Electronic Journal of Statistics
Vol. 0 (2006) 1–8
ISSN: 1935-7524

Consistent Algorithms for Multiclass Classification with an Abstain Option

Harish G. Ramaswamy

Department of Computer Science and Engineering,
Indian Institute of Technology Madras, Chennai, India
e-mail: hariguru@cse.iitm.ac.in

Ambuj Tewari *

Department of Statistics, and
Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, USA
e-mail: tewaria@umich.edu

Shivani Agarwal †

Department of Computer and Information Science,
University of Pennsylvania, Philadelphia, USA
e-mail: ashivani@seas.upenn.edu

Abstract: We consider the problem of n-class classification ($n \geq 2$), where the classifier can choose to abstain from making predictions at a given cost, say, a factor α of the cost of misclassification. Our goal is to design consistent algorithms for such n-class classification problems with a 'reject option'; while such algorithms are known for the binary ($n = 2$) case, little has been understood for the general multiclass case. We show that the well known Crammer-Singer surrogate and the one-vs-all hinge loss, albeit with a different predictor than the standard argmax, yield consistent algorithms for this problem when $\alpha = \frac{1}{2}$. More interestingly, we design a new convex surrogate, which we call the binary encoded predictions surrogate, that is also consistent for this problem when $\alpha = \frac{1}{2}$ and operates on a much lower dimensional space ($\log(n)$ as opposed to n). We also construct modified versions of all these three surrogates to be consistent for any given $\alpha \in [0, \frac{1}{2}]$.

MSC 2010 subject classifications: Primary 62H30; secondary 68T10.
Keywords and phrases: machine learning; classification; statistical consistency; surrogate loss; calibration; abstain loss..

1. Introduction

In classification problems, one often encounters cases where it would be better for the classifier to take no decision and abstain from predicting rather than make a wrong prediction. For example, in the problem of medical diagnosis with inexpensive tests as features, a conclusive decision is good, but in the face

*Partially supported by NSF CAREER grant IIS-1452099 and a Sloan Research Fellowship.
†Partially supported by NSF grant IIS-1717290.

of uncertainty, it is better to not make a prediction and instead go for costlier tests.

1.1. Binary Classification with an Abstain Option

For the case of binary classification, El-Yaniv and Wiener [7, 8] call this problem selective classification. They study the fundamental trade-off between abstaining and predicting and give theoretical results, but the algorithms suggested by their theory are not computationally tractable due to the usage of ERM oracles.

Another branch of work for the binary classification case [2, 28, 13] has roots in decision theory, where abstaining is just another decision that incurs a cost. The main idea here is to find appropriate computationally efficient optimization based algorithms that give the optimal answer in the limit of infinite data. Yuan and Wegkamp [28] show that many standard convex optimization based procedures for binary classification like logistic regression, least squares classification and exponential loss minimization (Adaboost) yield consistent algorithms for this problem. But as Bartlett and Wegkamp [2] show, the algorithm based on minimizing the hinge loss (SVM) requires a modification to be consistent. The suggested modification is rather simple: use a double hinge loss with three linear segments instead of the two segments in standard hinge loss, the ratio of slopes of the two non-flat segments depends on the cost of abstaining α. Cortes et al. [4] learn a separate "rejector" function, in addition to a classifier, for identifying instances to reject. They also show that such an algorithm is consistent for this problem. There have been several empirical studies [10, 11, 12, 9] as well on this topic.

1.2. Multiclass Classification with an Abstain Option

In the case of multiclass classification with an abstain option, there has been empirical work [31, 21, 27]. However, to the best of our knowledge, there exists very little theoretical work on this problem. Zhang et al. [29] define a new family of surrogates for this problem, but their family of surrogates are known to be not consistent for the decision theoretic version of the problem. There has also been work on learning separate thresholds for rejection per class [15], but such algorithms are also not known to be consistent for this problem.

We fill this gap in the literature by providing a formal treatment of the multiclass classification problem with an abstain option in the decision theoretic setting. Our work can also be seen to be in the statistical decision theoretic setting, and can be seen to generalize and extend the works of Bartlett and Wegkamp [2], Yuan and Wegkamp [28] and Grandvalet et al. [13] to the multiclass setting. In particular, we give consistent algorithms for this problem.

The reject option is accommodated into the problem of n-class classification through the evaluation metric. We seek a function $h : \mathcal{X} \to \{1, 2, \ldots, n, \perp\}$, where $\mathcal{X}$ is the instance space, and the n classes are denoted by $\{1, 2, \ldots, n\} = [n]$ and

$\perp$ denotes the action of abstaining or the 'reject' option. The loss incurred by such a function on an example (x, y) with $h(x) = t$ is given by

$$\ell^{\alpha}(y,t) = \begin{cases} 1 & \text{if } t \neq y \text{ and } t \neq \perp \\ \alpha & \text{if } t = \perp \\ 0 & \text{if } t = y \end{cases} \tag{1.1}$$

where $\alpha \in [0,1]$ denotes the cost of abstaining. We will call this loss the abstain(α) loss.

It can be easily shown that the Bayes optimal risk for the above loss is attained by the function $h^*_\alpha : \mathcal{X} \to [n] \cup \{\perp\}$ given by

$$h^*_\alpha(x) = \begin{cases} \text{argmax}_{y\in[n]} p_x(y) & \text{if } \max_{y\in[n]} p_x(y) \geq 1-\alpha \\ \perp & \text{otherwise} \end{cases} \tag{1.2}$$

where $p_x(y) = P(Y = y|X = x)$. The above is often called 'Chow's rule' [3]. It can also be seen that the interesting range of values for α is $[0, \frac{n-1}{n}]$ as for all $\alpha > \frac{n-1}{n}$ the Bayes optimal classifier for the abstain(α) loss never abstains. For example, in binary classification, only $\alpha \leq \frac{1}{2}$ is meaningful, as higher values of α imply it is never optimal to abstain.

For small α, the classifier h^*_α acts as a high-confidence classifier and would be useful in applications like medical diagnosis. For example, if one wishes to learn a classifier for diagnosing an illness with 80% confidence, and recommend further medical tests if it is not possible, the ideal classifier would be $h^*_{0.2}$, which is the minimizer of the abstain(0.2) loss. If $\alpha = \frac{1}{2}$, the Bayes classifier h^*_α has a very appealing structure: a class $y \in [n]$ is predicted only if the class y has a simple majority. The abstain(α) loss is also useful in applications where a 'greater than $1-\alpha$ conditional probability detector' can be used as a black box. For example a greater than $\frac{1}{2}$ conditional probability detector plays a crucial role in hierarchical classification [19].

Abstain(α) loss with $\alpha = \frac{1}{2}$ will be the main focus of our paper and will be the default choice when the abstain loss is referred to without any reference to α. This will be the case in Sections 3, 4, 5 and 7. In Section 6, we show how to extend our results to the case $\alpha \leq 1/2$. On the other hand, we leave the case $\alpha > 1/2$ to future work. We explain why this case might be fundamentally different in Section 1.4.

Since the Bayes classifier h^*_α depends only on the conditional distribution of $Y|X$, any algorithm that gives a consistent estimator of the conditional probability of the classes, e.g., minimizing the one vs all squared loss, [17, 25], can be made into a consistent algorithm (with a suitable change in the decision) for this problem. However, smooth surrogates that estimate the conditional probability do much more than what is necessary to solve this problem. Consistent piecewise linear surrogate minimizing algorithms, on the other hand, do only what is needed, in accordance with Vapnik's dictum [23]:

> When solving a given problem, try to avoid solving a more general problem as an intermediate step.

For example, least squares classification, logistic regression and SVM are all consistent for standard binary classification, but SVMs avoid the strictly harder conditional probability estimation problem as an intermediate problem. Piecewise linear surrogates (like the hinge loss used in SVM) have other advantages like easier optimization and sparsity (in the dual) as well, hence finding consistent piecewise linear surrogates for the abstain loss is an important and interesting task.

1.3. Contributions

We show that the n-dimensional multiclass surrogate of Crammer and Singer (CS) [5] and the simple one vs all hinge (OVA) surrogate loss [20] both yield consistent algorithms for the abstain$\left(\frac{1}{2}\right)$ loss. Both these surrogates are *not* consistent for the standard multiclass classification problem [22, 16, 30].

We then construct a new convex piecewise linear surrogate, which we call the *binary encoded predictions* (BEP) surrogate that operates on a $\log_2(n)$ dimensional space, and yields a consistent algorithm for the n-class abstain$\left(\frac{1}{2}\right)$ loss. When optimized over comparable function classes, this algorithm is more efficient than the Crammer-Singer and one vs all algorithms as it requires to only find $\log_2(n)$ functions over the instance space, as opposed to n functions. This result is surprising because, it has been shown that one needs to minimize at least a $n-1$ dimensional convex surrogate to get a consistent algorithm for the standard n-class problem, i.e., *without* the reject option [17]. Also, the only known generic way of generating consistent convex surrogate minimizing algorithms for an arbitrary loss [17, 18], when applied to the n-class abstain loss, yields an n-dimensional surrogate.

We also give modified versions of the CS, OVA and BEP surrogates that yield consistent algorithms for the abstain(α) loss for any given $\alpha \in \left[0, \frac{1}{2}\right]$.

1.4. The Role of α

Conditional probability estimation based surrogates can be used for designing consistent algorithms for the n-class problem with the reject option for any $\alpha \in (0, \frac{n-1}{n})$, but the Crammer-Singer surrogate, the one vs all hinge and the BEP surrogate and their corresponding variants all yield consistent algorithms only for $\alpha \in [0, \frac{1}{2}]$. While this may seem restrictive, we contend that these form an interesting and useful set of problems to solve. We also suspect that, abstain(α) problems with $\alpha > \frac{1}{2}$ are fundamentally more difficult than those with $\alpha \leq \frac{1}{2}$, for the reason that evaluating the Bayes classifier $h^*_\alpha(x)$ can be done for $\alpha \leq \frac{1}{2}$ without finding the maximum conditional probability – just check if any class has conditional probability greater than $(1-\alpha)$ as there can only be one. This is also evidenced by the more complicated partitions (more lines required to draw the partitions) of the simplex induced by the Bayes optimal classifier for $\alpha > \frac{1}{2}$ as shown in Figure 1.

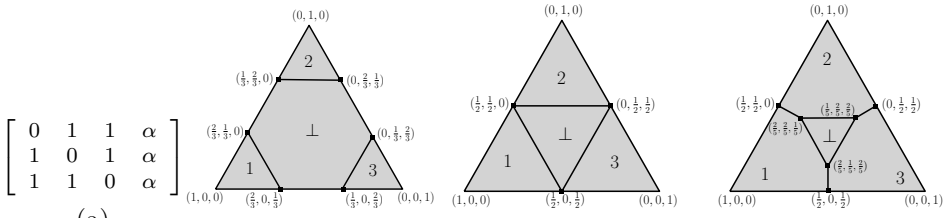

FIG 1. *(a) The abstain(α) loss with $n = 3$ as a matrix, where rows correspond to classes $\{1, 2, 3\}$ and columns correspond to predictions $\{1, 2, 3, \perp\}$; (b,c,d) the partition of the simplex Δ_3, depicting the optimal prediction for different conditional probabilities, induced by the Bayes classifier for the abstain($\frac{1}{3}$), abstain($\frac{1}{2}$) and abstain($\frac{3}{5}$) losses respectively.*

Notation: Throughout the paper, we let $\mathbb{R} = (-\infty, \infty)$ and $\mathbb{R}_+ = [0, \infty)$. Let $\mathbb{Z}, \mathbb{Z}_+$ denote the sets of all integers and non-negative integers, respectively. For $n \in \mathbb{Z}_+$, we let $[n] = \{1, \ldots, n\}$. For $z \in \mathbb{R}$, we let $z_+ = \max(0, z)$. We denote by Δ_n the probability simplex in $\mathbb{R}^n$: $\Delta_n = \{\mathbf{p} \in \mathbb{R}^n_+ : \sum_{i=1}^n p_i = 1\}$. For $n \in \mathbb{Z}_+$, we denote by $\mathbf{1}^n$ and $\mathbf{0}^n$ the n-dimensional all ones and all zeros vector, and for $i \in [n]$ we denote by $\mathbf{e}^n_i$ the n-dimensional vector with 1 in position i and 0 elsewhere. Often we omit the dimension n from $\mathbf{1}^n, \mathbf{0}^n, \mathbf{e}^n_i$ as it is clear from the context. For any vector $\mathbf{u}$, we denote by $u_{(i)}$ the i^{th} element of the components of $\mathbf{u}$ when sorted in descending order. We denote by $\text{sign}(u)$, the sign of a scalar u, with $\text{sign}(0) = 1$

2. Problem Setup

In this section, we formally set up the problem of multiclass classification with an abstain option and explain the notion of consistency for the problem.

Let the instance space be $\mathcal{X}$. Given training examples $(X_1, Y_1), \ldots, (X_m, Y_m)$ drawn i.i.d. from a distribution D on $\mathcal{X} \times [n]$, the goal is to learn a prediction function $h : \mathcal{X} \to [n] \cup \{\perp\}$.

For any given $\alpha \in [0, 1]$, the performance of a prediction function $h : \mathcal{X} \to [n] \cup \{\perp\}$ is measured via the abstain(α) loss ℓ^α from Equation (1.1). We denote the loss incurred on predicting t when the correct label is y by $\ell^\alpha(y, t)$. For any $t \in [n] \cup \{\perp\}$, we denote by $\boldsymbol{\ell}^\alpha_t$ the vector of losses $[\ell^\alpha(1, t), \ldots, \ell^\alpha(n, t)]^\top \in \mathbb{R}^n_+$. The abstain($\alpha$) loss and a schematic representation of the Bayes classifier for various values of α given by Equation (1.2) are given in Figure 1 for $n = 3$.

Specifically, the goal is to learn a function $h : \mathcal{X} \to [n] \cup \{\perp\}$ with low expected ℓ^α-error

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

Ideally, one wants the ℓ^α-error of the learned function to be close to the optimal ℓ^α-error

$$\text{er}^{\ell^\alpha,*}_D = \inf_{h:\mathcal{X}\to[n]\cup\{\perp\}} \text{er}^{\ell^\alpha}_D[h] \,.$$

An algorithm, which outputs a function $h_m : \mathcal{X} \to [n] \cup \{\perp\}$ on being given a random training sample as above, is said to be *consistent* w.r.t. ℓ^α if the ℓ^α-error of the learned function h_m converges in probability to the optimal for any distribution D: $\mathrm{er}_D^{\ell^\alpha}[h_m] \xrightarrow{\mathrm{P}} \mathrm{er}_D^{\ell,*}$. Here, the convergence in probability is over the learned classifier h_m as a function of the training sample distributed i.i.d. according to D.

However, minimizing the discrete ℓ^α-error directly is computationally difficult; therefore one uses instead a *surrogate loss function* $\psi : [n] \times \mathbb{R}^d \to \mathbb{R}_+$, for some $d \in \mathbb{Z}_+$, and learns a function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ by minimizing (approximately, based on the training sample) the ψ-error

$$\mathrm{er}_D^{\psi}[\mathbf{f}] = \mathbf{E}_{(X,Y)\sim D}[\psi(Y, \mathbf{f}(X))] \ .$$

Predictions on new instances $x \in \mathcal{X}$ are then made by applying the learned function $\mathbf{f}$ and mapping back to predictions in the target space $[n] \cup \{\perp\}$ via some mapping $\mathrm{pred} : \mathbb{R}^d \to [n] \cup \{\perp\}$, giving $h(x) = \mathrm{pred}(\mathbf{f}(x))$.

Under suitable conditions, algorithms that approximately minimize the ψ-error based on a training sample are known to be consistent with respect to ψ, i.e., to converge in probability to the optimal ψ-error

$$\mathrm{er}_D^{\psi,*} = \inf_{\mathbf{f}:\mathcal{X}\to\mathbb{R}^d} \mathrm{er}_D^{\psi}[\mathbf{f}] \ .$$

Also, when ψ is convex in its second argument, the resulting optimization problem is convex and can be efficiently solved.

Hence, we seek a surrogate and a predictor (ψ, pred), with ψ convex over its second argument, and satisfying a bound of the following form holding for all $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$

$$\mathrm{er}_D^{\ell^\alpha}[\mathrm{pred} \circ \mathbf{f}] - \mathrm{er}_D^{\ell^\alpha,*} \leq \xi\left(\mathrm{er}_D^{\psi}[\mathbf{f}] - \mathrm{er}_D^{\psi,*}\right)$$

where $\xi : \mathbb{R} \to \mathbb{R}$ is increasing, continuous at 0 and $\xi(0) = 0$. A surrogate and a predictor (ψ, pred), satisfying such a bound, known as an excess risk transform bound, would immediately give an algorithm consistent w.r.t. ℓ^α from an algorithm consistent w.r.t. ψ. We derive such bounds w.r.t. the $\ell^{\frac{1}{2}}$ loss for the Crammer-Singer surrogate, the one vs all hinge surrogate, and the BEP surrogate, with ξ as a linear function.

3. Excess Risk Bounds for the Crammer-Singer and One vs All Hinge Surrogates

In this section, we give an excess risk bound relating the abstain loss ℓ, and the Crammer-Singer surrogate ψ^{CS} [5] and also the one vs all hinge loss.

Define the Crammer-Singer surrogate $\psi^{\mathrm{CS}} : [n] \times \mathbb{R}^n \to \mathbb{R}_+$ and predictor $\mathrm{pred}_\tau^{\mathrm{CS}} : \mathbb{R}^n \to [n] \cup \{\perp\}$ as

$$\begin{aligned} \psi^{\mathrm{CS}}(y, \mathbf{u}) &= (\max_{j \neq y} u_j - u_y + 1)_+ \\ \mathrm{pred}_\tau^{\mathrm{CS}}(\mathbf{u}) &= \begin{cases} \mathrm{argmax}_{i\in[n]} u_i & \text{if } u_{(1)} - u_{(2)} > \tau \\ \perp & \text{otherwise} \end{cases} \end{aligned}$$

where $(a)_+ = \max(a, 0)$, $u_{(i)}$ is the i th element of the components of $\mathbf{u}$ when sorted in descending order and $\tau \in (0, 1)$ is a threshold parameter.

Similarly, define the one-vs-all surrogate $\psi^{\mathrm{OVA}} : [n] \times \mathbb{R}^n \to \mathbb{R}_+$ and predictor $\mathrm{pred}_\tau^{\mathrm{OVA}} : \mathbb{R}^n \to [n] \cup \{\perp\}$ as

$$\psi^{\mathrm{OVA}}(y, \mathbf{u}) = \sum_{i=1}^{n} \mathbf{1}(y = i)(1 - u_i)_+ + \mathbf{1}(y \neq i)(1 + u_i)_+$$

$$\mathrm{pred}_\tau^{\mathrm{OVA}}(\mathbf{u}) = \begin{cases} \mathrm{argmax}_{i \in [n]} u_i & \text{if } \max_j u_j > \tau \\ \perp & \text{otherwise} \end{cases}$$

where $(a)_+ = \max(a, 0)$ and $\tau \in (-1, 1)$ is a threshold parameter, and ties are broken arbitrarily, say, in favor of the label y with the smaller index.

The following is the main result of this section, the proof of which is in Section 8.

Theorem 3.1. *Let $n \in \mathbb{Z}_+$, $\tau_{\mathrm{CS}} \in (0, 1)$ and $\tau_{\mathrm{OVA}} \in (-1, 1)$. Then for all $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$*

$$\mathrm{er}_D^\ell[\mathrm{pred}_{\tau_{\mathrm{CS}}}^{\mathrm{CS}} \circ \mathbf{f}] - \mathrm{er}_D^{\ell,*} \leq \frac{\left(\mathrm{er}_D^{\psi^{\mathrm{CS}}}[\mathbf{f}] - \mathrm{er}_D^{\psi^{\mathrm{CS}},*}\right)}{2\min(\tau_{\mathrm{CS}}, 1 - \tau_{\mathrm{CS}})}$$

$$\mathrm{er}_D^\ell[\mathrm{pred}_{\tau_{\mathrm{OVA}}}^{\mathrm{OVA}} \circ \mathbf{f}] - \mathrm{er}_D^{\ell,*} \leq \frac{\left(\mathrm{er}_D^{\psi^{\mathrm{OVA}}}[\mathbf{f}] - \mathrm{er}_D^{\psi^{\mathrm{OVA}},*}\right)}{2(1 - |\tau_{\mathrm{OVA}}|)}$$

Remark 1: The form of the abstaining region for the CS and OVA predictors arise due to the properties of the surrogate. In particular, due to the fact that the CS surrogate is invariant to adding a constant to all coordinates of the surrogate prediction $\mathbf{u}$, the form of the CS abstaining region has to depend on the difference between two coordinates of $\mathbf{u}$.

Remark 2: It has been pointed out previously by Zhang [30], that if the data distribution D is such that $\max_y p_x(y) > 0.5$ for all $x \in \mathcal{X}$, the Crammer-Singer surrogate ψ^{CS} and the one vs all hinge loss are consistent with the zero-one loss when used with the standard argmax predictor. This conclusion also follows from the theorem above. However, our result yields more – in the case that the distribution satisfies the dominant class assumption only for some instances $x \in \mathcal{X}$, the function learned by using the surrogate and predictor $(\psi^{\mathrm{CS}}, \mathrm{pred}_\tau^{\mathrm{CS}})$ or $(\psi^{\mathrm{OVA}}, \mathrm{pred}_\tau^{\mathrm{OVA}})$ gives the right answer for such instances having a dominant class, and fails in a graceful manner by abstaining for other instances that do not have a dominant class.

4. Excess Risk Bounds for the BEP Surrogate

The Crammer-Singer surrogate and the one vs all hinge surrogate, just like surrogates designed for conditional probability estimation, are defined over an n-dimensional domain. Thus any algorithm that minimizes these surrogates must

learn n real valued functions over the instance space. In this section, we construct a $\lceil \log_2(n) \rceil$ dimensional convex surrogate, which we call the *binary encoded predictions* (BEP) surrogate, and give an excess risk bound relating this surrogate and the abstain loss. In particular these results show that the BEP surrogate is calibrated w.r.t. the abstain loss; this in turn implies that the *convex calibration dimension* (CC-dimension) [17] of the abstain loss is at most $\lceil \log_2(n) \rceil$.

The idea of learning $\log(n)$ predictors for an n-class classification problem has some precedent [1, 24], but their objectives are focussed on the multiclass 0-1 loss, and they are not concerned about consistency or calibration of surrogates.

For the purpose of simplicity let us assume $n = 2^d$ for some positive integer d.[1] Let $B : [n] \to \{+1, -1\}^d$ be any one-one and onto mapping, with an inverse mapping $B^{-1} : \{+1, -1\}^d \to [n]$. Define the BEP surrogate $\psi^{\text{BEP}} : [n] \times \mathbb{R}^d \to \mathbb{R}_+$ and its corresponding predictor $\text{pred}_\tau^{\text{BEP}} : \mathbb{R}^d \to [n] \cup \{\perp\}$ as

$$\psi^{\text{BEP}}(y, \mathbf{u}) = (\max_{j \in [d]} B_j(y) u_j + 1)_+$$

$$\text{pred}_\tau^{\text{BEP}}(\mathbf{u}) = \begin{cases} \perp & \text{if } \min_{i \in [d]} |u_i| \leq \tau \\ B^{-1}(\text{sign}(-\mathbf{u})) & \text{otherwise} \end{cases}$$

where $\text{sign}(u)$ is the sign of u, with $\text{sign}(0) = 1$ and $\tau \in (0, 1)$ is a threshold parameter.

To make the above definition clear, let us see what the surrogate and predictor look like for the case of $n = 4$ and $\tau = \frac{1}{2}$. We have $d = 2$. Let us fix the mapping B such that $B(y)$ is the standard d-bit binary representation of $(y - 1)$, with -1 in the place of 0. Then we have,

$$\begin{aligned} \psi^{\text{BEP}}(1, \mathbf{u}) &= (\max(-u_1, -u_2) + 1)_+ \\ \psi^{\text{BEP}}(2, \mathbf{u}) &= (\max(-u_1, u_2) + 1)_+ \\ \psi^{\text{BEP}}(3, \mathbf{u}) &= (\max(u_1, -u_2) + 1)_+ \\ \psi^{\text{BEP}}(4, \mathbf{u}) &= (\max(u_1, u_2) + 1)_+ \end{aligned}$$

$$\text{pred}_{\frac{1}{2}}^{\text{BEP}}(\mathbf{u}) = \begin{cases} 1 & \text{if } u_1 > \frac{1}{2}, u_2 > \frac{1}{2} \\ 2 & \text{if } u_1 > \frac{1}{2}, u_2 < -\frac{1}{2} \\ 3 & \text{if } u_1 < -\frac{1}{2}, u_2 > \frac{1}{2} \\ 4 & \text{if } u_1 < -\frac{1}{2}, u_2 < -\frac{1}{2} \\ \perp & \text{otherwise} \end{cases}$$

Figure 2 gives the partition induced by the predictor $\text{pred}_{\frac{1}{2}}^{\text{BEP}}$.

The following is the main result of this section, the proof of which is in Section 8.

[1] If n is not a power of 2, just add enough dummy classes that never occur.

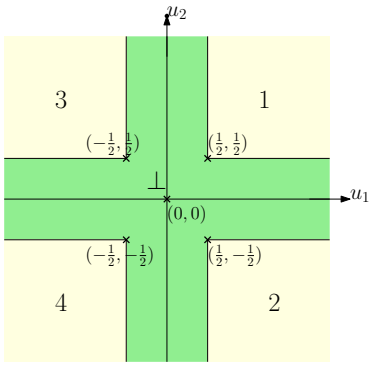

FIG 2. *The partition of* $\mathbb{R}^2$ *induced by* $\text{pred}^{\text{BEP}}_{\frac{1}{2}}$

Theorem 4.1. *Let* $n \in \mathbb{Z}_+$ *and* $\tau \in (0,1)$*. Let* $n = 2^d$*. Then, for all* $f : \mathcal{X} \to \mathbb{R}^d$

$$\text{er}^{\ell}_D[\text{pred}^{\text{BEP}}_{\tau} \circ \mathbf{f}] - \text{er}^{\ell,*}_D \quad \leq \quad \frac{\left(\text{er}^{\psi^{\text{BEP}}}_D[\mathbf{f}] - \text{er}^{\psi^{\text{BEP}},*}_D\right)}{2\min(\tau, 1-\tau)}.$$

Remark The excess risk bounds for the CS, OVA, and BEP surrogates suggest that $\tau = \frac{1}{2}$ is the best choice for CS and BEP surrogates, while $\tau = 0$ is the best choice for the OVA surrogate. However, intuitively τ is the threshold converting confidence values to predictions, and so it makes sense to use τ values closer to 0 (or -1 in the case of OVA) to predict aggressively in low-noise situations, and use larger τ to predict conservatively in noisy situations. Practically, it makes sense to choose the parameter τ via cross-validation.

5. BEP Surrogate Optimization Algorithm

In this section, we frame the problem of finding the linear (vector valued) function that minimizes the BEP surrogate loss over a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, with $\mathbf{x}_i \in \mathbb{R}^a$ and $y_i \in [n]$, as a convex optimization problem. Once again, for simplicity we assume that the size of the label space is $n = 2^d$ for some $d \in \mathbb{Z}_+$. The primal and dual versions of the resulting optimization problem with a norm squared regularizer are given below.

Primal problem:

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_d,\xi_1,\ldots,\xi_m} \sum_{i=1}^{m} \xi_i + \frac{\lambda}{2} \sum_{j=1}^{d} ||\mathbf{w}_j||^2$$

$$\text{such that} \quad \forall i \in [m], j \in [d]$$

$$\xi_i \geq B_j(y_i)\mathbf{w}_j^\top \mathbf{x}_i + 1$$

$$\xi_i \geq 0$$

Dual problem:

$$\max_{\boldsymbol{\beta}\in\mathbb{R}^{m\times(d+1)}} -\sum_{i=1}^{m}\beta_{i,0} - \frac{1}{2\lambda}\sum_{i=1}^{m}\sum_{i'=1}^{m}\langle \mathbf{x}_i, \mathbf{x}_{i'}\rangle \mu_{i,i'}(\boldsymbol{\beta})$$

$$\text{such that} \quad \forall i \in [m], j \in [d]\cup\{0\}$$

$$\beta_{i,j} \geq 0; \qquad \sum_{j'=0}^{d}\beta_{i,j'} = 1\ .$$

where $\mu_{i,i'}(\boldsymbol{\beta}) = \sum_{j=1}^{d} B_j(y_i)B_j(y_{i'})\beta_{ij}\beta_{i',j}$.

We optimize the dual as it can be easily extended to work with kernels. The structure of the constraints in the dual lends itself easily to a block coordinate ascent algorithm, where we optimize over $\{\beta_{i,j} : j \in \{0,\ldots,d\}\}$ and fix every other variable in each iteration. Such methods have been recently proven to have exponential convergence rate for SVM-type problems [26], and we expect results of those type to apply to our problem as well.

The problem to be solved at every iteration reduces to a l_2 projection of a vector $\mathbf{g}^i \in \mathbb{R}^d$ on to the set $\mathcal{S}_i = \{\mathbf{g} \in \mathbb{R}^d : \mathbf{g}^\top \mathbf{b}^i \leq 1\}$, where $\mathbf{b}^i \in \{\pm 1\}^d$ is such that $b^i_j = B_j(y_i)$. The projection problem is a simple variant of projecting a vector on the l_1 ball of radius 1, which can be solved efficiently in $O(d)$ time [6]. The vector $\mathbf{g}^i$ is such that for any $j \in [d]$,

$$g^i_j = \frac{\lambda}{\langle \mathbf{x}_i, \mathbf{x}_i\rangle}\left(\mathbf{b}^i_j - \frac{1}{\lambda}\left(\sum_{i'=1;i'\neq i}^{m}\langle \mathbf{x}_i, \mathbf{x}_{i'}\rangle \beta_{i',j}B_j(y_{i'})\right)\right).$$

6. Extension to Abstain(α) Loss for $\alpha \leq \frac{1}{2}$

The excess risk bounds derived for the CS, OVA hinge loss and BEP surrogates apply only to the abstain$\left(\frac{1}{2}\right)$ loss. But it is possible to derive such excess risk bounds for abstain(α) with $\alpha \in \left[0, \frac{1}{2}\right]$ with slight modifications to the CS, OVA and BEP surrogates.

Let $\gamma(a) = \max(a, -1)$ and $B : [n] \to \{-1, 1\}^d$ be any bijection. Define $\psi^{\mathrm{CS},\alpha} : [n] \times \mathbb{R}^n \to \mathbb{R}_+$, $\psi^{\mathrm{OVA},\alpha} : [n] \times \mathbb{R}^n \to \mathbb{R}_+$ and $\psi^{\mathrm{BEP},\alpha} : [n] \times \mathbb{R}^d \to \mathbb{R}_+$, with $n = 2^d$ as

$$\begin{aligned}
\psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) &= 2 \cdot \max\left(\alpha \max_{j\neq y}\gamma(u_j - u_y), (1-\alpha)\max_{j\neq y}\gamma(u_j - u_y)\right) + 2\alpha \\
\psi^{\mathrm{OVA},\alpha}(y, \mathbf{u}) &= 2 \cdot \left(\sum_{i=1}^{n}\left(\mathbf{1}(y = i)\alpha(1-u_i)_+ + \mathbf{1}(y \neq i)(1-\alpha)(1+u_i)_+\right)\right) \\
\psi^{\mathrm{BEP},\alpha}(y, \mathbf{u}) &= 2 \cdot \max\left(\alpha \max_{j\in[d]}\gamma(B_j(y)u_j), (1-\alpha)\max_{j\in[d]}\gamma(B_j(y)u_j)\right) + 2\alpha
\end{aligned}$$

Note that $\psi^{\mathrm{CS},\frac{1}{2}} = \psi^{\mathrm{CS}}$, $\psi^{\mathrm{OVA},\frac{1}{2}} = \psi^{\mathrm{OVA}}$ and $\psi^{\mathrm{BEP},\frac{1}{2}} = \psi^{\mathrm{BEP}}$.

For any $\mathbf{p} \in \Delta_n$, the $\mathbf{u} \in \mathbb{R}^n$ that optimises $\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA}}(.), \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS}}(.)$ and the $\mathbf{u} \in \mathbb{R}^d$ that optimises $\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP}}(.)$ takes one of $n+1$ possible values. The modifications to these surrogates change the optimal values in the exact way to ensure that the modified surrogates are optimal for the abstain(α) loss. See equations 8.15 and 8.16 for the optimal $\mathbf{u}$ values for the OVA surrogate, equations 8.1 and 8.2 for the CS surrogate and equation 8.28 and 8.29 for the BEP surrogate.

One can get similar excess risk bounds for these modified surrogates as shown in Theorem below, the proof of which is in Section 8.

Theorem 6.1. *Let $n \in \mathbb{Z}_+, \tau \in (0,1), \tau' \in (-1,1)$ and $\alpha \in \left[0, \frac{1}{2}\right]$. Let $n = 2^d$. Then, for all $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$, $\mathbf{g} : \mathcal{X} \to \mathbb{R}^n$,*

$$\begin{aligned}
\mathrm{er}_D^{\ell^\alpha}[\mathrm{pred}_\tau^{\mathrm{CS}} \circ \mathbf{g}] - \mathrm{er}_D^{\ell^\alpha,*} &\leq \frac{1}{2\min(\tau, 1-\tau)} \left(\mathrm{er}_D^{\psi^{\mathrm{CS},\alpha}}[\mathbf{g}] - \mathrm{er}_D^{\psi^{\mathrm{CS},\alpha},*} \right), \\
\mathrm{er}_D^{\ell^\alpha}[\mathrm{pred}_{\tau'}^{\mathrm{OVA}} \circ \mathbf{g}] - \mathrm{er}_D^{\ell^\alpha,*} &\leq \frac{1}{2(1-|\tau'|)} \left(\mathrm{er}_D^{\psi^{\mathrm{OVA},\alpha}}[\mathbf{g}] - \mathrm{er}_D^{\psi^{\mathrm{OVA},\alpha},*} \right), \\
\mathrm{er}_D^{\ell^\alpha}[\mathrm{pred}_\tau^{\mathrm{BEP}} \circ \mathbf{f}] - \mathrm{er}_D^{\ell^\alpha,*} &\leq \frac{1}{2\min(\tau, 1-\tau)} \left(\mathrm{er}_D^{\psi^{\mathrm{BEP},\alpha}}[\mathbf{f}] - \mathrm{er}_D^{\psi^{\mathrm{BEP},\alpha},*} \right).
\end{aligned}$$

Remark When $n = 2$, the Crammer-Singer surrogate, the one vs all hinge and the BEP surrogate all reduce to the hinge loss and α is restricted to be at most $\frac{1}{2}$ to ensure the relevance of the abstain option. Applying the above extension for $\alpha \leq \frac{1}{2}$ to the hinge loss, we get the 'generalized hinge loss' of Bartlett and Wegkamp [2].

7. Experimental Results

In this section, we give our experimental results for the proposed algorithms on both synthetic and real datasets. The synthetic data experiments illustrate the consistency of the three proposed algorithms for the abstain loss. The experiments on real data illustrate that one can achieve lower error rates on multiclass datasets if the classifier is allowed to abstain, and also show that the BEP algorithm has competitive performance with the other two algorithms

7.1. Synthetic Data

We optimize the Crammer-Singer surrogate, the one vs all hinge surrogate and the BEP surrogate, over appropriate kernel spaces on a synthetic data set and show that the abstain$\left(\frac{1}{2}\right)$ loss incurred by the trained model for all three algorithms approaches the Bayes optimal under various thresholds.

The dataset we used, with $n = 8$ classes and 2-dimensional features, was generated as follows. We randomly sample 8 prototype vectors $\mathbf{v}_1, \ldots, \mathbf{v}_8 \in \mathbb{R}^2$, with each $\mathbf{v}_y$ drawn independently from a zero mean unit variance 2D-Gaussian, $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$ distribution. These 8 prototype vectors correspond to the 8 classes.

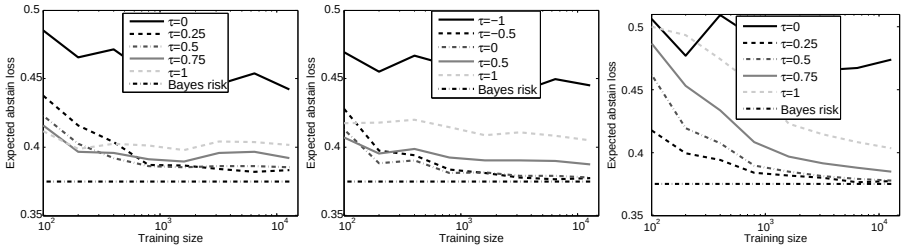

FIG 3. *Results on synthetic data: Performance of the CS (left), OVA (middle) and BEP (right) surrogates for various thresholds as a function of training sample size.*

Each example $(\mathbf{x}, y)$ is generated by first picking y from one of the 8 classes uniformly at random, and the instance $\mathbf{x}$ is set as $\mathbf{x} = \mathbf{v}_y + 0.65 \cdot \mathbf{u}$, where $\mathbf{u}$ is independently drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. We generated 12800 such $(\mathbf{x}, y)$ pairs for training, and another 10000 examples each for testing and hyper-parameter validation .

The CS, OVA, BEP surrogates were all optimized over a reproducing kernel Hilbert Space (RKHS) with a Gaussian kernel and the standard norm-squared regularizer. The kernel width parameter and the regularization parameter were chosen by grid search using the separate validation set.[2]

As Figure 3 indicates, the expected abstain risk incurred by the trained model approaches the Bayes risk with increasing training data for all three algorithms and intermediate τ values. The excess risk bounds in Theorems 3.1 and 4.1 break down when the threshold parameter τ lies in $\{0, 1\}$ for the CS and BEP surrogates, and in $\{-1, 1\}$ for the OVA surrogate. This is supported by the observation that, in Figure 3 the curves corresponding to these thresholds perform poorly. In particular, using $\tau = 0$ for the CS and BEP algorithms implies that the resulting algorithms never abstain.

Though all three surrogate minimizing algorithms we consider are consistent w.r.t. abstain loss, we find that the BEP and OVA algorithms use less computation time and samples than the CS algorithm to attain the same error. We note that for the BEP surrogate to perform well as above, it is critical to use a flexible function class (such as the RBF kernel induced RKHS as above). In particular, when optimized over a linear kernel function class the BEP surrogate performs poorly (experiments not shown here), due to its restricted representation power.

7.2. Real Data

We ran experiments on real multiclass datasets from the UCI repository, the details of which are in Table 1. In the `yeast`, `letter`, `vehicle` and `image`

[2]We used Joachims' SVM-light package [14] for the OVA and CS algorithms.

TABLE 1
Details of datasets used.

	# Train	# Test	# Feat	# Class
satimage	4,435	2,000	36	6
yeast	1,000	484	8	10
letter	16,000	4,000	16	26
vehicle	700	146	18	4
image	2,000	310	19	7
covertype	15,120	565,892	54	7

TABLE 2
Error percentages (as a fraction of all *test instances) of the three algorithms when the rejection percentage is fixed at 0%, 20% and 40%.*

Reject:	0%			20%			40%		
Algorithm:	CS	OVA	BEP	CS	OVA	BEP	CS	OVA	BEP
satimage	12.2	8.5	8.1	6.8	2.3	2.3	2.9	0.9	0.5
yeast	39.4	40.5	39.4	26.8	27.6	27.2	18.1	18.9	18.1
letter	4.2	2.5	4.6	1.2	0.1	0.6	0.3	0.0	0.0
vehicle	31.5	19.1	20.5	24.1	9.4	13.1	16.0	5.4	6.1
image	5.8	3.8	4.8	1.9	0.9	0.9	0.7	0.6	0.6
covertype	31.8	27.9	29.4	23.1	18.3	20.4	16.2	10.9	12.8

datasets, a standard train/test split is not indicated, hence we create a random split ourselves.

All three algorithms (CS, OVA and BEP) were optimized over an RKHS with a Gaussian kernel and the standard norm-squared regularizer. The kernel width and regularization parameters were chosen through validation – 10-fold cross-validation in the case of `satimage`, `yeast`, `vehicle` and `image` datasets, and a 75-25 split of the train set into train and validation for the `letter` and `covertype` datasets. For simplicity we set $\tau = 0$ (or $\tau = -1$ for OVA) during the validation phase in the first set of experiments. In the second set of experiments, we chose the value of τ along with the kernel width and regularisation parameters to optimise the abstain($\frac{1}{2}$) loss.

The results of the first set of experiments with the CS, OVA and BEP algorithms are given in Table 2. The rejection rate is fixed at some given level (0%, 20% and 40%) by choosing the threshold τ for each algorithm and dataset appropriately. As can be seen from the Table, the BEP algorithm's performance is comparable to the OVA, and is better than the CS algorithm. However, Table 4, which gives the training and testing times for the algorithms, reveals that the BEP algorithm runs the fastest, thus making the BEP algorithm a good option for large datasets. The main reason for the observed speedup of the BEP is that it learns only $\log_2(n)$ functions for a n-class problem and hence the speedup factor of the BEP over the OVA would potentially be better for larger n.

In the second set of experiments we fix the cost of abstaining α, to be equal to $\frac{1}{2}$. The kernel width, regularisation and threshold parameters are chosen to optimise the abstain($\frac{1}{2}$) loss in the validation phase. The abstain($\frac{1}{2}$) loss values

TABLE 3
The Abstain($\frac{1}{2}$) loss values for the CS, OVA, and BEP algorithms. The regularisation, kernel width and threshold parameters are tuned on the validation set.

Algorithm:	CS	OVA	BEP
satimage	0.122	0.085	0.080
yeast	0.376	0.361	0.382
letter	0.042	0.025	0.047
vehicle	0.328	0.184	0.201
image	0.058	0.039	0.051
covertype	0.319	0.275	0.294

TABLE 4
Total train time (total test time) in seconds.

Algorithm	CS	OVA	BEP
satimage	1582(49)	86(9)	42(5)
yeast	10(2)	6(1)	2(1)
letter	2527(180)	635(42)	220(13)
vehicle	5(0)	3(0)	1(0)
image	211(5)	16(1)	5(0)
covertype	9939(30721)	10563(13814)	502(3943)

for the CSA, OVA and BEP algorithm with tuned thresholds are given in Table 3. The most interesting values for this are on the `vehicle` and `yeast` dataset, where the final algorithms chose thresholds that abstain in the test set and perform marginally better than predicting some class on all instances, the loss values for which are simply given by the first three columns of Table 2.

8. Proofs

Both Theorems 3.1 and 4.1 follow from Theorem 6.1, whose proof we divide into three separate parts below.

8.1. Modified Crammer-Singer Surrogate

Let $\gamma(a) = \max(a, -1)$. We have,

$$\begin{aligned}
\psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) &= 2 \cdot \max\left(\alpha \max_{j \neq y} \gamma(u_j - u_y), (1-\alpha) \max_{j \neq y} \gamma(u_j - u_y)\right) + 2\alpha, \\
\mathrm{pred}_\tau^{\mathrm{CS}}(\mathbf{u}) &= \begin{cases} \mathrm{argmax}_{i \in [n]} u_i & \text{if } u_{(1)} - u_{(2)} > \tau \\ \perp & \text{otherwise} \end{cases}
\end{aligned}$$

Define the sets $\mathcal{U}_1, \ldots, \mathcal{U}_n, \mathcal{U}_\perp$ such that $\mathcal{U}_i$ is the set of vectors $\mathbf{u}$ in $\mathbb{R}^n$, for which $\mathrm{pred}_\tau^{\mathrm{CS}}(\mathbf{u}) = i$

$$\begin{aligned}
\mathcal{U}_y^\tau &= \{\mathbf{u} \in \mathbb{R}^n : u_y > u_j + \tau \text{ for all } j \neq y\}; \quad y \in [n] \\
\mathcal{U}_\perp^\tau &= \{\mathbf{u} \in \mathbb{R}^n : u_{(1)} \leq u_{(2)} + \tau\}.
\end{aligned}$$

The following lemma gives some crucial, but straightforward to prove, (in)equalities satisfied by the Crammer-Singer surrogate.

Lemma 8.1. *Let* $\alpha \in [0, \frac{1}{2}]$.

$$\begin{aligned} & \forall y \in [n], \quad \forall \mathbf{p} \in \Delta_n \\ \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{e}_y) &= 2(1-p_y), && (8.1)\\ \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{0}) &= 2\alpha, && (8.2)\\ & \forall \mathbf{u} \in \mathbb{R}^n, \forall y \in \mathrm{argmax}_i u_i, \forall y' \notin \mathrm{argmax}_i u_i \\ \psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) &= 2\alpha \cdot (u_{(2)} - u_{(1)} + 1)_+, && (8.3)\\ \psi^{\mathrm{CS},\alpha}(y', \mathbf{u}) &\geq 2(1-\alpha)(u_{(1)} - u_{(2)}) + 2\alpha, && (8.4) \end{aligned}$$

where $\mathbf{e}_y$ *is the vector in* $\mathbb{R}^n$ *with* 1 *in the* y^{th} *position and* 0 *everywhere else.*

We will prove the following theorem.

Theorem 8.2. *Let* $\alpha \in [0, \frac{1}{2}], n \in \mathbb{N}, \tau \in (0,1)$,. *Then for all* $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$

$$\mathrm{er}_D^{\ell^\alpha}[\mathrm{pred}_\tau^{\mathrm{CS}} \circ \mathbf{f}] - \mathrm{er}_D^{\ell^\alpha,*} \leq \frac{1}{2\min(\tau, 1-\tau)} \left(\mathrm{er}_D^{\psi^{\mathrm{CS},\alpha}}[\mathbf{f}] - \mathrm{er}_D^{\psi^{\mathrm{CS},\alpha},*} \right)$$

Proof. We will show that $\forall \mathbf{p} \in \Delta_n$ and all $\mathbf{u} \in \mathbb{R}^d$

$$\begin{aligned} & \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}) - \inf_{\mathbf{u}' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}') \\ &\geq 2\min(\tau, 1-\tau)\left(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}_\tau^{\mathrm{CS}}(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t \right). && (8.5) \end{aligned}$$

The Theorem simply follows from linearity of expectation.

Case 1: $p_y \geq 1 - \alpha$ for some $y \in [n]$.

We have that $y \in \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t$.

Case 1a: $\mathbf{u} \in \mathcal{U}^\tau_y$

The RHS of Equation (8.5) is zero, and hence becomes trivial.

Case 1b: $\mathbf{u} \in \mathcal{U}^\tau_\perp$

We have that $u_{(1)} - u_{(2)} \leq \tau$. Let $q = \sum_{i \in \mathrm{argmax}_j u_j} p_i$. We then have

$$\begin{aligned} & \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{e}_y) \\ &\overset{(8.1)}{=} \sum_{i: u_i = u_{(1)}} p_i \psi^{\mathrm{CS},\alpha}(i, \mathbf{u}) + \sum_{i: u_i < u_{(1)}} p_y \psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) - 2(1-p_y) \\ &\overset{(8.3),(8.4)}{\geq} 2q\alpha(u_{(2)} - u_{(1)} + 1) \\ &\quad +2(1-q)((1-\alpha)(u_{(1)} - u_{(2)}) + \alpha) - 2(1-p_y) \\ &= 2(\alpha + p_y - 1) + 2(u_{(2)} - u_{(1)})(\alpha + q - 1) \\ &\geq 2(p_y + \alpha - 1)(1-\tau) \quad . && (8.6) \end{aligned}$$

The last inequality follows from $u_{(2)} - u_{(1)} \geq -\tau$ because, if $q > p_y$ then $u_{(1)} = u_{(2)}$.

$$\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}_\tau^{\mathrm{CS}}(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}^\alpha_\perp - \mathbf{p}^\top \boldsymbol{\ell}^\alpha_y = p_y + \alpha - 1 \qquad (8.7)$$

From Equations (8.6) and (8.7) we have

$$\begin{aligned}\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}) - \inf_{\mathbf{u}' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}') \\ \geq \quad 2(1-\tau)\big(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t\big)\end{aligned} \tag{8.8}$$

Case 1c: $\mathbf{u} \in \mathbb{R}^n \setminus (\mathcal{U}^\tau_y \cup \mathcal{U}^\tau_\perp)$

We have $\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u}) = y' \neq y$. Also $p_{y'} \leq 1 - p_y \leq \alpha$ and $u_{(1)} = u_{y'} > u_{(2)} + \tau$.

Let $\mathbf{u}' \in \mathbb{R}^n$ be such that $u'_y = u_{y'}, u'_{y'} = u_y$ and $u_i = u'_i$ for all $i \notin \{y, y'\}$. We have

$$\begin{aligned}&\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}') \\ &= \quad p_y \psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) + p_{y'} \psi^{\mathrm{CS},\alpha}(y', \mathbf{u}) - \big(p_y \psi^{\mathrm{CS},\alpha}(y, \mathbf{u}') + p_{y'} \psi^{\mathrm{CS},\alpha}(y', \mathbf{u}')\big) \\ &= \quad p_y(\psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) - \psi^{\mathrm{CS},\alpha}(y, \mathbf{u}')) - p_{y'}(\psi^{\mathrm{CS},\alpha}(y', \mathbf{u}') - \psi^{\mathrm{CS},\alpha}(y', \mathbf{u})) \\ &= \quad (p_y - p_{y'})(\psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) - \psi^{\mathrm{CS},\alpha}(y, \mathbf{u}')) \\ &\overset{(8.3),(8.4)}{\geq} \quad (p_y - p_{y'})(2(1-\alpha)(u_{(1)} - u_{(2)}) + 2\alpha - 2\alpha(u_{(2)} - u_{(1)} + 1)_+) \\ &\geq \quad (p_y - p_{y'})(2(1-\alpha)(\tau) + 2\alpha - 2\alpha(-\tau + 1)) \\ &= \quad (p_y - p_{y'})(2\tau)\end{aligned} \tag{8.9}$$

The second inequality above follows from the reasoning that the term is minimized when $(u_{(1)} - u_{(2)})$ is as small as possible, which is τ in this case.

We also have that

$$\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}^\alpha_{y'} - \mathbf{p}^\top \boldsymbol{\ell}^\alpha_y = p_y - p_{y'} \tag{8.10}$$

From Equations (8.9) and (8.10) we have

$$\begin{aligned}\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}) - \inf_{\mathbf{u}' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}') \\ \geq \quad 2\tau\big(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t\big)\end{aligned} \tag{8.11}$$

Case 2: $p_{y'} < 1 - \alpha$ for all $y' \in [n]$

We have that $\perp \in \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t$

Case 2a: $\mathbf{u} \in \mathcal{U}^\tau_\perp$ (or $\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u}) = \perp$)

The RHS of Equation (8.5) is zero, and hence becomes trivial.

Case 2b: $\mathbf{u} \in \mathbb{R}^n \setminus \mathcal{U}^\tau_\perp$ (or $\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u}) \neq \perp$)

Let $\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u}) = \mathrm{argmax}_i u_i = y$. We have that $u_{(1)} = u_y > u_{(2)} + \tau$ and

$p_y < 1 - \alpha$.

$$\begin{aligned}
&\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{0}) \\
&\overset{(8.2)}{=} \left(\sum_{i=1; i \neq y}^{n} p_i \psi^{\mathrm{CS},\alpha}(i, \mathbf{u}) + p_y \psi^{\mathrm{CS},\alpha}(y, \mathbf{u}) \right) - 2\alpha \\
&\overset{(8.3),(8.4)}{\geq} 2(1 - p_y)(1 - \alpha)(u_{(1)} - u_{(2)}) + 2\alpha(1 - p_y) \\
&\qquad\qquad + 2p_y\alpha(u_{(2)} - u_{(1)} + 1) - 2\alpha \\
&= (u_{(1)} - u_{(2)})(2(1 - p_y)(1 - \alpha) - 2\alpha p_y) \\
&\geq 2(1 - p_y - \alpha)(\tau) \qquad (8.12)
\end{aligned}$$

We also have that

$$\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}^\alpha_y - \mathbf{p}^\top \boldsymbol{\ell}^\alpha_\perp = 1 - \alpha - p_y \qquad (8.13)$$

From Equations (8.12) and (8.13) we have

$$\begin{aligned}
&\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}) - \inf_{u' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{CS},\alpha}(\mathbf{u}') \\
&\geq 2\tau \big(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{CS}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t \big) \qquad (8.14)
\end{aligned}$$

Equation (8.5), and hence the theorem, follows from Equations (8.8), (8.11) and (8.14).

□

8.2. Modified One-vs-All Hinge

We have

$$\begin{aligned}
\psi^{\mathrm{OVA},\alpha}(y, \mathbf{u}) &= 2 \cdot \left(\sum_{i=1}^{n} \Big(\mathbf{1}(y = i)\alpha(1 - u_i)_+ + \mathbf{1}(y \neq i)(1 - \alpha)(1 + u_i)_+ \Big) \right) \\
\mathrm{pred}^{\mathrm{OVA}}_\tau(\mathbf{u}) &= \begin{cases} \mathrm{argmax}_{i \in [n]} u_i & \text{if } \max_j u_j > \tau \\ \perp & \text{otherwise} \end{cases}
\end{aligned}$$

Define the sets $\mathcal{U}^\tau_1, \ldots, \mathcal{U}^\tau_n, \mathcal{U}^\tau_\perp$ such that $\mathcal{U}_i$ is the set of vectors $\mathbf{u}$ in $\mathbb{R}^n$, for which $\mathrm{pred}^{\mathrm{OVA}}_\tau(\mathbf{u}) = i$

$$\begin{aligned}
\mathcal{U}^\tau_y &= \{\mathbf{u} \in \mathbb{R}^n : u_y > \tau, y = \mathrm{argmax}_{i \in [n]} u_i\}, \ y \in [n] \\
\mathcal{U}^\tau_\perp &= \{\mathbf{u} \in \mathbb{R}^n : u_j \leq \tau \text{ for all } j \in [n]\}.
\end{aligned}$$

The following lemma gives some crucial, but straightforward to prove, (in)equalities satisfied by the OVA hinge surrogate.

Lemma 8.3.

$$\begin{aligned} &\forall y \in [n], \forall \mathbf{p} \in \Delta_n\,, \forall \mathbf{u} \in [-1,1]^n \\ \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(2\mathbf{e}_y - \mathbf{1}) &= 4(1-p_y) && (8.15)\\ \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(-\mathbf{1}) &= 4\alpha && (8.16)\\ \psi^{\mathrm{OVA}}(y,\mathbf{u}) &= 2((1-\alpha)\sum_{j\neq y} u_j - \alpha u_y) + c && (8.17) \end{aligned}$$

where $c = 2((1-\alpha)(n-1)+\alpha)$.

Theorem 8.4. *Let* $n \in \mathbb{N}, \tau \in (0,1)$ *and* $\alpha \in \left[0, \frac{1}{2}\right]$. *Then for all* $\mathbf{f} : \mathcal{X} \to \mathbb{R}^n$,

$$\mathrm{er}_D^{\ell^\alpha}[\mathrm{pred}_\tau^{\mathrm{OVA}} \circ \mathbf{f}] - \mathrm{er}_D^{\ell^\alpha,*} \;\leq\; \frac{1}{2(1-|\tau|)}\left(\mathrm{er}_D^{\psi^{\mathrm{OVA},\alpha}}[\mathbf{f}] - \mathrm{er}_D^{\psi^{\mathrm{OVA},\alpha},*}\right).$$

Proof. We will show that $\forall \mathbf{p} \in \Delta_n$ and all $\mathbf{u} \in [-1,1]^n$

$$\begin{aligned} &\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(\mathbf{u}) - \inf_{u' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(\mathbf{u}') \\ &\geq\; 2(1-|\tau|)\big(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}_\tau^{\mathrm{OVA},\alpha}(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t\big), \qquad (8.18) \end{aligned}$$

the Theorem simply follows from the observation that for all $\mathbf{u} \in \mathbb{R}^n$ clipping the components of $\mathbf{u}$ to $[-1,1]$ does not increase $\psi^{\mathrm{OVA}}(y,\mathbf{u})$ for any y.

Case 1: $p_y \geq 1-\alpha$ for some $y \in [n]$.

We have that $y \in \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t$.

Case 1a: $\mathbf{u} \in [-1,1]^n \cap \mathcal{U}^\tau_y$

The RHS of Equation (8.18) is zero, and hence becomes trivial.

Case 1b: $\mathbf{u} \in [-1,1]^n \cap \mathcal{U}^\tau_\perp$

We have that $\max_j u_j \leq \tau$. And hence

$$\begin{aligned} &\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(2\mathbf{e}_y - 1) \\ &\overset{(8.15)}{=} \sum_{i=1}^n p_i \psi^{\mathrm{OVA},\alpha}(i,\mathbf{u}) - 4(1-p_y) \\ &\overset{(8.17)}{=} \sum_{i=1}^n 2p_i\left((1-\alpha)\sum_{j\neq i} u_j - \alpha u_i\right) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\ &= \sum_{i=1}^n 2u_i((1-\alpha)(1-p_i) - p_i\alpha) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\ &= \sum_{i=1}^n 2u_i(1-\alpha-p_i) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\ &\geq \sum_{i\in[n]\setminus\{y\}} 2(-1)(1-\alpha-p_i) + 2\tau(1-\alpha-p_y) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\ &= 2(1-p_y+(n-1)(\alpha-1)) + 2\tau(1-\alpha-p_y) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\ &= 2(1-\tau)(\alpha+p_y-1) \qquad (8.19) \end{aligned}$$

We also have

$$\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{OVA}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}^\alpha_\perp - \mathbf{p}^\top \boldsymbol{\ell}^\alpha_y = p_y + \alpha - 1 \tag{8.20}$$

From Equations (8.19) and (8.20) we have for all $\mathbf{u} \in [-1,1]^n \cap \mathcal{U}^\tau_\perp$

$$\begin{aligned} &\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(\mathbf{u}) - \inf_{u' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(\mathbf{u}') \\ &\geq \quad 2(1-\tau)\big(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{OVA}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t\big) \end{aligned} \tag{8.21}$$

Case 1c: $\mathbf{u} \in [-1,1]^n \setminus (\mathcal{U}^\tau_y \cup \mathcal{U}^\tau_\perp)$

We have $\mathrm{pred}^{\mathrm{OVA}}_\tau(\mathbf{u}) = y' \neq y$. Also $p_{y'} \leq \alpha$; $u_{y'} > \tau$ and $u_{y'} \geq u_y$.

$$\begin{aligned}
&\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA},\alpha}(2\mathbf{e}_y - 1) \\
&\overset{(8.15)}{=} \quad \sum_{i=1}^n p_i \psi^{\mathrm{OVA},\alpha}(i,\mathbf{u}) - 4(1-p_y) \\
&\overset{(8.17)}{=} \quad \sum_{i=1}^n 2p_i \left((1-\alpha)\sum_{j\neq i} u_j - \alpha u_i\right) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\
&= \quad \sum_{i=1}^n 2u_i((1-\alpha)(1-p_i) - p_i\alpha) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\
&= \quad \sum_{i=1}^n 2u_i(1-\alpha-p_i) + 2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\
&\geq \quad \sum_{i\in[n]\setminus\{y,y'\}} 2(-1)(1-\alpha-p_i) + 2u_{y'}(2-2\alpha-p_y-p_{y'}) \\
&\qquad\qquad +2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\
&\geq \quad 2(1-p_y-p_{y'}+(n-2)(\alpha-1)) + 2\tau(2-2\alpha-p_y-p_{y'}) \\
&\qquad\qquad +2((1-\alpha)(n-1)+\alpha) - 4(1-p_y) \\
&= \quad 2(-p_y-p_{y'}) + 2\tau(2-2\alpha-p_y-p_{y'}) + 4p_y \\
&= \quad 4\tau(1-\alpha) + (p_y+p_{y'})(-2-2\tau) + 4p_y \\
&= \quad 2p_y(1-\tau) + 4\tau(1-\alpha) - 2(1+\tau)p_{y'} \\
&= \quad 2p_y(1-\tau) + 4\tau(1-\alpha) - 2(1-\tau)p_{y'} - 4\tau p_{y'} \\
&= \quad 2(p_y-p_{y'})(1-\tau) + 4\tau(1-\alpha-p_{y'}) \\
&\geq \quad 2(p_y-p_{y'})(1-\tau)
\end{aligned} \tag{8.22}$$

We also have that

$$\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{OVA}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}^\alpha_{y'} - \mathbf{p}^\top \boldsymbol{\ell}^\alpha_y = p_y - p_{y'} \tag{8.23}$$

From Equations (8.22) and (8.23) we have for all $\mathbf{u} \in [-1,1]^n \setminus (\mathcal{U}^\tau_y \cup \mathcal{U}^\tau_\perp)$

$$\begin{aligned} &\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA}}(\mathbf{u}) - \inf_{u' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{OVA}}(\mathbf{u}') \\ &\geq \quad 2(1-\tau)\big(\mathbf{p}^\top \boldsymbol{\ell}_{\mathrm{pred}^{\mathrm{OVA}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}_t\big) \end{aligned} \tag{8.24}$$

Case 2: $p_{y'} < 1-\alpha$ for all $y' \in [n]$
We have that $\perp \in \operatorname{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t$
Case 2a: $\mathbf{u} \in \mathcal{U}^\tau_\perp$
The RHS of Equation (8.18) is zero, and hence becomes trivial.
Case 2b: $\mathbf{u} \in [-1,1]^n \setminus \mathcal{U}^\tau_\perp$
Let $\text{pred}^{\text{OVA}}_\tau(\mathbf{u}) = \operatorname{argmax}_i u_i = y$. We have that $u_y \geq \tau$.

$$
\begin{aligned}
&\mathbf{p}^\top \boldsymbol{\psi}^{\text{OVA},\alpha}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\text{OVA},\alpha}(-\mathbf{1}) \\
&\overset{(8.15)}{=} \sum_{i=1}^n p_i \psi^{\text{OVA},\alpha}(i,\mathbf{u}) - 4\alpha \\
&\overset{(8.17)}{=} \sum_{i=1}^n 2p_i \left((1-\alpha) \sum_{j \neq i} u_j - \alpha u_i \right) + 2((1-\alpha)(n-1)+\alpha) - 4\alpha \\
&= \sum_{i=1}^n 2u_i((1-\alpha)(1-p_i) - p_i\alpha) + 2((1-\alpha)(n-1)+\alpha) - 4\alpha \\
&= \sum_{i \in [n] \setminus \{y\}} 2u_i(1-\alpha-p_i) + 2u_y(1-\alpha-p_y) + 2((1-\alpha)(n-1)+\alpha) - 4\alpha \\
&\geq \sum_{i \in [n] \setminus \{y\}} 2(-1)(1-\alpha-p_i) + 2\tau(1-\alpha-p_y) + 2((1-\alpha)(n-1)+\alpha) - 4\alpha \\
&= 2(1-p_y+(n-1)(\alpha-1)) + 2\tau(1-\alpha-p_y) + 2((1-\alpha)(n-1)+\alpha) - 4\alpha \\
&= 2(1+\tau)(1-\alpha-p_y) \qquad (8.25)
\end{aligned}
$$

We also have that

$$
\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\text{pred}^{\text{OVA}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}^\alpha_y - \mathbf{p}^\top \boldsymbol{\ell}^\alpha_\perp = 1-\alpha-p_y \qquad (8.26)
$$

From Equations (8.25) and (8.26) we have for all $\mathbf{u} \in [-1,1]^n \setminus \mathcal{U}^\tau_\perp$

$$
\begin{aligned}
&\mathbf{p}^\top \boldsymbol{\psi}^{\text{OVA},\alpha}(\mathbf{u}) - \inf_{\mathbf{u}' \in \mathbb{R}^n} \mathbf{p}^\top \boldsymbol{\psi}^{\text{OVA},\alpha}(\mathbf{u}') \\
&\geq 2(1+\tau)\left(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\text{pred}^{\text{OVA}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t\right) \qquad (8.27)
\end{aligned}
$$

Equation (8.18), and hence the theorem, follows from Equations (8.21), (8.24) and (8.27).

□

8.3. Modified Binary Encoded Predictions Surrogate

Let $\gamma(a) = \max(a,-1)$, and $B : [n] \to \{-1,+1\}^d$ be any bijection, we then have

$$
\begin{aligned}
\psi^{\text{BEP},\alpha}(y,\mathbf{u}) &= 2 \cdot \max\left(\alpha \max_{j \in [d]} \gamma(B_j(y)u_j), (1-\alpha) \max_{j \in [d]} \gamma(B_j(y)u_j) \right) + 2\alpha \\
\text{pred}^{\text{BEP}}_\tau(\mathbf{u}) &= \begin{cases} \perp & \text{if } \min_{i \in [d]} |u_i| \leq \tau \\ B^{-1}(\text{sign}(-\mathbf{u})) & \text{Otherwise} \end{cases}
\end{aligned}
$$

Define the sets $\mathcal{U}_1^\tau, \ldots, \mathcal{U}_n^\tau, \mathcal{U}_\perp^\tau$, where $\mathcal{U}_k^\tau = \{\mathbf{u} \in \mathbb{R}^d : \text{pred}_\tau^{\text{BEP}}(\mathbf{u}) = k\}$. Which evaluates to

$$\begin{aligned}\mathcal{U}_y^\tau &= \{\mathbf{u} \in \mathbb{R}^d : \max_j B_j(y) u_j < -\tau\} \quad \text{for } y \in [n] \\ \mathcal{U}_\perp^\tau &= \{\mathbf{u} \in \mathbb{R}^d : \min_j |u_j| \leq \tau\}\end{aligned}$$

The following lemma gives some crucial, but straightforward to prove, (in)equalities satisfied by the BEP surrogate.

Lemma 8.5.

$$\begin{aligned}&\forall y, y' \in [n], \mathbf{p} \in \Delta_n, \mathbf{u} \in \mathbb{R}^d, y' \neq B^{-1}(\text{sign}(-\mathbf{u})) \\ \mathbf{p}^\top \boldsymbol{\psi}^{\text{BEP}}(-B(y)) &= 2(1-p_y) &(8.28)\\ \mathbf{p}^\top \boldsymbol{\psi}^{\text{BEP}}(\mathbf{0}) &= 2\alpha &(8.29)\\ \psi^{\text{BEP}}(B^{-1}(\text{sign}(-\mathbf{u})), \mathbf{u}) &\geq -2\alpha \min_j |u_j| + 2\alpha &(8.30)\\ \psi^{\text{BEP}}(y', \mathbf{u}) &\geq 2(1-\alpha) \min_j |u_j| + 2\alpha &(8.31)\end{aligned}$$

Theorem 8.6. *Let $n \in \mathbb{N}$ and $\tau \in (0,1)$. Let $n = 2^d$. Then for all $f : \mathcal{X} \to \mathbb{R}^d$*

$$\text{er}_D^\ell[\text{pred}_\tau^{\text{BEP}} \circ \mathbf{f}] - \text{er}_D^{\ell,*} \leq \frac{\left(\text{er}_D^{\psi^{\text{BEP}}}[\mathbf{f}] - \text{er}_D^{\psi^{\text{BEP}},*}\right)}{2\min(\tau, 1-\tau)}$$

Proof. We will show that $\forall \mathbf{p} \in \Delta_n$ and all $\mathbf{u} \in \mathbb{R}^d$

$$\begin{aligned}&\mathbf{p}^\top \boldsymbol{\psi}^{\text{BEP},\alpha}(\mathbf{u}) - \inf_{u' \in \mathbb{R}^d} \mathbf{p}^\top \boldsymbol{\psi}^{\text{BEP},\alpha}(\mathbf{u}') \\ &\geq 2\min(\tau, 1-\tau)(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\text{pred}_\tau^{\text{BEP}}(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}_t^\alpha) &(8.32)\end{aligned}$$

The theorem follows by linearity of expectation.

Case 1: $p_y \geq 1-\alpha$ for some $y \in [n]$

We have that $y \in \text{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}_t^\alpha$

Case 1a: $\mathbf{u} \in \mathcal{U}_y^\tau$ (or $\text{pred}_\tau^{\text{BEP}}(\mathbf{u}) = y$)

The RHS of Equation (8.32) is zero, and hence becomes trivial.

Case 1b: $\mathbf{u} \in \mathcal{U}_\perp^\tau$ (or $\text{pred}_\tau^{\text{BEP}}(\mathbf{u}) = \perp$)

Let $y' = B^{-1}(\text{sign}(-\mathbf{u}))$. We have $\min_j |u_j| \leq \tau$.

$$\begin{aligned}&\mathbf{p}^\top \boldsymbol{\psi}^{\text{BEP}}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\text{BEP}}(-B(y)) \\ &\overset{(8.28)}{=} p_{y'} \psi^{\text{BEP}}(y', \mathbf{u}) + \sum_{i \in [n] \setminus \{y'\}} p_i \psi^{\text{BEP}}(i, \mathbf{u}) - 2(1-p_y) \\ &\overset{(8.30),(8.31)}{\geq} 2\alpha p_{y'}(-\min_{j \in [d]} |u_j|) + 2(1-\alpha)(1-p_{y'})(\min_{j \in [d]} |u_j|) + 2\alpha - 2(1-p_y) \\ &= 2(1-\alpha-p_{y'}) \min_{j \in [d]} |u_j| + 2(p_y + \alpha - 1) \\ &\geq 2(p_y + \alpha - 1)(1-\tau) &(8.33)\end{aligned}$$

We also have that

$$\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{BEP}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}_\perp - \mathbf{p}^\top \boldsymbol{\ell}_y = p_y + \alpha - 1 \tag{8.34}$$

From Equations (8.33) and (8.34) we have that

$$\begin{aligned} &\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{u}) - \inf_{u' \in \mathbb{R}^d} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{u}') \\ \geq \quad & 2(1-\tau)(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{BEP}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t) \end{aligned} \tag{8.35}$$

Case 1c: $\mathbf{u} \in \mathbb{R}^d \setminus (\mathcal{U}^\tau_y \cup \mathcal{U}^\tau_\perp)$

Let $B^{-1}(\mathrm{sign}(-\mathbf{u})) = \mathrm{pred}(\mathbf{u}) = y'$ for some $y' \neq y$. We have $p_{y'} \leq 1 - p_y \leq \alpha \leq 1 - \alpha$, and $\min_j |u_j| > \tau$ and

$$\begin{aligned} &\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP}}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP}}(-B(y)) \\ \overset{(8.28)}{=} \quad & p_{y'} \psi^{\mathrm{BEP}}(y', \mathbf{u}) + \sum_{i=1; i \neq y'}^{n} p_i \psi^{\mathrm{BEP}}(i, \mathbf{u}) - 2(1 - p_y) \\ \overset{(8.30),(8.31)}{\geq} \quad & 2\alpha p_{y'}(-\min_{j \in [d]} |u_j|) + 2(1-\alpha)(1-p_{y'})(\min_{j \in [d]} |u_j|) + 2\alpha - 2(1-p_y) \\ = \quad & 2(1 - \alpha - p_{y'}) \min_{j \in [d]} |u_j| + 2(p_y + \alpha - 1) \\ \geq \quad & 2(1 - \alpha - p_{y'})\tau + 2(p_y + \alpha - 1) \\ = \quad & 2(1-\alpha)(\tau - 1) + 2p_y - 2\tau p_{y'} \\ = \quad & 2(1-\tau)(p_y + \alpha - 1) + 2\tau p_y - 2\tau p_{y'} \\ \geq \quad & 2\tau(p_y - p_{y'}) \end{aligned} \tag{8.36}$$

We also have that

$$\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{BEP}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t = \mathbf{p}^\top \boldsymbol{\ell}^\alpha_{y'} - \mathbf{p}^\top \boldsymbol{\ell}^\alpha_y = p_y - p_{y'} \tag{8.37}$$

From Equations (8.36) and (8.37) we have that

$$\begin{aligned} &\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{u}) - \inf_{u' \in \mathbb{R}^d} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{u}') \\ \geq \quad & 2(\tau)(\mathbf{p}^\top \boldsymbol{\ell}^\alpha_{\mathrm{pred}^{\mathrm{BEP}}_\tau(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t) \end{aligned} \tag{8.38}$$

Case 2: $p_y < 1 - \alpha$ for all $y \in [n]$

We have that $\perp \in \mathrm{argmin}_t \mathbf{p}^\top \boldsymbol{\ell}^\alpha_t$

Case 2a: $\mathbf{u} \in \mathcal{U}^\tau_\perp$

The RHS of Equation (8.32) is zero, and hence becomes trivial.

Case 2b: $\mathbf{u} \in \mathbb{R}^d \setminus \mathcal{U}^\tau_\perp$

Let $B^{-1}(\mathrm{sign}(-\mathbf{u})) = y' = \mathrm{pred}^{\mathrm{BEP}}_\tau(\mathbf{u})$ for some $y' \in [n]$. We have $p_{y'} < 1 - \alpha$

and $\min_j |u_j| > \tau$.

$$\begin{aligned}
&\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{u}) - \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{0}) \\
&\overset{(8.29)}{=} \quad p_{y'}\psi^{\mathrm{BEP}}(y', \mathbf{u}) + \sum_{i=1; i\neq y'}^{n} p_i \psi^{\mathrm{BEP}}(i, \mathbf{u}) - 2\alpha \\
&\overset{(8.30),(8.31)}{\geq} \quad 2\alpha p_{y'}(-\min_{j\in[d]} |u_j|) + 2(1-\alpha)(1-p_{y'})(\min_{j\in[d]} |u_j|) + 2\alpha - 2\alpha \\
&= \quad 2(1-\alpha-p_{y'})(\min_{j\in[d]} |u_j|) \\
&\geq \quad 2\tau(1-\alpha-p_{y'}) \qquad (8.39)
\end{aligned}$$

We also have that

$$\mathbf{p}^\top \boldsymbol{\ell}^{\alpha}_{\mathrm{pred}^{\mathrm{BEP}}_{\tau}(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^{\alpha}_t = \mathbf{p}^\top \boldsymbol{\ell}^{\alpha}_{y'} - \mathbf{p}^\top \boldsymbol{\ell}^{\alpha}_{\perp} = 1 - \alpha - p_{y'} \qquad (8.40)$$

From Equations (8.39) and (8.40) we have that

$$\begin{aligned}
&\mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{u}) - \inf_{u'\in\mathbb{R}^d} \mathbf{p}^\top \boldsymbol{\psi}^{\mathrm{BEP},\alpha}(\mathbf{u}') \\
&\geq \quad 2\tau(\mathbf{p}^\top \boldsymbol{\ell}^{\alpha}_{\mathrm{pred}^{\mathrm{BEP}}_{\tau}(\mathbf{u})} - \min_t \mathbf{p}^\top \boldsymbol{\ell}^{\alpha}_t) \qquad (8.41)
\end{aligned}$$

Equation (8.32), and hence the theorem, follows from Equations (8.35), (8.38) and (8.41). □

9. Conclusion

The multiclass classification problem with reject option, is a powerful abstraction that captures controlling the uncertainty of the classifier and is very useful in applications like medical diagnosis. We formalized this problem via an evaluation metric, called the abstain loss, and gave excess risk bounds relating the abstain loss to the Crammer-Singer surrogate, the one vs all hinge surrogate and also to the BEP surrogate which is a new surrogate and operates on a much smaller dimension. The resulting surrogate minimization algorithms perform well in experiments, allowing one to control the 'rejection' or 'abstention' rate while minimizing the misclassification error rate. Extending these results for other relevant evaluation metrics, in particular the abstain(α) loss for $\alpha > \frac{1}{2}$, is an interesting future direction.

References

[1] Allwein, E. L., Schapire, R. E. and Singer, Y. (2000). Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. *Journal of Machine Learning Research* **1** 113–141.

[2] Bartlett, P. L. and Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* **9** 1823–1840.

[3] Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* **16** 41–46.
[4] Cortes, C., DeSalvo, G. and Mohri, M. (2016). Learning with Rejection. In *Algorithmic Learning Theory*.
[5] Crammer, K. and Singer, Y. (2001). On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research* **2** 265–292.
[6] Duchi, J., Shalev-Shwartz, S., Singer, Y. and Chandra, T. (2008). Efficient Projections onto the l1 -Ball for Learning in High Dimensions. In *Proceedings of the 25th International Conference on Machine Learning*.
[7] el Yaniv, R. and Weiner, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research* **11** 1605–1641.
[8] el Yaniv, R. and Weiner, Y. (2011). Agnostic Selective Classification. In *Advances in Neural Information Processing Systems 24*.
[9] Fumera, G., Pillai, I. and Roli, F. (2003). Classification with reject option in text categorisation systems. In *IEEE International Conference on Image Analysis and Processing* 582–587.
[10] Fumera, G. and Roli, F. (2002). Suppport vector machines with embedded reject option. *Pattern Recognition with Support Vector Machines* 68–82.
[11] Fumera, G. and Roli, F. (2004). Analysis of error-reject trade-off in linearly combined multiple classifiers. *Pattern Recognition* **37** 1245–1265.
[12] Fumera, G., Roli, F. and Giacinto, G. (2000). Reject option with multiple thresholds. *Pattern Recognition* **33** 2099–2101.
[13] Grandvalet, Y., Rakotomamonjy, A., Keshet, J. and Canu, S. (2008). Support Vector Machines with a Reject Option. In *Advances in Neural Information Processing Systems 21*.
[14] Joachims, T. (1999). Making large-Scale SVM Learning Practical. In *Advances in Kernel Methods - Support Vector Learning* (B. Schölkopf, C. Burges and A. Smola, eds.) MIT-Press.
[15] Kummert, J., Paassen, B., Jensen, J., Goepfert, C. and Hammer, B. (2016). Local Reject Option for Deterministic Multi-class SVM. In *Artificial Neural Networks and Machine Learning - ICANN*.
[16] Lee, Y., Lin, Y. and Wahba, G. (2004). Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data. *Journal of the American Statistical Association* **99(465)** 67–81.
[17] Ramaswamy, H. G. and Agarwal, S. (2012). Classification Calibration Dimension for General Multiclass Losses. In *Advances in Neural Information Processing Systems 25*.
[18] Ramaswamy, H. G., Agarwal, S. and Tewari, A. (2013). Convex Calibrated Surrogates for Low-Rank Loss Matrices with Applications to Subset Ranking Losses. In *Advances in Neural Information Processing Systems 26*.
[19] Ramaswamy, H. G., Tewari, A. and Agarwal, S. (2015). Convex Calibrated Surrogates for Hierarchical Classification. In *Proceedings of The 32nd International Conference on Machine Learning*.

[20] Rifkin, R. and Klautau, A. (2004). In Defense of One-Vs-All Classification. *Journal of Machine Learning Research* **5** 101–141.

[21] Simeone, P., Marrocco, C. and Tortorella, F. (2012). Design of reject rules for ECOC classification systems. *Pattern Recognition* **45** 863–875.

[22] Tewari, A. and Bartlett, P. L. (2007). On the Consistency of Multiclass Classification Methods. *Journal of Machine Learning Research* **8** 1007–1025.

[23] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer.

[24] Varshney, K. R. and Willsky, A. S. (2010). Classification Using Geometric Level Sets. *Journal of Machine Learning Research* **11** 491–516.

[25] Vernet, E., Williamson, R. C. and Reid, M. D. (2011). Composite Multiclass Losses. In *Advances in Neural Information Processing Systems 24*.

[26] Wang, P.-W. and Lin, C.-J. (2014). Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research* **15** 1523–1548.

[27] Wu, Q., Jia, C. and Chen, W. (2007). A Novel Classification-Rejection Sphere SVMs for Multi-class Classification Problems. In *IEEE International Conference on Natural Computation*.

[28] Yuan, M. and Wegkamp, M. (2010). Classification Methods with Reject Option Based on Convex Risk Minimization. *Journal of Machine Learning Research* **11** 111–130.

[29] Zhang, C., Wang, W. and Qiao, X. (2017). On Reject and Refine Options in Multicategory Classification. *Journal of the American Statistical Association*.

[30] Zhang, T. (2004). Statistical Analysis of Some Multi-Category Large Margin Classification Methods. *Journal of Machine Learning Research* **5** 1225–1251.

[31] Zou, C., hui Zheng, E., wei Xu, H. and Chen, L. (2011). Cost-sensitive Multi-class SVM with Reject Option: A Method for Steam Turbine Generator Fault Diagnosis. *International Journal of Computer Theory and Engineering*.