Consistency versus Realizable H-Consistency for Multiclass Classification

Philip M. Long PLONG@MICROSOFT.COM
Microsoft, 1020 Enterprise Way, Sunnyvale, CA 94089

Rocco A. Servedio ROCCO@CS.COLUMBIA.EDU
Department of Computer Science, Columbia University, New York, NY 10027

Abstract

A *consistent* loss function for multiclass classification is one such that for any source of labeled examples, any tuple of scoring functions that minimizes the expected loss will have classification accuracy close to that of the Bayes optimal classifier. While consistency has been proposed as a desirable property for multiclass loss functions, we give experimental and theoretical results exhibiting a sequence of linearly separable data sources with the following property: a multiclass classification algorithm which optimizes a loss function due to Crammer and Singer (which is known *not* to be consistent) produces classifiers whose expected error goes to 0, while the expected error of an algorithm which optimizes a generalization of the loss function used by LogitBoost (a loss function which is known to be consistent) is bounded below by a positive constant.

We identify a property of a loss function, *realizable consistency with respect to a restricted class of scoring functions*, that accounts for this difference. As our main technical results we show that the Crammer–Singer loss function is realizable consistent for the class of linear scoring functions, while the generalization of LogitBoost is not. Our result for LogitBoost is a special case of a more general theorem that applies to several other loss functions that have been proposed for multiclass classification.

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Classification into $k > 2$ classes is often addressed by learning a real-valued scoring function h_z for each class $z \in \{1, \ldots, k\}$, and then classifying an item x as $\text{argmax}_z h_z(x)$. To learn the scoring functions, a popular approach is to minimize the average, over training examples (x, y), of $L(y, h_1(x), ..., h_k(x))$ where L is some loss function.

One very natural loss function (Zhang, 2004), which generalizes the loss function used by LogitBoost (Friedman et al., 2000), is

$$\begin{aligned} &L_{\text{logit}}(y, h_1(x), ..., h_k(x)) \\ &= \sum_{z=1}^{k} \log\left(\frac{1+\exp(h_z(x))}{1_{z\neq y} + 1_{z=y} \times \exp(h_z(x))}\right) \\ &= -h_y(x) + \sum_{z=1}^{k} \log\left(1+\exp(h_z(x))\right). \end{aligned}$$

Minimizing this loss function has been shown to be *consistent* (Zhang, 2004); informally, this means that minimizing this loss function results in a classifier whose accuracy is close to that of the Bayes optimal classifier.[1] (We give a precise definition of consistency in Section 2.2.) Such consistency has been cited as a hallmark of a principled algorithm (Harchaoui et al., 2012).

Another loss function, due to Crammer & Singer (2001), is

$$\begin{aligned} &L_{CS}(y, h_1(x), ..., h_k(x)) \\ &= \max\left\{0, 1 - h_y(x) + \max_{z\neq y} h_z(x)\right\}. \end{aligned}$$

[1] Other papers on consistency for multiclass classification include (Lee et al., 2004; Tewari & Bartlett, 2007; Liu, 2007).

If all other scores $h_z(x)$ are at least 1 less than $h_y(x)$ then this loss is 0; otherwise this loss is the maximum, over all other scores, of the amount by which that score fails to satisfy this constraint. L_{CS} is not consistent (Tewari & Bartlett, 2007).

A puzzling empirical finding. Consider the following, apparently easy, synthetic learning problem. Training and test examples are generated i.i.d. The source used to generate examples is linearly separable: there is a "target weight vector" $\mathbf{w}_z \in \mathbf{R}^{100}$ for each class $z \in \{1, \ldots, 10\}$, and a random example $(\mathbf{x}, y)$ is obtained by generating $\mathbf{x} \in \mathbf{R}^{100}$ according to the uniform distribution on the unit ball, and assigning a label y that maximizes $\mathbf{w}_y \cdot \mathbf{x}$. The target weight vectors are the rows of a 10×100 matrix W that is generated as follows: a 10×2 matrix A and a 2×100 matrix B are generated by i.i.d. sampling their components from the standard normal distribution, and W is set to equal AB.

We generated 10000 training and 10000 test examples from such a source, and used them to evaluate two algorithms. One algorithm learned the weights $\mathbf{v}_1, ..., \mathbf{v}_k$ to minimize $L_{\text{logit}}(y, \mathbf{v}_1 \cdot \mathbf{x}, ..., \mathbf{v}_k \cdot \mathbf{x})$ using 100 epochs of ASGD (Polyak & Juditsky, 1992; Xu, 2011; Bottou, 2010) and the other was the same, except with L_{CS}. (Both algorithms also used a mild Frobenius norm regularizer. Please see Section 5 for the details.) We repeated this process five times. The algorithm that minimized (the consistent) L_{logit} had average test accuracy 87.5%, while the algorithm that minimized (the inconsistent) L_{CS} had average test accuracy 94.4%. For one source, the accuracy obtained from L_{logit} was 82.6%, while the accuracy from L_{CS} was 94.2%. Thus, in these experiments on a seemingly easy synthetic learning problem, the consistent algorithm (minimizing L_{logit}) performs significantly less well than the algorithm (minimizing L_{CS}) which is known not to be consistent. Given how easy this learning problem seems to be, the outputs of the algorithm minimizing L_{logit} are surprisingly inaccurate.

Discussion of the experimental results. What happened? Because W has low rank, a given weight vector $\mathbf{w}_y$ tends to be similar to weight vectors for some other classes, and consequently a good classifier for a given example $(\mathbf{x}, y)$ will tend to also assign high scores to some classes other than y. Minimizing L_{CS} is compatible with this desideratum. During training with a stochastic gradient descent algorithm, once the weight vectors have sufficiently large magnitude, informally, L_{CS} does not mind classes other than y also having a high score, and a stochastic gradient update according to this loss function eventually will not change any weights. On the other hand, if L_{logit} is trained by stochastic gradient descent, it will keep "knocking down" the scores for classes other than y. Put another way, L_{CS} is only trying to classify the data correctly by a certain margin, while L_{logit} is trying to fit a model for the conditional probability that the class is y. The consistency of L_{logit} implies that minimizing this loss will lead to near-optimal accuracy *if this minimization is done over the space of all scoring functions*. However, in the above experiment, we performed the minimization only over the class of linear functions. (We note that this is a standard practice for large-scale multiclass image classification, see e.g. (Weston et al., 2010; Lin et al., 2011; Perronnin et al., 2012)). This provides a possible explanation of L_{logit}'s relatively poor performance.

Our results. As described above, in our experiments, using L_{CS} led to significantly better accuracy than L_{logit} even though L_{logit} is consistent and L_{CS} is not. To explain this phenomenon we introduce a new notion, which we term *realizable H-consistency*, where H is a restricted class of scoring functions. Informally, a loss function is realizable H-consistent if for any source that admits a zero-error classifier using scoring functions from H, any scoring functions from H which minimize the expected loss will have classification error close to zero. For learning algorithms that use a restricted class H of scoring functions, H-consistency may be more relevant than the original notion of consistency, which deals with all possible scoring functions, and realizable H-consistency is a more basic requirement.

We show that the loss function L_{CS} is realizable H-consistent for any class H of scoring functions that is closed under scaling (i.e. if $h \in H$ then $\alpha h \in H$ for all real α). This implies that L_{CS} is realizably consistent with respect to linear scoring functions (as well as with respect to deep nets, see the discussion after Theorem 9). We also show that L_{logit} is *not* realizably consistent with respect to the set of linear functions. These results together imply that the ratio between the error probabilities of an algorithm minimizing L_{logit} and an algorithm minimizing L_{CS} can be arbitrarily large, even if we restrict to linearly separable sources. Our lower bound for L_{logit} holds for any loss function satisfying some general conditions, and applies for several other loss functions previously proposed for multiclass classification.

Our results suggest that the refined notion of H-consistency may sometimes be a more useful and relevant one than the original notion of consistency. In settings where learning algorithms use a restricted class

H of scoring functions, H-consistency may be more closely linked to classification accuracy than general consistency.

2. Preliminaries

2.1. Basics

Throughout this paper Y denotes the set $[k] = \{1, \ldots, k\}$. A *source* is a joint distribution P over $X \times Y$. We write P_X to denote the marginal distribution of P over X.

Recall that the *Bayes optimal classifier for source P* is the mapping $f : X \to Y$ such that for each x' in the support of P_X, we have $\mathbf{Pr}_{(x,y)\sim P}[y = f(x')|x = x'] \geq \mathbf{Pr}_{(x,y)\sim P}[y = t|x = x']$ for every $t \neq f(x')$.

2.2. Different notions of consistency

We begin by recalling the standard notion of consistency.

For $\delta > 0$ and a source P, we say that a function $h = (h_1, \ldots, h_k)$ from X to $\mathbf{R}^k$ *δ-minimizes* $\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), \ldots, h_k(x))]$ if

$$|\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), \ldots, h_k(x))] - \inf_{a_1,\ldots,a_k : X\to\mathbf{R}} \mathbf{E}_{(x,y)\sim P}[L(y, a_1(x), \ldots, a_k(x))]| \leq \delta.$$

Note that the scoring functions $a_1, \ldots, a_k$ above may range over all measurable mappings from X to $\mathbf{R}$.

Definition 1. *A loss function L is* consistent *if for any source P and any $\epsilon > 0$ there exists some $\delta > 0$ such that if $h = (h_1, ..., h_k)$ δ-minimizes $\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), ..., h_k(x))]$, and $g : X \to Y$ satisfies $g(x) \in \operatorname{argmax}_{\hat{y}\in Y} h_{\hat{y}}(x)$ for all $x \in X$, then*

$$\mathbf{Pr}_{(x,y)\sim P}[g(x) \neq y] \leq \mathbf{Pr}_{(x,y)\sim P}[f(x) \neq y] + \epsilon,$$

where f is the Bayes optimal classifier.

The notion of a consistent loss function defined in Definition 1 was studied by (Zhang, 2004). A closely related notion was studied by Tewari and Bartlett (Tewari & Bartlett, 2007). Following those authors (and, others, including (Breiman, 2004; Long & Servedio, 2010)), we abstract away estimation error, and study the consequences of minimizing loss with respect to the underlying distribution.

Definition 2. *A source P is* realizable *if for each x' in the support of P_X, there is a y' such that $\mathbf{Pr}_{(x,y)\sim P}[y = y'|x = x'] = 1$.*

Definition 3. *A loss function L is* realizable consistent *if for any realizable source P, for any $\epsilon > 0$ there is a $\delta > 0$ such that if $h = (h_1, ..., h_k)$ δ-minimizes $\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), ..., h_k(x))]$, and g satisfies $g(x) \in \operatorname{argmax}_{\hat{y}\in Y} h_{\hat{y}}(x)$ for all $x \in X$, then $\mathbf{Pr}_{(x,y)\sim P}(g(x) \neq y) \leq \epsilon$.*

We now define a notion of consistency with respect to a set H of scoring functions. This formalizes the notion that choosing scoring functions from H so as to minimize the loss does nearly as well, in terms of classification error, as the best combination of scoring functions from H. (Very similar notions have long been studied; see (Vapnik, 1989; Haussler, 1992; Kearns et al., 1992).)

Let H be a set of functions mapping X to $\mathbf{R}$. For $\delta > 0$, we say that a function $h = (h_1, \ldots, h_k)$ from X to $\mathbf{R}^k$ *δ-minimizes* $\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), \ldots, h_k(x))]$ *with respect to H* if

$$|\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), \ldots, h_k(x))] - \inf_{a_1,\ldots,a_k \in H} \mathbf{E}_{(x,y)\sim P}[L(y, a_1(x), \ldots, a_k(x))]| \leq \delta.$$

Definition 4. *Let H be a set of measurable functions from X to $\mathbf{R}$. A loss function L is* H-consistent *if for any source P and any $\epsilon > 0$, there is a $\delta > 0$ such that if $h_1, \ldots, h_k \in H$ δ-minimize $\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), ..., h_k(x))]$ with respect to H and $g : X \to Y$ satisfies $g(x) \in \operatorname{argmax}_{\hat{y}\in Y} h_{\hat{y}}(x)$ for all $x \in X$, then for all, $a_1, \ldots, a_k \in H$, and any $f : X \to Y$ such that $f(x) \in \operatorname{argmax}_{\hat{y}\in Y} a_{\hat{y}}(x)$ for all $x \in X$, we have*

$$\mathbf{Pr}_{(x,y)\sim P}[g(x) \neq y] \leq \mathbf{Pr}_{(x,y)\sim P}[f(x) \neq y] + \epsilon.$$

Note that the original notion of consistency from Definition 1 corresponds to H-consistency when H is the class of all measurable functions.

Definition 5. *A source P is* realizable w.r.t. *H if there exist $h_1, \ldots, h_k \in H$ such that, for any $g : X \to [k]$ such that $g(x) \in \operatorname{argmax}_{\hat{y}\in Y} h_{\hat{y}}(x)$ for all $x \in X$, we have that $\mathbf{Pr}_{(x,y)\sim P}[y = g(x')|x = x'] = 1$ for any x' in the support of P_X.*

Realizable sources correspond to learning scenarios in which there is a "target function" that always provides the correct label for each example.

Definition 6. *A loss function L is* realizable H-consistent *if the following holds: for any source P that is realizable w.r.t. H and any $\epsilon > 0$, there is a $\delta > 0$ such that if $h_1, \ldots, h_k \in H$ δ-minimize $\mathbf{E}_{(x,y)\sim P}[L(y, h_1(x), \ldots, h_k(x))]$ with respect to H, and g satisfies $g(x) \in \operatorname{argmax}_{\hat{y}} h_{\hat{y}}(x)$ for all $x \in X$, then $\mathbf{Pr}_{(x,y)\sim P}[g(x) \neq y] \leq \epsilon$.*

3. The Crammer-Singer loss

The Crammer-Singer loss is defined as follows:

$$L_{CS}(y, s_1, ..., s_k) = \max\left\{0, 1 - s_y + \max_{j \neq y} s_j\right\}.$$

One may easily use a construction due to Tewari and Bartlett (Tewari & Bartlett, 2007) to prove that L_{CS} is not consistent (see also (Zhang, 2004)). Because of differences in the details of our definitions and theirs, we have included a proof of Theorem 7 in Appendix A.

Theorem 7. *L_{CS} is not consistent.*

While L_{CS} is not consistent, we now show that it is realizable H-consistent for any class H of scoring functions that satisfies the following natural scaling condition:

Definition 8. *A class H of scoring functions is* closed under scaling *if for any $h \in H$ and any real α, the function αh belongs to H.*

The set of linear functions is closed under scaling, and any set of scoring functions can be easily closed under scaling by adding all scalings of all of its members.

Theorem 9. *For any H that is closed under scaling, the Crammer-Singer loss L_{CS} is realizable consistent w.r.t. H.*

Proof: Let P be a realizable source w.r.t. H. Let $h_1, ..., h_k \in H$ be such that, if $\text{gap}(x) = h_y(x) - \max_{j \neq y} h_j(x)$, where $y = \text{argmax}_{\hat{y} \in Y} h_{\hat{y}}(x)$, we have

$$\mathbf{Pr}_{(x,y)\sim P}[\text{gap}(x) > 0] = 1. \tag{1}$$

Choose $\epsilon > 0$, and let $\kappa > 0$ be such that $\mathbf{Pr}_{x \sim P_X}[\text{gap}(x) \leq \kappa] \leq \epsilon/2$. Then, rescaling $h_1, ..., h_k$ by $1/\kappa$, we claim that

$$\mathbf{E}_{(x,y)\sim P}[L_{CS}(y, (1/\kappa)h_1(x), ..., (1/\kappa)h_k(x))] \leq \epsilon/2.$$

To see this, note that since $h_y(x) > h_j(x)$ for all x and all $j \neq y$, we have that L_{CS} is always at most 1, and hence

$$\begin{aligned}&\mathbf{E}[L_{CS}(y, (1/\kappa)h_1(x), ..., (1/\kappa)h_k(x))] \\ &\leq \mathbf{E}[L_{CS}(y, (1/\kappa)h_1(x), ..., (1/\kappa)h_k(x))|\text{gap}(x) > \kappa] \\ &\quad + \mathbf{Pr}[\text{gap}(x) \leq \kappa] \leq 0 + \epsilon/2. \end{aligned} \tag{2}$$

Now suppose that $g_1, ..., g_k \in H$ approximately minimize $\mathbf{E}_{(x,y)\sim P}[L_{CS}(y, g_1(x), ..., g_k(x))]$ to within $\epsilon/2$. Then (2) implies that

$$\mathbf{E}_{(x,y)\sim P}[L_{CS}(y, g_1(x), ..., g_k(x))] \leq \epsilon.$$

But $\mathbf{E}_{(x,y)\sim P}[L_{CS}(y, g_1(x), ..., g_k(x))]$ is an upper bound on the probability (over $(x, y) \sim P$) that $y \neq \text{argmax}_{\hat{y}} g_{\hat{y}}(x)$. This completes the proof. □

Note that the proof of Theorem 9 goes through almost without modification if L_{CS} is replaced with

$$L(y, s_1, ..., s_k) = \ell(-s_y + \max_{j \neq y} s_j),$$

for any continuous monotone function $\ell : \mathbf{R} \to \mathbf{R}^+$ such that $\lim_{x \to -\infty} \ell(x) = 0$, including the function ℓ defined by $\ell(x) = \ln(1 + e^x)$. The key is that these loss functions concern a competition among the scores, instead of evaluating the scores independently.

We also note that beyond the class H of linear functions, Theorem 9 also implies that L_{CS} is realizable H-consistent in the case that H is the set of functions computed by a deep network with a given architecture, or a convolutional network, if the squashing functions are left off of the output nodes. (The loss function can be viewed as taking on the role of the squashing function for the output nodes, so this is reasonable.)

4. A sufficient condition for a loss function to not be realizable consistent w.r.t. linear functions

Throughout this section we fix X to be the domain $\mathbf{R}^n$ and we let H denote the class of all linear functions $X \to \mathbf{R}$, i.e. $H = \{\mathbf{x} \to \mathbf{v} \cdot \mathbf{x}, \mathbf{v} \in \mathbf{R}^n\}$.

Our main result in this section is a proof that any loss function that satisfies some general conditions, detailed below, is not realizable consistent for the class of linear scoring functions:

Theorem 10. *Let $\ell : \mathbf{R} \to \mathbf{R}^{>0} = \{x \in \mathbf{R} : x > 0\}$ be any function such that (a) ℓ is twice continuously differentiable, (b) ℓ is strictly convex, (c) $\lim_{x \to +\infty} \ell(x) = +\infty$, and (d) $\inf_{x \in \mathbf{R}} \ell(x) - x > -\infty$. Then the loss function L_ℓ defined as*

$$L_\ell(y, h_1(\mathbf{x}), \ldots, h_k(\mathbf{x})) := -h_y(\mathbf{x}) + \sum_{z=1}^{k} \ell(h_z(\mathbf{x}))$$

is not realizable H-consistent, where H is the class of all linear functions.

It can be easily checked that this theorem implies that many of the "decoupled" loss functions of (Zhang, 2004) (see Section 4.4.2 of (Zhang, 2004)), shown there to be consistent, are in fact not realizable H-consistent. These include the L_{logit} function (for which $\ell(x) = \ln(1 + e^x)$) and the L_Q loss function defined by

$$L_Q(y, h_1(x), \ldots, h_k(x)) = -h_y(x) + \frac{1}{2}\sum_{z=1}^{k} h_z(x)^2$$

(for which $\ell(x) = x^2/2$). (Our proof will establish lower bounds of 1/5 on the error probability for algorithms minimizing L_{logit} and L_Q.)

The main steps in proving Theorem 10 are to (I) define a source P that is realizable w.r.t. H, and (II) prove that the optimal $(\mathbf{v}_1^\star, \ldots, \mathbf{v}_k^\star) \in H^k$ that minimizes $\mathbf{E}_{(\mathbf{x},y)\sim P}[L_\ell(y, \mathbf{v}_1^\star \cdot \mathbf{x}, \ldots, \mathbf{v}_k^\star \cdot \mathbf{x})]$ (over all k-tuples of functions in H as the final k arguments to L_ℓ) is such that for $g(\mathbf{x}) = \text{argmax}_{\hat{y}\in Y} h_{\hat{y}}(x)$, we have $\mathbf{Pr}_{(\mathbf{x},y)\sim P}[g(\mathbf{x}) \neq y]$ is bounded below by 4α where $\alpha > 0$ is a constant that depends only on ℓ (defined below). Once this is done Theorem 10 is an immediate consequence (taking $\epsilon = 4\alpha$ in Definition 6). We now turn to these two tasks.

(I): The source P. We now define the source P over $X \times Y$ that we shall consider. This source is very simple and in fact only uses the domain $X = \mathbf{R}^2$. The number k of classes is 8. The source depends on the loss ℓ through a parameter that is defined as follows. Since ℓ is twice continuously differentiable and strictly convex, there is a $\tau > 0$ such that

$$\forall x \in [-\tau, \tau],\ (1/\sqrt{2}) \leq \frac{\ell''(x)}{\ell''(0)} \leq \sqrt{2}. \tag{3}$$

Choose such a τ. (For L_Q, any positive constant will work, and for L_{logit}, we may choose $\tau = 1$.) Let $\alpha = \min\{\tau\ell''(0)/5, 1/20\}$. (For both L_Q and L_{logit}, α is 1/20.) Next, define the $k \times 2$ matrix W whose rows are $\mathbf{w}_1, \ldots, \mathbf{w}_k$: for $i = 1, \ldots, k$,

$$\mathbf{w}_i = (w_{i,1}, w_{i,2}) = (\cos(2\pi i/k), \sin(2\pi i/k)). \tag{4}$$

The marginal P_X is defined as follows. Each of $\mathbf{w}_1, \mathbf{w}_3, \mathbf{w}_5, \mathbf{w}_7$ have probability α (these are the "light points"), and each of $\mathbf{w}_2, \mathbf{w}_4, \mathbf{w}_6, \mathbf{w}_8$ have probability $1/4 - \alpha$ (these are the "heavy points"). The weight matrix W whose rows are $\mathbf{w}_1, ..., \mathbf{w}_8$ is then used to classify $\mathbf{x}$ as $y = \text{argmax}_z \mathbf{w}_z \cdot \mathbf{x}$, so P is realizable with respect to H.

(II): The vector of functions in H that minimizes L_ℓ has poor classification accuracy. Henceforth V shall denote a $k \times 2$ matrix whose rows are $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Let us define the function Ψ as follows:

Definition 11. $\Psi(V) = \mathbf{E}_{(\mathbf{x},y)\sim P}[L_\ell(y, \mathbf{v}_1 \cdot \mathbf{x}, \ldots, \mathbf{v}_k \cdot \mathbf{x})]$.

We want to show that Ψ has a unique minimum. As a step in this direction, let us first show that Ψ is strictly convex.

Directly expanding the definition of P, separating the expectation into the heavy points and the light points, we get that $\Psi(V)$ equals

$$\sum_{s=1}^{4}(1/4 - \alpha)\big(-\mathbf{v}_{2s} \cdot \mathbf{w}_{2s} + \sum_{z=1}^{8} \ell(\mathbf{v}_z \cdot \mathbf{w}_{2s})\big)$$
$$+ \quad \sum_{s=1}^{4} \alpha\big(-\mathbf{v}_{2s-1} \cdot \mathbf{w}_{2s-1} + \sum_{z=1}^{8} \ell(\mathbf{v}_z \cdot \mathbf{w}_{2s-1})\big).$$

Since

- Ψ is convex,
- ℓ is strictly convex (by property (b) of the theorem statement), and
- modifying any entry of V affects $\ell(\mathbf{v}_z \cdot \mathbf{w}_j)$ for at least one z and j,

we have that Ψ is strictly convex.

Now, let us eliminate the possibility that the value of Ψ can get arbirarily small. First, because $\ell(t) > 0$ for all t, for any example $(\mathbf{x}, y)$ we have

$$-\mathbf{v}_y \cdot \mathbf{x} + \sum_{z=1}^{k} \ell(\mathbf{v}_z \cdot \mathbf{x}) \geq -\mathbf{v}_y \cdot \mathbf{x} + \ell(\mathbf{v}_y \cdot \mathbf{x}) \geq \inf_{u\in\mathbf{R}} \ell(u) - u.$$

Since $\inf_{x\in\mathbf{R}} \ell(x) - x > -\infty$ (by property (d) of the theorem statement), and Ψ is an average of such losses, the function Ψ is lower bounded (there is some value $C \in \mathbf{R}$ such that $\Psi(V) > C$ for all V).

Next, let us eliminate the possibility that Ψ has no minimum; since Ψ is strongly convex, this can only occur if the value of Ψ gets smaller and smaller as V heads to infinity in some direction. We claim that any V with a very large entry must have $\Psi(V) > \Psi(V_{\text{zero}})$ where V_{zero} is the all-zero input. To see this, suppose $|V_{zi}| > M$ for some row z and column i. Choose some $y \in \{1, .., 8\}$ other than z such that the angle between $\mathbf{v}_z$ and $\mathbf{w}_y$ is at most $\pi/4$. (There must be at least one such y.) Then

$$\Psi(V) \geq \alpha\ell(\mathbf{v}_z \cdot \mathbf{w}_y) \geq \alpha\ell(||\mathbf{v}_z|| \cos(\pi/4)) \geq \alpha\ell(M/\sqrt{2}),$$

which is larger than $\Psi(V_{\text{zero}})$ for M sufficiently large, by our assumption that ℓ goes to $+\infty$ as $M \to +\infty$ (by property (c) of the theorem statement). So for the purpose of minimizing Ψ, we can assume without loss of generality that the domain of Ψ is $[-M, M]^{8\times 2}$ for some constant $M > 0$. Since Ψ is a strictly convex function defined on a bounded domain, there is a unique $V^\star$ at which it achieves its minimum value.

In the rest of the proof we will show that the error rate of $V^\star$ w.r.t. P is at least 4α. To prove this, first let us

characterize the form of $V^\star$. We claim that there are real values q and r such that

$$V^\star = \begin{pmatrix} r & 0 & -r & -q & -r & 0 & r & q \\ r & q & r & 0 & -r & -q & -r & 0 \end{pmatrix}^T. \quad (5)$$

First, let us prove that $V^\star_{2,1} = 0$. Suppose $V^\star_{2,1} \neq 0$. Then the symmetry of Ψ implies that, if we form V' by negating $V^\star_{2,1}$ in $V^\star$, then $V' \neq V^\star$ but $\Psi(V') = \Psi(V^\star)$. Since this contradicts the uniqueness of the minimizing point $V^\star$, it must be the case that $V^\star_{2,1} = 0$. The other 0-entries in (5) can be established in a similar way.

We can prove that

$$V^\star_{4,1} = -V^\star_{8,1}, \quad V^\star_{3,1} = -V^\star_{1,1}, \quad V^\star_{3,2} = V^\star_{1,2},$$
$$V^\star_{5,1} = -V^\star_{7,1}, \quad V^\star_{5,2} = V^\star_{7,2}$$

similarly by exploiting the symmetry of Ψ across the x_2 axis, and we can prove that

$$V^\star_{1,1} = V^\star_{1,2}, \quad V^\star_{2,2} = V^\star_{8,1}, \quad V^\star_{3,1} = V^\star_{7,2},$$
$$V^\star_{3,2} = V^\star_{7,1}, \quad V^\star_{4,1} = V^\star_{6,2}, \quad V^\star_{5,1} = V^\star_{5,2}$$

similarly by exploiting the symmetry across the line $x_1 = x_2$.

Below we will show that $q > 2r$; we claim that this implies $\mathbf{Pr}_{(x,y)\sim P}[g(\mathbf{x}) \neq y] \geq 4\alpha$ as desired. To see this, consider first the classification of $\mathbf{w}_1$. We have

$$\mathbf{v}^\star_1 \cdot \mathbf{w}_1 = (r, r) \cdot (1/\sqrt{2}, 1/\sqrt{2}) = \sqrt{2}r$$
$$\mathbf{v}^\star_2 \cdot \mathbf{w}_1 = (0, q) \cdot (1/\sqrt{2}, 1/\sqrt{2}) = q/\sqrt{2}.$$

So, if $q/r > 2$, then $V^\star$ misclassifies $\mathbf{w}_1$, and, similarly, all the other light points.

We now proceed to show that $q > 2r$. Our analysis will make use of the function $t : \mathbf{R} \to \mathbf{R}$ defined by $t(x) = \ell'(x) - \ell'(-x)$.

Calculating $\frac{\partial \Psi}{\partial v_{1,1}}\big|_{V^\star}$ by evaluating $\frac{\partial L(z, \mathbf{v}_1 \cdot \mathbf{w}_z, ..., \mathbf{v}_k \cdot \mathbf{w}_z)}{\partial v_{1,1}}$ for $z = 1, ..., 8$ in order, one per line, we get

$$\begin{aligned} &\alpha(-1/\sqrt{2} + \ell'(2r/\sqrt{2})(1/\sqrt{2})) \\ &+0 \\ &+\alpha\ell'(0)(-1/\sqrt{2}) \\ &+(1/4 - \alpha)\ell'(-r)(-1) \\ &+\alpha\ell'(-2r/\sqrt{2})(-1/\sqrt{2}) \\ &+0 \\ &+\alpha\ell'(0)(1/\sqrt{2}) \\ &+(1/4 - \alpha)\ell'(r). \end{aligned}$$

This simplifies to

$$-\alpha/\sqrt{2} + (\alpha/\sqrt{2})t(\sqrt{2}r) + (1/4 - \alpha)t(r).$$

Setting $\frac{\partial \Psi}{\partial v_{1,1}}\big|_{V^\star} = 0$, we get

$$\alpha/\sqrt{2} = (\alpha/\sqrt{2})t(\sqrt{2}r) + (1/4 - \alpha)t(r). \quad (6)$$

It is clear that $t(0) = 0$. Furthermore,

$$t'(u) = \ell''(u) - \ell''(-u)(-1) = \ell''(u) + \ell''(-u) > 0,$$

since ℓ is strictly convex. Thus t is increasing on all of $\mathbf{R}$. Hence by (6) we have that r must be positive, and moreover that

$$t(r) \leq \frac{\alpha/\sqrt{2}}{1/4 - \alpha} \leq \frac{5\alpha}{\sqrt{2}}. \quad (7)$$

Now we consider a different partial derivative, $\frac{\partial \Psi}{\partial v_{8,1}}\big|_{V^\star}$. Calculating $\frac{\partial \Psi}{\partial v_{8,1}}\big|_{V^\star}$ by evaluating $\frac{\partial L(z, \mathbf{v}_1 \cdot \mathbf{w}_z, ..., \mathbf{v}_k \cdot \mathbf{w}_z)}{\partial v_{8,1}}$ for $z = 1, ..., 8$ in order, one per line, we get

$$\begin{aligned} &\alpha\ell'(q/\sqrt{2})(1/\sqrt{2}) \\ &+0 \\ &+\alpha\ell'(q/\sqrt{2})(1/\sqrt{2}) \\ &+(1/4 - \alpha)\ell'(-q)(-1) \\ &+\alpha\ell'(-q/\sqrt{2})(-1/\sqrt{2}) \\ &+0 \\ &+\alpha\ell'(-q/\sqrt{2})(-1/\sqrt{2}) \\ &+(1/4 - \alpha)(-1 + \ell'(q)(1)) \end{aligned}$$

so setting $\frac{\partial \Psi}{\partial v_{8,1}}\big|_{V^\star} = 0$, we have

$$(1/4 - \alpha)t(q) + \sqrt{2}\alpha t(q/\sqrt{2}) = 1/4 - \alpha. \quad (8)$$

Note that from the above equation it is clear that q must be positive (recall that $t(0) = 0$ and t is increasing). Also, since t is increasing (8) implies

$$t(q) \geq \frac{1/4 - \alpha}{1/4 - \alpha + \sqrt{2}\alpha}. \quad (9)$$

We want to combine this constraint with (7) to prove that $q > 2r$.

Toward this end, we claim that $t(r) \leq 5\alpha/\sqrt{2}$, which we have from (7), implies that $r \leq \tau/2$. (Recall that τ was defined back at (3).) To see this, first assume for contradiction that $r > \tau/2$. Then,

$$t(r) = \int_{-r}^{r} \ell''(u)\, du > \int_{-\tau/2}^{\tau/2} \ell''(u)\, du$$
$$\geq \int_{-\tau/2}^{\tau/2} \ell''(0)/\sqrt{2}\, du = \tau\ell''(0)/\sqrt{2} \geq 5\alpha/\sqrt{2},$$

since $\alpha \leq \tau\ell''(0)/5$. Since this is a contradiction, we have $r \leq \tau/2$.

Since $r \leq \tau/2$, we have

$$t(2r) = \int_{-2r}^{2r} \ell''(u)\, du \leq \int_{-2r}^{2r} \sqrt{2}\ell''(0)\, du = 4\sqrt{2}r\ell''(0),$$

and, similarly $t(r) \geq 2r\ell''(0)/\sqrt{2}$, so that

$$\begin{aligned} t(2r) &\leq 4t(r) \\ &\leq 4\left(\frac{\alpha/\sqrt{2}}{1/4-\alpha}\right) \qquad \text{(by (7))} \\ &< \frac{1/4-\alpha}{1/4-\alpha+\sqrt{2}\alpha} \qquad \text{(since } \alpha \leq 1/20\text{)} \\ &\leq t(q), \end{aligned}$$

by (9). Since t is increasing, this implies $q > 2r$. This concludes the proof of Theorem 10.

5. Experiments

Experiments used a source P generated as follows. The domain X was $\mathbf{R}^d$ for $d = 100$. The number k of classes was 10. Class labels were assigned by applying a "target classifier" generated as follows. A weight matrix $W \in \mathbf{R}^{10\times 100}$ was generated by randomly generating a 10×2 matrix A and a 2×100 matrix B, and setting $W = AB$. The components of A and B were sampled i.i.d. from the standard normal distribution. If we refer to the rows of W as $\mathbf{w}_1, ..., \mathbf{w}_k$ as before, then the class y assigned to $\mathbf{x}$ was chosen to maximize $\mathbf{w}_y \cdot \mathbf{x}$. Elements of X were chosen uniformly at random from the surface of the unit ball.

We did experiments with three different loss functions: L_{CS}, L_{logit} and L_Q (Zhang, 2004). For each loss function L in this list, we experimented with the algorithm that, given $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)$, chooses V and $b_1, ..., b_k$ to minimize

$$\frac{\lambda}{2}||V||_F^2 + \frac{1}{m}\sum_{t=1}^{m} L(y_t, \mathbf{v}_1 \cdot \mathbf{x}_t + b_1, ..., \mathbf{v}_k \cdot \mathbf{x}_t + b_k),$$

where $||\cdot||_F$ is the Frobenius norm, and $\lambda = 10^{-6}$. The minimization was done using Averaged Stochastic Gradient Descent (Polyak & Juditsky, 1992; Xu, 2011; Bottou, 2010) (ASGD) with a step size of $\sqrt{1/t}$ on the tth update.

In our experiments, we used 10000 training examples and 10000 test examples. We ran each algorithm for 100 epochs. At the end of each epoch we evaluated the accuracy of the algorithm on the test data. We repeated this process, including the random generation of the target weight matrix W, five times. Results are plotted in Figure 1.

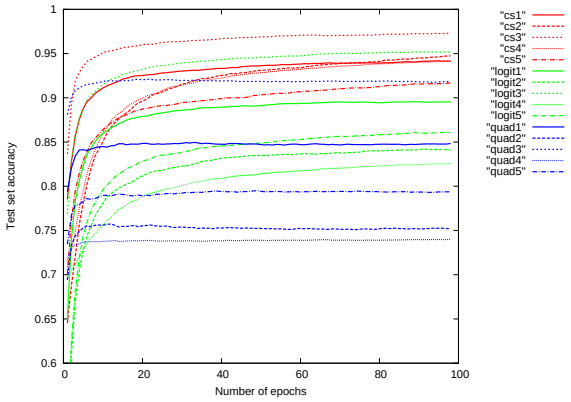

Figure 1. Comparison between the learning curves of classifiers trained using ASGD with different loss functions, run on data labeled with five different randomly generated target functions.

6. Conclusion

In this work we proposed the new notion of realizable H-consistency, which is a variant of the standard notion of consistency for loss functions. Our experimental and theoretical results indicate that for multiclass learning algorithms that work by minimizing a loss function over a restricted class H of scoring functions, realizable H-consistency may sometimes be more useful than consistency as a guide to classification performance.

Consideration of realizable H-consistency highlights circumstances where loss functions like L_{CS}, which involve a competition among the scores for different classes, may be preferred to generalizations of one-vs-rest, such as L_{logit}. As discussed after the proof of Theorem 9, L_{CS} is realizable H-consistent because it incorporates such a competition, and not, for example, because it generalizes the hinge loss. (The hinge loss was shown to be optimally noise-tolerant for binary classification, in a certain sense, in (Ben-David et al., 2012).)

In future work it would be interesting to extend our analyses to a non-realizable setting.

Acknowledgements

We thank Shenghuo Zhu, Anelia Angelova and Yuanqing Lin for valuable conversations.

A. Proof of Theorem 7

We will analyze the following source P. The domain consists of a single element x and the number k of

classes is 3.

$$P(x,1) = 3/7, P(x,2) = 2/7, P(x,3) = 2/7.$$

We first characterize the optimal value of P with respect to L_{CS}.

Lemma 12. $\mathbf{E}_{(x,y)\sim P}[L_{CS}(y, h_1(x), h_2(x), h_3(x)]$ *is minimized by setting* $h_1(x) = h_2(x) = h_3(x) = 1$, *where it takes the value* 1.

Before proving Lemma 12, we point out the following sanity check: one can easily verify that perturbing the solution $h_1(x) = h_2(x) = h_3(x) = 1$ by adding or subtracting a small positive constant to any of the variables while keeping the others constant makes the solution worse.

Now we proceed with the detailed proof.

Proof (of Lemma 12): For possible values u_1, u_2, u_3 of $h_1(x), h_2(x), h_3(x)$, if $[z]_+ \stackrel{\text{def}}{=} \max\{z, 0\}$, the quantity to be minimized is

$$\begin{aligned}\psi(u_1,u_2,u_3) \stackrel{\text{def}}{=} &(3/7)[1-u_1+\max\{u_2,u_3\}]_+ \\ &+(2/7)[1-u_2+\max\{u_1,u_3\}]_+ \\ &+(2/7)[1-u_3+\max\{u_1,u_2\}]_+.\end{aligned}$$

The minimimum of ψ can be equivalently represented as

$$\begin{aligned}&\min(3z_1+2z_2+2z_3)/7 \\ &\text{s.t.} \\ &z_1 \geq 1-u_1+u_2, z_1 \geq 1-u_1+u_3, z_1 \geq 0 \\ &z_2 \geq 1-u_2+u_1, z_2 \geq 1-u_2+u_3, z_2 \geq 0 \\ &z_3 \geq 1-u_3+u_1, z_3 \geq 1-u_3+u_2, z_3 \geq 0.\end{aligned}$$

A Lagrange multiplier formulation is

$$\begin{aligned}L \;=\; &(3z_1+2z_2+2z_3)/7 \\ &+\lambda_{1,2}(1-u_1+u_2-z_1) \\ &\quad +\lambda_{1,3}(1-u_1+u_3-z_1)+\lambda_{1,+}(-z_1) \\ &+\lambda_{2,1}(1-u_2+u_1-z_2) \\ &\quad +\lambda_{2,3}(1-u_2+u_3-z_2)+\lambda_{2,+}(-z_2) \\ &+\lambda_{3,1}(1-u_3+u_1-z_3) \\ &\quad +\lambda_{3,2}(1-u_3+u_2-z_3)+\lambda_{3,+}(-z_3).\end{aligned}$$

We claim that the following solution is optimal:

$$\begin{aligned}&u_1^* = u_2^* = u_3^* = 1 \\ &z_1^* = z_2^* = z_3^* = 1 \\ &\lambda_{1,+}^* = \lambda_{2,+}^* = \lambda_{3,+}^* = 0 \\ &\lambda_{1,2}^* = \lambda_{1,3}^* = \lambda_{2,1}^* = \lambda_{3,1}^* = 3/14 \\ &\lambda_{2,3}^* = \lambda_{3,2}^* = 1/14.\end{aligned}$$

Let us now check the KKT conditions. First, let's check partial derivatives with respect to $z_1, z_2, z_3, u_1, u_2, u_3$ respectively:

$$\begin{aligned}&3/7-\lambda_{1,2}^*-\lambda_{1,3}^*-\lambda_{1,+}^* \\ &\quad = 3/7-3/14-3/14-0=0 \\ &2/7-\lambda_{2,1}^*-\lambda_{2,3}^*-\lambda_{2,+}^* \\ &\quad = 2/7-3/14-1/14-0=0 \\ &2/7-\lambda_{3,1}^*-\lambda_{3,2}^*-\lambda_{2,+}^* \\ &\quad = 2/7-3/14-1/14-0=0 \\ &-\lambda_{1,2}-\lambda_{1,3}+\lambda_{2,1}+\lambda_{3,1} \\ &\quad = -3/14-3/14+3/14+3/14=0 \\ &-\lambda_{2,1}-\lambda_{2,3}+\lambda_{1,2}+\lambda_{3,2} \\ &\quad = -3/14-1/14+3/14+1/14=0 \\ &-\lambda_{3,1}-\lambda_{3,2}+\lambda_{1,3}+\lambda_{2,3} \\ &\quad = -3/14-1/14+3/14+1/14=0.\end{aligned}$$

All of the inequalities relating z_1, z_2, z_3 to u_1, u_2, u_3 are binding at $z_1^*, z_2^*, z_3^*, u_1^*, u_2^*, u_3^*$, so the complementary slackness conditions are satisfied for those constraints; since $\lambda_{i,+}^* = 0$ for all i, the complementary slackness constraints are satisfied for those constraints also. This completes the proof. □

Armed with Lemma 12, we are ready for the following.

Proof (of Theorem 7):

Choose $\delta > 0$, and consider $g_1, ..., g_3$ for which $g_1(x) = g_2(x) = 1$ and $g_3(x) = 1 + \delta/2$. We have

$$\begin{aligned}&\mathbf{E}_{(x,y)\sim P}[L_{CS}(y, h_1(x), h_2(x), h_3(x))] \\ &= (3/7)[1-1+\max\{1,1+\delta/2\}]_+ \\ &\quad +(2/7)[1-1+\max\{1,1+\delta/2\}]_+ \\ &\quad +(2/7)[1-(1+\delta/2)+\max\{1,1\}]_+ \\ &< 1+\delta.\end{aligned}$$

But, if $q(x) = \text{argmax}_{\hat{y}\in Y} g_{\hat{y}}(x)$, no matter how small δ is, we have $q(x) = 3$, so, for the Bayes optimal f, we have

$$\mathbf{Pr}_{(x,y)\sim P}[q(x) \neq y] = \mathbf{Pr}_{(x,y)\sim P}[f(x) \neq y] + 1/7,$$

completing the proof.

References

Ben-David, S., Loker, D., Srebro, N., and Sridharan, K. Minimizing the misclassification error rate using a surrogate convex loss. In *ICML*, 2012.

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pp. 177–187, 2010.

Breiman, L. Some infinity theory for predictor ensembles. *Annals of Statistics*, 32(1):1–11, 2004.

Crammer, K. and Singer, Y. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001.

Friedman, J., Hastie, T., and Tibshirani, R. Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 38(2):337–407, 2000.

Harchaoui, Zaïd, Douze, Matthijs, Paulin, Mattis, Dudík, Miroslav, and Malick, Jérôme. Large-scale image classification with trace-norm regularization. In *CVPR*, pp. 3386–3393, 2012.

Haussler, D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1): 78–150, 1992.

Kearns, M. J., Schapire, R. E., and Sellie, L. M. Toward efficient agnostic learning. *Proceedings of the 1992 Workshop on Computational Learning Theory*, pp. 341–352, 1992.

Lee, Y., Lin, Y., and Wahba, G. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Lin, Y., Lv, F., Zhu, S., Yang, M., Cour, T., Yu, K., Cao, L., and Huang, T. Large-scale image classification: fast feature extraction and SVM training. In *CVPR*, pp. 1689 – 1696, 2011.

Liu, Y. Fisher consistency of multicategory support vector machines. In *UAI*, pp. 291–298, 2007.

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. *Machine Learning*, 78(3):287–304, 2010.

Perronnin, F., Akata, Z., Harchaoui, Z., and Schmid, C. Towards good practice in large-scale learning for image classification. *CVPR*, pp. 3482–3489, 2012.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *JMLR*, 8:1007–1025, 2007.

Vapnik, V. N. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). *Proceedings of the 1989 Workshop on Computational Learning Theory*, 1989.

Weston, J., Bengio, S., and Usunier, N. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, 81(1):21–35, 2010.

Xu, W. Towards optimal one pass large scale learning with averaged stochastic gradient descent. *Arxiv*, 2011.

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