(x, y) \in S$ do
        $S_t = S_t \cup \{x, \mathbf{1}(y \ge t), \alpha\mathbf{1}(y \ge t) + (1-\alpha)\mathbf{1}(y < t)\}$
    $c_t = A(S_t)$
Return set of all $c_t$
```

Here, it is assumed that the support of the conditional distribution of the label is always on $[0, 1]$. In practice, one is unable to iterate through all $t \in [0, 1]$, so typically one iterates through the values $\{0, 1/n, \ldots, (n-1)/n, 1\}$ for some suitably large n.

To evaluate the quantile given the binary classifiers, Langford et al uses the following algorithm:

Algorithm 2 Quanting-Train (Set of classifiers $\{c_t\}$, test set S')

```
for each $x \in S'$ do
    $Q(x) = \mathbf{E}_{t \sim U(0,1)}[c_t(x)]$
```

In practice, one would approximate this expectation with a sum.

Langford et al showed strong experimental evidence showing that this approach to conditional quantile elicitaition, when n is large, has better minimization of the α-pinball loss than more common methods such as linear quantile regression. This can be attributed to the different function class that is used by Langford et al, as they do not restrict their final classifier to be a linear function of the instance space.

4.3 Theoretical Results

In this subsection, we show two results regarding the infinite threshold scoring rule in the context of conditional quantile elicitation. Then, in the next subsection, we provide an analysis of the Langford algorithm under our framework.

We first show that using the infinite threshold scoring rule as a loss function will elicit the conditional quantiles of a distribution:[4]

Theorem 9 (Calibration of Infinite Thresholds)**.** Consider a training set $((X_1, Y_1), \ldots, (X_m, Y_m))$ drawn i.i.d. from a distribution $\mathcal{D}$. Then, we have the following implication, where f_m is the function (which maps instances of $\mathcal{X}$ to monotonically increasing functions) learned by minimizing $\psi_{\phi,\alpha}$ over the sample of size m:

$$\mathrm{er}_{\mathcal{D}}^{\psi_{\phi,\alpha}}[f_m] \to^P \mathrm{er}_{\mathcal{D}}^{\psi,*} \quad \implies \quad \mathbf{E}_{(X,Y)\sim D}[\mathrm{dist}(\mathrm{pred}_w(f_m(X)), Q_\alpha(\mathbf{p}_X))] \to^P 0\,.$$

Here, pred_w is defined as in Theorem 4.

Proof. This result follows immediately from Theorem 4. We may consider the quantity:

$$\mathrm{er}_{\mathcal{D}}^{\psi_{\phi,\alpha}}[f_m] = \mathbf{E}_X\left[\mathbf{E}_{Y|X}[\psi_{\phi,\alpha}(Y, f_m(X))]\right]\,.$$

Because $\psi_{\phi,\alpha}$ weakly indirectly elicits $Q_\alpha(\mathbf{p}_x)$, where $\mathbf{p}_x$ is the conditional distribution of Y given x, we see that $\mathrm{er}_{\mathcal{D}}^{\psi_{\phi,\alpha}}[f_m] \to^P \mathrm{er}_{\mathcal{D}}^{\psi,*}$ implies that $\mathbf{E}_{Y|X}[\psi_{\phi,\alpha}(Y, f_m(X))]$ is minimized for almost all X under $\mathcal{D}_X$. Hence, by Theorem 4, it must be the case that, in probability,

[4] Here, we assume that the function class over which the minimization occurs is suitably large.

$\text{pred}_w(f_m(X)) \in Q_\alpha(\mathbf{p}_X)$ for almost all X under $\mathcal{D}_X$, implying

$$\mathbf{E}_{(X,Y)\sim D}[\text{dist}(\text{pred}_w(g_m(X)), Q_\alpha(\mathbf{p}_X))] \to^P 0\,,$$

by the definition of distance, as desired. □

In particular, for median estimation we can also show a regret bound for the infinite threshold surrogate, using techniques similar to Pedregosa et al [23], and combining this with the result shown by Steinwart et al [31]:

Theorem 10 (Regret Bound for Infinite Thresholds)**.** Let $\psi_{\phi,1/2}$ be the infinite threshold scoring rule as defined in Definition 1 with ϕ differentiable at 0, with $\phi'(0) < 0$ and $\phi(\beta) - \phi(-\beta)$ increasing in β, and let $p \in (0, \infty)$ and $q \in [1, \infty)$. Further, let $r = \frac{pq}{p+1}$. Let $\mathcal{D}$ have a median (1/2-quantile) of p-average type q. Then, for all $f : \mathcal{X} \to \mathcal{G}$ where $\mathcal{G}$ is the set of increasing functions, it is verified that:

$$||\text{dist}(\text{pred}_w \circ f, Q_{1/2})||_{L_r(D_X)} \leq 2^{1-1/q} q^{1/q} ||\tau^{-1}||_{L_p(D_X)} \xi \left(\text{er}_{\mathcal{D}}^{\psi_{\phi,1/2}}[f] - \text{er}_{\mathcal{D}}^{\psi_{\phi,1/2},*}\right)^{1/q}\,,$$

for a function ξ which is defined by ϕ.

Proof. For a particular $x \in \mathcal{X}$, let $f(x) = g$, a monotonically increasing function, and let $\mathbf{p}_x$ be the conditional distribution of Y given that $X = x$. Moreover, define $u = \text{pred}_w(g)$.

Then, we first note the following for the α-pinball loss ψ_α:

$$\begin{aligned}
\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,u)] &= \int_{-\infty}^{\infty} (\alpha(y-u)_+ + (1-\alpha)(u-y)_+)\, dF_{\mathbf{p}_x}(y) \\
&= \alpha \int_{-\infty}^{\infty}\int_u^y \mathbf{1}(u<y) d\kappa dF_{\mathbf{p}_x}(y) + (1-\alpha)\int_{-\infty}^{\infty}\int_y^u \mathbf{1}(u>y) d\kappa dF_{\mathbf{p}_x}(y) \\
&= \alpha \int_u^{\infty}\int_u^y d\kappa dF_{\mathbf{p}_x}(y) + (1-\alpha)\int_{-\infty}^{u}\int_y^u d\kappa dF_{\mathbf{p}_x}(y) \\
&= \alpha \int_u^{\infty}\int_u^{\infty} dF_{\mathbf{p}_x}(y) d\kappa + (1-\alpha)\int_{-\infty}^{u}\int_{-\infty}^u dF_{\mathbf{p}_x}(y) d\kappa \\
&= \alpha \int_u^{\infty} G_{\mathbf{p}_x}(\kappa) d\kappa + (1-\alpha)\int_{-\infty}^{u} F_{\mathbf{p}_x}(\kappa) d\kappa \,.
\end{aligned}$$

There are then 3 cases for u, as compared to an arbitrary $q \in Q_\alpha(\mathbf{p}_x)$, where we denote $Q_\alpha(\mathbf{p}_x) = [q_\alpha, q^\alpha]$.

Case 1: $u \in Q_\alpha(\mathbf{p}_x)$

In this case, $\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,u)] = \mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,q)]$, by Proposition 3.

Case 2: $u < q_\alpha$.

In this case, we may write:

$$\begin{aligned}
\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,u)] - \mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,q)] &= \alpha \int_u^{\infty} G_{\mathbf{p}_x}(\kappa) d\kappa + (1-\alpha)\int_{-\infty}^{u} F_{\mathbf{p}_x}(\kappa) d\kappa \\
&\quad - \left(\alpha \int_q^{\infty} G_{\mathbf{p}_x}(\kappa) d\kappa + (1-\alpha)\int_{-\infty}^{q} F_{\mathbf{p}_x}(\kappa) d\kappa\right) \\
&= \alpha \int_u^{q} G_{\mathbf{p}_x}(\kappa) d\kappa - (1-\alpha)\int_{u}^{q} F_{\mathbf{p}_x}(\kappa) d\kappa \,.
\end{aligned}$$

In the case of the median, this becomes:

$$\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,u)] - \mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,q)] = \frac{1}{2}\int_u^q (G_{\mathbf{p}_x}(\kappa) - F_{\mathbf{p}_x}(\kappa)) d\kappa \,.$$

Case 3: $u > q^\alpha$.

Similar to the above, in the case of the median we have:

$$\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,u)] - \mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,q)] = \frac{1}{2}\int_q^u (F_{\mathbf{p}_x}(\kappa) - G_{\mathbf{p}_x}(\kappa))d\kappa\,.$$

In any case, we may bound:

$$\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,u)] - \mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,q)] \leq \frac{1}{2}\int_q^u |F_{\mathbf{p}_x}(\kappa) - G_{\mathbf{p}_x}(\kappa)|d\kappa\,.$$

We define the ξ-transform of ϕ as was done in Bartlett et al [2] as follows:[5]

$$\xi(\theta) = \phi(0) - 2C^*_{1/2}\left(\frac{1+\theta}{2}\right)\,.$$

Hence, because we assume Y has support $[-1,1]$, we may then bound:

$$\begin{aligned}
\xi\left(\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,u)] - \mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_\alpha(Y,q)]\right) &\leq \xi\left(\frac{1}{2}\int_q^u |F_{\mathbf{p}_x}(\kappa) - G_{\mathbf{p}_x}(\kappa)|d\kappa\right)\\
&\leq \frac{1}{2}\int_q^u \xi\left(|F_{\mathbf{p}_x}(\kappa) - G_{\mathbf{p}_x}(\kappa)|\right)d\kappa \qquad \text{(Jensen's Inequality)}\\
&= \frac{1}{2}\int_q^u \xi\left(F_{\mathbf{p}_x}(\kappa) - G_{\mathbf{p}_x}(\kappa)\right)d\kappa \qquad \text{(Symmetry}^6\text{)}\\
&\leq \frac{1}{2}\int_q^u \xi\left(2F_{\mathbf{p}_x}(\kappa) - 1\right)d\kappa\\
&= \frac{1}{2}\int_q^u (\phi(0) - 2C^*_{1/2}(F_{\mathbf{p}_x}(\kappa))d\kappa\,.
\end{aligned}$$

[5] In Bartlett et l [2], this is referred to as the ψ-transform, but here we use ψ to represent scoring rules so we use the notation ξ instead.

We note that, if $\kappa \notin Q_{1/2}(\mathbf{p}_x)$, then $\phi(0) \leq 2C_{1/2}(g(\kappa), F_{\mathbf{p}_x}(\kappa))$. This is clear, as:

$$\begin{aligned}
2C_{1/2}(g(\kappa), F_{\mathbf{p}_x}(\kappa)) &= F_{\mathbf{p}_x}\phi(g(\kappa)) + (1 - F_{\mathbf{p}_x})\phi(-g(\kappa)) \\
&\geq \phi(F_{\mathbf{p}_x}g(\kappa) - (1 - F_{\mathbf{p}_x})g(\kappa)) \qquad \text{(Jensen's inequality)} \\
&\geq \phi(g(\kappa)(2F_{\mathbf{p}_x}(\kappa) - 1)) \\
&\geq \phi(0) + (g(\kappa)(2F_{\mathbf{p}_x}(\kappa) - 1))\phi'(0) \\
&\geq \phi(0)\,,
\end{aligned}$$

where the last inequality comes from the fact that if $\kappa \notin Q_{1/2}(\mathbf{p}_x)$, then $g(\kappa)$ an $(2F_{\mathbf{p}_x}(\kappa)-1)$ must differ in sign. Hence,

$$\begin{aligned}
\frac{1}{2}\int_q^u (\phi(0) - 2C^*_{1/2}(F_{\mathbf{p}_x}(\kappa))d\kappa &\leq \frac{1}{2}\int_q^u (2C_{1/2}(g(\kappa), F_{\mathbf{p}_x}(\kappa)) - 2C^*_{1/2}(F_{\mathbf{p}_x}(\kappa))d\kappa \\
&\leq \frac{1}{2}\int_q^u (2C_{1/2}(g(\kappa), F_{\mathbf{p}_x}(\kappa)) - 2C^*_{1/2}(F_{\mathbf{p}_x}(\kappa))d\kappa \\
&= 2\left(\mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_{\phi,\alpha}(Y, g)] - \mathbf{E}_{Y\sim\mathbf{p}_x}[\psi_{\phi,\alpha}(Y, g^*)]\right)\,,
\end{aligned}$$

where g^* is a minimizer of $\psi_{\phi,\alpha}$. Because the above holds for all $x \in X$, we see that it holds over the expectation by monotonicity of expectation. Hence, we can write:

$$\mathrm{er}^{\psi_\alpha}_{\mathcal{D}}[f] - \mathrm{er}^{\psi_\alpha,*}_{\mathcal{D}} \leq \xi\left(\mathrm{er}^{\psi_{\phi,1/2}}_{\mathcal{D}}[f] - \mathrm{er}^{\psi_{\phi,1/2},*}_{\mathcal{D}}\right)\,.$$

Combining this with Theorem 8, we get the desired bound

$$||\mathrm{dist}(\mathrm{pred}_w \circ f, Q_{1/2})||_{L_r(D_X)} \leq 2^{1-1/q}q^{1/q}||\tau^{-1}||_{L_p(D_X)}\xi\left(\mathrm{er}^{\psi_{\phi,1/2}}_{\mathcal{D}}[f] - \mathrm{er}^{\psi_{\phi,1/2},*}_{\mathcal{D}}\right)^{1/q}\,.$$

□

[6] The reason that this argument does not generalize to the α-quantile is because this symmetry is required for this specific regret bound.

4.3.1 Reduction Approach via Riemann Sums

Now, suppose that for all $x \in \mathcal{X}$, $\mathcal{Y}$ has support on $[0,1]$, as in [19]. Similar to the context in Steinwart et al [31], so long as the conditional support of Y is bounded we can generally normalize our data such that this statement holds. We may then consider a relaxation of the infinite thresholds surrogate, which discretizes the integral:

Definition 20 (n-Finite Threshold Surrogate)**.** The n-Finite Threshold Surrogate, $\psi_{n,\phi,\alpha} : \mathcal{Y} \times \mathbb{R}^{n+1} \to \mathbb{R}_+$ is defined as the following summation:

$$\psi_{n,\phi,\alpha}(y, \mathbf{v}) = \alpha \sum_{t=0}^{y} \frac{1}{n}\phi(-v_t) + (1-\alpha) \sum_{t=y+1}^{n} \frac{1}{n}\phi(v_t)\,, \tag{4}$$

Remark. If the prediction space is restricted to $\mathcal{A} \subseteq \mathbb{R}^{n+1}$, defined:

$$\mathcal{A} = \{\mathbf{v} \mid v_0 \leq v_1 \leq \cdots \leq v_n\}\,.$$

Then as $n \to \infty$, the n-Finite Threshold Surrogate converges to the Infinite Threshold surrogate.

We claim that, from the loss minimization perspective, Langford et al are effectively minimizing the n-Finite Threshold Surrogate:

Proposition 4. If a binary classification algorithm is use in Quanting-Train which minimizes a margin loss ϕ, then Quanting-Train minimizes the n-Finite Threshold Surrogate

Proof. Consider the generalization error of $\psi_{n,\phi,\alpha}$:

$$\begin{aligned}\mathbf{E}_{Y\sim\mathbf{p}}[\psi_{n,\phi,\alpha}(y,\mathbf{v})] &= \int_0^1 \left(\alpha \sum_{t=0}^{y} \frac{1}{n}\phi(-v_t) + (1-\alpha) \sum_{t=y+1}^{n} \frac{1}{n}\phi(v_t) \right) dF(y) \\ &= \alpha \sum_{t=0}^{y} \int_0^1 \frac{1}{n}\phi(-v_t)dF(y) + (1-\alpha) \sum_{t=y+1}^{n} \int_0^1 \frac{1}{n}\phi(v_t)dF(y)\,.\end{aligned}$$

Here, we note that for an arbitrary but particular t, when $y = t$ the first summation contributes to the loss with probability $1 - F(t)$, and the second contributes to the loss with probability $F(t)$. Hence,

$$\mathbf{E}_{Y \sim \mathbf{p}}[\psi_{n,\phi,\alpha}(y, \mathbf{v})] = \sum_{t=0}^{n} \alpha(1 - F(t))\phi(-v_t) + (1 - \alpha)F(t)\phi(v_t)\,.$$

To minimize this term, one can simply minimize each term of the summand, as the sum is separable. This is exactly what would be done by Quanting-Train when it performs the importance-weighted binary classification, if the classification is done using a margin loss. □

Remark. It is worth noting that Langford et al [19] use a distinctly different pred function in Quanting-Test than pred_w as defined in Theorem 4. It appears to be close to pred_w, as one would expect that if the minimization was done over $\mathcal{A}$ as opposed to $\mathbb{R}^{n+1}$, the increasing function would intersect 0 at precisely $\mathbf{E}_{t \sim U(0,1)}[c_t(x)]$— this is exactly the point at which the margin would go from being negative to becoming positive. However, because Langford et al. do not enforce this strictly increasing assumption, their effective pred function is markedly different from pred_w.

5 Application to Multiclass Classification

We turn our attention to the application of *multiclass classification*, which is a commonly studied problem in machine learning [7, 8, 9, 12, 20, 22, 1]. Through the use of quantiles, and coarse probability estimation, we present results for traditionally "difficult" losses, namely the 0-1 loss and the 0-1 loss with a reject option. Here again, we reuse notation from previous sections to reflect the similarity between these results and previous results.

5.1 Setting

In the problem of multiclass classification, there is an instance space $\mathcal{X}$, a finite label space $\mathcal{Y} = [n] = \{1, \ldots, n\}$, and a finite prediction space $\widehat{\mathcal{Y}} = [k] = \{1, \ldots, k\}$ (usually $k = n$, although that is not always the case.) We are given training examples $S = \big((X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)\big)$ drawn i.i.d. from a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$.

Given the training set S, the goal is to generate a function $h : \mathcal{X} \to \mathcal{Y}$, which performs well according to some loss function $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$. Specifically, we wish to minimize the *ℓ-generalization error* of h, which we recall is definied $\mathrm{er}^{\ell}_{\mathcal{D}}[h] = \mathbf{E}_{(X,Y)\sim\mathcal{D}}\left[\ell(Y, h(X))\right]$. Let $\mathrm{er}^{\ell,*}_{\mathcal{D}} = \inf_{\widehat{h}:\mathcal{X}\to\mathcal{Y}} \mathbf{E}_{(X,Y)\sim\mathcal{D}}[\ell(Y, \widehat{h}(X))]$ be the minimum possible error achievable by any classifier, i.e. the Bayes error. Then, we wish to design an algorithm that, given the random training sample S, generates a classifier $h_m : \mathcal{X} \to \mathcal{Y}$ such that $\mathrm{er}^{\ell}_{\mathcal{D}}[h_m] \to^{P} \mathrm{er}^{\ell,*}_{\mathcal{D}}$.

In other words, the algorithm should have the property:

$$\forall \epsilon > 0, \quad \mathbf{P}\left(\left|\mathrm{er}^{\ell}_{\mathcal{D}}[h_m] - \mathrm{er}^{\ell,*}_{\mathcal{D}}\right| > \epsilon\right) \to 0 \quad (m \to \infty)$$

Such algorithms are defined as *consistent* with respect to ℓ. Unfortunately, it can be shown that finding such a classifier h directly is NP-Hard [10].

Instead, a common approach is to instead minimize a convex *surrogate loss* $\psi : \mathcal{Y} \times \mathcal{C} \to \mathbb{R}_+$, where $\mathcal{C} \subseteq \mathbb{R}^d$ (for some suitable number d) is the surrogate prediction space. Then, as opposed to learning h directly, one can instead perform some convex optimization task on the sample S to learn a function $f_m : \mathcal{X} \to \mathcal{C}$, as well as some suitable mapping $\text{pred} : \mathcal{C} \to \widehat{\mathcal{Y}}$ to convert from the surrogate prediction space to the original prediction space. We can similarly define a notion of algorithms consistent with respect to ψ, for which $\text{er}_{\mathcal{D}}^{\psi}[f_m] \to^P \text{er}_{\mathcal{D}}^{\psi,*}$. In order to determine whether minimization of the surrogate loss leads to useful minimization of the original loss, we define a notion of *calibration*. Because we are dealing specifically in a discrete space now, define $\Delta_n = \{\mathbf{p} \in \mathbb{R}_+^n \,|\, \sum_i \mathbf{p}_i = 1\}$ to be the probability simplex in n dimensions.

5.2 Review of Literature

We present several results from the multiclass classification literature. First, we provide results showing why specific loss functions may be "difficult" to construct algorithms for. Then, we present a general approach, *coarse probability estimation*, as a way to work around such "difficult" loss functions.

5.2.1 Convex Calibration Dimension

One goal in designing efficient algorithms for multiclass classification for specific losses would be to minimize the dimension d of the prediction space, so that there space of possible learned functions is smaller and thus the optimization problem becomes tractable. Ramaswamy and Agarwal [24] quantify exactly this with their definition of the *convex calibration dimension*:

Definition 21 (Convex Calibration Dimension [24])**.** Consider a loss $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$. The convex calibration dimension (CC Dimension) of the loss is:

$$\begin{aligned} \text{CCDim}(\ell) \;\triangleq\; & \min\{d \in \mathbb{Z}_+ : \exists \text{ a convex set } \mathcal{C} \subseteq \mathbb{R}^d \text{ and a convex surrogate loss} \\ & \psi : \mathcal{Y} \times \mathcal{C} \to \mathbb{R}_+ \text{ that is } \ell\text{-calibrated.}\}\,. \end{aligned}$$

Here, ℓ-calibrated is the same as in Definition 13. Using proper scoring rules, one can estimate conditional class probabilities for $n-1$ classes to derive a consistent classifier which shows $\text{CCDim}(\ell) \leq n-1$ for any loss on n classes [1, 24]. Many loss functions admit surrogate losses with significantly lower dimensions for prediction spaces, such as the ordinal regression loss, which uses a surrogate prediction space in one dimension as opposed to $n-1$ [23].

Of importance in this paper are lower bounds on the convex calibration dimension of specific losses. In particular, we note the following result for the 0-1 loss, $\ell_{\text{0-1}}(y, \widehat{y}) = \mathbf{1}(y \neq \widehat{y})$.

Theorem 11 (Convex calibration dimension of $\ell_{\text{0-1}}$ [24])**.** Consider the 0-1 loss $\ell_{\text{0-1}}(y, \widehat{y})$ defined on a label and prediction space containing n classes. Then,

$$\text{CCDim}(\ell_{\text{0-1}}) \geq n-1$$

This, along with the upper bound using conditional class probabilities implies $\text{CCDim}(\ell_{\text{0-1}}) = n-1$, which is not a very promising result for multiclass classification under the 0-1 loss if the number of classes is large.

5.2.2 Coarse Probability Estimation

In order to work around some of the issues that occur when the Convex Calibration Dimension of a loss function is high, we introduce the idea of coarse probability estimation. In particular, we learn a vector-valued property of quantiles, and then use this vector to approximate which class has highest probability. In particular, for an integer s, we elicit the vector valued property

$$\Gamma(\mathbf{p}) = Q_{1/s}(\mathbf{p}) \times \cdots \times Q_{(s-1)/s}(\mathbf{p}) \in \mathcal{I}^{s-1}\,.$$

This is demonstrated in the following example:

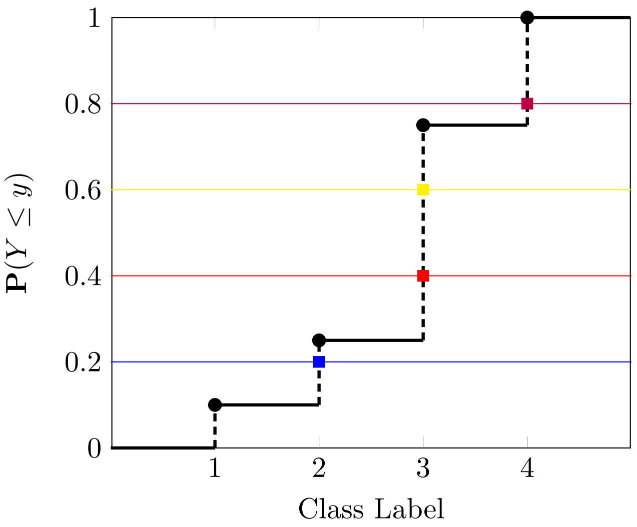

Figure 1: Quantiles overlain on cumulative distribution function of labels where $Y \sim \mathbf{p}$.

Example 7 (Coarse Probability Estimation from Quantiles)**.** Consider a probability distribution on $n = 4$ classes, $\mathbf{p} = [0.1, 0.15, 0.5, 0.25]^T \in \Delta_4$, and note the diagram of the CDF presented in Figure 1, in which the $\frac{1}{5}, \frac{2}{5}, \frac{3}{5}$ and $\frac{4}{5}$ quantiles are shown. From the diagram, we have

$$\left(Q_{\frac{1}{5}}(\mathbf{p}), Q_{\frac{2}{5}}(\mathbf{p}), Q_{\frac{3}{5}}(\mathbf{p}), Q_{\frac{4}{5}}(\mathbf{p})\right) = (2, 3, 3, 4)$$

From these data points, we can infer that if Y is drawn according to $\mathbf{p}$, then
$\mathbf{P}(Y = 1) \in [0, 0.2], \mathbf{P}(Y = 2) \in [0, 0.4], \mathbf{P}(Y = 3) \in [0.2, 0.6]$, and $\mathbf{P}(Y = 4) \in [0, 0.4]$.

5.3 Theoretical Results

In this section, we present several results detailing how quantile elicitation can be applied to multiclass classification. We begin by formalizing a framework that quantifies what we refer to as “approximate consistency,” and then go on to discuss applications to two known losses.

5.3.1 Approximately Consistent Loss Functions

We begin with defining *unavoidable* error, which quantifies the extent to which an approach can have error within an optimal range:

Definition 22 (Unavoidable Error)**.** Consider a loss function $\ell : [n] \times [k] \to \mathbb{R}_+$, a surrogate loss $\psi : [n] \times \mathcal{C} \to \mathbb{R}_+$ with a mapping $\text{pred} : \mathcal{C} \to [k]$. Consider an algorithm that learns functions f_m from a training sample S of m pairs (X_i, Y_i) drawn i.i.d. from some distribution $\mathcal{D}$. Then, we define the unavoidable error er_U^ℓ of the pair (ψ, pred) to be:

$$\begin{aligned}\text{er}_U^\ell(\psi, \text{pred}) = \inf \Big\{ \epsilon \mid \epsilon \geq 0 \text{ and } \forall \mathcal{D} \text{ over } \mathcal{X} \times \mathcal{Y},\ \exists \delta \leq \epsilon \ s.t. \\ \text{er}_{\mathcal{D}}^\psi[f_m] - \text{er}_{\mathcal{D}}^{\psi,*} \to^P 0 \quad \implies \quad \text{er}_{\mathcal{D}}^\ell[\text{pred} \circ f_m] - \text{er}_{\mathcal{D}}^{\ell,*} \to^P \delta \Big\}\,.\end{aligned}$$

Note that, if (ψ, pred) is (ℓ, Δ_n) calibrated, then $\text{er}_U^\ell(\psi, \text{pred}) = 0$ by Theorem 6. If $\text{er}_U^\ell(\psi, \text{pred}) > 0$, we call the pair (ψ, pred) an *approximate surrogate loss.* Moreover, to compare surrogate prediction spaces of different dimensions, we define:

Definition 23. d-Unavoidable Error Consider a loss function $\ell : [n] \times [k] \to \mathbb{R}_+$, and let $\Psi(d)$ be the set of all pairs (ψ, pred) such that $\psi : [n] \times \mathcal{C} \to \mathbb{R}_+$ and $\text{pred} : \mathcal{C} \to [k]$ with $\mathcal{C} \subseteq \mathbb{R}^d$ Then, the d-unavoidable error $\text{er}_{U,d}^\ell$ of ℓ is:

$$\text{er}_{U,d}^\ell = \inf_{(\psi,\text{pred}) \in \Psi(d)} \text{er}_U^\ell(\psi, \text{pred})$$

This d-unavoidable error is simply the best possible unavoidable error of any surrogate loss with surrogate prediction space $\mathcal{C} \subseteq \mathbb{R}^d$. This leads us to the following chain of inequalities similar to MAP estimation:

Proposition 5 (Surrogate Loss "Hierarchy")**.** For any loss $\ell : [n] \times [k] \to \mathbb{R}_+$, we have:

$$\text{er}_{U,1}^\ell \geq \text{er}_{U,2}^\ell \geq \cdots \geq \text{er}_{U,n-1}^\ell = 0\,.$$

If ℓ has convex calibration dimension d', then:

$$\mathrm{er}^{\ell}_{U,d'} = 0\,.$$

Proof. First, note that the final equality comes directly from [25], as we can always use $n-1$ proper loss functions to generate class probability estimates to produce a (ℓ, Δ_n)-consistent surrogate loss.

Next, assume towards contradiction that for some integer $r \geq 1$, we have:

$$\mathrm{er}^{\ell}_{U,r+1} > \mathrm{er}^{\ell}_{U,r}\,.$$

In other words, $\mathrm{er}^{\ell}_{U,r+1} = \mathrm{er}^{\ell}_{U,r} + \delta$ for some value $\delta > 0$.

By the definition of infimum, we know that for any $\epsilon > 0$, there exists some surrogate loss pair (ψ, pred) such that $\mathrm{er}^{\ell}_{U,r} > \mathrm{er}^{\ell}_{U}(\psi, \mathrm{pred}) - \epsilon$. In particular, there is some pair (ψ, pred) for which this holds when $\epsilon = \frac{\delta}{2} < \delta$. Denote this pair $(\psi_{\delta/2}, \mathrm{pred}_{\delta/2})$.

Next, we append a "dummy dimension" on $(\psi_{\delta/2}, \mathrm{pred}_{\delta/2})$. Let $\psi_{\delta/2} : \mathcal{Y} \times \mathcal{C} \to \mathbb{R}_+$ and $\mathrm{pred}_{\delta/2} : \mathcal{C} \to [k]$. Then, we can define the following surrogate loss with $r+1$ dimensional surrogate space - let $u \in \mathcal{C} \subseteq \mathbb{R}^r$, $x \in \mathbb{R}$, so that (u, x) is an element in an $\mathbb{R}$ dimensional space:

$$\psi'_{\delta/2}(y, (u, x)) = \psi_{\delta/2}(u)$$
$$\mathrm{pred}'_{\delta/2}((u, x)) = \mathrm{pred}_{\delta/2}(u)$$

Since x is not used by these functions, the optimal classifier from the pair $(\psi'_{\delta/2}, \mathrm{pred}'_{\delta/2})$ behaves exactly as that of the pair $(\psi_{\delta/2}, \mathrm{pred}_{\delta/2})$. However, this implies there exists a surrogate loss with an $r+1$-dimensional surrogate space which has unavoidable error $\mathrm{er}^{\ell}_{U,r} + \delta/2 < \mathrm{er}^{\ell}_{U,r} + \delta$, contradicting the assumption that $\mathrm{er}^{\ell}_{U,r+1} = \mathrm{er}^{\ell}_{U,r} + \delta$.

Finally, we show that if ℓ has convex calibration dimension d', then:

$$\mathrm{er}^{\ell}_{U,d'} = 0$$

Because, by definition of convex calibration dimension, there exists a surrogate loss function with a d'-dimensional surrogate space that is consistent with ℓ, and thus this surrogate loss function will have unavoidable error 0, the lowest possible unavoidable error. □

Thus, we have quantified a relaxation hierarchy for surrogate risk minimization analogous to that of linear programming hierarchies in MAP estimation. In the next section, we provide upper bounds on the d-unavoidable error of the 0-1 loss.

5.3.2 Multiclass Classification with the 0-1 Loss

We present a variation on the quantile elicitation ideas presented by Agarwal and Agarwal [1]. Using $\psi_{\phi,\alpha}$ and pred_w, we may elicit the α-quantile(s) of the conditional distribution of labels given an instance. We may then let $\boldsymbol{\psi}_{c,\phi}$ be the loss which elicits the quantiles $\frac{1}{c}, \frac{2}{c}, \ldots, \frac{c-1}{c}$. Moreover, once the values $\mathbf{u} = (u_1, \ldots, u_{c-1})$, which represent the values of the quantiles, are learned, we define $N_y(\mathbf{u})$ as the number of times that the value y appears in $\mathbf{u}$. Then, our function pred_c is as follows:

$$\mathrm{pred}_c(\mathbf{u}) = \mathrm{argmax}_{y' \in [n]} N_{y'}(\mathbf{u})$$

With this algorithm, we can show the following result:

Theorem 12. Let $\ell_{\text{0-1}}$ be the 0-1 multiclass loss, $\ell_{\text{0-1}}(y, \widehat{y}) = \mathbf{1}(y \neq \widehat{y})$. Then,

$$\mathrm{er}^{\ell_{\text{0-1}}}_{U,c-1} \leq \frac{2}{c} \, .$$

Proof. We must simply show that a Bayes optimal classifier with respect to $\boldsymbol{\psi}_c$ has regret at most $\frac{2}{c}$ with respect to $\ell_{\text{0-1}}$. That is, if we let f^* be a Bayes optimal classifier with respect to $\boldsymbol{\psi}_c$, we wish to show that for any distribution $\mathcal{D}$, we have $\text{er}_{\mathcal{D}}^{\ell_{\text{0-1}}}[\text{pred}_c \circ f^*] - \text{er}_{\mathcal{D}}^{\ell_{\text{0-1}},*} \leq \frac{2}{c}$. Additionally, let h^* be a Bayes optimal classifier with respect to $\ell_{\text{0-1}}$.

Next, we note:

$$\begin{aligned}\text{er}_{\mathcal{D}}^{\ell_{\text{0-1}}}[\text{pred}_c \circ f^*] - \text{er}_{\mathcal{D}}^{\ell_{\text{0-1}},*} &= \mathbf{E}_{(X,Y)\sim\mathcal{D}}\left[\ell_{\text{0-1}}(Y, \text{pred}_c(f^*(X))) - \ell_{\text{0-1}}(Y, h^*(X))\right] \\ &= \mathbf{E}_X\left[\mathbf{E}_{Y|X=x}\left[\ell_{\text{0-1}}(Y, \text{pred}_c(f^*(x))) - \ell_{\text{0-1}}(Y, h^*(x))\right]\right] .\end{aligned}$$

Hence, it is sufficient to show $\mathbf{E}_{Y|X=x}\left[\ell_{\text{0-1}}(Y, \text{pred}_c(f^*(x))) - \ell_{\text{0-1}}(Y, h^*(x))\right] \leq \frac{2}{c}\left(\frac{n-1}{n}\right)$, as $\text{er}_{\mathcal{D}}^{\ell_{\text{0-1}}}[\text{pred}_c \circ f^*] - \text{er}_{\mathcal{D}}^{\ell_{\text{0-1}},*}$ is a weighted sum of the inner term (which is also known as the inner risk).

Thus, more generally, since Y is a random variable taking values 1 through n, if we show that for all $\mathbf{p} \in \Delta_n$,[7] if $\mathbf{u}_c$ represents the quantiles $\frac{1}{c}, \ldots, \frac{c-1}{c}$, and $y^* = \text{argmax}_{y\in[n]}\, p_y$, we have $\mathbf{E}_{Y\sim\mathbf{p}}\left[\ell_{\text{0-1}}(Y, \text{pred}_c(\mathbf{u})) - \ell_{\text{0-1}}(Y, y^*)\right] \leq \frac{2}{c}$, we will have shown the claim, as the conditional distribution of Y given $x \in \mathcal{X}$ is will lie in Δ_n, regardless of the choice of $\mathcal{D}$, and the functions f^* and h^* will output $\mathbf{u}_c$ and y^* by definition as Bayes optimal classifiers [1].

Case 1: $\mathbf{p} \in \{\mathbf{q} \in \Delta_n \mid \exists y \textit{ s.t. } q_y > q_{y'} + \frac{2}{c}\ \forall y' \neq y\}$

> This case has actually been shown by Agarwal and Agarwal [1], as we know that quantile elicitation method to be calibrated with respect to $\ell_{\text{0-1}}$ under this "low-noise" condition, and our randomized pred_c classifier will always only output 1 class as having the maximum probability, making the two algorithms equivalent in this case. Thus,
>
> $$\begin{aligned}\mathbf{E}_{Y\sim\mathbf{p}}\left[\ell_{\text{0-1}}(Y, \text{pred}_c(\mathbf{u}_c)) - \ell_{\text{0-1}}(Y, y^*)\right] &= \mathbf{E}_{Y\sim\mathbf{p}}\left[\ell_{\text{0-1}}(Y, y^*) - \ell_{\text{0-1}}(Y, \widehat{y})\right] \\ &= 0 \leq \frac{2}{c}\,,\end{aligned}$$

[7] For notation, if $\mathbf{p} \in \Delta_n$, then let p_y be the probability of the outcome y.

as desired.

Case 2: $\mathbf{p} \notin \{\mathbf{q} \in \Delta_n \mid \exists y \; s.t. \; q_y > q_{y'} + \frac{2}{c} \; \forall y' \neq y\}$

There are two subcases here:

Case 2a: $\mathbf{p} \in \operatorname{relint}(\Delta_n)$

Let us consider the set $A = \operatorname{argmax}_{y' \in [n]} N_{y'}(\mathbf{u})$. Consider the value p_{y^*}, as well as the set $B = \{y | p_y > p_{y^*} - \frac{2}{c}\}$. We claim that $A \subseteq B$. Suppose not, and there exists some class y' such that $p_y < p_{y^*} - \frac{2}{c}$. Then, it must be the case $N_{y'}(\mathbf{u}) \leq N_y(\mathbf{u}) - 1$, as we measure quantiles in intervals of $\frac{1}{c}$. Thus, the algorithm will have the following regret:

$$\begin{aligned}
\mathbf{E}_{Y \sim \mathbf{p}} \left[\ell_{0\text{-}1}(Y, \operatorname{pred}_c(\mathbf{u}_c)) - \ell_{0\text{-}1}(Y, \widehat{y})\right] &= \mathbf{E}_{Y \sim \mathbf{p},\ \operatorname{pred}_c} \left[\ell_{0\text{-}1}(Y, \operatorname{pred}_c(\mathbf{u}_c))\right] \\
&\quad - (1 - p_{y^*}) \\
&\leq \left(1 - \left(p_{y^*} - \frac{2}{c}\right)\right) - (1 - p_{y^*}) \\
&= \frac{2}{c},
\end{aligned}$$

as desired.

Case 2b: $\mathbf{p} \notin \operatorname{relint}(\Delta_n)$

If $\max N_y(\mathbf{u}) \geq 2$, then the argument is the same as that above, as no class with probability 0 will be seen in more than one quantile. Otherwise, if $\max N_y(\mathbf{u}) = 1$, it is possible that we select a class with 0 probability. However, if $\max N_y(\mathbf{u}) = 1$, then $p_{y^*} \leq \frac{2}{c}$, and so no matter what class is selected, we will have:

$$\begin{aligned}\mathbf{E}_{Y\sim\mathbf{p}}\left[\ell_{0\text{-}1}(Y,\mathrm{pred}_c(\mathbf{u}_c))-\ell_{0\text{-}1}(Y,\widehat{y})\right] &\le \mathbf{E}_{Y\sim\mathbf{p},\ \mathrm{pred}_c}\left[\ell_{0\text{-}1}(Y,\mathrm{pred}_c(\mathbf{u}_c))\right]-\left(1-\frac{2}{c}\right)\\ &\le (1-0)-\left(1-\frac{2}{c}\right)\\ &=\frac{2}{c}\,,\end{aligned}$$

as desired.

Thus, for all possible conditional distributions of the label, we have

$$\mathbf{E}_{Y\sim\mathbf{p}}\left[\ell_{0\text{-}1}(Y,\mathrm{pred}_c(\mathbf{u}_c))-\ell_{0\text{-}1}(Y,\widehat{y})\right]\le\frac{2}{c}\,,$$

showing that $\mathrm{er}_U^{\ell_{0\text{-}1}}(\boldsymbol{\psi}_c,\mathrm{pred}_c)\le\frac{2}{c}$, and therefore

$$\mathrm{er}_{U,c-1}^{\ell_{0\text{-}1}}\le\mathrm{er}_U^{\ell_{0\text{-}1}}(\boldsymbol{\psi}_c,\mathrm{pred}_c)\le\frac{2}{c}\,,$$

as desired. □

In particular, if we let $s=\lceil\lg n\rceil$ as in [1], then this error is goes to 0 as n increases, which is desirable in the case of multiclass classification when n is large.

5.3.3 Multiclass Classification with a Reject Option

Now, we turn our attention to the loss $\ell^\beta:[n]\times[n+1]$, defined as follows:

$$\ell^\beta(y,\widehat{y})=\begin{cases}\mathbf{1}(y=\widehat{y}) & \widehat{y}\in[n]\\ \beta & \widehat{y}=n+1\end{cases}\,.$$

This loss function has been studied, particularly in the cases where $\beta\le\frac{1}{2}$, for which surrogate losses with prediction space in $O(\lg n)$ dimensions are known [26]. In particular, it is noted in Ramaswamy et al [26] that classification when $\beta > \frac{1}{2}$ is thought to be in some sense hard due to the form of the optimal classifier. Thus, it is a natural loss for us to try to develop an approximate surrogate loss.

Let us use $\boldsymbol{\psi}_{2\lceil \frac{1}{1-\beta} \rceil}$ to elicit quantiles $\frac{1}{2\lceil \frac{1}{1-\beta} \rceil}, \ldots, \frac{2\lceil \frac{1}{1-\beta} \rceil - 1}{2\lceil \frac{1}{1-\beta} \rceil}$. Moreover, let us use the following function $\text{pred}_\beta(\mathbf{u})$:

$$\text{pred}_\beta(\mathbf{u}) = \begin{cases} y \in \text{argmax}_{y' \in [n]} N_{y'}(\mathbf{u}) & \max_{y' \in [n]} N_{y'}(\mathbf{u}) \geq 2 \\ n+1 & \forall y' \in [n], N_{y'}(\mathbf{u}) \leq 1 \end{cases} .$$

Then, we have the following:

Theorem 13. For the multiclass loss with a reject option $\beta \geq \frac{1}{2}$,

$$\text{er}^{\ell^\beta}_{U, 2\lceil \frac{1}{1-\beta} \rceil - 1} \leq \frac{1}{\lceil \frac{1}{1-\beta} \rceil} .$$

Proof. Here, we wish to show that

$$\text{er}^{\ell^\beta}_{U, 2\lceil \frac{1}{1-\beta} \rceil - 1} \leq \frac{1}{\lceil \frac{1}{1-\beta} \rceil} .$$

For simplicity, let us write $c = 2\lceil \frac{1}{1-\beta} \rceil$, so that we can write the problem statement as:

$$\text{er}^{\ell^\beta}_{U, c-1} \leq \frac{2}{c} .$$

This is similar to proposition 3, and as such we approach the problem in a similar fashion.

As was done before, it is sufficient to show

$$\mathbf{E}_{Y \sim \mathbf{p}} \left[\ell^\beta(Y, \text{pred}_c(\mathbf{u}_c)) - \ell_\beta(Y, y^*) \right] \leq \frac{2}{c} .$$

Where $\mathbf{u}_c$ are again the $\frac{1}{c}, \ldots, \frac{c-1}{c}$ quantiles, and now we let

$$y^* = \begin{cases} \text{argmax}_{y' \in [n]} p_{y'} & \max_{y' \in [n]} p_{y'} > 1 - \beta \\ n+1 & o.w. \end{cases} .$$

We observe to following cases:

Case 1: $\mathbf{p} \in \{\mathbf{q} \in \Delta_n \mid \max_{y'} p_{y'} \notin [1 - \beta - \frac{1}{c}, 1 - \beta + \frac{1}{c}]\}$

Note that $c = 2\lceil \frac{1}{1-\beta} \rceil$, and so if the above case occurs, then if the maximum class probability lies below $1-\beta-\frac{1}{c} \leq \frac{1}{2(1-\beta)}$, it must be the case that $N_y(\mathbf{u}) \leq 1$ for all i, and therefore $\text{pred}_c(\mathbf{u}) = n+1$, which is optimal resulting in an inner risk of 0.

On the other hand, if $\max_{y'} p_{y'} \geq \frac{3}{2(1-\beta)}$, it must be the case that $\max_y N_y(\mathbf{u}) \geq 2$, as at least one class must cross two quantiles. Because it is possible that more than one quantile has probability at least $\frac{3}{2(1-\beta)}$, we see that, because the reject option is automatically not considered, all of the remaining cases for class probabilities are encompassed in the 0-1 loss proof of proposition 3. As such, the proofs are omitted.

Case 2: $\mathbf{p} \notin \{\mathbf{q} \in \Delta_n \mid \max_{y'} p_{y'} \notin [1 - \beta - \frac{1}{c}, 1 - \beta + \frac{1}{c}]\}$

In this case, $\max_{y'} p_{y'} \in [1 - \beta - \frac{1}{c}, 1 - \beta + \frac{1}{c}]$, and there are two cases: If $N_y(\mathbf{u}) \leq 1$ for all y, then the algorithm will choose $\text{pred}_c(\mathbf{u}) = n+1$, which has:

$$\begin{aligned} \mathbf{E}_{Y \sim \mathbf{p}} \left[\ell^\beta(Y, \text{pred}_c(\mathbf{u}_c)) - \ell^\beta(Y, \widehat{y}) \right] &= \mathbf{E}_{Y \sim \mathbf{p}} \left[\ell^\beta(Y, n+1) - \ell^\beta(Y, \widehat{y}) \right] \\ &\leq 1 - \beta - \left(1 - \beta - \frac{1}{c}\right) \qquad = \frac{1}{c} . \end{aligned}$$

Otherwise, if $N_y(\mathbf{u}) \geq 2$ for some y (at least one of which is selected by the algorithm), it must be the case that $p_y \geq \frac{1}{c}$, and so the regret is at most $\frac{2}{c}$, as we have:

$$\begin{aligned}\mathbf{E}_{Y\sim\mathbf{p}}\left[\ell^{\beta}(Y,\mathrm{pred}_c(\mathbf{u}_c))-\ell^{\beta}(Y,\widehat{y})\right] &\leq \left(1-\frac{1}{c}\right)-\mathbf{E}_{Y\sim\mathbf{p}}\left[\ell^{\beta}(Y,\widehat{y})\right]\\ &\leq \left(1-\beta+\frac{1}{c}\right)-\left(1-\beta-\frac{1}{c}\right) \quad =\frac{2}{c}\,.\end{aligned}$$

Thus, in all cases, the inner risk is at most $\frac{2}{c}$, implying that for the 0-1 loss with a reject option, the unavoidable risk is at most $\frac{2}{c}$. □

Through this application of quantiles, we present a constant dimensional surrogate for the 0-1 multiclass loss with a reject option - however, the unavoidable error is potentially quite high depending on the value of β, and again we see a tradeoff between accuracy and computational complexity in terms of the number of quantiles needed and the unavoidable error bound. Note that, this problem is only interesting if $\beta < \frac{n-1}{n}$, as otherwise the reject option is never optimal to use - this realization bounds the computational complexity of the task as needing to learn $O(n)$ quantiles. Moreover, the unavoidable error bound is not different than that of the original 0-1 loss, so we are not fully utilizing the effect of the reject option (although, as can be seen in the proof, the reject option is still partially utilized in the analysis and can be developed into a tighter bound).

More generally, as a framework to generate quantile based approximation surrogate losses to other loss functions, one can first determine at what probabilities an optimal classifier would need to threshold, and then use the coarse probability estimates to observe this error within a margin. The margin of error can then be used to provide an upper bound on the unavoidable error.

5.4 Experimental Results

We now present some experimental results regarding the effectiveness of quantile elicitation methods on multiclass classification.

5.4.1 Synthetic Experiments

Using code derived from Pedregosa et al. [23], we were able to run multiple experiments on synthetic data sets, showing the effectiveness of quantile regression for multiclass classification.

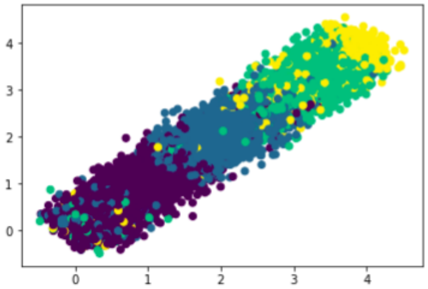

Figure 2: Visualization of Synthetic Data, $\alpha = 2$, $n = 4$

We performed the task of multiclass classification on two types of synthetic data sets with varying number of classes. For each data set, the instance $\mathcal{X} = \mathbb{R}^2$ was generated on the line $y = x$, with Gaussian noise added. Additionally, each class $y \in [n]$ was assigned a point on the line c_y, so that the conditional distribution was defined according to $\mathbf{P}(Y = y|x) \propto (||x - c_y||_2)^{-\alpha}$. For one type of data set, we set $\alpha = \frac{1}{2}$ resulting in very noisy data, and for the other type of data set, for example as shown in Figure 3, we set $\alpha = 2$.

To learn quantiles, we minimized $\psi_{\phi,\alpha}$ using ϕ as the logistic loss, and minimized over step functions with $n - 1$ steps - with this approach, the integral form of $\psi_{\phi,\alpha}$ becomes a sum which is tractable for optimization using modified code from [23]. In each trial, we varied the number of classes in the data set, learning $3\lceil \lg_2 n \rceil - 1$ quantiles over $10,000$ data points and testing on another 10,000 data points generated with the same distribution independently. For comparison, we used the sklearn implementation of logistic regression, which also uses linear classifiers. Our results are summarized in the following table:

Data Set	Algorithm	$n = 4$	$n = 8$	$n = 16$	$n = 32$	$n = 64$
$\alpha = 0.5$	Quantile Elicitation	0.6337	0.7731	0.9076	0.9515	0.9738
$\alpha = 0.5$	Logistic Regression	0.5896	0.7595	0.8746	0.9399	0.9713
$\alpha = 2$	Quantile Elicitation	0.2014	0.2404	0.3697	0.5669	0.7589
$\alpha = 2$	Logistic Regression	0.1963	0.2691	0.4588	0.7687	0.9024

Table 2: 0-1 Loss of Algorithms on Synthetic Multiclass Data Sets, varying number of classes n

We see that, for the noisy data set, the performance of the algorithms is comparable, with both having fairly poor accuracy. In contrast, the quantile elicitation algorithm outperforms logistic regression in the less noisy data set.

5.4.2 Experiments on Real World Datasets

We also used the quantile elicitation method of section 5.3.2 to provide experimental results on real world data. In particular, we run the method on the MNIST dataset and the CIFAR-10 dataset. In each dataset, the features provided are that of the image, and the classification task is to predict one of 10 classes.

For the MNIST dataset, we trained two neural networks. The first network had two convolutional layers followed by two linear layers, with an output layer with 5 nodes, each measuring a quantile in multiples of 1/6. This neural network was trained with the loss on each node as the α-pinball loss for the corresponding quantile. We then compared the result of this neural network to that of a second neural network with the same structure, but with a softmax layer as the output layer instead of only 5 nodes. Each neural network was trained This allows us to compare the approximation of our method to a close method that is commonly used in multiclass classification and has a similar function class.

Similarly, for the CIFAR-10 dataset, we again trained three neural networks. Again, the first network had two convolutional layers followed by two linear layers, with an output layer with 5 nodes, each measuring a quantile in multiples of 1/6 using the α-pinball loss. This was again compared to a second neural network with the same structure but a softmax

Dataset	Algorithm	0-1 Loss
MNIST	Quantile Elicitation	0.1597
MNIST	Softmax Classification	0.0135
CIFAR10	Quantile Elicitation	0.5960
CIFAR10	Softmax Classification	0.4217

Table 3: 0-1 Loss for real world dataset experiments

layer as the final layer. The results of both experiments are summarized in Table 3.

In both cases, we see that the algorithm based on quantile elicitation has a higher loss than the softmax algorithm, as it is only “approximately” consistent.

6 Conclusions

In this thesis, we have developed and discussed applications of the infinite threshold scoring rule for quantile elicitation. This scoring rule allows for the development of a quantile elicitation algorithm simply by having a binary classification algorithm, and hence allows for the implementation of many new scoring rules for quantile elicitation. We show that the scoring rule can be used for conditional quantile elicitation, and provide experimental evidence showing that it can result in efficient algorithms particularly for median elicitation. We then provide an application in multiclass classification, where we are able to develop a theory of approximately consistent surrogate losses, and use quantile elicitation as a means to perform multiclass classification when the number of labels is large.

References

[1] Arpit Agarwal and Shivani Agarwal. On consistent surrogate risk minimization and property elicitation. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 4–22, Paris, France, 03–06 Jul 2015. PMLR.

[2] Peter L. Bartlett, Michael Jordan, and Jon McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[3] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex Analysis and Optimization.* Athena Scientific, 2003.

[4] David Buffoni, Clément Calauzènes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *International Conference on Machine Learning*, 2011.

[5] Anna Choromanska, Alekh Agarwal, and John Langford. Extreme multi class classification. *Advances in Neural Information Processing Systems*, 2013.

[6] Andreas Christmann and Ingo Steinwart. How svms can estimate quantiles and the median. In *Advances in Neural Information Processing Systems 20*, pages 305–312. 2008.

[7] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

[8] Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 2001.

[9] Thomas Dietterich and G Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[10] Vitaly Feldman, Venkatesan Guruswami, Prasad Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *CoRR*, abs/1012.0729, 2010.

[11] Rafael M. Frongillo, Ian A. Kash, and Stephen Becker. Open problem: Property elicitation and elicitation complexity. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 1655–1658, 2016.

[12] Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, 2011.

[13] Tilmann Gneiting. Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27(2):197–207, 2011.

[14] Yacine Jernite, Anna Choromanska, and David Sontag. Simultaneous learning of trees and representations for extreme classification and density estimation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1665–1674, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[15] Roger Koenker. *Quantile Regression.* Econometric Society Monographs. Cambridge University Press, 2005.

[16] Roger Koenker and Kevin Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001.

[17] Nicolas Lambert and Yoav Shoham. Eliciting truthful answers to multiple-choice questions. In *ACM Conference on Electronic Commerce*, 2009.

[18] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings 9th ACM Conference on Electronic Commerce*, 2008.

[19] John Langford, Roberto Oliveira, and Bianca Zadrozny. Predicting conditional quantiles via reduction to classification. *Uncertainty in Artificial Intelligence*, 06 2012.

[20] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

[21] John McCarthy. Measures of the value of information. *Proceedings of the National Academy of Sciences*, 42(9):654–655, 1956.

[22] Deirdre O'Brien, Maya Gupta, and Robert Gray. Cost-sensitive multi-class classification from probability estimates. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[23] Fabian Pedregosa, Francis Bach, and Alexandre Gramfort. On the consistency of ordinal regression methods. *The Journal of Machine Learning Research*, 18(1):1769–1803, 2017.

[24] Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *arXiv*, 1408.2764, August 2014.

[25] Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research*, 17:1–45, 2016.

[26] Harish G. Ramaswamy, Ambuj Tewari, and Shivani Agarwal. Consistent algorithms for multiclass classification with a reject option. *CoRR*, abs/1505.04137, 2015.

[27] Mark D. Reid and Robert C. Williamson. Composite binary losses. *Journal of Machine Learning Research*, 11:2387–2422, 2010.

[28] Saharon Rosset, Claudia Perlich, Bianca Zadrozny, Srujana Merugu, Sholom Weiss, and Rick Lawrence. Wallet estimation models. *Proceedings of the International Workshop on Customer Relationship Management: Data Mining Meets Marketing (CRM Workshop)*, 01 2005.

[29] Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.

[30] Mark J. Schervish, Joseph B. Kadane, and Teddy Seidenfeld. Characterization of proper and strictly proper scoring rules for quantiles. *Preprint, Carnegie Mellon University*, March 2012.

[31] Ingo Steinwart and Andreas Christmann. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17:211–225, 2011.

[32] Ingo Steinwart, Chloé Pasin, Robert Williamson, and Siyu Zhang. Elicitation and identification of properties. In *Proceedings of the 27th Annual Conference on Learning Theory*, 2014.

[33] V. Vapnik. Principles of risk minimization for learning theory. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 831–838. Morgan-Kaufmann, 1992.

[34] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–134, 2004.