Lower bounds on convex surrogates for abstain loss

author names withheld

Editor: Under Review for COLT 2021

Abstract

We study the problem of constructing consistent surrogates that abstain from classification when given data points with low confidence over the likelihood of the outcomes. Particularly, we review lower bounds on the dimensionality required to learn construct a *consistent, convex* loss function that abstains in low-confidence settings.

1. Introduction

Supervised machine learning algorithms typically learn to make predictions about future inputs by *empirical risk minimization*, where one learns a hypothesis function that minimizes a given loss function over a labeled training dataset. The loss function used often naturally corresponds to the question at hand; 0-1 loss is used for multiclass classification, squared loss is used to predict the expected value, and pinball loss is used to learn quantiles. Here, we study the *abstain problem*, where we want our machine learning algorithm to abstain and defer classification to a human when the data distribution has low confidence over the labels on a given input. We present and review some of the previous consistent surrogates for the abstain problem, and discuss the minimum dimension of a convex loss that learns this task in a statistically *consistent* manner.

We consider surrogates $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ for a given discrete loss corresponding to the task of interest. Here, we call d the dimension of the surrogate, and we aim to understand how to construct surrogates of low dimension (in the number of outcomes), as this contributes to lowering the complexity of the optimization problem as a whole. However, when we require surrogate to be consistent, this imposes a natural tension with finding a low-dimensional surrogate, as increasing the dimensional allows for more degrees of freedom to link surrogate predictions back to their discrete analogous predictions.

2. The abstain loss

In this work, we particularly focus on the *abstain loss*. Given a pre-determined parameter α, the abstain loss is a variation of cost-sensitive classification which allows for a constant punishment $\alpha \in (0, 1)$ for abstaining, represented here by the report $\bot$.

$$\ell^\alpha(r, y) = \begin{cases} 0 & r = y \\ \alpha & r = \bot \\ 1 & r \notin \{y, \bot\} \end{cases} \qquad \text{Abstain loss}$$

Ideally, one wants to learn a hypothesis function h to minimize this loss in expectation for any distribution D over $\mathcal{X} \times \mathcal{Y}$. However, if the hypothesis class $\mathcal{H}$ is sufficiently rich to contain the Bayes optimal classifier, we can focus on what the optimal classifier $h^*(x)$ should predict in order to minimize $\mathbb{E}_{Y\sim D_x} L(\cdot, Y)$ for all $x \in \mathcal{X}$. Therefore, for this paper, we focus on the marginal distributions $D_x =: p \in \Delta_{\mathcal{Y}}$ for any $x \in \mathcal{X}$, rather than taking the expectation over D. This allows us to study the abstain property, which yields the prediction that an optimal classifier should yield conditioned on any input x.

However, since it is typically computationally hard to optimize discrete losses, one typically wants to optimize "nicer" losses (i.e. continuous, convex, piecewise linear, etc.) In particular, we are interested in constructing convex surrogate losses that yield statistical guarantees known as *consistency.* In essence, a surrogate loss is consistent with respect to an original loss if there exists a link function from surrogate to original space so that, when one approaches the optimal hypothesis in surrogate space as they tend to infinite data, linking the surrogate hypothesis yields the optimal hypothesis in the original report space.

Definition 1 (Consistent) *A surrogate $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ is* consistent *with respect to a target loss $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$ if there exists a link function $\psi : \mathbb{R}^d \to \mathcal{R}$ such that for all distributions D on $\mathcal{X} \times \mathcal{Y}$ and all sequences of (vector) functions $f_m : \mathcal{X} \to \mathbb{R}^d$,*

$$\mathbb{E}_D L(f_m(X), Y) \to \inf_f \mathbb{E}_P L(f(X), Y) \implies \mathbb{E}_D \ell(\psi \circ f_m(X), Y) \to \inf_f \mathbb{E}_D \ell(\psi \circ f(X), Y) \ .$$

In discrete prediction settings, such as the abstain problem, previous work has repeatedly shown that consistency and calibration are equivalent (Zhang, 2004; Bartlett et al., 2006; Tewari and Bartlett, 2007; Ramaswamy and Agarwal, 2016).

Definition 2 (Calibrated) *Let $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ be a discrete loss eliciting the property γ. A surrogate loss $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}_+$ is* calibrated *with respect to ℓ if there exists a link function $\psi : \mathbb{R}^d \to \mathcal{R}$ such that*

$$\forall p \in \Delta_{\mathcal{Y}} : \inf_{u \in \mathbb{R}^d : \psi(u) \not\in \gamma(p)} L(u; p) > \inf_{u \in \mathbb{R}^d} L(u; p) \ . \tag{1}$$

JF: Fix notation γ

Calibration is generally an easier condition to work with than consistency, and it allows us to abstract from distributions over $\mathcal{X} \times \mathcal{Y}$ to distributions over $\mathcal{Y}$, denoted $p \in \Delta_{\mathcal{Y}}$.

3. Calibrated surrogates for the abstain loss

3.1. One-vs-all (oVa)

3.2. Crammer-Singer

3.3. Binary Encoded Prediction (BEP)

Ramaswamy et al. (2018) introduces the Binary Encoded Prediction (BEP) surrogate that is calibrated with respect to the abstain loss. For this loss, we assign each outcome into a corner of the $\{-1, +1\}^d$ hypercube by a bijection $B : [n] \to \{-1 + 1\}^d$, where $d := \lceil \lg(n) \rceil$.

$$L^{BEP}(u, y) = \left(\max_{j \in [d]} B_j(y) u_j + 1 \right)_+ \tag{BEP}$$

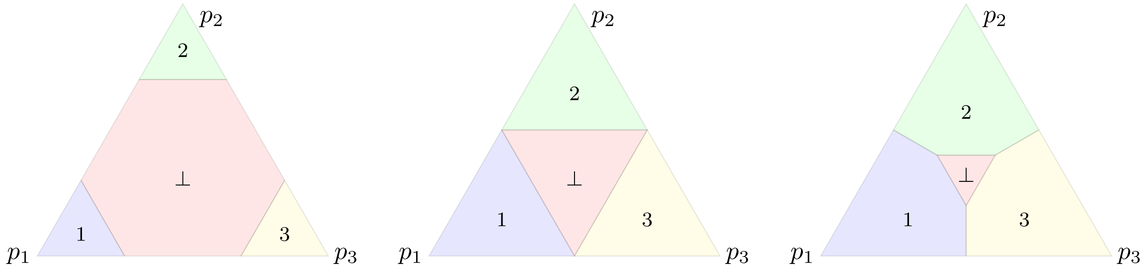

Figure 1: With $n = 3$, we can visualize values of $\gamma^\alpha(p)$ as one varies p over $\Delta_\mathcal{Y}$. Here, we vary α in each figure. $\alpha = 3/10$ (L), $\alpha = 1/2$ (M), $\alpha = 3/5$ (R).

3.3.1. [JF: Links for BEP?]

One of the primary strengths of this surrogate is the exponential reduction in the dimension d of the surrogate report u, while still being calibrated over the entire simplex. The One-vs-all [JF: cite if no subsection] and Crammer-Singer [JF: cite if no subsection] surrogates both require the dimension of u to grow linearly, rather than logarithmically, with the number of outcomes. Significant improvements to this dimensionality while the surrogate is consistent can lead to improvements in the efficiency of the optimization problem. This begs the following question:

> What is the *lowest* dimension in which we can construct a consistent (convex) surrogate for the abstain problem?

4. Bounds on consistent surrogates

Definition 3 *The* convex calibration dimension $\texttt{cc dim}(\ell)$ *of a discrete loss* $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ *is the minimum dimension* d *such that there is a convex surrogate* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ *such that* L *is calibrated with respect to* ℓ.

Definition 4 (Embeds) *A convex surrogate loss* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ *eliciting* $\Gamma : \Delta_\mathcal{Y} \rightrightarrows \mathbb{R}^d$ embeds *a discrete loss* $\ell : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}_+$ *eliciting* $\gamma : \Delta_\mathcal{Y} \rightrightarrows \mathcal{R}$ *if there exists an injective embedding* $\varphi : \mathcal{R} \to \mathbb{R}^d$ *such that (i) for all* $r \in \mathcal{R}$*, we have* $L(\varphi(r)) = \ell(r)$*, and (ii) for all* $p \in \Delta_\mathcal{Y}$ *and* $r \in \mathcal{R}$*, we have* $r \in \gamma(p) \iff \varphi(r) \in \Gamma(p)$.

Definition 5 *The* embedding dimension $\texttt{embed}(\ell)$ *of a discrete loss* ℓ *is given by the minimum dimension* d *such that there exists a surrogate loss* $L : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ *embedding* ℓ.

[JF: Can verify this easily enough, so maybe don't have proof?]

Proposition 6 *The BEP surrogate [JF: or an α-appropriate variation] embeds ℓ^α for $\alpha \in (0, 1/2]$.*

Proposition 7 $\texttt{cc dim}(\ell^\alpha) \leq \texttt{embed}(\ell^\alpha)$.

Proof If L embeds ℓ, then it is calibrated with respect to ℓ by Finocchiaro et al. (2019) Theorem 3. ■

4.1. Upper bounds

Theorem 8 $\texttt{cc dim}(\ell^\alpha) \leq \lceil \lg(n) \rceil$ *for values of* $\alpha \in (0, 1/2]$.

The upper bound is constructive and is true by the consistency of the BEP surrogate.

4.2. Lower bounds

Theorem 9 ((?) Proposition 1) *For $\alpha \leq 1/2$ with $n = 3$, we observe* $\texttt{embed}(\ell^\alpha) = 2$.

Theorem 10 ((?) Corollary 22) *For $\alpha = 1/2$ with $n \geq 5$, we observe* $\texttt{embed}(\ell^\alpha) \geq 3$.

Open Question 1 *Is* $\texttt{cc dim}(\ell^{1/2}) = \Theta(\lg(n))$*?*

Open Question 2 *Given $\alpha \in (0, 1/2]$, and n, what is* $\texttt{cc dim}(\ell^\alpha)$*?*

5. Future work and Conclusions

[JF: Note that actually strictly more efficient than learning the mode] While bounds on the convex calibration dimension of the abstain loss exist, they are far from tight. For $\alpha < 1/2$ and $n \geq 3$, the bounds we know simply state $2 \leq \texttt{cc dim}(\ell^\alpha) \leq \texttt{embed}(\ell^\alpha) \leq \lceil \lg(n) \rceil$. As n grows large, as in extreme classification, the gap provided by this bound is very large.

The next conjecture suggests that embedding is a sufficient framework to study convex calibration dimension.

Open Question 3 *For any $\alpha \in (0, 1/2]$, is* $\texttt{cc dim}(\ell^\alpha) = \texttt{embed}(\ell^\alpha)$*?*

If true, this would suggest that we can use tools from embedding techniques to characterize the convex calibration dimension. Some of these tools include the use of Minkowski sums to calculate the subgradient sets of the expected loss of embedded reports at various distributions and a quadratic feasibility program whose solution yields subgradient sets at embedded points for a convex surrogate loss embedding the discrete loss.

However, it is not clear, if the lower bound on constructing a calibrated surrogate for the abstain loss is $\lceil \lg(n) \rceil$, why that is. Current lower bounds come from the optimality condition that a report is in the property value at a distribution p if and only if $\mathbf{0}$ is in the subgradient set of the expected loss over p. However, if tighter lower bounds come from monotonicity conditions concerned with the adjacency of level sets, we need new tools to study these bounds.

References

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214505000000907.

Jessie Finocchiaro, Rafael Frongillo, and Bo Waggoner. An embedding framework for consistent polyhedral surrogates. In *Advances in neural information processing systems*, 2019.

Harish G Ramaswamy and Shivani Agarwal. Convex calibration dimension for multiclass loss matrices. *The Journal of Machine Learning Research*, 17(1):397–441, 2016.

Harish G Ramaswamy, Ambuj Tewari, Shivani Agarwal, et al. Consistent algorithms for multiclass classification with an abstain option. *Electronic Journal of Statistics*, 12(1): 530–554, 2018.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *The Journal of Machine Learning Research*, 8:1007–1025, 2007. URL http://dl.acm.org/citation.cfm?id=1390325.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.