JMLR: Workshop and Conference Proceedings vol 40:1–19, 2015

On Consistent Surrogate Risk Minimization and Property Elicitation

Arpit Agarwal ARPIT.AGARWAL@CSA.IISC.ERNET.IN
Shivani Agarwal SHIVANI@CSA.IISC.ERNET.IN
Department of Computer Science and Automation
Indian Institute of Science, Bangalore 560012, India

Abstract

Surrogate risk minimization is a popular framework for supervised learning; property elicitation is a widely studied area in probability forecasting, machine learning, statistics and economics. In this paper, we connect these two themes by showing that calibrated surrogate losses in supervised learning can essentially be viewed as eliciting or estimating certain properties of the underlying conditional label distribution that are sufficient to construct an optimal classifier under the target loss of interest. Our study helps to shed light on the design of convex calibrated surrogates. We also give a new framework for designing convex calibrated surrogates under low-noise conditions by eliciting properties that allow one to construct 'coarse' estimates of the underlying distribution.

Keywords: Surrogate risk minimization, convex calibrated surrogates, multiclass classification, property elicitation, proper scoring rules, proper losses.

1. Introduction

Surrogate risk minimization is one of the most popular algorithmic frameworks for supervised learning, and there has been much interest in the machine learning and learning theory community in recent years in designing convex calibrated surrogates for various multiclass learning problems, including 0-1 classification, subset ranking, multilabel classification and others, which lead to consistent learning algorithms (Bartlett et al., 2006; Zhang, 2004a,b; Tewari and Bartlett, 2007; Steinwart, 2007; Cossock and Zhang, 2008; Xia et al., 2008; Duchi et al., 2010; Buffoni et al., 2011; Ravikumar et al., 2011; Calauzènes et al., 2012; Lan et al., 2012; Ramaswamy and Agarwal, 2012; Ramaswamy et al., 2013). In this paper, we seek to understand at a fundamental level what properties of the underlying distribution these calibrated surrogates aim to estimate.

We turn to tools from the areas of property elicitation and proper scoring rules, which have a long history in the probability forecasting literature and have recently received renewed interest in the machine learning, statistics, and economics communities (Savage, 1971; Schervish, 1989; Gneiting and Raftery, 2007; Lambert et al., 2008; Lambert and Shoham, 2009; Vernet et al., 2011; Abernethy and Frongillo, 2012; Steinwart et al., 2014). In particular, we show that calibrated surrogates for a supervised learning problem can essentially be viewed as eliciting a property of the conditional label distribution that is sufficient to construct an optimal classifier for the given loss.

We connect the two themes of this paper by defining the notion of *calibrated properties* for any given loss, which are properties of the conditional distribution from which one can construct an optimal prediction under that loss. We show that any strictly proper scoring rule for a calibrated property forms a calibrated surrogate. We use this framework to study the design of convex calibrated surrogates using both linear and nonlinear properties. We show how the standardization functions studied by Buffoni et al. (2011) for subset ranking losses, as well as the general least-squares type surrogates

studied by Ramaswamy et al. (2013), effectively amount to estimating linear properties of the distribution. We then show how using nonlinear properties can allow for the design of lower-dimensional convex calibrated surrogates. One offshoot of our work is a new framework for studying low-noise conditions; we show that eliciting a vector of quantiles allows one to obtain interval estimates of the label probabilities, based on which one can construct calibrated surrogates under any such condition where such a coarse probability estimate suffices to find an optimal classifier.

Notation. For $n \in \mathbb{Z}_+$, denote $[n] = \{1, \dots, n\}$ and $\Delta_n = \{\mathbf{p} \in \mathbb{R}^n_+ : \sum_{i=1}^n p_i = 1\}$. Denote by $\mathcal{S}_n$ the set of permutations on n objects. For $\mathbf{u} \in \mathbb{R}^n$, denote $\operatorname{argsort}(\mathbf{u}) = \{\sigma \in \mathcal{S}_n : \mathbf{u}_i > \mathbf{u}_j \implies \sigma(i) < \sigma(j)\,,\ \forall i, j \in [n]\}$. For a set $A \subseteq \mathbb{R}^n$, denote by $\operatorname{relint}(A)$ the relative interior of A, by $\operatorname{bndry}(A)$ the boundary of A, and by $\dim(A)$ the dimension of the affine extension of A. For a matrix $\mathbf{L} \in \mathbb{R}^{n \times k}$, denote by $\operatorname{col}(\mathbf{L})$ the column-space of $\mathbf{L}$, and by $\operatorname{affdim}(\mathbf{L})$ the affine dimension of the set of columns of $\mathbf{L}$. For a strictly convex function $\phi : \mathbb{R}^n \to \mathbb{R}$, denote by B_ϕ the Bregman divergence with respect to ϕ, defined as $B_\phi(\mathbf{u}_1, \mathbf{u}_2) = \phi(\mathbf{u}_1) - \phi(\mathbf{u}_2) - \partial\phi^\top_{\mathbf{u}_2}(\mathbf{u}_1 - \mathbf{u}_2)$ where $\partial\phi_{\mathbf{u}_2}$ denotes a subderivative of ϕ at $\mathbf{u}_2$.

2. Preliminaries and Background

We give background and set up terminology related to surrogate risk minimization in Section 2.1 and property elicitation in Section 2.2; the rest of the paper will then connect these two themes.

2.1. Surrogate Risk Minimization and Calibrated Surrogates

We consider supervised learning problems with instance space $\mathcal{X}$, finite label space $\mathcal{Y} = [n]$, and finite prediction space $\widehat{\mathcal{Y}} = [k]$ (often $\widehat{\mathcal{Y}} = \mathcal{Y}$, but this need not always be the case). Given training examples $(X_1, Y_1), \dots, (X_m, Y_m)$ drawn i.i.d. from some underlying distribution D on $\mathcal{X} \times [n]$, the goal is to learn a function $h : \mathcal{X} \to [k]$ with good performance according to some *loss function* $\ell : [n] \times [k] \to \mathbb{R}_+$, or equivalently, according to some *loss matrix* $\mathbf{L} \in \mathbb{R}^{n \times k}_+$ (we will use these two notions interchangeably, with the understanding that $L_{yt} = \ell(y, t)\ \forall y \in [n], t \in [k]$). In particular, the goal is to learn a function h with small *ℓ-generalization error* w.r.t. D, defined as $\operatorname{er}^\ell_D[h] = \mathbf{E}_{(X,Y)\sim D}[\ell(Y, h(X))]$; an algorithm that given m random examples learns a (random) function h_m is *ℓ-consistent* w.r.t. D if $\operatorname{er}^\ell_D[h_m] \xrightarrow{P} \inf_{h:\mathcal{X}\to[k]} \operatorname{er}^\ell_D[h]$ (as $m \to \infty$). For any $x \in \mathcal{X}$, we will denote $p_y(x) = \mathbf{P}(Y = y | X = x)\ \forall y \in [n]$ (under D) and $\mathbf{p}(x) = (p_1(x), \dots, p_n(x))^\top$. For $\mathbf{p} \in \Delta_n$, we will find it convenient to define $\operatorname{Opt}(\ell, \mathbf{p}) = \operatorname{argmin}_{t\in[k]} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y, t)]$. Clearly, any classifier h that satisfies $h(x) \in \operatorname{Opt}(\ell, \mathbf{p}(x))\ \forall x \in \mathcal{X}$ achieves the optimal ℓ-error under D.

Surrogate risk minimization algorithms. Since minimizing the discrete loss ℓ directly is hard, a common algorithmic approach is to minimize a *surrogate loss* $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ for some suitable $d \in \mathbb{Z}_+$. In particular, one learns a function $\mathbf{f}_m : \mathcal{X} \to \mathbb{R}^d$ by solving

$$\min_{\mathbf{f}} \sum_{i=1}^m \psi(Y_i, \mathbf{f}(X_i))$$

over a suitably rich class of functions $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$; and then returns $h_m = \text{pred} \circ \mathbf{f}_m$ for some suitable mapping $\text{pred} : \mathbb{R}^d \to [k]$ (for example, for multiclass 0-1 classification, where $k = n$ and $\ell_{0\text{-}1}(y, t) = \mathbf{1}(t \neq y)$, many common algorithms such as those considered by Zhang (2004b) and Tewari and Bartlett (2007) learn a function $\mathbf{f}_m : \mathcal{X} \to \mathbb{R}^n$ and then return a classifier $h_m = \operatorname{argmax} \circ \mathbf{f}_m$). In practice, the surrogate ψ is often chosen to be convex in its second argument to enable efficient minimization. It is known that if the minimization is performed over a universal

function class (with suitable regularization), then the resulting algorithm is ψ-consistent w.r.t. D, i.e. that the *ψ-generalization error* of $\mathbf{f}_m$ w.r.t. D, defined for a function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ as $\mathrm{er}_D^\psi[\mathbf{f}] = \mathbf{E}_{(X,Y)\sim D}[\psi(Y, \mathbf{f}(X))]$, converges to the optimal: $\mathrm{er}_D^\psi[\mathbf{f}_m] \xrightarrow{P} \inf_{\mathbf{f}:\mathcal{X}\to\mathbb{R}^d} \mathrm{er}_D^\psi[\mathbf{f}]$. There has been much work over the last several years on understanding when ψ-consistency (of $\mathbf{f}_m$) also implies ℓ-consistency (of h_m), and how to design surrogates satisfying this property; in particular, this has led to the study of surrogates that are *calibrated* with respect to the target loss ℓ (Bartlett et al., 2006; Zhang, 2004a,b; Tewari and Bartlett, 2007; Steinwart, 2007; Ramaswamy and Agarwal, 2012).

Calibrated surrogates. A pair (ψ, pred) is said to be *ℓ-calibrated* over $\mathcal{P} \subseteq \Delta_n$ if

$$\forall \mathbf{p} \in \mathcal{P} : \quad \inf_{\mathbf{u}\in\mathbb{R}^d : \mathrm{pred}(\mathbf{u}) \notin \mathrm{Opt}(\ell,p)} \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\mathbf{u})] \; > \; \inf_{\mathbf{u}\in\mathbb{R}^d} \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\mathbf{u})] \,.$$

It is known that (ψ, pred) is ℓ-calibrated over $\mathcal{P}$ if and only if ψ-consistency (of $\mathbf{f}_m$) implies ℓ-consistency (of $h_m = \mathrm{pred} \circ \mathbf{f}_m$) for all distributions D for which $\mathbf{p}(x) \in \mathcal{P}\ \forall x$ (Bartlett et al., 2006; Zhang, 2004b; Tewari and Bartlett, 2007; Ramaswamy and Agarwal, 2012, 2015). Thus, given a target loss ℓ, in order to design a surrogate risk minimization algorithm that is ℓ-consistent w.r.t. some class of distributions D, one needs to design (ψ, pred) that is ℓ-calibrated over the corresponding set of conditional distributions $\mathcal{P}$. As noted above, one is often interested in *convex* calibrated surrogates, for which ψ is convex in its second argument, to enable efficient minimization.

2.2. Property Elicitation and Proper Scoring Rules/Losses

When the goal is to elicit a full distribution $\mathbf{p} \in \Delta_n$, it is well known that one can use a (strictly) proper scoring rule/loss. A scoring rule/loss in this context is a function $\psi : [n] \times \Delta_n \to \mathbb{R}_+$ that assigns a 'penalty' $\psi(y, \mathbf{p}')$ to an estimate/report $\mathbf{p}' \in \Delta_n$ when an outcome $y \in [n]$ is observed, and is said to be *proper* over $\mathcal{P} \subseteq \Delta_n$ if

$$\forall \mathbf{p} \in \mathcal{P} : \quad \mathbf{p} \;\in\; \mathrm{argmin}_{\mathbf{p}'\in\Delta_n} \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y,\mathbf{p}')] \,,$$

and *strictly proper* over $\mathcal{P}$ if the above minimizer is unique for all $\mathbf{p} \in \mathcal{P}$.[1] In probability forecasting and economics, where the goal is to elicit the distribution from an agent, the agent reports a distribution $\mathbf{p}'$, and on observing an outcome y drawn from the true distribution $\mathbf{p}$, receives a reward (or in our setting, incurs a loss) given by the scoring rule, namely $\psi(y, \mathbf{p}')$; a strictly proper scoring rule ensures that truthful reporting maximizes the agent's expected reward. In machine learning and statistics, where the goal is to estimate the distribution from random observations $y_1, \ldots, y_m$ sampled from $\mathbf{p}$, one estimates $\mathbf{p}'$ to minimize the average value of the scoring rule on the observed sample, $\frac{1}{m}\sum_{i=1}^m \psi(y_i, \mathbf{p}')$; here a strictly proper scoring rule yields a consistent estimator.

Proper (and strictly proper) scoring rules/losses for eliciting full probability distributions are fairly well characterized (Savage, 1971; Schervish, 1989; Gneiting and Raftery, 2007; Vernet et al., 2011). More recently, there has been much interest in understanding what types of scoring rules/losses can be used when the goal is to elicit not the full probability distribution $\mathbf{p}$, but rather some *property* of $\mathbf{p}$ of interest (Lambert et al., 2008; Lambert and Shoham, 2009; Abernethy and Frongillo, 2012; Steinwart et al., 2014; Frongillo and Kash, 2015).

Property of a distribution. In general, a property is any 'statistic' of a distribution. Formally, for $\mathcal{P} \subseteq \Delta_n$ and $d \in \mathbb{Z}_+$, we will define a (d-dimensional) *property* over $\mathcal{P}$ as any function $\Gamma : \mathcal{P} \to \mathbb{R}^d$

1. Note that we use the terms scoring rule and loss here interchangeably; in the literature, scoring rules usually assign a 'utility' to an estimate $\mathbf{p}'$ that needs to be maximized, while losses assign a 'penalty' that needs to be minimized. We will use the latter interpretation for both (in general, one can be obtained from the other simply by switching signs).

that maps each distribution $\mathbf{p} \in \mathcal{P}$ to a (d-dimensional) statistic $\Gamma(\mathbf{p}) \in \mathbb{R}^d$. One such example is the mean: $\Gamma(\mathbf{p}) = \mu(\mathbf{p}) = \mathbf{E}_{Y\sim\mathbf{p}}[Y]$. Other examples of one-dimensional properties include the median, generalized quantiles, and many others. An example of a d-dimensional property is the vector of the first d moments: $\Gamma(\mathbf{p}) = (\mu_1(\mathbf{p}), \ldots, \mu_d(\mathbf{p}))^\top$, where $\mu_i(\mathbf{p}) = \mathbf{E}_{Y\sim\mathbf{p}}[Y^i]\ \forall i \in [d]$; more generally, a d-dimensional property is any vector of d one-dimensional properties.

Proper scoring rules/losses for eliciting properties of a distribution. Clearly, a (strictly) proper scoring rule that elicits the full distribution can be used to elicit any property of the distribution. However, this involves estimating an $(n-1)$-dimensional property, which can be expensive for large n and may not always be necessary, We will define a *d-dimensional scoring rule/loss* as a function $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$, and will say it is *proper for a property* $\Gamma : \mathcal{P} \to \mathbb{R}^d$ if

$$\forall \mathbf{p} \in \mathcal{P}: \quad \Gamma(\mathbf{p}) \in \operatorname{argmin}_{\mathbf{u}\in\mathbb{R}^d} \mathbf{E}_{Y\sim\mathbf{p}}[\psi(Y, \mathbf{u})]\,,$$

and *strictly proper for* Γ if the above minimizer is unique for all $\mathbf{p} \in \mathcal{P}$. We will say a d-dimensional property $\Gamma : \mathcal{P} \to \mathbb{R}^d$ is *directly elicitable* if there exists a strictly proper d-dimensional scoring rule for Γ. Further, if for some $d' \geq d$, there is a directly elicitable d'-dimensional property $\Gamma' : \mathcal{P} \to \mathbb{R}^{d'}$ which can be used to recover Γ, i.e. for which there exists a mapping $\boldsymbol{\pi} : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that $\boldsymbol{\pi}(\Gamma'(\mathbf{p})) = \Gamma(\mathbf{p})\ \forall \mathbf{p} \in \mathcal{P}$, then we will say that Γ is *d'-elicitable*. Clearly, every property is $(n-1)$-elicitable, and a d-dimensional property that is directly elicitable is d-elicitable.

Linear properties. A class of properties that are relatively better understood are *linear* properties. Specifically, a property $\Gamma : \mathcal{P} \to \mathbb{R}^d$ is said to be *linear* if it can be written as a vector of expectations, i.e. if there exists a function $\boldsymbol{\rho} : [n] \to \mathbb{R}^d$ such that $\Gamma(\mathbf{p}) = \mathbf{E}_{Y\sim\mathbf{p}}[\boldsymbol{\rho}(Y)]\ \forall \mathbf{p} \in \mathcal{P}$. It is known that linear properties are directly elicitable; moreover, as shown by Abernethy and Frongillo (2012), all strictly proper scoring rules for a linear property have the form of a Bregman divergence:

Theorem 1 (Abernethy and Frongillo (2012)) *Let $\mathcal{P} \subseteq \Delta_n$ and $\boldsymbol{\rho} : [n] \to \mathbb{R}^d$, and let $\Gamma : \mathcal{P} \to \mathbb{R}^d$ be a linear property defined as $\Gamma(\mathbf{p}) = \mathbf{E}_{Y\sim\mathbf{p}}[\boldsymbol{\rho}(Y)]\ \forall \mathbf{p} \in \mathcal{P}$. Then a scoring rule $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ is strictly proper for Γ if and only if there is a strictly convex function $\phi : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\psi(y, \mathbf{u}) = B_\phi(\boldsymbol{\rho}(y), \mathbf{u}) \quad \forall y \in [n], \mathbf{u} \in \mathbb{R}^d\,.$$

3. Calibrated Properties

We now make a connection between the two main themes of this paper by defining the notion of a *calibrated property* for a given loss ℓ. As we will see, any strictly proper scoring rule for an ℓ-calibrated property will yield an ℓ-calibrated surrogate loss.

Specifically, recall that given a loss $\ell : [n] \times [k] \to \mathbb{R}_+$, the goal is to learn a classifier that approaches the optimal ℓ-error under D, and that this is achieved by classifying according to $h(x) \in \mathrm{Opt}(\ell, \mathbf{p}(x))$ for all x. This means that for any $\mathbf{p} \in \Delta_n$ (or more generally, $\mathbf{p} \in \mathcal{P}$ for some suitable $\mathcal{P} \subseteq \Delta_n$), one is simply interested in finding an ℓ-optimal prediction $t^*(\mathbf{p}) \in [k]$, i.e. any $t^*(\mathbf{p})$ that satisfies $t^*(\mathbf{p}) \in \operatorname{argmin}_{t\in[k]} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y, t)]$. While we could consider the property $t^*(\mathbf{p})$ directly, this is a discrete-valued property that is generally hard to estimate directly.[2] Instead, we

2. Note that in the probability forecasting/mechanism design setting, where there is an agent who holds information about the probability distribution and the goal is to elicit this information from him by assigning a suitable reward/loss using a scoring rule, eliciting a discrete-valued property poses no problem. However in the learning/statistics setting that we consider here, where one gets random observations from the underlying distribution and the goal is to estimate the property of interest from these observations by minimizing/maximizing a scoring rule, a discrete-valued property leads to a discrete optimization problem that in general can be hard.

will consider properties $\Gamma : \mathcal{P} \to \mathbb{R}^d$ that map $\mathbf{p} \in \mathcal{P}$ to a real number or vector $\Gamma(\mathbf{p}) \in \mathbb{R}^d$ from which one can *recover* an ℓ-optimal prediction $t^*(\mathbf{p}) \in [k]$ using a suitable mapping $\text{pred} : \mathbb{R}^d \to [k]$; we will refer to such properties as *ℓ-calibrated* properties:

Definition 2 (ℓ-calibrated property) *Let $\mathcal{P} \subseteq \Delta_n$, $\Gamma : \mathcal{P} \to \mathbb{R}^d$, and $\text{pred} : \mathbb{R}^d \to [k]$. We will say (Γ, pred) is* ℓ-calibrated *over $\mathcal{P}$ if for all $\mathbf{p} \in \mathcal{P}$ and all sequences $\{\mathbf{u}_m\}$ in $\mathbb{R}^d$,*

$$\mathbf{u}_m \to \Gamma(\mathbf{p}) \implies \mathbf{E}_{Y \sim \mathbf{p}}[\ell(Y, \text{pred}(\mathbf{u}_m)] \to \min_{t \in [k]} \mathbf{E}_{Y \sim \mathbf{p}}[\ell(Y, t)] \,.$$

Note in particular this implies that if (Γ, pred) is ℓ-calibrated over $\mathcal{P}$, then we have that for all $\mathbf{p} \in \mathcal{P}$, $\text{pred}(\Gamma(\mathbf{p})) \in \text{Opt}(\ell, \mathbf{p})$. The sequence convergence condition is stronger and is needed in the proof of the following result, which tells us that the problem of designing an ℓ-calibrated surrogate loss in d dimensions can be reduced to finding an ℓ-calibrated property in d dimensions that is (directly) elicitable, together with any strictly proper scoring rule for it:

Theorem 3 (ℓ-calibrated surrogates via elicitable ℓ-calibrated properties) *Let $\ell : [n] \times [k] \to \mathbb{R}_+$ and $\mathcal{P} \subseteq \Delta_n$. Let $\Gamma : \mathcal{P} \to \mathbb{R}^d$ and $\text{pred} : \mathbb{R}^d \to [k]$ be such that Γ is directly elicitable and (Γ, pred) is ℓ-calibrated over $\mathcal{P}$. Let $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ be any strictly proper scoring rule for Γ. Then (ψ, pred) forms an ℓ-calibrated surrogate over $\mathcal{P}$.*

Proof Let $\mathbf{p} \in \mathcal{P}$. By strict properness of ψ for Γ, we have that $\Gamma(\mathbf{p})$ is the unique minimizer of $\mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})]$ over $\mathbf{u} \in \mathbb{R}^d$; for convenience, denote this unique minimizer by $\mathbf{u}^*$. Now, for each $t \in [k]$, define

$$\text{regret}^\ell_{\mathbf{p}}(t) := \mathbf{E}_{Y \sim \mathbf{p}}[\ell(Y, t)] - \min_{t \in [k]} \mathbf{E}_{Y \sim \mathbf{p}}[\ell(Y, t)] \,.$$

Since (Γ, pred) is ℓ-calibrated over $\mathcal{P}$, we have $\text{pred}(\mathbf{u}^*) = \text{pred}(\Gamma(\mathbf{p})) \in \text{Opt}(\ell, \mathbf{p})$, and therefore $\text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u}^*)) = 0$. Let

$$\epsilon = \min_{t \in [k] : \text{regret}^\ell_{\mathbf{p}}(t) > 0} \text{regret}^\ell_{\mathbf{p}}(t) \,.$$

Then we have

$$\begin{aligned} \inf_{\mathbf{u} \in \mathbb{R}^d : \text{pred}(\mathbf{u}) \notin \text{Opt}(\ell, \mathbf{p})} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] &= \inf_{\mathbf{u} \in \mathbb{R}^d : \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u})) \geq \epsilon} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] \\ &= \inf_{\mathbf{u} \in \mathbb{R}^d : \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u})) \geq \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u}^*)) + \epsilon} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] \,. \end{aligned}$$

Now, we claim that the mapping $\mathbf{u} \mapsto \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u}))$ is continuous at $\mathbf{u} = \mathbf{u}^*$. To see this, note that since (Γ, pred) is ℓ-calibrated over $\mathcal{P}$, for all sequences $\{\mathbf{u}_m\}$ in $\mathbb{R}^d$ such that $\mathbf{u}_m \to \mathbf{u}^*$, we have $\text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u}_m)) \to 0 = \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u}^*))$. In particular, this implies that $\exists \delta > 0$ such that

$$\|\mathbf{u} - \mathbf{u}^*\|_2 < \delta \implies \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u})) - \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u}^*)) < \epsilon \,.$$

This gives

$$\begin{aligned} \inf_{\mathbf{u} \in \mathbb{R}^d : \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u})) \geq \text{regret}^\ell_{\mathbf{p}}(\text{pred}(\mathbf{u}^*)) + \epsilon} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] &\geq \inf_{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u} - \mathbf{u}^*\|_2 \geq \delta} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] \\ &> \inf_{\mathbf{u} \in \mathbb{R}^d} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] \,. \end{aligned}$$

where the last inequality follows from the fact that $\mathbf{u}^*$ is the unique minimizer of $\mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})]$. Since $\mathbf{p} \in \mathcal{P}$ was arbitrary, the result follows. ■

As a simple example, it is easy to see that $(n-1)$-dimensional properties that preserve the full probability structure (also called 'link' functions) are ℓ-calibrated for any loss ℓ, and that the corresponding strictly proper rules lead to class probability estimation (CPE) algorithms that estimate the full conditional distribution $\mathbf{p}(x)$ (and are consistent for any loss ℓ):

Example 1 (Link functions and class probability estimation (CPE)) *Let $\boldsymbol{\lambda} : \Delta_n \to \mathbb{R}^{n-1}$ be a bijective mapping (sometimes called a multiclass 'link' function) with a continuous inverse $\boldsymbol{\lambda}^{-1}$. Then the property $\Gamma : \Delta_n \to \mathbb{R}^{n-1}$ defined as $\Gamma(\mathbf{p}) = \boldsymbol{\lambda}(\mathbf{p})$ is trivially ℓ-calibrated over Δ_n for any loss $\ell : [n] \times [k] \to \mathbb{R}_+$; to see this, take any mapping $\text{pred}_\ell : \mathbb{R}^{n-1} \to [k]$ that satisfies $\text{pred}_\ell(\mathbf{u}) \in \text{Opt}(\ell, \boldsymbol{\lambda}^{-1}(\mathbf{u}))$ $\forall \mathbf{u} \in \mathbb{R}^{n-1}$. This property is also trivially elicitable; indeed, this is the property effectively elicited by class probability estimation algorithms using a multiclass proper composite surrogate loss with link $\boldsymbol{\lambda}$ (Vernet et al., 2011).*

While estimating the full conditional distribution $\mathbf{p}(x)$ clearly yields consistent algorithms for any loss ℓ, this requires $n-1$ dimensions and is not always needed. Indeed, for many losses ℓ, finding an optimal classifier requires estimating only a restricted, lower-dimensional property of $\mathbf{p}(x)$. In such cases, one can use a strictly proper scoring rule for the corresponding property to design a calibrated surrogate loss operating in a smaller number of dimensions. We shall see several examples of such surrogates below. In particular, in Section 4 we shall see examples of calibrated surrogate losses that effectively elicit low-dimensional linear properties of $\mathbf{p}(x)$. In Section 5 we will consider how to exploit low-dimensional nonlinear calibrated properties. In both cases, we will be particularly interested in *convex* scoring rules that lead to convex calibrated surrogates.

4. Calibrated Surrogates via Calibrated Linear Properties

In this section we show that some recent works that have proposed general frameworks for obtaining convex calibrated surrogates effectively amount to using proper scoring rules for calibrated linear properties. In particular, we start by showing that the notion of 'standardization function' used to obtain calibrated surrogates for certain subset ranking losses (Buffoni et al., 2011) corresponds to a calibrated linear property (Section 4.1). We then show that the general framework described recently by Ramaswamy et al. (2013) for obtaining convex calibrated surrogates for any loss ℓ in $d = \text{affdim}(\mathbf{L})$ dimensions also amounts to using a calibrated linear property (Section 4.2). Finally, we show that for any loss ℓ, the number of dimensions d needed to construct an ℓ-calibrated linear property is fundamentally lower bounded by $\text{affdim}(\mathbf{L}) - 1$ (Section 4.3), making the construction of Ramaswamy et al. (2013) essentially unimprovable as far as linear properties are concerned.

4.1. Subset Ranking Losses and Standardization Functions

Subset ranking refers to ranking problems such as those that arise in information retrieval, where each instance $x \in \mathcal{X}$ consists of a query with say r associated documents, and a label $y \in \mathcal{Y}$ represents some 'preference' or 'relevance' information about these documents in relation to the query; for example a label could be a (possibly weighted) directed acyclic graph (DAG) on r nodes indicating which of the r documents are more relevant to the query than others ($\mathcal{Y} = \mathcal{G}_r$ for some finite set $\mathcal{G}_r$ of possibly weighted DAGs on r nodes, with $n = |\mathcal{G}_r|$), or simply a vector of r binary or multi-valued relevance judgments for the documents ($\mathcal{Y} = \{0,1\}^r$ with $n = 2^r$ or $\mathcal{Y} = [q]^r$ for some $q \in \mathbb{Z}_+$ with $n = q^r$). In most such settings, given a new query with r documents,

the goal is to rank the documents by relevance to the query, i.e. the prediction space is the set of permutations of r objects, $\widehat{\mathcal{Y}} = \mathcal{S}_r$ (thus $k = r!$). There has been much work in recent years on understanding how to design convex calibrated surrogates for various subset ranking losses used in practice, such as the (normalized) discounted cumulative gain ((N)DCG), pairwise disagreement (PD), mean average precision (MAP), etc (Cossock and Zhang, 2008; Xia et al., 2008; Duchi et al., 2010; Ravikumar et al., 2011; Buffoni et al., 2011; Calauzènes et al., 2012; Lan et al., 2012).

In particular, Buffoni et al. (2011) introduced the notion of 'standardization function', and showed that many previous results on calibrated surrogates for subset ranking could be explained through this notion. Specifically, let $\mathcal{Y}$ be one of the label spaces above and $\widehat{\mathcal{Y}} = \mathcal{S}_r$, and let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$ be any subset ranking loss. A *standardization function* for ℓ over $\mathcal{P} \subseteq \Delta_\mathcal{Y}$ is defined as any function $\mathbf{s} : \mathcal{Y} \to \mathbb{R}^r$ such that

$$\forall \mathbf{p} \in \mathcal{P} : \quad \operatorname{argsort}\left(\mathbf{E}_{Y\sim\mathbf{p}}[\mathbf{s}(Y)]\right) \subseteq \operatorname{argmin}_{\sigma\in\mathcal{S}_r} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,\sigma)]\,. \tag{1}$$

We show below that if such a function $\mathbf{s}$ exists, then the r-dimensional linear property $\Gamma : \mathcal{P} \to \mathbb{R}^r$ defined as $\Gamma(\mathbf{p}) = \mathbf{E}_{Y\sim\mathbf{p}}[\mathbf{s}(Y)]$ is ℓ-calibrated over $\mathcal{P}$ (see Appendix A for a proof):

Theorem 4 (Standardization functions yield calibrated linear properties) *Let $\ell : \mathcal{Y} \times \mathcal{S}_r \to \mathbb{R}_+$ be a subset ranking loss for some suitable $\mathcal{Y}$ as above, and let $\mathcal{P} \subseteq \Delta_\mathcal{Y}$. Let $\mathbf{s} : \mathcal{Y} \to \mathbb{R}^r$ be a standardization function for ℓ over $\mathcal{P}$. Let $\Gamma : \mathcal{P} \to \mathbb{R}^r$ be the linear property defined as*

$$\Gamma(\mathbf{p}) = \mathbf{E}_{Y\sim\mathbf{p}}[\mathbf{s}(Y)]\,,$$

and let $\operatorname{pred} : \mathbb{R}^r \to \mathcal{S}_r$ *be any mapping that satisfies* $\operatorname{pred}(\mathbf{u}) \in \operatorname{argsort}(\mathbf{u})\ \forall \mathbf{u} \in \mathbb{R}^r$. *Then* $(\Gamma, \operatorname{pred})$ *is ℓ-calibrated over $\mathcal{P}$.*

Thus, if a subset ranking loss ℓ has a standardization function over $\mathcal{P}$, then one can construct an r-dimensional convex calibrated surrogate for ℓ over $\mathcal{P}$ by constructing a convex strictly proper scoring rule for the calibrated linear property Γ above (e.g. by using $\phi(\mathbf{u}) = \frac{1}{2}\|\mathbf{u}\|_2^2$ in Theorem 1). Note that this is a huge savings over the naïve CPE approach of Example 1, which would use $|\mathcal{Y}| - 1$ dimensions (for most subset ranking settings, $|\mathcal{Y}|$ is exponential in r). The following example illustrates one application of the above result:

Example 2 (Discounted cumulative gain (DCG) loss for subset ranking) *The DCG loss for multi-valued relevance vector labels ($\mathcal{Y} = [q]^r$ for some $q \in \mathbb{Z}_+$), $\ell_{\text{DCG@}\tau} : [q]^r \times \mathcal{S}_r \to \mathbb{R}_+$ (where $\tau \in [r]$ is a cut-off value), is widely used in information retrieval and is defined as*

$$\ell_{\text{DCG@}\tau}(\mathbf{y}, \sigma) \;=\; Z - \sum_{i=1}^{\tau} \frac{2^{y_{\sigma^{-1}(i)}} - 1}{\log_2(i+1)} \quad \forall \mathbf{y} \in [q]^r, \sigma \in \mathcal{S}_r$$

for a suitable constant Z that ensures non-negativity of the loss. As shown by Buffoni et al. (2011), the function $\mathbf{s} : [q]^r \to \mathbb{R}^r$ defined as $s_i(\mathbf{y}) = 2^{y_{\sigma^{-1}(i)}} - 1\ \forall i \in [r]$ is a standardization function for $\ell_{\text{DCG@}\tau}$ over $\Delta_\mathcal{Y}$, and therefore it follows from Theorem 4 that any strictly proper scoring rule for the corresponding linear property $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^r$ given by $\Gamma_i(\mathbf{p}) = \mathbf{E}_{\mathbf{Y}\sim\mathbf{p}}[2^{Y_{\sigma^{-1}(i)}} - 1]\ \forall i \in [r], \mathbf{p} \in \Delta_\mathcal{Y}$ yields an $\ell_{\text{DCG@}\tau}$-calibrated surrogate over $\Delta_\mathcal{Y}$. In particular, using $\phi(\mathbf{u}) = \frac{1}{2}\|\mathbf{u}\|_2^2$ in Theorem 1, one gets the convex $\ell_{\text{DCG@}\tau}$-calibrated surrogate used by Cossock and Zhang (2008).

Another example of an application of Theorem 4 involves the weighted pairwise disagreement (WPD) loss for subset ranking (Duchi et al., 2010). In particular, Duchi et al. (2010) proposed a convex r-dimensional surrogate for subset ranking which they showed to be calibrated w.r.t. the WPD loss under a certain low-noise condition; this surrogate can also be viewed as a strictly proper scoring rule for a linear property, composed with a link function (see Appendix B for details).

4.2. Affdim(L)-Dimensional Surrogates of Ramaswamy et al. (2013)

Recently, Ramaswamy et al. (2013) gave a very general framework for constructing a convex calibrated surrogate (over the full simplex Δ_n) for any given loss $\ell : [n] \times [k] \to \mathbb{R}_+$ in $d = \text{affdim}(\mathbf{L})$ dimensions. In particular, they gave the following result:

Theorem 5 (Ramaswamy et al. (2013)) *Let $\ell : [n] \times [k] \to \mathbb{R}^k_+$ be such that $\mathbf{L} = \mathbf{AB} + c$ for some $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{B} \in \mathbb{R}^{d \times k}$, and $c \in \mathbb{R}$. Let $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ and $\text{pred} : \mathbb{R}^d \to [k]$ be defined as follows:*

$$\psi(y, \mathbf{u}) = \sum_{i=1}^d (u_i - A_{yi})^2 , \quad \text{pred}(\mathbf{u}) \in \text{argmin}_{t \in [k]} \sum_{i=1}^d B_{it} u_i .$$

Then (ψ, pred) is ℓ-calibrated over Δ_n.

The proof of the above result (Ramaswamy et al., 2013) can be re-interpreted as showing that the linear property $\Gamma : \Delta_n \to \mathbb{R}^d$ (where $d = \text{affdim}(\mathbf{L})$) given by $\Gamma_i(\mathbf{p}) = \mathbf{E}_{Y \sim \mathbf{p}}[A_{Yi}] \; \forall i \in [d]$ is ℓ-calibrated over Δ_n via the above mapping pred; the convex least-squares type surrogate loss ψ defined above is then simply the strictly proper scoring rule for this property resulting from using $\phi(\mathbf{u}) = \frac{1}{2}\|\mathbf{u}\|_2^2$ in Theorem 1. For completeness, we state this below and give a self-contained proof in Appendix C. Note also that this implies that any other strictly proper scoring rule for this linear property (such as those obtained by using Bregman divergences associated with other convex functions ϕ in Theorem 1) will also lead to an ℓ-calibrated surrogate over Δ_n.

Theorem 6 (Affdim(L)-dimensional calibrated linear properties) *Let $\ell : [n] \times [k] \to \mathbb{R}^k_+$ be such that $\mathbf{L} = \mathbf{AB} + c$ for some $\mathbf{A} \in \mathbb{R}^{n \times d}$, $\mathbf{B} \in \mathbb{R}^{d \times k}$, and $c \in \mathbb{R}$. Let $\Gamma : \Delta_n \to \mathbb{R}^d$ be the linear property defined as*

$$\Gamma_i(\mathbf{p}) = \mathbf{E}_{Y \sim \mathbf{p}}[A_{Yi}] \quad \forall i \in [d],$$

and let $\text{pred} : \mathbb{R}^d \to [k]$ be defined as in Theorem 5. Then (Γ, pred) is ℓ-calibrated over Δ_n.

Ramaswamy et al. (2013) also applied Theorem 5 to obtain low-dimensional convex calibrated surrogates for several subset ranking losses. For subset ranking losses with $\text{affdim}(\mathbf{L}) = r$ (such as the DCG@r loss), the linear property constructed by the above result effectively provides a standardization function over $\Delta_\mathcal{Y}$. For other subset ranking losses, the two approaches can give complementary results. For example, for the WPD and MAP losses, which have affine dimensions $\Theta(r^2)$ (Ramaswamy and Agarwal, 2015), it is known that there is no standardization function over $\Delta_\mathcal{Y}$ (Buffoni et al., 2011), and that there is no convex calibrated surrogate over $\Delta_\mathcal{Y}$ in r dimensions (Calauzènes et al., 2012; Ramaswamy and Agarwal, 2015). On the other hand, by Theorem 5, there do exist $\Theta(r^2)$-dimensional calibrated linear properties and therefore $\Theta(r^2)$-dimensional convex calibrated surrogates for these losses over $\Delta_\mathcal{Y}$; moreover, as demonstrated in Example 8, one can construct standardization functions for these losses over restricted sets of distributions $\mathcal{P} \subset \Delta_\mathcal{Y}$, allowing for r-dimensional convex calibrated surrogates over such restricted sets $\mathcal{P}$.

The following example illustrates a different application of the above result:

Example 3 (Hamming loss for sequence prediction) *Consider a sequence prediction task with $\mathcal{Y} = \widehat{\mathcal{Y}} = \{0, 1\}^r$ (thus $n = k = 2^r$). A widely used loss in this setting is the Hamming loss $\ell_{\text{Ham}} : \{0, 1\}^r \times \{0, 1\}^r \to \mathbb{R}_+$ given by*

$$\ell_{\text{Ham}}(\mathbf{y}, \mathbf{t}) = \sum_{i=1}^r \mathbf{1}(t_i \neq y_i) \quad \forall \mathbf{y}, \mathbf{t} \in \{0, 1\}^r .$$

As shown by Ramaswamy and Agarwal (2012), $\text{affdim}(\mathbf{L}^{\text{Ham}}) \leq r$, and therefore by Theorem 6, one can construct an r-dimensional linear property $\Gamma : \Delta_\mathcal{Y} \to \mathbb{R}^r$ that is ℓ_{Ham}-calibrated over $\Delta_\mathcal{Y}$. Any strictly proper scoring rule for Γ then forms an r-dimensional ℓ_{Ham}-calibrated surrogate over $\Delta_\mathcal{Y}$; in particular, using $\phi(\mathbf{u}) = \frac{1}{2}\|\mathbf{u}\|_2^2$ in Theorem 1, one gets the surrogate given by Theorem 5.

4.3. Lower Bound on Dimension of Calibrated Linear Properties

Theorem 6 shows that for any loss ℓ, there is a linear property in $d = \mathrm{affdim}(\mathbf{L})$ dimensions that is ℓ-calibrated over Δ_n. In the following result, we show that this is essentially the best one can do with linear properties (see Appendix D for a proof):

Theorem 7 (Lower bound on dimension of calibrated linear properties) *Let $\ell : [n] \times [k] \to \mathbb{R}_+$. Let $\Gamma : \Delta_n \to \mathbb{R}^d$ be a linear property. If there exists a mapping* pred $: \mathbb{R}^d \to [k]$ *such that* (Γ, pred) *is ℓ-calibrated over Δ_n, then*

$$d \;\geq\; \mathrm{affdim}(\mathbf{L}) - 1\,.$$

5. Calibrated Surrogates via Calibrated Nonlinear Properties

We now consider settings where one can exploit calibrated *nonlinear* properties to design convex calibrated surrogates in an even smaller number of dimensions than is possible via linear properties. We start by considering quantiles, which are 1-dimensional nonlinear (possibly interval-valued) properties; quantiles can be directly elicited via convex strictly proper scoring rules and lead to calibrated 1-dimensional surrogates for certain ordinal regression type losses (Section 5.1). We then develop a general framework for designing low-dimensional convex calibrated surrogates under 'low-noise' conditions by eliciting vectors of quantiles that yield 'coarse' information about a distribution (Section 5.2). We conclude with a result that gives a necessary condition for a general nonlinear property to be directly elicitable via a convex strictly proper scoring rule (Section 5.3).

5.1. Quantiles and Interval-Valued Properties

Quantiles and generalized quantiles have recently received significant attention in the property elicitation literature (Kiefer, 2010; Gneiting, 2011; Schervish et al., 2012; Grant and Gneiting, 2013; Steinwart et al., 2014). These are nonlinear properties; moreover, for discrete distributions, these properties can take a range of values over an interval. Therefore we will need to allow for interval-valued properties Γ that map each distribution $\mathbf{p} \in \Delta_n$ (or more generally, each $\mathbf{p} \in \mathcal{P}$ for some $\mathcal{P} \subseteq \Delta_n$) to a vector of *intervals*, $\Gamma(\mathbf{p}) \in \mathcal{I}^d$, where $\mathcal{I}$ denotes the set of all intervals on the real line. In this case, we will say a scoring rule $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ is proper for $\Gamma : \mathcal{P} \to \mathcal{I}^d$ if

$$\forall \mathbf{p} \in \mathcal{P} : \quad \Gamma(\mathbf{p}) \subseteq \mathrm{argmin}_{\mathbf{u} \in \mathbb{R}^d} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})]\,,$$

and strictly proper for Γ if the above holds with equality (i.e. no value $\mathbf{u} \notin \Gamma(\mathbf{p})$ is a minimizer).

Given a loss $\ell : [n] \times [k] \to \mathbb{R}_+$, we will say an interval-valued property $\Gamma : \mathcal{P} \to \mathcal{I}^d$ is ℓ-calibrated over $\mathcal{P}$ if $\exists$ pred $: \mathbb{R}^d \to [k]$ such that for all $\mathbf{p} \in \mathcal{P}$ and all convergent sequences $\{\mathbf{u}_m\}$ in $\mathbb{R}^d$,

$$\lim_{m \to \infty} \mathbf{u}_m \;\in\; \Gamma(\mathbf{p}) \;\implies\; \mathbf{E}_{Y \sim \mathbf{p}}[\ell(Y, \text{pred}(\mathbf{u}_m)] \to \min_{t \in [k]} \mathbf{E}_{Y \sim \mathbf{p}}[\ell(Y, t)]\,.$$

Again, it can be shown that a strictly proper scoring rule ψ for an ℓ-calibrated interval-valued property $\Gamma : \mathcal{P} \to \mathcal{I}^d$ forms an ℓ-calibrated surrogate over $\mathcal{P}$.

Quantiles. For $\alpha \in (0, 1)$, the α-quantile of $\mathbf{p} \in \Delta_n$ is defined as the interval

$$Q_\alpha(\mathbf{p}) = \left\{u \in \mathbb{R} : \mathbf{P}_{Y \sim \mathbf{p}}(Y \leq u) \geq \alpha \text{ and } \mathbf{P}_{Y \sim \mathbf{p}}(Y \geq u) \geq 1 - \alpha\right\} \in \mathcal{I}\,. \tag{2}$$

It is known that the scoring rule $\psi : [n] \times \mathbb{R} \to \mathbb{R}_+$ defined as

$$\psi(y, u) = (1 - \alpha) \cdot (u - y)_+ + \alpha \cdot (y - u)_+ \tag{3}$$

is a convex strictly proper scoring rule for the α-quantile, i.e. for the property $\Gamma : \Delta_n \to \mathcal{I}$ defined as $\Gamma(\mathbf{p}) = Q_\alpha(\mathbf{p})$. For the median $\Gamma(\mathbf{p}) = Q_{\frac{1}{2}}(\mathbf{p})$, the above scoring rule becomes $\psi(y, u) = \frac{1}{2}|u - y|$.

Example 4 (Generalized ordinal regression loss) *Let $k = n$ and $\alpha \in (0,1)$, and consider the generalized ordinal regression loss $\ell : [n] \times [n] \to \mathbb{R}_+$ defined as*

$$\ell_{\text{ord}(\alpha)}(y,t) = (1-\alpha)(t-y)_+ + \alpha(y-t)_+ .$$

It is easy to see that the α-quantile $\Gamma(\mathbf{p}) = Q_\alpha(\mathbf{p})$ is an $\ell_{\text{ord}(\alpha)}$-calibrated nonlinear property over Δ_n; the scoring rule ψ in Eq. (3) is therefore a 1-dimensional convex calibrated surrogate for $\ell_{\text{ord}(\alpha)}$ over Δ_n. Note that this is a significant improvement over what can be achieved with linear properties for these losses, e.g. for $\alpha = \frac{1}{2}$, the loss matrix $\mathbf{L}^{\text{ord}(\alpha)}$ has affine dimension $n-1$, and thus by Theorem 7, any calibrated linear property for this loss must have dimension at least $n-2$.

5.2. Calibrated Surrogates under Low-Noise Conditions Using Vectors of Quantiles

We now give a general framework for constructing low-dimensional convex calibrated surrogates under suitable 'low-noise' conditions by eliciting a vector of quantiles that forms a calibrated nonlinear property under such conditions.

The broad idea is to estimate 'coarse' information about a distribution $\mathbf{p} \in \Delta_n$ using a vector of quantiles. Specifically, for any integer $s \in \mathbb{Z}_+$ ($s \geq 2$) and for a suitable set of distributions $\mathcal{P} \subseteq \Delta_n$, we define an $(s-1)$-dimensional interval-valued property $\Gamma_s : \mathcal{P} \to \mathcal{I}^{s-1}$ as follows:

$$\Gamma_s(\mathbf{p}) \;=\; Q_{\frac{1}{s}}(\mathbf{p}) \times \ldots \times Q_{\frac{s-1}{s}}(\mathbf{p}) \;\in\; \mathcal{I}^{s-1} . \tag{4}$$

From the discussion in Section 5.1, it follows that the scoring rule $\psi_s : [n] \times \mathbb{R}^{s-1} \to \mathbb{R}_+$ defined as

$$\psi_s(y, \mathbf{u}) = \sum_{i=1}^{s-1} \left(\left(1 - \frac{i}{s}\right) \cdot (u_i - y)_+ + \left(\frac{i}{s}\right) \cdot (y - u_i)_+ \right) \tag{5}$$

is a convex strictly proper scoring rule for Γ_s.

In order to design calibrated surrogates using the above vector-of-quantiles property Γ_s, we will find it convenient to define for each $y \in [n]$ a function $N_y : \mathbb{R}^{s-1} \to \mathbb{Z}_+$, which for each $\mathbf{u} \in \mathbb{R}^{s-1}$ counts how many times the label y appears in the vector $\lfloor \mathbf{u} \rfloor$ (where $\lfloor \mathbf{u} \rfloor = \left(\lfloor u_1 \rfloor, \ldots, \lfloor u_{s-1} \rfloor \right)^\top$):

$$N_y(\mathbf{u}) = \sum_{i=1}^{s-1} \mathbf{1}(y = \lfloor u_i \rfloor) \quad \forall \mathbf{u} \in \mathbb{R}^{s-1} .$$

The following lemma shows that eliciting any $\mathbf{u} \in \Gamma_s(\mathbf{p})$ allows one to elicit for each $y \in [n]$ an interval of width at most $\frac{2}{s}$ containing p_y:

Lemma 8 (Vectors of quantiles give interval estimates for probabilities) *Let $\mathcal{P} \subseteq \Delta_n$ and $\mathbf{p} \in \mathcal{P}$. Let $\Gamma_s : \mathcal{P} \to \mathcal{I}^{s-1}$ be defined as in Eq. (4) above, and let $\mathbf{u} \in \Gamma_s(\mathbf{p})$. Then for each $y \in [n]$, we have*

$$p_y \in \begin{cases} \left[\frac{N_y(\mathbf{u})-1}{s}, \frac{N_y(\mathbf{u})+1}{s} \right] & \text{if } N_y(\mathbf{u}) \geq 1 \\ \left[0, \frac{1}{s}\right] & \text{if } N_y(\mathbf{u}) = 0 . \end{cases}$$

Proof Let $y \in [n]$. If $N_y(\mathbf{u}) = 0$, then no quantile in $\Gamma_s(\mathbf{p})$ consists of the singleton interval $\{y\}$, and consequently, we must have $p_y \leq \frac{1}{s}$. Now suppose $N_y(\mathbf{u}) \geq 1$. Then the number of quantiles in $\Gamma_s(\mathbf{p})$ that consist of the singleton interval $\{y\}$ is at least $N_y(\mathbf{u}) - 2$ and at most $N_y(\mathbf{u})$, and therefore we must have $\frac{N_y(\mathbf{u})-1}{s} \leq p_y \leq \frac{N_y(\mathbf{u})+1}{s}$. ■

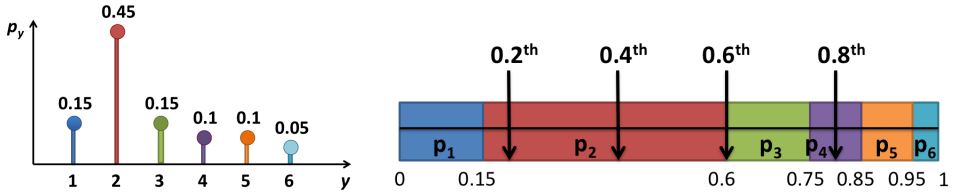

Figure 1: Illustration of quantile vector property $\Gamma_s(\mathbf{p})$ used to elicit coarse information about a distribution $\mathbf{p} \in \Delta_n$ (here $n = 6$, $s = 5$). See Example 5 for details.

Example 5 (Quantile vectors and probability interval estimates) *Consider the example shown in Figure 1 ($n = 6$, $s = 5$). The figure shows the $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$ and $\frac{4}{5}$-quantiles of the probability vector $\mathbf{p} = (0.15, 0.45, 0.15, 0.1, 0.1, 0.05)^\top \in \Delta_6$. Here $Q_{\frac{1}{5}}(\mathbf{p}) = \{2\}$, $Q_{\frac{2}{5}}(\mathbf{p}) = \{2\}$, $Q_{\frac{3}{5}}(\mathbf{p}) = [2, 3]$, and $Q_{\frac{4}{5}}(\mathbf{p}) = \{4\}$, and so $\Gamma_5(\mathbf{p}) = \{2\} \times \{2\} \times [2, 3] \times \{4\}$. Consider $\mathbf{u} = (2,\ 2,\ 2.5,\ 4)^\top \in \Gamma_5(\mathbf{p})$. As can be seen, here $N_1(\mathbf{u}) = N_3(\mathbf{u}) = N_5(\mathbf{u}) = N_6(\mathbf{u}) = 0$; $N_2(\mathbf{u}) = 3$; and $N_4(\mathbf{u}) = 1$. Therefore by Lemma 8, we obtain the following interval estimates for elements of $\mathbf{p}$ from $\mathbf{u}$: $p_1,\ p_3,\ p_5,\ p_6 \in [0, 0.2]$; $p_2 \in [0.4, 0.8]$; and $p_4 \in [0, 0.4]$. Similarly, consider $\mathbf{u}' = (2,\ 2,\ 3,\ 4)^\top$, which also lies in $\Gamma_5(\mathbf{p})$. In this case, we would have $N_1(\mathbf{u}') = N_5(\mathbf{u}') = N_6(\mathbf{u}') = 0$; $N_2(\mathbf{u}') = 2$; and $N_3(\mathbf{u}') = N_4(\mathbf{u}') = 1$, and therefore we would get the following interval estimates for elements of $\mathbf{p}$ from $\mathbf{u}'$: $p_1,\ p_5,\ p_6 \in [0, 0.2]$; $p_2 \in [0.2, 0.6]$; and $p_3,\ p_4 \in [0, 0.4]$.*

Thus vectors of quantiles give coarse information about the probability distribution $\mathbf{p} \in \Delta_n$, and can be useful wherever it is sufficient to elicit not $\mathbf{p}$ exactly, but rather some intervals in which p_y lie. In particular, this can be useful for designing low-dimensional convex surrogates that are calibrated for a loss over a suitable set of 'low-noise' distributions. We give two such examples below, one for the multiclass 0-1 loss, and one for multiclass classification with a reject option.

Example 6 (O(log(n))-dimensional convex surrogate calibrated for 0-1 loss under low-noise condition) *Let $k = n$ and consider the multiclass 0-1 loss $\ell_{0\text{-}1} : [n] \times [n] \to \mathbb{R}_+$ defined as*

$$\ell_{0\text{-}1}(y, t) = \mathbf{1}(y \neq t)\,.$$

Consider the following 'low-noise' condition, under which the highest-probability element is separated from the next highest-probability element by a probability difference of at least $\frac{2}{\lceil \log_2(n) \rceil}$:

$$\mathcal{P}^{0\text{-}1}_{\text{LN}} \;=\; \left\{ \mathbf{p} \in \Delta_n : \exists y \in [n] \text{ such that } p_y > p_{y'} + \frac{2}{\lceil \log_2(n) \rceil}\ \forall y' \neq y \right\}.$$

Then it follows from Lemma 8 that for any $\mathbf{p} \in \mathcal{P}^{0\text{-}1}_{\text{LN}}$, by estimating a vector $\mathbf{u} \in \Gamma_{\lceil \log_2(n) \rceil}(\mathbf{p})$, one can accurately identify the largest-probability element under $\mathbf{p}$, $\operatorname{argmax}_{y \in [n]} p_y$ (and make an optimal prediction under $\ell_{0\text{-}1}$). Therefore the $(\lceil \log_2(n) \rceil - 1)$-dimensional property $\Gamma_{\lceil \log_2(n) \rceil}$ is $\ell_{0\text{-}1}$-calibrated over $\mathcal{P}^{0\text{-}1}_{\text{LN}}$ using $\text{pred}^{0\text{-}1} : \mathbb{R}^{\lceil \log_2(n) \rceil - 1} \to [n]$ satisfying

$$\text{pred}^{0\text{-}1}(\mathbf{u}) \;\in\; \operatorname{argmax}_{y \in [n]} N_y(\mathbf{u})\,.$$

For large n, for which the above low-noise condition is quite broad,[3] this construction gives a significant improvement over the $n - 1$ dimensions needed for a convex surrogate to be calibrated for $\ell_{0\text{-}1}$ over Δ_n (Ramaswamy and Agarwal, 2012).

3. Indeed, the low-noise condition $\mathcal{P}^{0\text{-}1}_{\text{LN}}$ here includes many probability distributions that are excluded from the commonly studied 'dominant-label' condition $\mathcal{P}^{0\text{-}1}_{\text{DL}} = \{\mathbf{p} \in \Delta_n : \max_{y \in [n]} p_y > \frac{1}{2}\}$, which is required for example for the common (n-dimensional) Crammer-Singer surrogate to be $\ell_{0\text{-}1}$-calibrated.

Example 7 (O(log(n))-dimensional convex surrogate calibrated for multiclass classification with a reject option under low-noise condition) *Consider now a multiclass classification problem with a reject option. Here $k = n + 1$, with the prediction $(n + 1)$ corresponding to the 'reject' option; a common loss in this setting is the loss $\ell_{\text{reject}} : [n] \times [n + 1] \to \mathbb{R}_+$ defined as*

$$\ell_{\text{reject}}(y, t) = \begin{cases} \mathbf{1}(y \neq t) & \text{if } t \in [n] \\ \frac{1}{2} & \text{if } t = n + 1. \end{cases}$$

Consider the following 'low-noise' condition, under which each probability element is separated from $\frac{1}{2}$ by at least $\frac{1}{\lceil \log_2(n) \rceil}$:

$$\mathcal{P}_{\text{LN}}^{\text{reject}} = \left\{ \mathbf{p} \in \Delta_n : p_y \notin \left[\frac{1}{2} - \frac{1}{\lceil \log_2(n) \rceil}, \frac{1}{2} + \frac{1}{\lceil \log_2(n) \rceil} \right] \forall y \in [n] \right\}.$$

Then it follows from Lemma 8 that for any $\mathbf{p} \in \mathcal{P}_{\text{LN}}^{\text{reject}}$, by estimating a vector $\mathbf{u} \in \Gamma_{\lceil \log_2(n) \rceil}(\mathbf{p})$, one can accurately identify whether any label has probability greater than $\frac{1}{2}$ under $\mathbf{p}$ (and make an optimal prediction under ℓ_{reject}). Therefore the $(\lceil \log_2(n) \rceil - 1)$-dimensional property $\Gamma_{\lceil \log_2(n) \rceil}$ is ℓ_{reject}-calibrated over $\mathcal{P}_{\text{LN}}^{\text{reject}}$ using $\text{pred}^{\text{reject}} : \mathbb{R}^{\lceil \log_2(n) \rceil - 1} \to [n]$ defined as follows:

$$\text{pred}^{\text{reject}}(\mathbf{u}) = \begin{cases} \text{argmax}_{y \in [n]} N_y(\mathbf{u}) & \text{if } \exists y \in [n] \text{ such that } N_y(\mathbf{u}) \geq \frac{\lceil \log_2(n) \rceil}{2} \\ n + 1 & \text{otherwise.} \end{cases}$$

To our knowledge, the above approach gives the first general framework for designing low-noise conditions together with convex surrogates that are calibrated under these conditions for different losses. In particular, the framework allows one to develop convex calibrated surrogates under any low-noise condition where a coarse estimate of the underlying probability vector suffices to make an optimal prediction under the loss of interest.

5.3. Necessary Condition for Convex Elicitability

As we have seen, linear properties and quantile-based properties are always directly elicitable by a convex strictly proper scoring rule. For general nonlinear properties, the following result gives a necessary condition for convex elicitability (see Appendix E for a proof):

Theorem 9 (Necessary condition for convex elicitability of a property over Δ_n) *Let $\Gamma : \Delta_n \to \mathbb{R}^d$. If Γ is directly elicitable via a convex proper scoring rule, then*

$$\dim(\Gamma^{-1}(\mathbf{u})) \geq n - d - 1 \quad \forall \mathbf{u} \in \Gamma(\text{relint}(\Delta_n)).$$

Corollary 10 *Let $\Gamma : \Delta_n \to \mathbb{R}^d$ be d'-elicitable via a convex proper scoring rule in $d' \geq d$ dimensions. Then*

$$d' \geq n - \dim(\Gamma^{-1}(\mathbf{u})) - 1 \quad \forall \mathbf{u} \in \Gamma(\text{relint}(\Delta_n)).$$

Acknowledgments

Thanks to Harish G. Ramaswamy for related discussions and to the anonymous reviewers for helpful comments. AA thanks Google for a travel grant to present this work at COLT. SA thanks DST for support under a Ramanujan Fellowship, and Yahoo! Labs for an unrestricted grant.

References

Jacob D. Abernethy and Rafael M. Frongillo. A characterization of scoring rules for linear properties. In *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

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

David Buffoni, Clément Calauzènes, Patrick Gallinari, and Nicolas Usunier. Learning scoring functions with order-preserving losses and standardized supervision. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

Clément Calauzènes, Nicolas Usunier, and Patrick Gallinari. On the (non-)existence of convex, calibrated surrogate losses for ranking. In *Advances in Neural Information Processing Systems*, 2012.

David Cossock and Tong Zhang. Statistical analysis of Bayes optimal subset ranking. *IEEE Transactions on Information Theory*, 54(11):5140–5154, 2008.

John Duchi, Lester Mackey, and Michael Jordan. On the consistency of ranking algorithms. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

Rafael Frongillo and Ian Kash. Vector-valued property elicitation. In *Proceedings of the 28th Annual Conference on Learning Theory*, 2015.

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

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Kyrill Grant and Tilmann Gneiting. Consistent scoring functions for quantiles. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 163–173. Institute of Mathematical Statistics, 2013.

Nicholas M. Kiefer. Incentive-compatible elicitation of quantiles, 2010. URL https://www.american.edu/cas/economics/info-metrics/pdf/upload/Working-Paper-Kiefer.pdf.

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

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

Yanyan Lan, Jiafeng Guo, Xueqi Cheng, and Tie-Yan Liu. Statistical consistency of ranking methods in a rank-differentiable probability space. In *Advances in Neural Information Processing Systems*, 2012.

Harish G. Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Processing Systems*, 2012.

Harish G. Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *Journal of Machine Learning Research. To appear*, 2015.

Harish G. Ramaswamy, Shivani Agarwal, and Ambuj Tewari. Convex calibrated surrogates for low-rank loss matrices with applications to subset ranking losses. In *Advances in Neural Information Processing Systems*, 2013.

Pradeep Ravikumar, Ambuj Tewari, and Eunho Yang. On NDCG consistency of listwise ranking methods. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011.

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

Mark J. Schervish. A general method for comparing probability assessors. *The Annals of Statistics*, 17(4):1856–1879, 1989.

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.

Ingo Steinwart. How to compare different loss functions and their risks. *Constructive Approximation*, 26:225–287, 2007.

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.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

Elodie Vernet, Robert C. Williamson, and Mark D. Reid. Composite multiclass losses. In *Advances in Neural Information Processing Systems*, 2011.

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

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

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004b.

Appendix A. Proof of Theorem 4

Proof Let $\mathbf{p} \in \mathcal{P}$, and let $\{\mathbf{u}_m\}$ be any sequence in $\mathbb{R}^r$ such that $\mathbf{u}_m \to \Gamma(\mathbf{p})$. We will show that $\mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y, \text{pred}(\mathbf{u}_m))] \to \min_{\sigma\in S_r} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,\sigma)]$.

Let $\delta := \min_{i,j\in[r]:|\Gamma_i(\mathbf{p})-\Gamma_j(\mathbf{p})|>0} |\Gamma_i(\mathbf{p}) - \Gamma_j(\mathbf{p})|$. Since $\mathbf{u}_m \to \Gamma(p)$, we have $\exists M$ such that

$$\forall m \geq M: \quad \|\mathbf{u}_m - \Gamma(\mathbf{p})\|_2 < \delta\,.$$

Now clearly, for all $m \geq M$ and $i, j \in [r]$, we must have $\Gamma_i(\mathbf{p}) > \Gamma_j(\mathbf{p}) \implies u_{mi} > u_{mj}$ (else the L_2-distance between $\mathbf{u}_m$ and $\Gamma(\mathbf{p})$ would exceed δ). Therefore, for all $m \geq M$, we have $\text{argsort}(\mathbf{u}_m) \subseteq \text{argsort}(\Gamma(\mathbf{p}))$, and thus $\mathbf{E}_{Y\sim\mathbf{p}}\big[\ell(Y, \text{pred}(\mathbf{u}_m))] = \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y, \text{pred}(\Gamma(\mathbf{p})))]$. Also, by construction of pred, we know that $\mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y, \text{pred}(\Gamma(\mathbf{p})))] = \min_{\sigma\in S_r} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,\sigma)]$. This implies that for all $m \geq M$, $\mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y, \text{pred}(\mathbf{u}_m))] = \min_{\sigma\in S_r} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,\sigma)]$.

Since $\mathbf{p} \in \mathcal{P}$ was arbitrary, this proves the result. ■

Appendix B. Additional Example of Application of Theorem 4: Viewing Subset Ranking Surrogate of Duchi et al. (2010) as a Strictly Proper Scoring Rule for a Linear Property

Example 8 (Weighted pairwise disagreement (WPD) loss for subset ranking) *Another popular subset ranking loss is the WPD loss for weighted preference graph labels, $\ell_{\text{WPD}} : \mathcal{Y} \times \mathcal{S}_r \to \mathbb{R}_+$, where $\mathcal{Y}$ is some finite set of weighted DAGs on r nodes; for a weighted DAG $G = ([r], E^G, \mathbf{W}^G) \in \mathcal{Y}$, where $E^G \subset [r] \times [r]$ denotes the set of edges of G and $\mathbf{W}^G \in \mathbb{R}_+^{r\times r}$ denotes the edge weights with $W^G_{ij} > 0$ iff $(i,j) \in E^G$, and for a permutation $\sigma \in \mathcal{S}_r$, this loss is defined as*

$$\ell_{\text{WPD}}(G, \sigma) \;=\; \sum_{i,j} W^G_{ij} \Big(\mathbf{1}(\sigma(i) > \sigma(j)) + \frac{1}{2}\mathbf{1}(\sigma(i) = \sigma(j)) \Big)\,.$$

For any $\mathbf{p} \in \Delta_\mathcal{Y}$, define $W^\mathbf{p}_{ij} = \mathbf{E}_{G\sim\mathbf{p}}[W^G_{ij}]$ and $E^\mathbf{p} = \big\{(i,j) \in [r] \times [r] : W^\mathbf{p}_{ij} > W^\mathbf{p}_{ji}\big\}$. Duchi et al. (2010) considered the following set of 'low-noise' distributions $\mathbf{p} \in \Delta_\mathcal{Y}$:

$$\mathcal{P}^{\text{WPD}}_{\text{LN}} \;=\; \Big\{ \mathbf{p} \in \Delta_\mathcal{Y} : \text{the unweighted graph } G^\mathbf{p} = ([r], E^\mathbf{p}) \text{ is a DAG, and} \\ \forall i,k \in [r]:\; W^\mathbf{p}_{ik} > W^\mathbf{p}_{ki} \implies \textstyle\sum_{j=1}^r \big(W^\mathbf{p}_{ij} - W^\mathbf{p}_{ji}\big) > \sum_{j=1}^r \big(W^\mathbf{p}_{kj} - W^\mathbf{p}_{jk}\big) \Big\}\,.$$

It is easy to see that the function $\mathbf{s} : \mathcal{Y} \to \mathbb{R}^r$ defined as $s_i(G) = \sum_{j=1}^r (W^G_{ij} - W^G_{ji})\; \forall i \in [r]$ is a standardization function for ℓ_{WPD} over $\mathcal{P}^{\text{WPD}}_{\text{LN}}$, and therefore by Theorem 4, any strictly proper scoring rule for the corresponding linear property $\Gamma : \mathcal{P}^{\text{WPD}}_{\text{LN}} \to \mathbb{R}^r$ given by $\Gamma_i(\mathbf{p}) = \sum_{j=1}^r (W^\mathbf{p}_{ij} - W^\mathbf{p}_{ji})\; \forall i \in [r], \mathbf{p} \in \mathcal{P}^{\text{WPD}}_{\text{LN}}$ yields an ℓ_{WPD}-calibrated surrogate over $\mathcal{P}^{\text{WPD}}_{\text{LN}}$. The convex r-dimensional surrogate shown to be ℓ_{WPD}-calibrated over $\mathcal{P}^{\text{WPD}}_{\text{LN}}$ by Duchi et al. (2010) can be viewed as a strictly proper scoring rule for this property composed with a link function.

Appendix C. Proof of Theorem 6

Proof Note first that for any $\mathbf{p} \in \Delta_n$ and $t \in [k]$, we have

$$\begin{aligned}
\mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,t)] &= \sum_{y=1}^{d} p_y \left(\sum_{i=1}^{d} A_{yi} B_{it} + c \right) \\
&= \sum_{y=1}^{d} \sum_{i=1}^{d} p_y A_{yi} B_{it} + c \\
&= \sum_{i=1}^{d} B_{it} \sum_{y=1}^{d} p_y A_{yi} + c \\
&= \sum_{i=1}^{d} B_{it} \mathbf{E}_{Y\sim\mathbf{p}}[A_{Yi}] + c \;=\; \sum_{i=1}^{d} B_{it} \Gamma_i(\mathbf{p}) + c \,. \qquad (6)
\end{aligned}$$

Now, let $\mathbf{p} \in \Delta_n$, and let $\{\mathbf{u}_m\}$ be any sequence in $\mathbb{R}^d$ such that $\mathbf{u}_m \to \Gamma(\mathbf{p})$. For each m, define $t_m := \text{pred}(\mathbf{u}_m) \in [k]$. Then we have

$$\begin{aligned}
&\mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,t_m)] - \min_{t\in[k]} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,t)] \\
&= \sum_{i=1}^{d} B_{it_m} \Gamma_i(\mathbf{p}) - \min_{t\in[k]} \sum_{i=1}^{d} B_{it} \Gamma_i(\mathbf{p}) \,, \qquad \text{by Eq. (6)} \\
&= \sum_{i=1}^{d} B_{it_m} (\Gamma_i(\mathbf{p}) - u_{mi}) + \sum_{i=1}^{d} B_{it_m} u_{mi} - \min_{t\in[k]} \sum_{i=1}^{d} B_{it} \Gamma_i(\mathbf{p}) \\
&= \sum_{i=1}^{d} B_{it_m} (\Gamma_i(\mathbf{p}) - u_{mi}) + \min_{t\in[k]} \sum_{i=1}^{d} B_{it} u_{mi} - \min_{t\in[k]} \sum_{i=1}^{d} B_{it} \Gamma_i(\mathbf{p}) \,,
\end{aligned}$$

where the last equality holds due to the definition of pred. It is easy to see that the term on the right hand side goes to zero as $m \to \infty$. Thus we get that $\mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,t_m)] \to \min_{t\in[k]} \mathbf{E}_{Y\sim\mathbf{p}}[\ell(Y,t)]$. Since $\mathbf{p} \in \Delta_n$ was arbitrary, this proves the result. ■

Appendix D. Proof of Theorem 7

For each $t \in [k]$, denote $\boldsymbol{\ell}_t = (\ell(1,t), \cdots, \ell(n,t))^\top$. We will need the following definition:

Definition 11 (Trigger Probabilities Ramaswamy and Agarwal (2012)) *Let $\ell : [n] \times [k] \to \mathbb{R}_+$. For each $t \in [k]$, the set of trigger probabilities of t with respect to ℓ is defined as*

$$\mathcal{Q}_t^\ell := \left\{ \mathbf{p} \in \Delta_n : \mathbf{p}^\top (\boldsymbol{\ell}_t - \boldsymbol{\ell}_{t'}) \leq 0 \;\; \forall t' \in [k] \right\} = \left\{ \mathbf{p} \in \Delta_n : t \in \text{Opt}(\ell, \mathbf{p}) \right\}.$$

Proof Suppose $\exists \text{pred} : \mathbb{R}^d \to [k]$ such that (Γ, pred) is ℓ-calibrated over Δ_n. We will show that $d \geq \text{affdim}(\mathbf{L}) - 1$.

Suppose for the sake of contradiction that $d < \text{affdim}(\mathbf{L}) - 1$. Let $\mathbf{s} : [n] \to \mathbb{R}^d$ be such that $\Gamma(\mathbf{p}) = \mathbf{E}_{Y\sim\mathbf{p}}[\mathbf{s}(Y)]$ $\forall \mathbf{p} \in \Delta_n$, and define $\mathbf{U} \in \mathbb{R}^{d\times n}$ as $u_{iy} := s_i(y)$ $\forall i \in [d], y \in [n]$. Observe

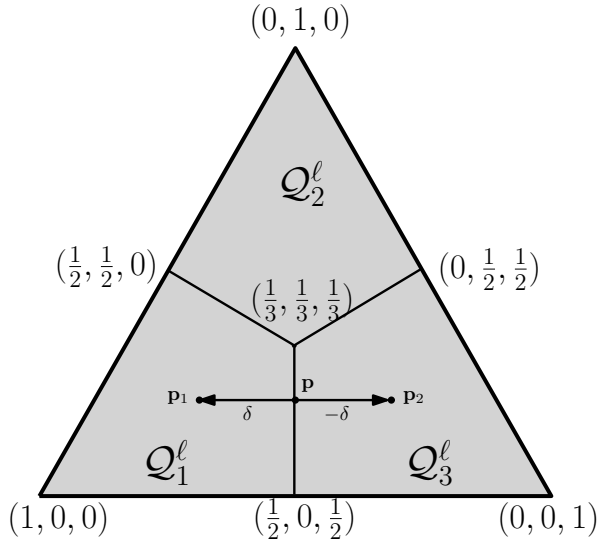

Figure 2: Illustration of steps in the proof of Theorem 7. We first find $\mathbf{p} \in \mathcal{Q}_1^\ell \cap \mathcal{Q}_3^\ell$, and then perturb $\mathbf{p}$ along $\boldsymbol{\delta}$ and $-\boldsymbol{\delta}$ to find $\mathbf{p}_1$ and $\mathbf{p}_2$.

that $\Gamma(\mathbf{p}) = \mathbf{U}\mathbf{p}$. For each $i \in [d]$, let $\mathbf{u}_i \in \mathbb{R}^n$ denote the i-th row vector of $\mathbf{U}$, so that $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_d]^\top$. Define $\widetilde{\mathbf{U}} := [\mathbf{u}_1 \cdots \mathbf{u}_d\ \mathbf{1}]^\top$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector.

The main idea of the proof is to find $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_n$ such that $\mathbf{U}\mathbf{p}_1 = \mathbf{U}\mathbf{p}_2$ but $\text{Opt}(\ell, \mathbf{p}_1) \cap \text{Opt}(\ell, \mathbf{p}_2) = \emptyset$; this will contradict the fact that (Γ, pred) is ℓ-calibrated over Δ_n. We find such $\mathbf{p}_1, \mathbf{p}_2$ by first finding $\mathbf{p} \in \Delta_n$ that lies at the intersection of two trigger probability sets, and then perturbing it along suitable directions $\boldsymbol{\delta}, -\boldsymbol{\delta}$ (see Figure 2). The following steps give more details.

Step 1: Let $i, j \in [k]$ be such that $\boldsymbol{\ell}_i - \boldsymbol{\ell}_j \notin \text{col}(\widetilde{\mathbf{U}}^\top)$ and $\mathcal{Q}_i^\ell \cap \mathcal{Q}_j^\ell \neq \emptyset$. To see that such i, j always exist, note that by our assumption that $d + 1 < \text{affdim}(\mathbf{L})$, $\exists i', j' \in [k]$ such that $\boldsymbol{\ell}_{i'} - \boldsymbol{\ell}_{j'} \notin \text{col}(\widetilde{\mathbf{U}}^\top)$. If $\mathcal{Q}_{i'}^\ell \cap \mathcal{Q}_{j'}^\ell \neq \emptyset$, define $i := i'$ and $j := j'$ and we are done. Suppose that $\mathcal{Q}_{i'}^\ell \cap \mathcal{Q}_{j'}^\ell = \emptyset$. Consider a sequence of neighboring trigger probability sets $\mathcal{Q}_{i_1}^\ell, \mathcal{Q}_{i_2}^\ell, \cdots, \mathcal{Q}_{i_m}^\ell$ such that $i_1 = i'$, $i_m = j'$, and $\mathcal{Q}_{i_r}^\ell \cap \mathcal{Q}_{i_{r+1}}^\ell \neq \emptyset$ for all $r \in [m-1]$. We can write $\boldsymbol{\ell}_{i'} - \boldsymbol{\ell}_{j'} = (\boldsymbol{\ell}_{i_1} - \boldsymbol{\ell}_{i_2}) + (\boldsymbol{\ell}_{i_2} - \boldsymbol{\ell}_{i_3}) + \cdots + (\boldsymbol{\ell}_{i_{m-1}} - \boldsymbol{\ell}_{i_m})$. Since $\boldsymbol{\ell}_{i'} - \boldsymbol{\ell}_{j'} \notin \text{col}(\widetilde{\mathbf{U}}^\top)$, $\exists r \in [m-1]$ such that $\boldsymbol{\ell}_{i_r} - \boldsymbol{\ell}_{i_{r+1}} \notin \text{col}(\widetilde{\mathbf{U}}^\top)$. Define $i := r$ and $j := r + 1$. Then we have $\boldsymbol{\ell}_i - \boldsymbol{\ell}_j \notin \text{col}(\widetilde{\mathbf{U}}^\top)$ and $\mathcal{Q}_i^\ell \cap \mathcal{Q}_j^\ell \neq \emptyset$.

Step 2: Fix i, j as above, and let $\mathbf{p} \in \mathcal{Q}_i^\ell \cap \mathcal{Q}_j^\ell \cap \text{relint}(\Delta_n)$ such that $\mathbf{p} \notin \mathcal{Q}_t^\ell\ \forall t \neq i, j$ (which means that $\mathbf{p}^\top \boldsymbol{\ell}_i = \mathbf{p}^\top \boldsymbol{\ell}_j < \mathbf{p}^\top \boldsymbol{\ell}_t\ \forall t \neq i, j$). The trigger probability sets form a power diagram of the probability simplex, which implies that $\mathcal{Q}_i^\ell \cap \mathcal{Q}_j^\ell \not\subset \text{bndry}(\Delta_n)$ and $\mathcal{Q}_i^\ell \cap \mathcal{Q}_j^\ell \not\subset \mathcal{Q}_t^\ell\ \forall t \neq i, j$; therefore, such a point $\mathbf{p}$ always exists.

Step 3: Let $\boldsymbol{\delta} \in \mathbb{R}^n$ such that $\widetilde{\mathbf{U}}\boldsymbol{\delta} = \mathbf{0}$ and $(\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta} \neq 0$. To see that such a $\boldsymbol{\delta}$ always exists, let $p = \text{rank}(\widetilde{\mathbf{U}})$. Observe that $p < n - 1$ as $d < \text{affdim}(\mathbf{L}) - 1$ and $p \leq d$. Let $\mathbf{v}_1, \cdots, \mathbf{v}_{n-p} \in \mathbb{R}^n$ be an orthonormal basis of the null space of $\widetilde{\mathbf{U}}$. Clearly, $\text{span}(\mathbf{u}_1, \cdots, \mathbf{u}_d, \mathbf{1}, \mathbf{v}_1, \cdots, \mathbf{v}_{n-p}) = \mathbb{R}^n$, and therefore, $\exists \alpha_1, \cdots, \alpha_{d+1}, \beta_1, \cdots, \beta_{n-p}$ such that $\boldsymbol{\ell}_i - \boldsymbol{\ell}_j = \sum_{r=1}^d \alpha_r \mathbf{u}_r + \alpha_{d+1}\mathbf{1} + \sum_{r=1}^{n-p} \beta_r \mathbf{v}_r$. Since $\boldsymbol{\ell}_i - \boldsymbol{\ell}_j \notin \text{col}(\widetilde{\mathbf{U}}^\top)$, $\exists q \in [n-p]$ such that $\beta_q \neq 0$. Take $\boldsymbol{\delta} = \mathbf{v}_q$. By construction, $\widetilde{\mathbf{U}}\boldsymbol{\delta} = \mathbf{0}$.

Moreover,

$$\begin{aligned}(\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta} &= \sum_{r=1}^{d} \alpha_r \mathbf{u}_r^\top \mathbf{v}_q + \alpha_{d+1} \mathbf{1}^\top \mathbf{v}_q + \sum_{r=1}^{n-p} \beta_r \mathbf{v}_r^\top \mathbf{v}_q \\ &= \beta_q ||\mathbf{v}_q||_2^2 \,, \qquad \text{since } \widetilde{\mathbf{U}} \mathbf{v}_q = 0 \text{ and } \mathbf{v}_r^\top \mathbf{v}_q = 0 \; \forall r \neq q \\ &\neq 0 \,.\end{aligned}$$

Thus we have shown that $\exists \boldsymbol{\delta} \in \mathbb{R}^n$ such that $\widetilde{\mathbf{U}} \boldsymbol{\delta} = \mathbf{0}$ and $(\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta} \neq 0$. In the remainder of the proof we will assume without loss of generality that $(\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta} < 0$ (the case $(\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta} > 0$ can be treated similarly as below).

Step 4: This is the most crucial step in the proof in which we find $\mathbf{p}_1, \mathbf{p}_2$ by perturbing $\mathbf{p}$ along $\boldsymbol{\delta}$ as shown in Figure 2. We have to ensure: (1) This perturbation leads to valid probability vectors; (2) One of the perturbed vectors lands in $\mathcal{Q}_i^\ell$ and the other one lands in $\mathcal{Q}_j^\ell$.

Let a be the least positive integer such that $\forall r \in [n], |\delta_r / a| \leq \min(p_r, 1 - p_r)$, and let $\boldsymbol{\delta}' := \boldsymbol{\delta}/a$. Next, let b be the least positive integer such that $\forall t \neq i, j$,

$$\mathbf{p}^\top (\boldsymbol{\ell}_t - \boldsymbol{\ell}_i) > (\boldsymbol{\delta}'/b)^\top (\boldsymbol{\ell}_i - \boldsymbol{\ell}_t) \,, \tag{7}$$

$$\mathbf{p}^\top (\boldsymbol{\ell}_t - \boldsymbol{\ell}_j) > (\boldsymbol{\delta}'/b)^\top (\boldsymbol{\ell}_t - \boldsymbol{\ell}_j) \,, \tag{8}$$

and define $\boldsymbol{\delta}'' := \boldsymbol{\delta}'/b$. Now, $\widetilde{\mathbf{U}} \boldsymbol{\delta}'' = 0$ and $(\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta}'' \neq 0$. Define $\mathbf{p}_1 := \mathbf{p} + \boldsymbol{\delta}''$ and $\mathbf{p}_2 := \mathbf{p} - \boldsymbol{\delta}''$. We can see that $p_{1r} \geq 0$ and $p_{2r} \geq 0 \; \forall r \in [n]$. Also,

$$\begin{aligned}\mathbf{1}^\top \mathbf{p}_1 &= \mathbf{1}^\top \mathbf{p} + \mathbf{1}^\top \boldsymbol{\delta}'' \\ &= 1 + 0 \,, \qquad \text{since } \widetilde{\mathbf{U}} \boldsymbol{\delta}'' = 0 \text{ and } \mathbf{1} \in \text{col}(\widetilde{\mathbf{U}}^\top) \\ &= 1 \,.\end{aligned}$$

Similarly, $\mathbf{1}^\top \mathbf{p}_2 = 1$. Therefore, $\mathbf{p}_1$ and $\mathbf{p}_2$ are valid probability vectors in Δ_n.

Now, we claim that $\mathbf{p}_1 \in \mathcal{Q}_i^\ell$ and $\mathbf{p}_1 \notin \mathcal{Q}_t^\ell \; \forall t \neq i$. We have,

$$\begin{aligned}(\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \mathbf{p}_1 &= (\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \mathbf{p} + (\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta}'' \\ &= 0 + (\boldsymbol{\ell}_i - \boldsymbol{\ell}_j)^\top \boldsymbol{\delta}'' \,, \qquad \text{since } \mathbf{p} \in \mathcal{Q}_i^\ell \cap \mathcal{Q}_j^\ell \\ &< 0 \,.\end{aligned}$$

This gives $\mathbf{p}_1 \notin \mathcal{Q}_j^\ell$. Moreover, $\forall t \neq i, j$, we have

$$\begin{aligned}(\boldsymbol{\ell}_i - \boldsymbol{\ell}_t)^\top \mathbf{p}_1 &= \mathbf{p}^\top (\boldsymbol{\ell}_i - \boldsymbol{\ell}_t) + \boldsymbol{\delta}''^\top (\boldsymbol{\ell}_i - \boldsymbol{\ell}_t) \\ &< 0 \,, \qquad \text{by Eq. (7) .}\end{aligned}$$

Thus $\mathbf{p}_1 \in \mathcal{Q}_i^\ell$ and $\mathbf{p}_1 \notin \mathcal{Q}_t^\ell \; \forall t \neq i$. Similarly, $\mathbf{p}_2 \in \mathcal{Q}_j^\ell$ and $\mathbf{p}_2 \notin \mathcal{Q}_t^\ell \; \forall t \neq j$. Therefore, $\text{Opt}(\ell, \mathbf{p}_1) \cap \text{Opt}(\ell, \mathbf{p}_2) = \emptyset$. Moreover,

$$\begin{aligned}\mathbf{U} \mathbf{p}_1 &= \mathbf{U} \mathbf{p} + \mathbf{U} \boldsymbol{\delta}'' \\ &= \mathbf{U} \mathbf{p} \,, \qquad \text{since } \mathbf{U} \boldsymbol{\delta}' = 0 \\ &= \mathbf{U} \mathbf{p}_2 \,.\end{aligned}$$

This gives us a contradiction since Γ will not be able to differentiate between $\mathbf{p}_1$ and $\mathbf{p}_2$, even though the optimal predictions for them with respect to ℓ are different; in particular, we get $\text{pred}(\Gamma(\mathbf{p}_1)) = \text{pred}(\mathbf{U}\mathbf{p}_1) = \text{pred}(\mathbf{U}\mathbf{p}_2) = \text{pred}(\Gamma(\mathbf{p}_2))$, and so we cannot have $\text{pred}(\Gamma(\mathbf{p}_1)) \in \text{Opt}(\ell, \mathbf{p}_1)$ and $\text{pred}(\Gamma(\mathbf{p}_2)) \in \text{Opt}(\ell, \mathbf{p}_2)$, i.e. (Γ, pred) cannot be ℓ-calibrated over Δ_n. Therefore we must have $d > \text{affdim}(\mathbf{L}) - 1$. ■

Appendix E. Proof of Theorem 9

Proof (of Theorem 9) Suppose Γ is directly elicitable via a convex proper scoring rule, and let $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ be a convex strictly proper scoring rule for Γ. We will show that $\dim(\Gamma^{-1}(\mathbf{u})) \geq n - d - 1 \; \forall \mathbf{u} \in \Gamma(\text{relint}(\Delta_n))$.

Let $\mathbf{p} \in \text{relint}(\Delta_n)$, and let $\mathbf{u}^* = \Gamma(\mathbf{p})$. Since ψ is strictly proper for Γ, we have

$$\mathbf{u}^* = \text{argmin}_{\mathbf{u} \in \mathbb{R}^d} \mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u})] \,.$$

Moreover, since ψ is convex, we have

$$\mathbf{0} \in \partial(\mathbf{E}_{Y \sim \mathbf{p}}[\psi(Y, \mathbf{u}^*)]) = \sum_{y=1}^{n} p_y \partial\psi(y, \mathbf{u}^*) \,,$$

where $\partial\psi(y, \mathbf{u}^*)$ denotes the set of subdifferentials of $\psi(y, \mathbf{u})$ at $\mathbf{u}^*$ (if $\psi(y, \cdot)$ is differentiable, each such set is a singleton). Therefore for each $y \in [n]$, $\exists \mathbf{w}_y \in \partial\psi(y, \mathbf{u}^*)$ such that $\sum_{y=1}^{n} p_y \mathbf{w}_y = \mathbf{0}$. Let $\mathbf{A} = [\mathbf{w}_1 \cdots \mathbf{w}_n] \in \mathbb{R}^{d \times n}$, and let

$$\mathcal{H} = \{\mathbf{q} \in \Delta_n : \mathbf{A}\mathbf{q} = 0\} = \{\mathbf{q} \in \mathbb{R}^n : \mathbf{A}\mathbf{q} = \mathbf{0}, \mathbf{1}^\top \mathbf{q} = 1, -\mathbf{q} \leq \mathbf{0}\} \,,$$

where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector. We have $\mathbf{p} \in \mathcal{H}$, and also $-\mathbf{p} < \mathbf{0}$. Therefore, by Lemma 14 of Ramaswamy and Agarwal (2012), we have

$$\mu_{\mathcal{H}}(\mathbf{p}) \geq n - (d + 1) \,,$$

where $\mu_{\mathcal{H}}(\mathbf{p})$ is the feasible subspace dimension of $\mathcal{H}$.[4] Now,

$$\begin{aligned} \mathbf{q} \in \mathcal{H} \implies \mathbf{A}\mathbf{q} = \mathbf{0} &\implies \mathbf{0} \in \sum_{y=1}^{n} q_y \partial\psi(y, \mathbf{u}^*) \\ &\implies \mathbf{u}^* = \text{argmin}_{\mathbf{u} \in \mathbb{R}^d} \mathbf{E}_{Y \sim \mathbf{q}}[\psi(Y, \mathbf{u})] \\ &\implies \Gamma(\mathbf{q}) = \mathbf{u}^* \,, \end{aligned}$$

which gives $\mathcal{H} \subseteq \Gamma^{-1}(\mathbf{u}^*)$, and therefore,

$$\dim(\Gamma^{-1}(\mathbf{u}^*)) \;\geq\; \mu_{\Gamma^{-1}(\mathbf{u}^*)}(\mathbf{p}) \;\geq\; \mu_{\mathcal{H}}(\mathbf{p}) \;\geq\; n - (d + 1) \,.$$

Since $\mathbf{p} \in \text{relint}(\Delta_n)$ was arbitrary, the result follows. ■

4. The feasible subspace dimension of a convex set $\mathcal{C}$ at $\mathbf{p} \in \mathcal{C}$ is defined as the dimension of the subspace $\mathcal{F}_{\mathcal{C}}(\mathbf{p}) \cup (-\mathcal{F}_{\mathcal{C}}(\mathbf{p}))$, where $\mathcal{F}_{\mathcal{C}}(\mathbf{p})$ is the cone of feasible directions of $\mathcal{C}$ at $\mathbf{p}$ (Ramaswamy and Agarwal, 2012).