JMLR: Workshop and Conference Proceedings vol 35:1–18, 2014

On the Consistency of Output Code Based Learning Algorithms for Multiclass Learning Problems

Harish G. Ramaswamy HARISH_GURUP@CSA.IISC.ERNET.IN
Balaji Srinivasan Babu BALAJI@ECE.IISC.ERNET.IN
Shivani Agarwal SHIVANI@CSA.IISC.ERNET.IN
Indian Institute of Science, Bangalore 560012, India

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Australian National University and National ICT Australia, Canberra, ACT 2601, Australia

Abstract

A popular approach to solving multiclass learning problems is to reduce them to a set of binary classification problems through some output code matrix: the widely used one-vs-all and all-pairs methods, and the error-correcting output code methods of Dietterich and Bakiri (1995), can all be viewed as special cases of this approach. In this paper, we consider the question of statistical consistency of such methods. We focus on settings where the binary problems are solved by minimizing a binary surrogate loss, and derive general conditions on the binary surrogate loss under which the one-vs-all and all-pairs code matrices yield consistent algorithms with respect to the multiclass 0-1 loss. We then consider general multiclass learning problems defined by a general multiclass loss, and derive conditions on the output code matrix and binary surrogates under which the resulting algorithm is consistent with respect to the target multiclass loss. We also consider *probabilistic* code matrices, where one reduces a multiclass problem to a set of *class probability labeled* binary problems, and show that these can yield benefits in the sense of requiring a smaller number of binary problems to achieve overall consistency. Our analysis makes interesting connections with the theory of proper composite losses (Buja et al., 2005; Reid and Williamson, 2010); these play a role in constructing the right 'decoding' for converting the predictions on the binary problems to the final multiclass prediction. To our knowledge, this is the first work that comprehensively studies consistency properties of output code based methods for multiclass learning.

Keywords: Multiclass learning, output codes, consistency, one-versus-all, all-pairs, error-correcting output codes, proper composite losses.

1. Introduction

Output code based methods are a popular approach to multiclass learning (Sejnowski and Rosenberg, 1987; Dietterich and Bakiri, 1995; Schapire, 1997; Schapire and Singer, 1999; Guruswami and Sahai, 1999; Allwein et al., 2000; Crammer and Singer, 2002; Langford and Beygelzimer, 2005). Given a multiclass problem with label space $\mathcal{Y} = \{1, \ldots, n\}$ and prediction space $\widehat{\mathcal{Y}} = \{1, \ldots, k\}$, an output code method constructs a set of some d binary classification problems via a code matrix $\mathbf{M} \in \{-1, 0, +1\}^{n \times d}$: for each $j \in [d]$, the j-th binary problem is constructed by replacing multiclass labels $y \in \mathcal{Y}$ with binary labels $M_{yj} \in \{\pm 1\}$ (examples with labels y such that $M_{yj} = 0$ are ignored). A binary classifier is then trained for each of these d binary problems; assuming use of a real-valued classification algorithm that provides confidence-rated predictions, for any new instance x, this yields a vector of d real-valued predictions $\mathbf{f}(x) = (f_1(x), \ldots, f_d(x)) \in \mathbb{R}^d$. As a final step, this vector of d predictions on the binary problems is 'decoded' via a mapping decode : $\mathbb{R}^d \to \widehat{\mathcal{Y}}$ into

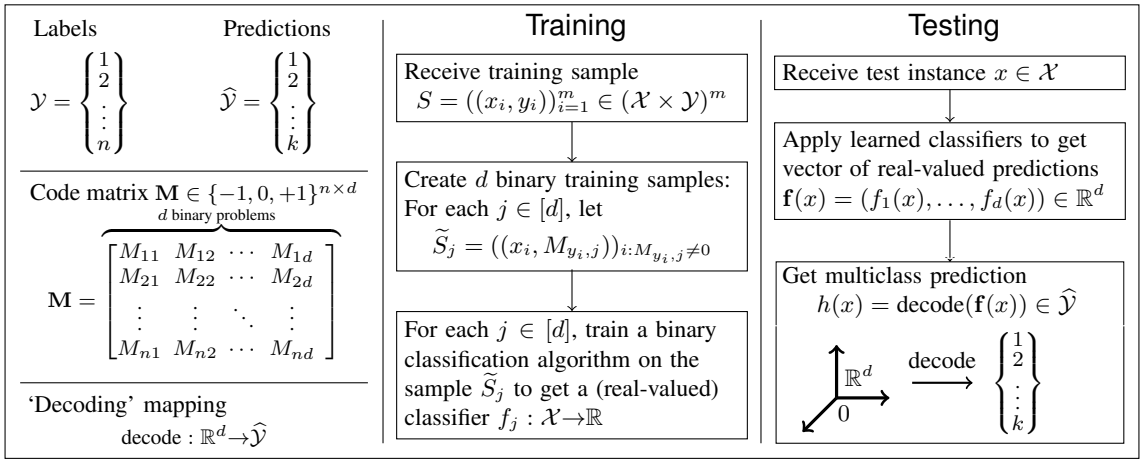

Figure 1: Operation of a typical output code based algorithm on a multiclass learning problem with n labels and k possible predictions (often, $k = n$, but this is not always the case). The code matrix $\mathbf{M} \in \{-1, 0, 1\}^{n \times d}$ is used to reduce the problem into d binary problems; for a new instance x, the outputs of the d binary classifiers are 'decoded' into one of the k possible multiclass predictions.

a multiclass prediction: $h(x) = \text{decode}(\mathbf{f}(x)) \in \widehat{\mathcal{Y}}$. See Figure 1 for a summary. As an example, for the multiclass 0-1 problem with $\widehat{\mathcal{Y}} = \mathcal{Y} = \{1, \ldots, n\}$, the widely-used *one-vs-all* method constructs n binary problems using an $n \times n$ code matrix $\mathbf{M}$ with $M_{yy} = 1 \; \forall y$ and $M_{yj} = -1 \; \forall j \neq y$; here one typically decodes via $h(x) \in \text{argmax}_{j \in [n]} f_j(x)$.

A fundamental question that arises is: under what conditions is such an output code based method guaranteed to be statistically consistent for the target multiclass problem, i.e. to converge to a Bayes optimal model in the limit of infinite data? Statistical consistency is a basic requirement for a good learning algorithm, and there has been much interest and progress in recent years in understanding consistency properties of various commonly used algorithms that minimize various types of (usually convex) surrogate losses, for problems ranging from binary and multiclass 0-1 classification to subset ranking and multilabel classification, and to some extent, general multiclass problems (Zhang, 2004a,b; Bartlett et al., 2006; Tewari and Bartlett, 2007; Steinwart, 2007; Cossock and Zhang, 2008; Xia et al., 2008; Duchi et al., 2010; Ravikumar et al., 2011; Buffoni et al., 2011; Gao and Zhou, 2011; Calauzènes et al., 2012; Lan et al., 2012; Ramaswamy and Agarwal, 2012). To the best of our knowledge, however, consistency properties of output code based learning algorithms are not yet understood, except in a few special cases such as for the one-vs-all code based method with binary classification algorithms minimizing margin-based surrogates (Zhang, 2004b), and for methods using Hadamard codes together with reductions to (implicitly) an uncountably infinite number of importance-weighted binary classification problems (Langford and Beygelzimer, 2005).

In this paper, we study consistency properties of output code based methods for general multiclass learning problems defined by a general loss matrix. We focus on settings where the induced binary problems are solved by minimizing a binary surrogate loss. This allows us to view the overall method as minimizing a certain multiclass surrogate loss composed of the code matrix and the binary surrogate, and to then appeal to the framework of calibrated surrogates for studying consistency of surrogate-minimizing algorithms (Bartlett et al., 2006; Tewari and Bartlett, 2007; Ramaswamy

and Agarwal, 2012). The notion of proper composite losses that have been studied recently in the context of class probability estimation Buja et al. (2005); Reid and Williamson (2010) plays an important role in our study: in particular, we show consistency of output code methods using strictly proper composite binary surrogates. Indeed, the decoding that achieves consistency for general multiclass losses effectively uses the link function associated with the binary proper composite surrogate, allowing one to effectively estimate class probabilities for each of the binary problems; these are then used to construct the right multiclass prediction for the target problem.

1.1. Related Work

As noted above, there has been limited work on studying consistency of output code based methods. We summarize here some of the main studies that have looked at obtaining various types of formal guarantees on the performance of such methods, and point out differences from our work.

Early work on analyzing output code based methods focused on bounding the training error of the overall multiclass algorithm in terms of the training errors of the underlying binary algorithm on the component binary problems; this was done for both error-correcting code matrices, and code matrices effectively produced by certain boosting methods (Schapire, 1997; Guruswami and Sahai, 1999; Allwein et al., 2000). For the boosting methods, by standard VC-dimension/margin-based arguments, these results could then also be used to bound the multiclass generalization error in terms of binary training errors. In most of these studies, the focus was on the multiclass 0-1 loss; here we study general multiclass losses, and consider the question of whether the corresponding multiclass generalization error converges to the Bayes optimal.

Crammer and Singer (2002) considered a converse type of problem: given a fixed set of d binary classifiers $h_j : \mathcal{X} \to \{\pm 1\}$ ($j \in [d]$), how should one design a binary code matrix $\mathbf{M} \in \{\pm 1\}^{n \times d}$ so that the result of a Hamming decoding, given by $h(x) \in \operatorname{argmin}_{y \in [n]} \sum_{j=1}^{d} \mathbf{1}(h_j(x) \neq M_{yj})$, minimizes the multiclass 0-1 training error on a given training sample? They showed this problem is NP-hard, and then considered a continuous analogue with real-valued classifiers $f_j : \mathcal{X} \to \mathbb{R}$, where the goal is to design a continuous code matrix $\mathbf{M} \in \mathbb{R}^{n \times d}$ such that again a Hamming-type decoding (using an appropriate similarity measure between real vectors) minimizes the multiclass 0-1 training error; in this case, they gave efficient algorithms for designing such a code matrix. In our case, we do not fix the classifiers, but rather assume these are to be learned by some binary algorithm; we also do not fix the decoding, and allow this to be selected based on the target multiclass loss (which need not be the 0-1 loss). The question of interest to us is then the following: what types of code matrices $\mathbf{M}$, together with a suitable binary algorithm and suitable decoding, will allow us to obtain statistical convergence guarantees with respect to the target multiclass loss?

Langford and Beygelzimer (2005) considered a very general type of multiclass problem, where the loss vector that assigns losses for various predictions on any example is randomly drawn; this contains the multiclass problems we consider, which are defined by a loss matrix, as a special case. They considered the use of binary Hadamard code matrices together with a Hamming decoding to reduce such a general multiclass problem to a set of d parameterized families of importance-weighted binary classification problems, each family effectively containing an uncountably infinite number of such individual problems. Their work focussed on bounding the multiclass regret of the overall algorithm in terms of the average binary regret; their result implies in particular that if the binary algorithms used are consistent, then the overall algorithm is consistent. The actual reduction and analysis are quite complex; as we shall see, for the types of problems we are interested in,

by allowing flexibility in designing the code matrix and corresponding decoding, and focusing our attention on surrogate-minimizing binary algorithms, we get intuitive and easy-to-use reductions, with correspondingly simple analyses that build on recent work on calibrated surrogates and proper composite losses.

1.2. Our Contributions

We derive several new results; our main contributions are as follows:

- We give general conditions for the one-vs-all and all-pairs methods with binary surrogates to be consistent for the multiclass 0-1 loss. The one-vs-all result generalizes a result of Zhang (2004b), which was restricted to margin-based surrogates.
- We show that given an arbitrary target multiclass loss matrix $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ (whose (y,t)-th element prescribes the loss incurred on predicting $t \in [k]$ when the true label is $y \in [n]$), any code matrix $\mathbf{M} \in \{\pm 1\}^{n \times d}$ such that the column space of $\mathbf{L}$ is contained in the column space of $\left[\mathbf{M}, \mathbf{e}_n\right]$ (where $\mathbf{e}_n$ is the $n \times 1$ all-ones vector), together with suitable binary surrogates and suitable decoding, yields an output code based algorithm that is consistent for $\mathbf{L}$.
- In general, given a loss matrix $\mathbf{L} \in \mathbb{R}_+^{n \times k}$, finding a binary code matrix satisfying the above condition can require a large number of columns. We introduce the notion of *probabilistic* code matrices, where one constructs *class probability labeled* binary problems, and show that one can always construct a probabilistic code matrix $\mathbf{M} \in [0,1]^{n \times d}$ for which the number of columns (and induced binary problems) d is at most the *rank* of $\mathbf{L}$, and which together with suitable binary surrogates and suitable decoding, also yields a consistent algorithm for $\mathbf{L}$.[1]

Organization. Section 2 contains preliminaries. Section 3 considers consistency of one-vs-all and all-pairs methods for the multiclass 0-1 loss. Section 4 considers consistency of output code based methods for general multiclass losses. Section 5 concludes with a brief discussion.

2. Preliminaries and Background

Setup. We are interested in multiclass learning problems with an instance space $\mathcal{X}$, label space $\mathcal{Y} = [n] = \{1, \ldots, n\}$, and prediction space $\widehat{\mathcal{Y}} = [k] = \{1, \ldots, k\}$. Often, $\widehat{\mathcal{Y}} = \mathcal{Y}$, but this is not always the case (e.g. in some subset ranking problems, $\mathcal{Y}$ contains preference graphs over a set of objects, while $\widehat{\mathcal{Y}}$ contains permutations over those objects). Given a training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$, where examples (x_i, y_i) are drawn i.i.d. from some underlying probability distribution D on $\mathcal{X} \times \mathcal{Y}$, the goal is to learn a prediction model $h : \mathcal{X} \to \widehat{\mathcal{Y}}$.

The performance of a model $h : \mathcal{X} \to \widehat{\mathcal{Y}}$ is measured via a *loss function* $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$ (where $\mathbb{R}_+ = [0, \infty)$), which assigns a penalty $\ell(y,t)$ for predicting $t \in \widehat{\mathcal{Y}}$ when the true label is $y \in \mathcal{Y}$; or equivalently, via a *loss matrix* $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ with elements $\ell_{yt} = \ell(y,t)\ \forall y \in [n], t \in [k]$. For each $t \in [k]$, we will denote the t-th column of the loss matrix $\mathbf{L}$ by $\boldsymbol{\ell}_t \in \mathbb{R}_+^n$. The goal is to learn a model h with low expected loss or ℓ-error: $\mathrm{er}_D^\ell[h] = \mathbf{E}_{(x,y)\sim D}[\ell(y, h(x))]$. The best possible error one can hope to achieve is the Bayes optimal ℓ-error: $\mathrm{er}_D^{\ell,*} = \inf_{h:\mathcal{X}\to\widehat{\mathcal{Y}}} \mathrm{er}_D^\ell[h]$.

An ideal learning algorithm will satisfy the property that as it receives increasingly large training samples, the error of the prediction models it learns converges in probability to the Bayes optimal error. Formally, let $h_S : \mathcal{X} \to \widehat{\mathcal{Y}}$ denote the prediction model learned by an algorithm from a training

1. Note that a probabilistic code matrix here is not a random or stochastic matrix, but simply a matrix with $[0,1]$ valued entries.

sample $S \sim D^m$; then the learning algorithm is *consistent* w.r.t. ℓ and D if $\mathrm{er}_D^{\ell}[h_S] \xrightarrow{P} \mathrm{er}_D^{\ell,*}$, i.e. if $\forall \epsilon > 0$, $\mathbf{P}_{S\sim D^m}\left(\mathrm{er}_D^{\ell}[h_S] \geq \mathrm{er}_D^{\ell,*} + \epsilon\right) \to 0$ as $m \to \infty$. If this holds for all distributions D, we will say the algorithm is (universally) ℓ-consistent. It is known that algorithms minimizing the ℓ-error on the training sample, $\sum_{i=1}^m \ell(y_i, h(x_i))$, over a suitable class of functions $\mathcal{H}_m$, are ℓ-consistent; however this is typically a computationally hard optimization problem due to the discrete nature of the loss. Consequently, one often minimizes a continuous (convex) *surrogate* loss instead.

Multiclass Surrogates and Calibration. Let $\mathcal{C} \subseteq \overline{\mathbb{R}}^d$ for some $d \in \mathbb{Z}_+$ (where $\overline{\mathbb{R}} = [-\infty, \infty]$). Consider a multiclass surrogate loss $\psi : \mathcal{Y} \times \mathcal{C} \to \overline{\mathbb{R}}_+$ (where $\overline{\mathbb{R}}_+ = [0, \infty]$) and a mapping decode : $\mathcal{C} \to \widehat{\mathcal{Y}}$. Given a training sample S as above, a surrogate loss minimizing algorithm using the pair (ψ, decode) learns a function $\mathbf{f}_S : \mathcal{X} \to \mathcal{C}$ by minimizing $\sum_{i=1}^m \psi(y_i, \mathbf{f}(x_i))$ over some suitable class of functions $\mathcal{F}_m$, and returns a prediction model $h_S : \mathcal{X} \to \widehat{\mathcal{Y}}$ defined by $h_S(x) = \text{decode}(\mathbf{f}_S(x))$. Usually, one selects $\mathcal{C}$ to be a convex set and $\{\mathcal{F}_m\}$ to be a sequence of convex class of functions to facilitate efficient minimization. Under suitable conditions on $\{\mathcal{F}_m\}$, such an algorithm is ψ-*consistent*, i.e. the ψ-errors of the models $\mathbf{f}_S$ it learns, defined for a model $\mathbf{f} : \mathcal{X} \to \mathcal{C}$ as $\mathrm{er}_D^{\psi}[\mathbf{f}] = \mathbf{E}_{(x,y)\sim D}[\psi(y, \mathbf{f}(x))]$, converge in probability to the Bayes ψ-error $\mathrm{er}_D^{\psi,*} = \inf_{\mathbf{f}:\mathcal{X}\to\mathcal{C}} \mathrm{er}_D^{\psi}[\mathbf{f}]$. The notion of *calibration* links this property to consistency w.r.t. the target loss ℓ.

Specifically, let Δ_n denote the probability simplex in $\mathbb{R}^n$, i.e. $\Delta_n = \{\mathbf{p} \in \mathbb{R}_+^n : \sum_{i=1}^n p_i = 1\}$, and let $L_\psi : \Delta_n \times \mathcal{C} \to \overline{\mathbb{R}}_+$ be defined as

$$L_\psi(\mathbf{p}, \mathbf{u}) = \mathbf{E}_{y\sim\mathbf{p}}\left[\psi(y, \mathbf{u})\right] = \sum_{y=1}^n p_y\, \psi(y, \mathbf{u}) = \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u})\,,$$

where we have denoted $\boldsymbol{\psi}(\mathbf{u}) = \left(\psi(1, \mathbf{u}), \ldots, \psi(n, \mathbf{u})\right)^\top \in \overline{\mathbb{R}}_+^n$. Then the pair (ψ, decode) is said to be ℓ-*calibrated* if $\forall \mathbf{p} \in \Delta_n$,

$$\inf_{\mathbf{u}\in\mathcal{C}:\text{decode}(\mathbf{u})\notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} L_\psi(\mathbf{p}, \mathbf{u}) \;>\; \inf_{\mathbf{u}\in\mathcal{C}} L_\psi(\mathbf{p}, \mathbf{u})\,.$$

If this does not hold for any mapping decode, then we simply say the surrogate ψ is not ℓ-calibrated. The following result explains why calibration is a useful property:

Theorem 1 ((Tewari and Bartlett, 2007; Zhang, 2004b; Ramaswamy and Agarwal, 2012)) *Let $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, $\psi : \mathcal{Y} \times \mathcal{C} \to \overline{\mathbb{R}}_+$, and* decode $: \mathcal{C} \to \widehat{\mathcal{Y}}$. *Then (ψ, decode) is ℓ-calibrated iff for all distributions D on $\mathcal{X} \times \mathcal{Y}$ and all sequences of (vector) functions $\mathbf{f}_m : \mathcal{X} \to \mathcal{C}$*

$$\mathrm{er}_D^{\psi}[\mathbf{f}_m] \to \mathrm{er}_D^{\psi,*} \quad \textit{implies} \quad \mathrm{er}_D^{\ell}[\text{decode} \circ \mathbf{f}_m] \to \mathrm{er}_D^{\ell,*}\,.$$

Thus, if (ψ, decode) is ℓ-calibrated, then a surrogate-minimizing algorithm using (ψ, decode) as above (with suitable function classes $\mathcal{F}_m$) is also ℓ-consistent.

Binary Surrogates and Proper Composite Losses. A binary surrogate loss operating on $\mathcal{V} \subseteq \overline{\mathbb{R}}$ can be represented by its *partial losses* $\phi_1 : \mathcal{V} \to \overline{\mathbb{R}}_+$ and $\phi_{-1} : \mathcal{V} \to \overline{\mathbb{R}}_+$. Given a training sample with binary labels, $\widetilde{S} = ((x_1, \widetilde{y}_1), \ldots, ((x_m, \widetilde{y}_m)) \in (\mathcal{X} \times \{\pm 1\})^m$, a surrogate-minimizing binary algorithm using (ϕ_1, ϕ_{-1}) learns a function $f_{\widetilde{S}} : \mathcal{X} \to \mathcal{V}$ by minimizing $\sum_{i=1}^m \phi_{\widetilde{y}_i}(f(x_i))$ over a suitable class of functions; if a binary classification is desired, one usually decodes this to a binary classification in $\{\pm 1\}$ via the sign$(\cdot)$ function: $h_{\widetilde{S}}(x) = \text{sign}(f_{\widetilde{S}}(x))$. A binary surrogate (ϕ_1, ϕ_{-1}) is said to be *margin-based* if $\exists \phi : \mathcal{V} \to \overline{\mathbb{R}}_+$ such that $\phi_1(v) = \phi(v)$ and $\phi_{-1}(v) = \phi(-v)$ $\forall v \in \mathcal{V}$. Common examples of margin-based binary surrogates include the hinge, logistic, and exponential losses, defined over $\mathcal{V} = \overline{\mathbb{R}}$ via

Loss	$\mathcal{V}$	$\phi_1(v)$	$\phi_{-1}(v)$	$\lambda(\eta)$	$\lambda^{-1}(v)$
Logistic	$\overline{\mathbb{R}}$	$\ln(1+e^{-v})$	$\ln(1+e^{v})$	$\ln\left(\frac{\eta}{1-\eta}\right)$	$\frac{1}{1+e^{-v}}$
Exponential	$\overline{\mathbb{R}}$	e^{-v}	e^{v}	$\frac{1}{2}\ln\left(\frac{\eta}{1-\eta}\right)$	$\frac{1}{1+e^{-2v}}$
Least-squares	$[-1,1]$	$(v-1)^2$	$(v+1)^2$	$2\eta-1$	$\frac{v+1}{2}$

Table 1: Some well-known strictly proper composite binary surrogates, with their partial losses ϕ_1, ϕ_{-1}, link functions λ and corresponding inverse links λ^{-1}.

$$\phi^{\text{hinge}}(v) = \max(1-v, 0)\,; \qquad \phi^{\log}(v) = \ln(1+e^{-v})\,; \qquad \phi^{\exp}(v) = e^{-v}\,;$$

and the least-squares loss, which we shall define on $\mathcal{V} = [-1, 1]$ via

$$\phi^{\text{sq}}(v) = (1-v)^2\,.$$

Note that non-margin-based binary surrogates allow for the possibility of asymmetric penalization with respect to positive and negative labels.

A binary surrogate (ϕ_1, ϕ_{-1}) is said to be *proper composite* if there exists a strictly increasing *link function* $\lambda : [0,1] \to \mathcal{V}$ such that $\forall \eta \in [0,1]$,

$$\lambda(\eta) \in \operatorname{argmin}_{v \in \mathcal{V}} \; \eta\,\phi_1(v) + (1-\eta)\,\phi_{-1}(v)\,,$$

and *strictly proper composite* if in addition the above minimizer is unique for all η. As discussed in (Buja et al., 2005; Reid and Williamson, 2010), strictly proper composite losses have the property that their minimization yields accurate (Fisher consistent) class probability estimates; specifically, if $f_{\widetilde{S}} : \mathcal{X} \to \mathcal{V}$ is obtained by minimizing (ϕ_1, ϕ_{-1}) as above, then accurate class probability estimates are obtained via $\widehat{\eta}_{\widetilde{S}}(x) = \lambda^{-1}(f_{\widetilde{S}}(x))$. The logistic, exponential and least-squares losses above are all strictly proper composite and have a continuous inverse link function (see Table 1); the hinge loss is not proper composite.

Output Codes and Code-Based Multiclass Surrogates. Given a multiclass learning problem with label space $\mathcal{Y} = [n]$ and $\widehat{\mathcal{Y}} = [k]$ as above, an output code based method constructs a set of d binary problems, for some $d \in \mathbb{Z}_+$, via a code matrix $\mathbf{M} \in \{-1, 0, +1\}^{n \times d}$. Specifically, given a multiclass training sample $S = ((x_1, y_1), \ldots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$, one uses $\mathbf{M}$ to construct d binary training samples as follows: for each $j \in [d]$, $\widetilde{S}_j = ((x_i, M_{y_i,j}))_{i:M_{y_i,j} \neq 0}$. Using these binary training samples, one learns d binary classifiers; assuming use of a real-valued binary classification algorithm that provides confidence-rated predictions, this gives d real-valued classifiers $f_j : \mathcal{X} \to \mathcal{V}$, $j \in [d]$. Given a new instance x, one then uses a mapping decode : $\overline{\mathbb{R}}^d \to \widehat{\mathcal{Y}}$ to decode the d real-valued predictions on x into a multiclass prediction: $h(x) = \text{decode}(f_1(x), \ldots, f_d(x))$.

Now, if the binary classification algorithm used minimizes a binary surrogate (ϕ_1, ϕ_{-1}) defined on some space $\mathcal{V} \subseteq \overline{\mathbb{R}}$, then the resulting output code based algorithm can be viewed as learning a function $\mathbf{f}_S : \mathcal{X} \to \mathcal{V}^d$ by minimizing the multiclass surrogate loss $\psi : \mathcal{Y} \times \mathcal{V}^d \to \overline{\mathbb{R}}_+$ defined as

$$\psi(y, \mathbf{u}) = \sum_{j=1}^{d} \Big(\mathbf{1}(M_{yj} = 1)\,\phi_1(u_j) + \mathbf{1}(M_{yj} = -1)\,\phi_{-1}(u_j) \Big) \tag{1}$$

on the original multiclass training sample S. Thus, given a target loss $\ell : \mathcal{Y} \times \widehat{\mathcal{Y}} \to \mathbb{R}_+$, the overall output code based algorithm using code matrix $\mathbf{M}$, binary surrogate (ϕ_1, ϕ_{-1}), and mapping decode

as above (assuming suitable function classes when minimizing the binary surrogate) is ℓ-consistent if and only if the pair (ψ, decode) above is ℓ-calibrated! In what follows, we will refer to the surrogate loss ψ in Eq. (1) as the $(\mathbf{M}, \phi_1, \phi_{-1})$ *code-based surrogate*; when (ϕ_1, ϕ_{-1}) is a margin-based surrogate defined via $\phi : \mathcal{V} \to \overline{\mathbb{R}}_+$, we will simply refer to it as the $(\mathbf{M}, \phi)$ *code-based surrogate*.

3. Consistency of One-vs-All and All-Pairs Methods for Multiclass 0-1 Loss

We start by considering a setting where $\widehat{\mathcal{Y}} = \mathcal{Y} = [n]$, and the target loss of interest is the multiclass 0-1 loss $\ell^{0\text{-}1} : [n] \times [n] \to \mathbb{R}_+$ given by

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

We consider consistency of the widely used one-vs-all and all-pairs code matrices in this setting.

3.1. One-vs-All Code Matrix

The one-vs-all method uses an $n \times n$ binary code matrix $\mathbf{M}^{\text{OvA}} \in \{\pm 1\}^{n \times n}$ defined as follows:

$$M^{\text{OvA}}_{yj} = \begin{cases} 1 & \text{if } y = j \\ -1 & \text{otherwise} \end{cases} \qquad \forall y, j \in [n] \,.$$

The following result establishes $\ell^{0\text{-}1}$-consistency of the one-vs-all method with any strictly proper composite binary surrogate with a continuous inverse link.

Theorem 2 *Let (ϕ_1, ϕ_{-1}) be a strictly proper composite binary surrogate defined over an interval $\mathcal{V} \subseteq \overline{\mathbb{R}}$, with a continuous inverse link function $\lambda^{-1} : \mathcal{V} \to [0, 1]$. Define* decode $: \mathcal{V}^n \to [n]$ *as*

$$\text{decode}(\mathbf{u}) \in \text{argmax}_{j \in [n]} \, \lambda^{-1}(u_j) \,.$$

Then the $(\mathbf{M}^{\text{OvA}}, \phi_1, \phi_{-1})$ code-based surrogate together with the mapping decode *is $\ell^{0\text{-}1}$-calibrated.*

The proof of Theorem 2 follows directly from a more general result we prove later (Theorem 6). Intuitively, by construction of $\mathbf{M}^{\text{OvA}}$ and strict properness of (ϕ_1, ϕ_{-1}), as the sample size m increases, for any instance x the estimate $\widehat{\eta}_{\widetilde{S}_j}(x) = \lambda^{-1}(f_{\widetilde{S}_j}(x))$ converges to the true probability $\mathbf{P}(y = j \,|\, x)$ of the label being j; the above decoding effectively selects the most probable class, which minimizes the 0-1 loss.

A similar result for margin-based binary surrogates was given by Zhang (2004b); Theorem 2 partly generalizes Zhang's result to non-margin-based binary surrogates. As with Zhang's result, Theorem 2 applies to the logistic, exponential and least-squares losses, but not to the hinge loss. The one-vs-all method with hinge loss is in fact *not* (universally) consistent for the multiclass 0-1 loss (Lee et al., 2004); we state this explicitly below (see Appendix A for a self-contained proof):

Theorem 3 *The $(\mathbf{M}^{\text{OvA}}, \phi^{\text{hinge}})$ code-based surrogate is not $\ell^{0\text{-}1}$-calibrated.*

3.2. All-Pairs Code Matrix

The all-pairs method uses an $n \times \binom{n}{2}$ code matrix $\mathbf{M}^{\text{all-pairs}} \in \{-1, 0, 1\}^{n \times \binom{n}{2}}$ defined as follows:

$$M^{\text{all-pairs}}_{y,(i,j)} = \begin{cases} 1 & \text{if } y = i \\ -1 & \text{if } y = j \\ 0 & \text{if } y \neq i \text{ and } y \neq j \end{cases} \qquad \forall y \in [n], \; (i, j) \in [n] \times [n] : i < j \,.$$

The following result establishes $\ell^{0\text{-}1}$-consistency of the all-pairs method with any strictly proper composite binary surrogate with a continuous inverse link (and suitable decoding).

Theorem 4 *Let (ϕ_1, ϕ_{-1}) be a strictly proper composite binary surrogate defined over an interval $\mathcal{V} \subseteq \overline{\mathbb{R}}$, with a continuous inverse link $\lambda^{-1} : \mathcal{V} \to [0,1]$. Define* $\text{win} : \mathcal{V} \to \{0,1\}$ *as*

$$\text{win}(v) = \mathbf{1}\left(\lambda^{-1}(v) > \frac{1}{2}\right)$$

and for each $i \in [n]$, define $\text{score}_i : \mathcal{V}^{\binom{n}{2}} \to \mathbb{R}_+$ *as*

$$\text{score}_i(\mathbf{u}) = \sum_{j:i<j} \text{win}(u_{i,j}) + \sum_{j:j<i} (1 - \text{win}(u_{j,i}))\,.$$

Define $\text{decode} : \mathcal{V}^{\binom{n}{2}} \to [n]$ *as*

$$\text{decode}(\mathbf{u}) \in \text{argmax}_{i \in [n]}\, \text{score}_i(\mathbf{u})\,.$$

Then the $(\mathbf{M}^{\text{all-pairs}}, \phi_1, \phi_{-1})$ code-based surrogate together with the mapping decode *is $\ell^{0\text{-}1}$-calibrated.*

A proof of Theorem 4 is given in Appendix B. The result clearly yields consistency of the all-pairs method with the logistic, exponential and least-squares losses. In this case, a similar consistency result also holds for the hinge loss:

Theorem 5 *Define* $\text{win}^{\text{hinge}} : \overline{\mathbb{R}} \to \{0,1\}$ *as*

$$\text{win}^{\text{hinge}}(v) = \mathbf{1}(v > 0)\,,$$

and for each $i \in [n]$, define $\text{score}_i^{\text{hinge}} : \overline{\mathbb{R}}^{\binom{n}{2}} \to \mathbb{R}_+$ *as*

$$\text{score}_i^{\text{hinge}}(\mathbf{u}) = \sum_{j:i<j} \text{win}^{\text{hinge}}(u_{i,j}) + \sum_{j:j<i} (1 - \text{win}^{\text{hinge}}(u_{j,i}))\,.$$

Define $\text{decode} : \overline{\mathbb{R}}^{\binom{n}{2}} \to [n]$ *as*

$$\text{decode}(\mathbf{u}) \in \text{argmax}_{i \in [n]}\, \text{score}_i^{\text{hinge}}(\mathbf{u})\,.$$

Then the $(\mathbf{M}^{\text{all-pairs}}, \phi^{\text{hinge}})$ code-based surrogate together with the mapping decode *is $\ell^{0\text{-}1}$-calibrated.*

The proof of Theorem 5 is similar to that of Theorem 4; details are provided in Appendix C for completeness.

4. Consistency of Output Code Based Methods for General Multiclass Losses

We now consider consistency of output code based methods for more general multiclass problems, defined by an arbitrary multiclass loss $\ell : [n] \times [k] \to \mathbb{R}_+$, or equivalently, a loss matrix $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ with (y,t)-th element $\ell_{yt} = \ell(y,t)$. Such losses can be used to describe a wide range of learning problems in practice, ranging from multiclass 0-1 classification and ordinal regression to structured prediction problems such as sequence prediction, multi-label classification, and subset ranking (Ramaswamy and Agarwal, 2012; Ramaswamy et al., 2013). We start by considering binary code matrices in Section 4.1; we then introduce *probabilistic* code matrices in Section 4.2.

4.1. Binary ($\{\pm 1\}$-Valued) Code Matrices

The following result shows that for any target multiclass loss, with suitable binary surrogates and suitable decoding, one can always design a binary code matrix such that the resulting output code based method is consistent for the target loss. Recall that given a loss matrix $\mathbf{L} \in \mathbb{R}_+^{n \times k}$, we denote by $\boldsymbol{\ell}_t \in \mathbb{R}_+^n$ the t-th column of $\mathbf{L}$.

Theorem 6 *Let $\ell : [n] \times [k] \to \mathbb{R}_+$ be any target multiclass loss, and $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ the corresponding loss matrix. Let $d \in \mathbb{Z}_+$, and let $\mathbf{M} \in \{\pm 1\}^{n \times d}$ be any code matrix such that the column space of $\mathbf{L}$ is contained in the column space of $[\mathbf{M}, \mathbf{e}_n]$ (where $\mathbf{e}_n$ is the $n \times 1$ all-ones vector). Let (ϕ_1, ϕ_{-1}) be a strictly proper composite binary surrogate defined over an interval $\mathcal{V} \subseteq \overline{\mathbb{R}}$, with a continuous inverse link $\lambda^{-1} : \mathcal{V} \to [0,1]$. Define* decode $: \mathcal{V}^d \to [k]$ *as*

$$\text{decode}(\mathbf{u}) \in \text{argmin}_{t \in [k]} \left(\lambda^{-1}(\mathbf{u})\right)^\top \boldsymbol{\beta}_t + \gamma_t \,,$$

where $\lambda^{-1}(\mathbf{u}) \in [0,1]^d$ is a vector with j-th component $\lambda^{-1}(u_j)$, and $\boldsymbol{\beta}_t \in \mathbb{R}^d$, $\gamma_t \in \mathbb{R}$ are such that [2]

$$\boldsymbol{\ell}_t = \tfrac{1}{2}(\mathbf{M} + 1)\, \boldsymbol{\beta}_t + \gamma_t \, \mathbf{e}_n \,.$$

Then the $(\mathbf{M}, \phi_1, \phi_{-1})$ code-based surrogate together with the mapping decode *is ℓ-calibrated.*

Proof Let $\widetilde{\mathbf{M}} = \frac{1}{2}(\mathbf{M} + 1) \in [0,1]^{n \times d}$. Let $\psi : [n] \times \mathcal{V}^d \to \mathbb{R}_+$ be the $(\mathbf{M}, \phi_1, \phi_{-1})$ code-based surrogate:

$$\begin{aligned} \psi(y, \mathbf{u}) &= \sum_{j=1}^{d} \Big(\mathbf{1}(M_{yj} = 1)\, \phi_1(u_j) + \mathbf{1}(M_{yj} = -1)\, \phi_{-1}(u_j) \Big) \\ &= \sum_{j=1}^{d} \Big(\widetilde{M}_{yj}\, \phi_1(u_j) + (1 - \widetilde{M}_{yj})\, \phi_{-1}(u_j) \Big) \,. \end{aligned}$$

For each $t \in [k]$, let $\boldsymbol{\beta}_t \in \mathbb{R}^d, \gamma_t \in \mathbb{R}$ be such that

$$\boldsymbol{\ell}_t = \widetilde{\mathbf{M}} \boldsymbol{\beta}_t + \gamma_t \mathbf{e}_n \,;$$

these exist since the columns of $\mathbf{L}$ are contained in the column space of $[\mathbf{M}, \mathbf{e}_n]$, which is the same as the column space of $[\widetilde{\mathbf{M}}, \mathbf{e}_n]$.

Fix any $\mathbf{p} \in \Delta_n$. Then we have

$$\mathbf{p}^\top \boldsymbol{\ell}_t = \mathbf{p}^\top (\widetilde{\mathbf{M}} \boldsymbol{\beta}_t + \gamma_t \mathbf{e}_n) = \mathbf{p}^\top \widetilde{\mathbf{M}} \boldsymbol{\beta}_t + \gamma_t \,.$$

Now, denote by $\widetilde{\mathbf{c}}_1, \ldots, \widetilde{\mathbf{c}}_d$, the columns of $\widetilde{\mathbf{M}}$. Then we have

$$\begin{aligned} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) &= \sum_{y=1}^{n} p_y\, \psi(y, \mathbf{u}) \\ &= \sum_{y=1}^{n} p_y \Big(\sum_{j=1}^{d} \widetilde{M}_{yj}\, \phi_1(u_j) + (1 - \widetilde{M}_{yj})\, \phi_{-1}(u_j) \Big) \\ &= \sum_{j=1}^{d} \Big((\mathbf{p}^\top \widetilde{\mathbf{c}}_j)\, \phi_1(u_j) + (1 - \mathbf{p}^\top \widetilde{\mathbf{c}}_j)\, \phi_{-1}(u_j) \Big) \,. \end{aligned}$$

2. Note that such $\boldsymbol{\beta}$ and γ always exist due to the column space condition on $\mathbf{M}$ and $\mathbf{L}$.

Because ϕ_1, ϕ_{-1} are strictly proper composite, the minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\cdot)$ is unique. Let $\mathbf{u}^\mathbf{p} \in \mathcal{V}^d$ be the unique minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u})$ over $\mathcal{V}^d$.

By strict properness of ϕ_1, ϕ_{-1} we have for all $j \in [d]$

$$\lambda^{-1}(u_j^\mathbf{p}) = \mathbf{p}^\top \widetilde{\mathbf{c}}_j$$

Thus we have

$$\widetilde{\mathbf{M}}^\top \mathbf{p} = \lambda^{-1}(\mathbf{u}^\mathbf{p}) \,,$$

and therefore

$$\mathbf{p}^\top \boldsymbol{\ell}_t = (\lambda^{-1}(\mathbf{u}^\mathbf{p}))^\top \boldsymbol{\beta}_t + \gamma_t \,.$$

In particular, this gives

$$\text{decode}(\mathbf{u}^\mathbf{p}) \in \text{argmin}_{t \in [k]} \, \mathbf{p}^\top \boldsymbol{\ell}_t \,.$$

Moreover, since λ^{-1} is a continuous function, we have $\exists \delta > 0$ such that for any $\mathbf{u} \in \mathcal{V}^d$,

$$\|\mathbf{u} - \mathbf{u}^\mathbf{p}\| < \delta \implies \text{decode}(\mathbf{u}) \in \text{argmin}_{t \in [k]} \, \mathbf{p}^\top \boldsymbol{\ell}_t \,.$$

Thus, we have

$$\begin{aligned} \inf_{\mathbf{u} \in \mathcal{V}^d : \text{pred}(\mathbf{u}) \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) &\geq \inf_{\mathbf{u} \in \mathcal{V}^d : \|\mathbf{u} - \mathbf{u}^\mathbf{p}\| \geq \delta} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \\ &> \inf_{\mathbf{u} \in \mathcal{V}^d} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \,, \end{aligned}$$

where the last inequality follows since $\mathbf{u}^\mathbf{p}$ is the unique minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\cdot)$, and the set $\{\mathbf{u} \in \mathcal{V}^d : \|\mathbf{u} - \mathbf{u}^\mathbf{p}\| \geq \delta\}$ is closed.

Since the above holds for all $\mathbf{p} \in \Delta_n$, we have that (ψ, decode) is ℓ-calibrated. ■

It is easy to verify that the result in Theorem 2 on consistency of one-vs-all methods with respect to the multiclass 0-1 loss follows as a direct consequence of Theorem 6. More generally, applying Theorem 6 to the multiclass 0-1 loss immediately yields the following corollary, which shows that for the 0-1 loss on an n-class problem, it suffices to use a binary code matrix with $n-1$ columns:

Corollary 7 *Let $\mathbf{M} \in \{\pm 1\}^{n \times (n-1)}$ be any code matrix with $n-1$ linearly independent columns whose span does not contain the all-ones vector $\mathbf{e}_n$. Let (ϕ_1, ϕ_{-1}) be a strictly proper composite binary surrogate defined over an interval $\mathcal{V} \subseteq \overline{\mathbb{R}}$, with a continuous inverse link $\lambda^{-1} : \mathcal{V} \to [0,1]$. Define* decode $: \mathcal{V}^{n-1} \to [n]$ *as*

$$\text{decode}(\mathbf{u}) \in \text{argmin}_{t \in [n]} \, (\lambda^{-1}(\mathbf{u}))^\top \boldsymbol{\beta}_t + \gamma_t \,,$$

where $\lambda^{-1}(\mathbf{u}) \in [0,1]^{n-1}$ is a vector with j-th component $\lambda^{-1}(u_j)$, and $\boldsymbol{\beta}_t \in \mathbb{R}^{n-1}$, $\gamma_t \in \mathbb{R}$ are such that

$$\boldsymbol{\ell}_t^{\text{0-1}} = \tfrac{1}{2}(\mathbf{M} + 1) \, \boldsymbol{\beta}_t + \gamma_t \, \mathbf{e}_n \,.$$

Then the $(\mathbf{M}, \phi_1, \phi_{-1})$ code-based surrogate together with the mapping decode *is $\ell^{\text{0-1}}$-calibrated.*

For the 0-1 loss on an n-class problem, it is known that any convex, calibrated surrogate must operate on a space of dimension $d \geq n-1$ (Ramaswamy and Agarwal, 2012), and therefore the above result shows that in this case, one does not lose anything in terms of the dimension d of the multiclass surrogate when using a code-based surrogate with a binary code matrix. However, this

is not always the case: for a general loss matrix $\mathbf{L}$, it is known that there exist convex, calibrated surrogates operating on a space of dimension $d \leq \text{rank}(\mathbf{L})$ (Ramaswamy and Agarwal, 2012; Ramaswamy et al., 2013), but it is easy to construct examples of problems where the number of binary columns needed to span the column space of $\mathbf{L}$ far exceeds this number (see Example 1). To this end, we consider using output codes defined via more flexible *probabilistic* code matrices below.

4.2. Probabilistic ($[0,1]$-Valued) Code Matrices

Given a probabilistic code matrix $\mathbf{M} \in [0,1]^{n \times d}$ and binary surrogate (ϕ_1, ϕ_{-1}) defined on $\mathcal{V} \subseteq \overline{\mathbb{R}}$, we define the $(\mathbf{M}, \phi_1, \phi_{-1})$ *probabilistic* code-based surrogate $\psi : [n] \times \mathcal{V}^d \to \overline{\mathbb{R}}_+$ as

$$\psi(y, \mathbf{u}) = \sum_{j=1}^{d} \Big(M_{yj}\, \phi_1(u_j) + (1 - M_{yj})\, \phi_{-1}(u_j) \Big) .$$

This amounts to constructing d binary *class probability labeled* problems where for the j-th binary problem, examples with multiclass label $y \in [n]$ are transformed to binary examples with probability of a positive label given as $M_{yj} \in [0,1]$. Any binary surrogate-minimizing algorithm can be easily adapted to incorporate such class probability labels.[3]

The following result shows that for any loss matrix $\mathbf{L}$ of rank d, it suffices to construct a probabilistic code matrix with at most d columns to achieve consistency with respect to $\mathbf{L}$:

Theorem 8 *Let $\ell : [n] \times [k] \to \mathbb{R}_+$ be any target multiclass loss, and $\mathbf{L} \in \mathbb{R}_+^{n \times k}$ the corresponding loss matrix. Suppose $\exists d \in \mathbb{Z}_+$, vectors $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n \in \mathbb{R}^d$ and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k \in \mathbb{R}^d$, and constants $\gamma_1, \ldots, \gamma_k \in \mathbb{R}$ such that*

$$\ell(y, t) = \boldsymbol{\alpha}_y^\top \boldsymbol{\beta}_t + \gamma_t \qquad \forall y \in [n], t \in [k] .$$

Define

$$\alpha_{\min} = \min_{y \in [n], j \in [d]} \alpha_{yj} ; \qquad Z = \sum_{y=1}^{n} \sum_{j=1}^{d} (\alpha_{yj} - \alpha_{\min}) ;$$

and define $\mathbf{M} \in [0,1]^{n \times d}$ as

$$M_{yj} \;=\; \frac{\alpha_{yj} - \alpha_{\min}}{Z} \;\in\; [0,1] \qquad \forall y \in [n], j \in [d] .$$

Let (ϕ_1, ϕ_{-1}) be a strictly proper composite binary surrogate defined over an interval $\mathcal{V} \subseteq \overline{\mathbb{R}}$, with a continuous inverse link $\lambda^{-1} : \mathcal{V} \to [0,1]$. Define decode $: \mathcal{V}^d \to [k]$ *as*

$$\text{decode}(\mathbf{u}) \in \text{argmin}_{t \in [k]} \, (\lambda^{-1}(\mathbf{u}))^\top \widetilde{\boldsymbol{\beta}}_t + \widetilde{\gamma}_t ,$$

where $\lambda^{-1}(\mathbf{u}) \in [0,1]^d$ is a vector with j-th component $\lambda^{-1}(u_j)$, and

$$\widetilde{\boldsymbol{\beta}}_t = Z\, \boldsymbol{\beta}_t ; \qquad \widetilde{\gamma}_t = \gamma_t + \alpha_{\min} |\boldsymbol{\beta}_t| ,$$

where $|\boldsymbol{\beta}_t| = \sum_{j=1}^{d} \beta_{tj}$. Then the $(\mathbf{M}, \phi_1, \phi_{-1})$ probabilistic code-based surrogate together with the mapping decode *is ℓ-calibrated.*

3. We note that our use of probabilistic codes here differs from that of Dekel and Singer (2002), who considered losses based on KL-divergence and decoding based on expected Hamming distance.

The proof of Theorem 8 is a straightforward extension of that of Theorem 6; details are provided in Appendix D for completeness. For many multiclass losses, the use of probabilistic code matrices yields consistent output code based algorithms with a strictly smaller number of columns in the code matrix (and therefore smaller number of binary component problems) than is possible using binary code matrices. This is illustrated in the following example:

Example 1 *Consider a loss matrix* $\mathbf{L} \in \mathbb{R}_+^{5\times 5}$ *given by*

$$\mathbf{L} = \begin{bmatrix} 1.0 & 2.1 & 2.4 & 3.3 & 5.0 \\ 2.0 & 2.4 & 2.4 & 3.0 & 4.0 \\ 3.0 & 2.7 & 2.4 & 2.7 & 3.0 \\ 4.0 & 3.0 & 2.4 & 2.4 & 2.0 \\ 5.0 & 3.3 & 2.4 & 2.1 & 1.0 \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \\ 0.4 \\ 0.5 \end{bmatrix} \begin{bmatrix} 10 & 3 & 0 & -3 & -10 \end{bmatrix} + \begin{bmatrix} 0 & 1.8 & 2.4 & 3.6 & 6 \\ 0 & 1.8 & 2.4 & 3.6 & 6 \\ 0 & 1.8 & 2.4 & 3.6 & 6 \\ 0 & 1.8 & 2.4 & 3.6 & 6 \\ 0 & 1.8 & 2.4 & 3.6 & 6 \end{bmatrix}.$$

If we wish to use Theorem 6 to construct an $\mathbf{L}$*-calibrated binary code-based surrogate, then we must construct a binary code matrix* $\mathbf{M} \in \{\pm 1\}^{n\times d}$ *such that the column space of* $[\mathbf{M}, \mathbf{e}_n]$ *contains the column space of* $\mathbf{L}$*; it can be verified that this requires* $d \geq 3$*, which would require solving at least 3 binary problems. On the other hand, we can clearly apply Theorem 8, and get an* $\mathbf{L}$*-calibrated probabilistic code-based surrogate using a code matrix* $\mathbf{M} \in [0,1]^{n\times d}$ *with* $d = 1$ *column, which requires solving only 1 binary class probability labeled problem.*

The following examples illustrate how Theorem 8 can be used to design consistent probabilistic code-based methods for specific multiclass losses:

Example 2 (Discounted Cumulative Gain (DCG) Loss for Label/Subset Ranking) *The DCG loss applies to subset ranking settings and is widely used in information retrieval. Here each instance contains a query with* r *documents; a label assigns (say binary) relevances to these r documents* ($n = 2^r$)*. Predictions are permutations of the r documents* ($k = r!$)*. This defines a* $2^r \times r!$ *multiclass loss matrix:*

$$\mathbf{L}^{\text{DCG}}_{y,\sigma} = C - \sum_{i=1}^{r} \frac{2^{y_i} - 1}{\log(1 + \sigma(i))} \qquad \forall y \in \{0,1\}^r, \sigma \in \Pi_r,$$

where C *is a constant ensuring the loss is non-negative. From the definition above, it is clear the DCG loss has rank at most* $r + 1$*, and it is easy to express the loss in terms of* r*-dimensional vectors* $\boldsymbol{\alpha}_y$ *and* $\boldsymbol{\beta}_\sigma$ *and constants* γ_σ *as in Theorem 8. Applying Theorem 8 gives a* $2^r \times r$ *probabilistic code matrix* $\mathbf{M}$ *(thus* r *binary problems), such that solving the resulting binary problems using a suitable binary surrogate and then decoding as in the theorem (which in this case amounts to sorting the* r *documents according to certain scores obtained from the solutions to the r binary problems) yields a consistent output code based method for the DCG loss.*

Example 3 (Hamming Loss for Sequence Prediction) *The Hamming loss is widely used in sequence prediction problems, including multi-label prediction problems where labels can be viewed as binary vectors. Here each label as well as prediction is a sequence of (say binary) tags of some length* r ($n = k = 2^r$)*. This defines a* $2^r \times 2^r$ *loss matrix:*

$$\mathbf{L}^{\text{Ham}}_{y,t} = \sum_{i=1}^{r} \mathbf{1}(y_i \neq t_i) = \sum_{i=1}^{r} y_i(1 - 2t_i) + \sum_{i=1}^{r} t_i \qquad \forall y \in \{0,1\}^r, t \in \{0,1\}^r.$$

From the definition above, it is clear the DCG loss has rank at most $r+1$, and it is easy to express the loss in terms of r-dimensional vectors $\boldsymbol{\alpha}_y$ and $\boldsymbol{\beta}_t$ and constants γ_t as in Theorem 8. Applying Theorem 8 gives us a $2^r \times r$ probabilistic code matrix $\mathbf{M}$ (thus only r binary problems), such that solving the resulting binary problems using a suitable binary surrogate and then decoding as in the theorem (which in this case amounts to thresholding the r tags according to certain scores obtained from the solutions to the r binary problems) yields a consistent output code based method for the Hamming loss.

5. Conclusion and Future Directions

Output code based methods are a widely used approach to solving multiclass learning problems. One of the primary reasons for their appeal is that they reduce a complex multiclass problem into a set of simple binary classification problems, which can be solved easily in practice using a variety of standard learning algorithms. However, it is important to ask what types of performance guarantees such an approach provides for the target multiclass problem, since this is ultimately the problem one wants to perform well on. In particular, an ideal learning algorithm should be consistent for the target multiclass loss, i.e. should converge to the Bayes optimal error in the limit of infinite data. Surprisingly, there has been relatively little work on understanding consistency properties of output code based methods as a general methodology for solving general multiclass learning problems.

In this paper, we have considered consistency properties of output code based methods that solve the binary component problems by minimizing a binary surrogate loss. This allows us to view the overall code based method as minimizing a certain multiclass surrogate loss, and to appeal to the framework of calibrated surrogates for studying consistency. One of the main insights from our work is that if the binary surrogate minimized by the binary algorithm is a proper composite loss (and satisfies a few other conditions), then one can leverage the fact that minimizing such a surrogate yields not only accurate binary predictions, but also accurate class probability estimates; this insight plays a critical role in designing the right 'decoding' to achieve consistency with respect to the target loss. To our knowledge, the importance of selecting a good 'decoding' step for output code based methods has not received much attention previously. Additionally, we show that one can gain significant benefits in terms of the number of binary problems that need to be created by providing the binary algorithm with class probability labels in $[0, 1]$ rather than just binary labels.

A natural direction to further build on the results in the current work would be to obtain quantitative regret transfer bounds, using for example tools of Steinwart (2007) and Pires et al. (2013).

Acknowledgments

HR was supported by a TCS PhD Fellowship, and also thanks Google for a travel award. BSB thanks his advisor P. Vijay Kumar for his support. SA acknowledges support from the Department of Science & Technology (DST) of the Indian Government under a Ramanujan Fellowship, from the Indo-US Science & Technology Forum (IUSSTF), and from Yahoo in the form of an unrestricted grant. RW was supported by the Australian Research Council and NICTA, both of which are funded by the Australian Government.

References

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1:113–141, 2000.

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 *International Conference on Machine Learning*, 2011.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation: Structure and applications. Technical report, University of Pennsylvania, November 2005.

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 25*, pages 197–205. 2012.

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

Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. *Machine Learning*, 47:201–233, 2002.

Ofer Dekel and Yoram Singer. Multiclass learning by probabilistic embeddings. In *Neural Information Processing Systems*, 2002.

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

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

Wei Gao and Zhi-Hua Zhou. On the consistency of multi-label learning. In *Conference on Learning Theory*, 2011.

Venkatesan Guruswami and Amit Sahai. Multiclass learning, boosting and error-correcting codes. In *Conference on Learning Theory*, 1999.

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 25*, pages 1241–1249. 2012.

John Langford and Alina Beygelzimer. Sensitive error correcting output codes. In *Conference on Learning Theory*, 2005.

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.

Bernardo Á. Pires, Csaba Szepesvari, and Mohammad Ghavamzadeh. Cost-sensitive multiclass classification risk bounds. In *International Conference on Machine Learning*, 2013.

Harish G. Ramaswamy and Shivani Agarwal. Classification calibration dimension for general multiclass losses. In *Advances in Neural Information Processing Systems 25*, pages 2087–2095. 2012.

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 26*. 2013.

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

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

Robert E. Schapire. Using output codes to boost multiclass learning problems. In *International Conference on Machine Learning*, 1997.

Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

Terrence J. Sejnowski and Charles R. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168, 1987.

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

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

Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise approach to learning to rank: Theory and algorithm. In *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 3

Proof Let $n = 3$, and let $\psi : [3] \times \overline{\mathbb{R}}^3 \to \overline{\mathbb{R}}_+$ denote the $(\mathbf{M}^{\text{OvA}}, \phi^{\text{hinge}})$ code-based surrogate:

$$\psi(y, \mathbf{u}) = (1 - u_y)_+ + \sum_{y' \neq y} (1 + u_{y'})_+ .$$

Then for any $\mathbf{p} \in \Delta_3$ and $\mathbf{u} \in \overline{\mathbb{R}}^3$, we have

$$\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) = \sum_{y=1}^{3} p_y(1-u_y)_+ + (1-p_y)(1+u_y)_+ \,.$$

Consider the two probability vectors

$$\mathbf{p}^1 = \left(\tfrac{1}{3}+\epsilon, \tfrac{1}{3}-\tfrac{\epsilon}{2}, \tfrac{1}{3}-\tfrac{\epsilon}{2}\right)^\top$$
$$\mathbf{p}^2 = \left(\tfrac{1}{3}-\tfrac{\epsilon}{2}, \tfrac{1}{3}+\epsilon, \tfrac{1}{3}-\tfrac{\epsilon}{2}\right)^\top$$

For small enough $\epsilon > 0$, we have that the unique values $\mathbf{u}^1$ and $\mathbf{u}^2$ minimizing ${\mathbf{p}^1}^\top \boldsymbol{\psi}(\mathbf{u})$ and ${\mathbf{p}^2}^\top \boldsymbol{\psi}(\mathbf{u})$ are both equal to $\mathbf{u}' = (-1,-1,-1)^\top$. It is also clear that

$$\operatorname{argmin}_t {\mathbf{p}^1}^\top \boldsymbol{\ell}_t^{\text{0-1}} = \{1\} \qquad \text{and} \qquad \operatorname{argmin}_t {\mathbf{p}^2}^\top \boldsymbol{\ell}_t^{\text{0-1}} = \{2\}\,.$$

This implies that for any decode : $\mathbb{R}^3 \to \{1,2,3\}$, one of the following two statements must fail:

$$\inf_{\mathbf{u}\in\mathbb{R}^3, \text{decode}(\mathbf{u})\neq 1} {\mathbf{p}^1}^\top \boldsymbol{\psi}(\mathbf{u}) > {\mathbf{p}^1}^\top \boldsymbol{\psi}(\mathbf{u}') = \inf_{\mathbf{u}} {\mathbf{p}^1}^\top \boldsymbol{\psi}(\mathbf{u})\,;$$
$$\inf_{\mathbf{u}\in\mathbb{R}^3, \text{decode}(\mathbf{u})\neq 2} {\mathbf{p}^2}^\top \boldsymbol{\psi}(\mathbf{u}) > {\mathbf{p}^2}^\top \boldsymbol{\psi}(\mathbf{u}') = \inf_{\mathbf{u}} {\mathbf{p}^2}^\top \boldsymbol{\psi}(\mathbf{u})\,.$$

Thus the $(\mathbf{M}^{\text{OvA}}, \phi^{\text{hinge}})$ code-based surrogate is not $\ell^{\text{0-1}}$-calibrated. ■

Appendix B. Proof of Theorem 4

Proof Let $\psi : [n] \times \mathcal{V}^{\binom{n}{2}} \to \overline{\mathbb{R}}_+$ denote the $(\mathbf{M}^{\text{all-pairs}}, \phi_1, \phi_{-1})$ code based surrogate:

$$\psi(y, \mathbf{u}) \;\; = \;\; \sum_{i<j} \mathbf{1}(y=i)\phi_1(u_{i,j}) + \mathbf{1}(y=j)\phi_{-1}(u_{i,j})$$

Fix any $\mathbf{p} \in \Delta_n$. Then

$$\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) = \sum_{i<j} p_i\phi_1(u_{i,j}) + p_j\phi_{-1}(u_{i,j})$$

Let $\mathbf{u}^{\mathbf{p}} \in \mathcal{V}^{\binom{n}{2}}$ be any minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\cdot)$ over $\mathcal{V}^{\binom{n}{2}}$.
Note that for $i < j$ such that $p_i + p_j > 0$, we have that $u^{\mathbf{p}}_{i,j}$ is uniquely determined by the strict proper composite property as follows:

$$\lambda^{-1}(u^{\mathbf{p}}_{i,j}) = \frac{p_i}{p_i + p_j}$$

which gives

$$p_i \geq p_j \iff \lambda^{-1}(u^{\mathbf{p}}_{i,j}) \geq \frac{1}{2}$$

Now, clearly,

$$\begin{aligned} \text{decode}(\mathbf{u}^{\mathbf{p}}) = i &\Rightarrow \text{score}_i(\mathbf{u}^{\mathbf{p}}) \geq \text{score}_j(\mathbf{u}^{\mathbf{p}}) \qquad \forall j \in [n] \\ &\Rightarrow p_i \geq p_j \qquad \forall j \in [n] \\ &\Rightarrow i \in \text{argmin}_{t\in[n]} \mathbf{p}^\top \boldsymbol{\ell}_t^{\text{0-1}} . \end{aligned}$$

Thus

$$\text{decode}(\mathbf{u}^{\mathbf{p}}) \in \text{argmin}_{t\in[n]} \mathbf{p}^\top \boldsymbol{\ell}_t^{\text{0-1}} .$$

Also for any sequence $\mathbf{u}_m$ such that $\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_m)$ converges to $\inf_{\mathbf{u}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u})$, and any i, j such that that $p_i + p_j > 0$, we have that $u_{m,(i,j)}$ converges to $u^{\mathbf{p}}_{i,j}$. The previous statement follows due to the strict proper compositeness and continuity of (ϕ_1, ϕ_{-1}) . Now, using the continuity of λ^{-1} we have $\forall \mathbf{p} \in \Delta_n$

$$\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_m) \to \inf_u \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \Longrightarrow \text{decode}(\mathbf{u}_m) \in \text{argmin}_{t\in[n]} \mathbf{p}^\top \boldsymbol{\ell}_t^{\text{0-1}} \text{ for large enough } m$$

It can easily be seen that the above statement is equivalent to (ψ, decode) being $\ell^{\text{0-1}}$-calibrated. ■

Appendix C. Proof of Theorem 5

Proof Let $\psi : [n] \times \overline{\mathbb{R}}^{\binom{n}{2}} \to \overline{\mathbb{R}}_+$ denote the $(\mathbf{M}^{\text{all-pairs}}, \phi^{\text{hinge}})$ code based surrogate:

$$\psi(y, \mathbf{u}) \quad = \quad \sum_{i<j} \mathbf{1}(y = i)(1 - u_{i,j})_+ + \mathbf{1}(y = j)(1 + u_{i,j})_+$$

Fix any $\mathbf{p} \in \Delta_n$. Then

$$\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) = \sum_{i<j} p_i(1 - u_{i,j})_+ + p_j(1 + u_{i,j})_+$$

Let $\mathbf{u}^{\mathbf{p}}$ be a minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\cdot)$. Let $u^{\mathbf{p}}_{i,j} = -u^{\mathbf{p}}_{i,j}$ if $i > j$. It is clear that $u^{\mathbf{p}}_{i,j} = 1$ if $p_i > p_j$, and $u^{\mathbf{p}}_{i,j} = -1$ if $p_i < p_j$. In the case that $p_i = p_j$, $u^{\mathbf{p}}_{i,j}$ can be any element in $[-1, 1]$. Thus $\forall i, j \in [n]$,

$$p_i > p_j \Longrightarrow \text{score}_i^{\text{hinge}}(\mathbf{u}^{\mathbf{p}}) > \text{score}_j^{\text{hinge}}(\mathbf{u}^{\mathbf{p}}) .$$

Hence

$$\text{decode}(\mathbf{u}^{\mathbf{p}}) \in \text{argmax}_y \, p_y = \text{argmin}_t \, \mathbf{p}^\top \boldsymbol{\ell}_t^{\text{0-1}} .$$

For any sequence $\{\mathbf{u}_m\}$ such that $\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}_m) \to \inf_{\mathbf{u}} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u})$ and for all indices (i, j) such that $p_i \neq p_j$, we have that $u_{m,(i,j)}$ converges to $u^{\mathbf{p}}_{i,j}$. Thus we have $\text{decode}(\mathbf{u}_m) \in \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t^{\text{0-1}}$ for large enough m. Since this holds for all $\mathbf{p} \in \Delta_n$, we have that the $(\phi^{\text{hinge}}, \mathbf{M}^{\text{all-pairs}})$ code-based surrogate is $\ell^{\text{0-1}}$-calibrated. ■

Appendix D. Proof of Theorem 8

Proof Let $\psi : [n] \times \mathcal{V}^d \to \mathbb{R}_+$ denote the $(\mathbf{M}, \phi_1, \phi_{-1})$ probabilistic code-based surrogate:

$$\psi(y, \mathbf{u}) = \sum_{j=1}^{d} M_{yj}\, \phi_1(u_j) + (1 - M_{yj})\, \phi_{-1}(u_j)\,.$$

Fix any $\mathbf{p} \in \Delta_n$. Then we have

$$\mathbf{p}^\top \boldsymbol{\ell}_t = \mathbf{p}^\top (\mathbf{M}\widetilde{\boldsymbol{\beta}}_t + \widetilde{\gamma}_t \mathbf{e}_n) = \mathbf{p}^\top \mathbf{M}\widetilde{\boldsymbol{\beta}}_t + \widetilde{\gamma}_t\,.$$

Now, denote by $\mathbf{c}_1, \ldots, \mathbf{c}_d$ the columns of $\mathbf{M}$. Then we have

$$\begin{aligned}
\mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) &= \sum_{y=1}^{n} p_y \Bigg(\sum_{j=1}^{d} M_{yj}\, \phi_1(u_j) + (1 - M_{yj})\, \phi_{-1}(u_j) \Bigg) \\
&= \sum_{j=1}^{d} \Big((\mathbf{p}^\top \mathbf{c}_j) \phi_1(u_j) + (1 - \mathbf{p}^\top \mathbf{c}_j) \phi_{-1}(u_j) \Big)\,.
\end{aligned}$$

Because ϕ_1, ϕ_{-1} are strictly proper composite, the minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\cdot)$ is unique. Let $\mathbf{u}^{\mathbf{p}} \in \mathcal{V}^d$ be the unique minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\cdot)$ over $\mathcal{V}^d$.

By strict properness of ϕ_1, ϕ_{-1} we have for all $j \in [d]$

$$\lambda^{-1}(u_j^{\mathbf{p}}) = \mathbf{p}^\top \mathbf{c}_j$$

Thus we have

$$\mathbf{M}^\top \mathbf{p} = \lambda^{-1}(\mathbf{u}^{\mathbf{p}})\,,$$

and therefore

$$\mathbf{p}^\top \boldsymbol{\ell}_t = (\lambda^{-1}(\mathbf{u}^{\mathbf{p}}))^\top \widetilde{\boldsymbol{\beta}}_t + \widetilde{\gamma}_t\,.$$

In particular, this gives

$$\text{decode}(\mathbf{u}^{\mathbf{p}}) \in \text{argmin}_{t \in [k]}\, \mathbf{p}^\top \boldsymbol{\ell}_t\,.$$

Moreover, since λ^{-1} is a continuous function, we have $\exists \delta > 0$ such that for any $\mathbf{u} \in \mathcal{V}^d$,

$$\|\mathbf{u} - \mathbf{u}^{\mathbf{p}}\| < \delta \implies \text{decode}(\mathbf{u}) \in \text{argmin}_{t \in [k]}\, \mathbf{p}^\top \boldsymbol{\ell}_t\,.$$

Thus, we have

$$\begin{aligned}
\inf_{\mathbf{u} \in \mathcal{V}^d : \text{pred}(\mathbf{u}) \notin \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) &\geq \inf_{\mathbf{u} \in \mathcal{V}^d : \|\mathbf{u} - \mathbf{u}^{\mathbf{p}}\| \geq \delta} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u}) \\
&> \inf_{\mathbf{u} \in \mathcal{V}^d} \mathbf{p}^\top \boldsymbol{\psi}(\mathbf{u})\,,
\end{aligned}$$

where the last inequality follows since $\mathbf{u}^{\mathbf{p}}$ is the unique minimizer of $\mathbf{p}^\top \boldsymbol{\psi}(\cdot)$, and the set $\{\mathbf{u} \in \mathcal{V}^d : \|\mathbf{u} - \mathbf{u}^{\mathbf{p}}\| \geq \delta\}$ is closed.

Since the above holds for all $\mathbf{p} \in \Delta_n$, we have that (ψ, decode) is ℓ-calibrated. ■