On the Consistency of Top-k Surrogate Losses

Forest Yang [1 2] **Sanmi Koyejo** [1 3]

Abstract

The top-k error is often employed to evaluate performance for challenging classification tasks in computer vision as it is designed to compensate for ambiguity in ground truth labels. This practical success motivates our theoretical analysis of consistent top-k classification. Surprisingly, it is not rigorously understood when taking the k-argmax of a vector is guaranteed to return the k-argmax of another vector, though doing so is crucial to describe Bayes optimality; we do both tasks. Then, we define top-k calibration and show it is necessary and sufficient for consistency. Based on the top-k calibration analysis, we propose a class of top-k calibrated Bregman divergence surrogates. Our analysis continues by showing previously proposed hinge-like top-k surrogate losses are not top-k calibrated and suggests no convex hinge loss is top-k calibrated. On the other hand, we propose a new hinge loss which is consistent. We explore further, showing our hinge loss remains consistent under a restriction to linear functions, while cross entropy does not. Finally, we exhibit a differentiable, convex loss function which is top-k calibrated for specific k.

1. Introduction

Consider a multiclass classifier which is granted k guesses, so its prediction is declared error-free only if any one of the guesses is correct. This conceptually defines the top-k error (Akata et al., 2012). Top-k error[1] is popular in computer vision, natural language processing, and other applied problems where there are a large number of possible classes, along with potential ambiguity regarding the label of a sample and/or when a sample may correspond to multiple labels, e.g., when an image of a park containing a pond may be correctly labeled as either a park or a pond Russakovsky et al. (2015); Xiao et al. (2010); Zhou et al. (2018).

Like the zero-one loss for binary classification, the top-k error is computationally hard to minimize directly because it is discontinuous and only has zero gradients. Instead, practical algorithms depend on minimizing a surrogate loss, often a convex upper bound (Lapin et al., 2015; 2016). To this end, the corresponding predictive model is most often trained to output a continuous-valued score vector, and the classes corresponding to the top k entries of the score vector constitute the classification prediction (Lapin et al., 2018). While popular in practice, there is limited work on the theoretical properties of top-k error and its surrogate losses. We are particularly interested in the *consistency* of surrogate losses, which states whether the learned classifier converges to the population optimal prediction (commonly known as the Bayes optimal) in the infinite sample limit.

Main Contributions. Our contributions are primarily theoretical, and are outlined as follows:

- We characterize Bayes-optimal scorers for the weighted top-k error, i.e., a slight generalization top-k error with class-specific weights. The scorers are functions which predict continuous vectors, so the k maximum arguments define the prediction. Our analysis highlights the top-k preserving property as fundamental to top-k consistency, then outlines the notion of calibration which is necessary and sufficient to construct consistent top-k surrogate losses.
- We propose a family of consistent (weighted) top-k surrogate losses based on Bregman divergences. We show the inconsistency of previously proposed top-k hinge-like surrogate losses and propose new ones, one of which is (weighted) top-k consistent. Since any convex hinge loss must have form similar to the ones proved inconsistent, this suggests that consistent hinge losses must be nonconvex.
- We further prove the consistency of the new hinge loss when given top-k separable data and restricted to linear predictors. On the other hand, we also show

[1]Google Research Accra [2]University of California, Berkeley [3]University of Illinois at Urbana-Champaign. Correspondence to: Forest Yang <forestyang@berkeley.edu>, Sanmi Koyejo <sanmi@illinois.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]The top-k error is simply 1 - top-k accuracy, thus the metrics are equivalent.

that cross entropy, while being top-k consistent in the unrestricted setting, is not consistent when restricted to linear models.

- A loss being convex and differentiable can often lead to strong guarantees. Investigating this, we find that while a convex and differentiable top-k calibrated loss function must also be calibrated for $k' \leq k$, but by exhibiting a counterexample we show that it need not be calibrated for $k' > k$.

- We employ these losses in synthetic experiments, observing aspects of their behavior which reflect our theoretical analysis.

Taken together, our results contribute to the fundamental understanding of top-k error and its (in)consistent surrogates.

1.1. Notation

For any $N \in \mathbb{Z}^+$, we use the notation $[N] = \{1, \ldots, N\}$. We assume there are M classes and denote the input space as $\mathcal{X}$. We also denote the ith coordinate basis vector as e_i; the dimension should be clear from context. $\mathcal{Y} = [M]$ is the discrete label space. The data is assumed to be generated i.i.d. from some distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$.

Define the probability simplex $\Delta_M := \{v \in \mathbb{R}^M \mid \forall m \in [M],\ v_m \geq 0,\ \sum_{m=1}^M v_m = 1\}$, and let $\eta(x) \in \Delta_M$ be the conditional distribution of $y \in \mathcal{Y}$ given $x \in \mathcal{X}$, i.e. $\eta(x)_m = P(y = m \mid X = x)$. Furthermore, given a vector $v \in \mathbb{R}^m$, let $v_{[j]}$ denote the jth greatest entry of v. For example, if $v = (1, 4, 4, 2)$, then $v_{[1]} = 4, v_{[2]} = 4, v_{[3]} = 2, v_{[4]} = 1$.

1.2. Related Work

The statistical properties of surrogates for binary classification are well-studied (Zhang, 2004b; Bartlett et al., 2003a). Furthermore, many of these results have been extended to multiclass classification with the accuracy metric (Zhang, 2004a; Tewari & Bartlett, 2005). Usually, $y \in \{1, \ldots, M\}$, $s \in \mathbb{R}^M$ is a vector-valued score, and the prediction is the index of the entry of s with the highest value. There have also been recent studies on a general framework for consistent classification with more general concave and fractional linear multiclass metrics (Narasimhan et al., 2015). In the realm of multilabel classification, there is work on extending multiclass algorithms to multilabel classification (Lapin et al., 2018), characterizing consistency for multilabel classification (Gao & Zhou, 2013), and constructing a general framework for consistent classification with multilabel metrics (Koyejo et al., 2015).

On the other hand, statistical properties such as consistency of surrogate loss functions for the top-k error are not so thoroughly characterized. It is known that softmax loss $-\log\left(\frac{e^{s_y}}{\sum_{m=1}^M e^{s_m}}\right)$ is top-k consistent and that the multiclass hinge loss $\max_{m\in[M]}\{\mathbb{1}[m \neq y] + s_m - s_y\}$ proposed by Crammer & Singer (2001) is top-k inconsistent (Zhang, 2004a). However, the consistency of recently proposed improved top-k surrogates such as proposals in Berrada et al. (2018); Lapin et al. (2015; 2016; 2018) has so far remained unresolved. Our work resolves some of these open questions by showing their inconsistency, in addition to providing a more robust framework for top-k consistency.

2. Top-k consistency

We begin by formally defining the top-k error.

Definition 2.1 (Top-k error)**.** Given label vector $y \in \mathcal{Y}$ with $y_l = 1$ and prediction $s \in \mathbb{R}^M$, the top-k error is defined as

$$\text{err}_k(s, y) = \mathbb{1}[l \notin r_k(s)], \tag{1}$$

where $r_k : \mathbb{R}^M \to \{J : J \subset [M], |J| = k\}$ is a top-k selector which selects the k indices of the greatest entries of the input, breaking ties arbitrarily. Different r_k's correspond to different ways of breaking ties; we will take a worst-case perspective for ensuring Bayes optimality.

In general, s is the output of some predictor θ given a sample $x \in \mathcal{X}$. The goal of a classification algorithm under the top-k metric is to learn a predictor $\theta : \mathcal{X} \to \mathbb{R}^M$ that minimizes the risk

$$L_{\text{err}_k}(\theta) := \mathbb{E}_{(x,y)\sim\mathbb{P}}[\text{err}_k(\theta(x), y)].$$

Given $s \in \mathbb{R}^M$ and $\eta \in \Delta_M$, we may define the conditional risk

$$L_{\text{err}_k}(s, \eta) := \mathbb{E}_{y\sim\eta}[\text{err}_k(s, y)].$$

Furthermore, we define optimal risk and conditional risk

$$L^*_{\text{err}_k} := \inf_{\theta:\mathcal{X}\to\mathbb{R}^M} L_{\text{err}_k}(\theta),$$
$$L^*_{\text{err}_k}(\eta) := \inf_{s\in\mathbb{R}^M} L_{\text{err}_k}(s, \eta).$$

Analogous population statistics for arbitrary loss functions $\psi : \mathbb{R}^M \times \mathcal{Y} \to \mathbb{R}$ are denoted by swapping the metrics, e.g. ψ risk is defined as $L_\psi(\theta) := \mathbb{E}_{(x,y)\sim\mathbb{P}}[\psi(\theta(x), y)]$.

2.1. Bayes Optimality

Here we define and characterize Bayes optimal predictors for the top-k error.

Definition 2.2 (Top-k Bayes optimal)**.** The predictor $\theta^* : \mathcal{X} \to \mathbb{R}^M$ is top-k Bayes optimal if

$$L_{\text{err}_k}(\theta^*) = L^*_{\text{err}_k}.$$

We remark that it is much less obvious which s, given η, are optimal for (minimize) the top-k conditional risk $L_{\mathrm{err}_k}(s,\eta)$ than for the binary conditional risk, where $s \in \mathbb{R}$ is optimal (for a worst case selector) iff $\eta > 1/2 \implies s > 0$ and $\eta < 1/2 \implies s < 0$. This has led to seemingly natural but incorrect statements in prior work. For example, Lapin et al. (2016; 2018) write

$$s \in \arg\min_s L_{\mathrm{err}_k}(s,\eta) \iff$$
$$\{y \mid s_y \geq s_{[k]}\} \subseteq \{y \mid \eta_y \geq \eta_{[k]}\},$$

which says that the top-k indices of s are contained in the top-k indices of η. However, consider the following counter-example. Let $s = (0,1,1)$, $\eta = (1,0,0)$ and $k = 2$. Note $s_{[k]} = 1, \eta_{[k]} = 0$. Then, $\{y \mid s_y \geq s_{[k]}\} = \{2,3\} \subseteq \{y \mid \eta_y \geq \eta_{[k]}\} = \{1,2,3\}$. By the above definition, s is considered optimal. Yet, it is not, because for any top 2-selector $r_2(s) = \{2,3\}$, which has 100% top-k error. On the other hand, $s^* = (1,0,0)$ has 0 top-k error.

One of our main contributions is to define the top-k preserving property, a necessary and sufficient property for top-k optimality that solves this difficulty.

Definition 2.3 (Top-k preserving property)**.** Given $x \in \mathbb{R}^M$ and $y \in \mathbb{R}^M$, we say that y is *top-k preserving with respect to* x, denoted $\mathsf{P}_k(y,x)$, if for all $m \in [M]$,

$$x_m > x_{[k+1]} \implies y_m > y_{[k+1]}$$
$$x_m < x_{[k]} \implies y_m < y_{[k]}.$$

The negation of this statement is $\neg\mathsf{P}_k(y,x)$.

This is not a symmetric condition. For example, although $y = (4,3,2,1)$ is top-2 preserving with respect to $x = (4,2,2,1)$, x is not top-2 preserving with respect to y. The following proposition and its proof illuminate the connection between top-k preserving and top-k optimality.

Proposition 2.1. *$\theta : \mathcal{X} \to \mathbb{R}^M$ is top-k Bayes optimal for any top-k selector r_k if and only if $\theta(X)$ is top-k preserving with respect to $\eta(X)$ almost surely.*

Proof. Fix $x \in \mathcal{X}$ and $s \in \mathbb{R}^M$, with $\eta = \eta(x)$. We have

$$L_{\mathrm{err}_k}(s,\eta) = \mathbb{E}_{y\sim\eta}[\mathrm{err}_k(s,y)] = \sum_{m\in[M]\setminus r_k(s)} \eta_m$$
$$= 1 - \sum_{m\in r_k(s)} \eta_m \geq 1 - \sum_{m=1}^{k} \eta_{[m]}.$$

The last inequality holds because $|r_k(s)| = k$, so $\sum_{m\in r_k(s)} \eta_m \leq \sum_{m=1}^k \eta_{[m]}$. Equality occurs if and only if $\sum_{m\in r_k(s)} \eta_m = \sum_{m=1}^k \eta_{[m]}$. If equality does not hold, there exists $i \in r_k(s)$, $j \in [M] \setminus r_k(s)$ such that $\eta_j > \eta_i$. If $\eta_j > \eta_{[k+1]}$, then since $s_j \notin r_k(s)$, $s_j \not> s_{[k+1]}$. If $\eta_j \leq \eta_{[k+1]}$, then $\eta_i < \eta_{[k+1]} \leq \eta_{[k]}$. However, $s_i \not< s_{[k]}$, because $i \in r_k(s)$. Either way, $\neg\mathsf{P}_k(s,\eta)$.

If $\neg\mathsf{P}_k(s,\eta)$, then there exists $i \in [M]$ such that $\eta_i > \eta_{[k+1]}$ but $s_i \leq s_{[k+1]}$, or $\eta_i < \eta_{[k]}$ but $s_i \geq s_{[k]}$. In the first case, there is an r_k such that $i \notin r_k(s)$, because there are at least k indices $j \in [M]$, $j \neq i$ such that $s_j \geq s_i$. In the second case, there is an r_k such that $i \in r_k(s)$, because s_i is one of the top k values of s. In either case, there is an r_k such that $\sum_{m\in r_k(s)} \eta_m < \sum_{m=1}^k \eta_{[m]}$. Thus, $L_{\mathrm{err}_k}(s,\eta)$ is optimal for any selector r_k if and only if $\mathsf{P}_k(s,\eta)$, i.e. s is top-k preserving with respect to η.

Finally, we note that

$$L_{\mathrm{err}_k}(\theta) = \mathbb{E}_{X\sim\mu}[L_{\mathrm{err}_k}(\theta(X),\eta(X))],$$

where μ is the conditional distribution of X. It follows that θ minimizes $L_{\mathrm{err}_k}(\theta)$ if and only if $\theta(X)$ minimizes $L_{\mathrm{err}_k}(\theta(X),\eta(X))$ almost surely. In other words, θ is a Bayes optimal predictor for any r_k if and only if $\mathsf{P}_k(\theta(X),\eta(X))$ almost surely. □

2.2. Top-k calibration

Top-k calibration characterizes when minimizing ψ for a fixed x leads to the Bayes decision for that x. Analogous notions have been defined for binary classification, (Bartlett et al., 2003a) multiclass classification, (Zhang, 2004a), and ranking (Calauzènes et al., 2013).

Definition 2.4 (Top-k calibration)**.** A loss function $\psi : \mathbb{R}^M \times \mathcal{Y} \to \mathbb{R}$ is *top-k calibrated* if for all $\eta \in \Delta_M$,

$$\inf_{s\in\mathbb{R}^M : \neg\mathsf{P}_k(s,\eta)} L_\psi(s,\eta) > \inf_{s\in\mathbb{R}^M} L_\psi(s,\eta) = L^*_\psi(\eta).$$

If a minimizer s^* of $L_\psi(s,\eta)$ exists, this implies that s^* must be top-k preserving with respect to η. By Proposition 2.1, top-k calibration is necessary for minimizing L_ψ to guarantee minimizing L_{err_k}.

More generally, if $\{s^{(n)}\}$ is a sequence such that $L_\psi(s^{(n)},\eta) \to \inf_s L_\psi(s,\eta)$, then it is eventually top-k preserving, i.e. for all n greater than some N, $\mathsf{P}_k(s^{(n)},\eta)$.

2.3. Obtaining consistency

We can convert top-k calibration into top-k consistency for all lower bounded loss functions. By Corollary 4.5 of Calauzènes et al. (2013), since minimizing err_k is equivalent to maximizing recall at k, and $|\mathcal{Y}| = M$ is finite, if ψ is continuous and nonnegative then top-k calibration implies uniform calibration, which implies the existence of a surrogate regret bound $L_{\mathrm{err}_k}(f) - L^*_{\mathrm{err}_k} \leq \Gamma(L_\psi(f) - L^*_\psi)$, where $\Gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is continuous at 0, and $\Gamma(0) = 0$. Then continuity of Γ at 0 implies consistency: $L_\psi(f^{(n)}) \to L^*_\psi \implies L_{\mathrm{err}_k}(f^{(n)}) \to L^*_{\mathrm{err}_k}$. As an aside, we note

that before we were aware of Calauzènes et al. (2013), we proved a slightly generalized version of this result without the additional assumption that ψ is continuous. Details are included in the appendix for completeness.

Theorem 2.2. *Suppose ψ is a nonnegative top-k calibrated loss function. Then ψ is top-k consistent, i.e., for any sequence of measurable functions $f^{(n)} : \mathcal{X} \to \mathbb{R}^M$, we have*

$$L_\psi(f^{(n)}) \to L^*_\psi \implies L_{\mathrm{err}_k}(f^{(n)}) \to L^*_{\mathrm{err}_k}.$$

Proof. See appendix. □

3. Bregman Divergence Top-k Consistent Surrogates

Next, we outline top-k consistent surrogates based on Bregman divergences. Given a convex, differentiable function $\phi : \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$, define the Bregman divergence D_ϕ by

$$D_\phi(s,t) = \phi(t) - \phi(s) - \nabla\phi(s)^\top(t-s). \qquad (2)$$

$D_\phi(s,\cdot)$ can be interpreted as the error when approximating $\phi(\cdot)$ by the first order Taylor expansion of ϕ centered at s. Bregman divergences include squared loss and KL divergence as special cases.

Here, we present the result that any Bregman divergence composed with an *inverse* top-k preserving function is top-k calibrated. First we define inverse top-k preserving functions, then give the theorem.

Definition 3.1 (Inverse top-k preserving function.)**.** Given $A \subseteq \mathbb{R}^M$ and $B \subseteq \mathbb{R}^M$, $f : A \to B$ is *inverse top-k preserving* if $\forall x \in A$, $\mathsf{P}_k(x, f(x))$.

Theorem 3.1. *Suppose $\phi : \mathbb{R}^M \to \mathbb{R}^M$ is strictly convex and differentiable. If $g : \mathbb{R}^M \to \mathbb{R}^M$ is inverse top-k preserving, continuous, and $\Delta_M \subseteq \mathrm{range}(g)$, then $\psi : \mathbb{R}^M \times \mathcal{Y} \to \mathbb{R}$ defined by*

$$\psi(s,y) = D_\phi(g(s), e_y)$$

is top-k calibrated.

Proof. See Appendix. □

Theorem 3.1 is similar to one of the main results (Theorem 8) in Ravikumar et al. (2011), except inverse order-preserving is relaxed to inverse top-k preserving, the above is only a sufficient condition for top-k calibration, and we make no invertibility assumptions.

3.1. Cross entropy is top-k calibrated

By Theorem 3.1, the commonly used softmax with cross-entropy loss is top-k calibrated:

$$\mathrm{Ent}(s,y) = -\ln\left(\frac{e^{s_y}}{\sum_{m=1}^M e^{s_m}}\right)$$

can be rewritten as $\mathrm{Ent}(s,y) = D_\phi(g(s), e_y)$ with $\phi(x) = \sum_{m=1}^M x_m \ln x_m$ and $g(s)_m = \frac{e^{s_m}}{\sum_{i=1}^M e^{s_i}}$. ϕ is strictly convex and differentiable, and g satisfies the assumptions of Theorem 3.1. In fact, g satisfies the stronger *rank preserving* condition,

$$\forall i,j \in [M],\ s_i > s_j \iff g(s)_i > g(s)_j.$$

As a result, $\mathrm{Ent}(s,y)$ is top-k calibrated for every k, i.e. *rank consistent*. An interesting question is whether there is a surrogate loss which does not satisfy such a strong property, and is top-k calibrated for just a specific k. We answer in the affirmative in the sequel.

4. Top-k hinge-like losses

Hinge-like losses for top-k classification have been proposed by Lapin et al. (2015; 2016), inspired by ranking losses in Usunier et al. (2009), and minimized via SDCA. They note that cross entropy is competitive across datasets and values of k, but slight improvement is attainable with hinge losses. We list these losses as well as new ones we propose, ψ_4, ψ_5, in Table 1.

Table 1. Discussed hinge-like top-k loss functions along with whether they are top-k calibrated. We use the notation $(x)_+ = \max\{x, 0\}$.

Loss fn.	Loss eqn.	Ref.	Calib.
ψ_1	$\left(1 + (s_{\setminus y})_{[k]} - s_y\right)_+$	10; 5	No
ψ_2	$\left(\frac{1}{k}\sum_{i=1}^k (s + \bar{1}(y))_{[i]} - s_y\right)_+$	10; 11; 12	No
ψ_3	$\frac{1}{k}\sum_{i=1}^k \left[(s + \bar{1}(y))_{[i]} - s_y\right]_+$	10; 11; 12	No
ψ_4	$\left(\frac{1}{k}\sum_{i=1}^k (1 + (s_{\setminus y})_{[i]}) - s_y\right)_+$.	New	No
ψ_5	$\left(1 + s_{[k+1]} - s_y, 0\right)_+$	New	Yes

The motivation of these losses is as follows. ψ_1 is a generalization of multiclass SVM (Crammer & Singer, 2001). ψ_2 and ψ_3 are convex upper bounds on ψ_1.
We propose ψ_4 as a tighter convex upper bound on ψ_1 and ψ_5 as the tightest bound on err_k of all, and the only top-k calibrated loss. Next, we show that ψ_5 is top-k calibrated and the rest, $\psi_1, \psi_2, \psi_3, \psi_4$, are not top-$k$ calibrated. These facts are not in previous literature.

4.1. Characterization of hinge-like losses

We compute the minimizers of the expected loss $L_{\psi_1}(s,\eta) = \mathbb{E}_{y\sim\eta}[\psi_1(s,y)]$ given a conditional distribution $\eta \in \Delta_M$. Though we arrive at inconsistency, our results also indicate that if η is from the restricted probability simplex $\{\eta \in \Delta_M \mid \eta_{[k]} > \sum_{i=k+1}^M \eta_{[i]}\}$, ψ_1 is calibrated/consistent.

Theorem 4.1 (Abridged)**.** *Let $\eta \in \Delta_M$ and suppose $\eta_1 \geq$*

$\eta_2 \geq \ldots \geq \eta_M$. *Then,*

$$\eta_k \geq \sum_{i=k+1}^{M} \eta_i \implies [\overbrace{1\,1\,\ldots\,1}^{k-1}\;0\,0\,\ldots\,0] \in \arg\min_s L_{\psi_1}(s,\eta)$$

$$\eta_k \leq \sum_{i=k+1}^{M} \eta_i \implies [\overbrace{1\,1\,\ldots\,1\,1}^{k}\;0\,\ldots\,0] \in \arg\min_s L_{\psi_1}(s,\eta).$$

Proof. See appendix for the exact set of minimizers when η has no zero entries, and proof. □

This implies that ψ_1 is not top-k calibrated: if $\eta_1 > \ldots > \eta_M$ then in the first case of the above theorem, s^* is not top-k preserving with respect to η: for any $m \in \{k+1, \ldots, M\}$, $\eta_m < \eta_k$, and yet $s_m \not< s_{[k]} = 0$. Yet, s^* is a minimizer of $L_{\psi_1}(s,\eta)$, so ψ_1 is not top-k calibrated.

The following proposition implies that $\{\psi_2, \psi_3, \psi_4\}$ are not top-k calibrated, and are thus inconsistent.

Proposition 4.2. *For any $\psi \in \{\psi_2, \psi_3, \psi_4\}$, if $\sum_{m=k+1}^{M} \eta_{[m]} > \frac{k}{k+1}$, we have $0 \in \arg\min_s L_\psi(s,\eta)$, and thus $L^*_\psi(\eta) = \min_s L_\psi(s,\eta) = L_\psi(0,\eta) = 1$.*

Proof. See Appendix. □

To show this leads to inconsistency, take $\eta = (1/8, 1/8, 1/12, 1/12, \ldots, 1/12) \in \Delta_{11}$ with $k = 2$. η satisfies $\sum_{i=k+1}^{M} \eta_{[i]} = \frac{3}{4} > \frac{2}{3} = \frac{k}{k+1}$, so the optimal is $s^* = 0$. But, s^* is not top-k preserving wrt η. This implies that $\psi \in \{\psi_2, \psi_3, \psi_4\}$ is not top-k calibrated.

Proposition 4.3. *$\psi_5 : \mathbb{R}^M \times \mathcal{Y} \to \mathbb{R}$ is top-k calibrated.*

Proof. See Appendix. Note since ψ_5 is bounded below, by Theorem 2.2, it is top-k consistent. □

4.2. Conjecture on the lack of convex hinge losses

Generally, a hinge loss can be considered to have the form

$$\psi(s,y) = \max\{w^\top P(f(s))f(s - s_y\mathbf{1}), 0\}$$

where f is an affine function (or may also contain a $(\cdot)_+$) and P is a permutation matrix depending on $f(s)$. w is a fixed vector. For example, for ψ_2, we have $f(s) = s + \bar{\mathbf{1}}(y)$, P the sort matrix, and $w = 1/k$ for the first k entries. ψ_3 and ψ_4 are similar. If we assume ψ is convex, we must have P be the sorting matrix and w's entries in decreasing order (Usunier et al., 2009). Intuitively, the closest we can get to being top-k calibrated is when w's nonzero entries are equal; this leads to essentially the existing hinge loss surrogates, which are uncalibrated. Thus, we conjecture that no convex, piecewise affine loss is top-k calibrated.

5. Linear (in)consistency

Until now, we have been discussing consistency with respect to all measurable functions, as is standard. We may instead consider consistency with respect to a restricted function class $\mathcal{F}$. This type of consistency was explored for $k = 1$ in Long & Servedio (2013). Time of

Definition 5.1 ($\mathcal{F}$-consistency)**.** $\psi : \mathbb{R}^m \times \mathcal{Y} \to \mathbb{R}$ is $\mathcal{F}$ top-k consistent (or $\mathcal{F}$-consistent) if

$$L_\psi(f_n) \to \inf_{f' \in \mathcal{F}} L_\psi(f') \implies L_{\mathrm{err}_k}(f_n) \to \inf_{f' \in \mathcal{F}} L_{\mathrm{err}_k}(f'),$$

where $(f_n)_{n=1}^\infty$ is a sequence of functions $\mathcal{X} \to \mathbb{R}^M$ in $\mathcal{F}$. If no conditions or set of distributions are specified, $\mathcal{F}$-consistent means the above holds for every probability distribution over $\mathcal{X} \times \mathcal{Y}$.

Previously, the infimum with respect to the scoring function was over all measurable functions, but in practice, we minimize using some function class, e.g., functions computed by a neural net architecture.

$\mathcal{F}$-consistency seems much more difficult to analyze than consistency because we may no longer decompose the risk into $L(f(x), \eta(x))$ for each x, as f cannot vary its outputs arbitrarily. Furthermore, if $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F}$ consists of linear functions, $\mathcal{F}$-consistency of a convex ψ suggests $\mathsf{P} = \mathsf{NP}$, due to the efficiency of convex minimization and the NP-hardness of finding a linear separator which maximizes accuracy (Ben-David et al., 2003).

On the other hand, letting $L^*(\mathcal{F}) = \inf_{f' \in \mathcal{F}} L(f')$, as long as $L^*_\psi(\mathcal{F}) = L^*_\psi$ and $L^*_{\mathrm{err}_k}(\mathcal{F}) = L^*_{\mathrm{err}_k}$, top-$k$ consistency implies $\mathcal{F}$-top-k consistency because

$$\begin{aligned} L_\psi(f_n) \to L^*_\psi(\mathcal{F}) &\implies L_\psi(f_n) \to L^*_\psi \\ \implies L_\psi(f_n) \to L^*_{\mathrm{err}_k} &\implies L_\psi(f_n) \to L^*_{\mathrm{err}_k}(\mathcal{F}). \quad (3) \end{aligned}$$

Furthermore, we can answer easier questions about $\mathcal{F}$-consistency by making additional assumptions, e.g. top-k separability. If there is a top-k separator, i.e. a predictor with perfect top-k accuracy, then does our algorithm (i.e., minimizing a surrogate loss) find it? Despite NP-hardness in general, if a linear separator exists for a binary classification problem, one can be found efficiently, so it seems appropriate to ask an analogous question for top-k separability in the context of surrogate losses.

Proposition 5.1. *Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F} = \{x \mapsto Wx : W \in \mathbb{R}^{M \times d}\}$. Then if we consider top-k separable probability distributions over $\mathcal{X} \times \mathcal{Y}$, i.e. $L^*_{\mathrm{err}_k}(\mathcal{F}) = 0 = L^*_{\mathrm{err}_k}$, then:*

1. *If $k = 1$,* Ent *is $\mathcal{F}$-consistent.*
2. *If $d \geq 3, M \geq 3$, and $k = 2$,* Ent *is not $\mathcal{F}$-consistent.*

3. *ψ_1 and ψ_5 are $\mathcal{F}$-consistent.*

The above proposition says the answer is yes for ψ_1 and ψ_5, and generally no for Ent unless $k = 1$. To see Propopsition 5.1.1, note that top-1 separability means $\exists W \in \mathbb{R}^{M\times d}$ where $\Pr[(Wx)_y > (Wx)_{[2]}] = 1$. Then, w.p. 1 over x, y,

$$\begin{aligned}\mathrm{Ent}(cWx, y) &= \log\left(1 + \sum_{m\neq y} e^{c((Wx)_m - (Wx)_y)}\right)\\ &\xrightarrow{c\to\infty} \log(1) = 0.\end{aligned}$$

Thus, $\mathrm{Ent}^*(\mathcal{F}) = 0 = \mathrm{Ent}^*$ and we have $\mathcal{F}$ consistency by (3). Note we cannot spply this "scaling to 0 loss" argument for Ent when $k \geq 2$. The rest of the proof is in the appendix.

6. A convex, differentiable loss function

While we achieved top-k calibration for a specific k with the ψ_5 loss, one might wonder whether this is possible with a convex, differentiable loss function. In some sense, because of the case of $M = 2$, one would expect that if a convex, differentiable loss function is top-k calibrated for some $k < M$, then it is top-k calibrated for all k'. In Bartlett et al. (2003b), it was proven that a convex margin function just needs to have negative derivative at 0 to be binary consistent, raising the question of whether a similar claim can be made when the number of labels increases. The increase in number of directions the score vector can travel makes the question much harder to answer.
It turns out that this is partially true, and partially untrue. It is true in the sense of the following theorem:

Theorem 6.1. *Suppose $\psi(s, y)$ is convex and differentiable for each $y \in [M]$, and moreover if we think of $\Psi(s)$ as the M length vector whose entries are $\psi(s, y)$, symmetric in the sense of $\Psi(Ps) = P\Psi(s)$ for all permutation matrices P. Then, if ψ is top-k calibrated for some $k < M$, it is top-k' calibrated for all $k' \leq k$.*

Proof. Let e_i denote the ith coordinate basis vector. Suppose that s^* minimizes $L_\psi(s, \eta) = \langle \eta, \Psi(s)\rangle$. Suppose that i, j are in the arguments of the top-k entries of η, and $\eta_i > \eta_j$. Define $\eta^\flat$ as η but with η_i and η_j replaced with their average. For a large enough $\delta > 0$, $\tilde{\eta} = \eta^\flat + \delta(e_i - e_j) \in \Delta_M$ has j no longer in the top-k entries of $\tilde{\eta}$. Suppose $\tilde{s}$ minimizes $\langle \tilde{\eta}, \Psi(s)\rangle$ and $s^\flat$ minimizes $\langle \eta^\flat, \Psi(s)\rangle$. We have

$$\begin{aligned}0 &> \langle \tilde{\eta}, \Psi(\tilde{s}) - \Psi(s^\flat)\rangle\\ &\geq \tilde{\eta}^\top \nabla\Psi(s^\flat)(\tilde{s} - s^\flat)\\ &= \delta(e_i - e_j)^\top \nabla\Psi(s^\flat)(\tilde{s} - s^\flat)\\ &= \delta(\nabla\psi(s^\flat, i) - \nabla\psi(s^\flat, j))^\top(\tilde{s} - s^\flat)\\ &= \delta(a - b)(\tilde{s}_i - \tilde{s}_j).\end{aligned}$$

The second line is by convexity of Ψ, the third line is by optimality of $s^\flat$ for $\eta^\flat$. The last line uses symmetry of Ψ: since $s^\flat_i = s^\flat_j$ (follows from convexity of Ψ and $\eta^\flat_i = \eta^\flat_j$), the ith and jth gradients are equal to each other, except their ith and jth entries, a and b, are swapped.
Since ψ is top-k preserving and j is no longer in the top-k entries of $\tilde{\eta}$, we have $\tilde{s}_i > \tilde{s}_j$. Thus, $a < b$. Notice that we can replace $\tilde{s}$ with s^* and everything in the chain holds – $s^\flat$ is not a minimizer of $\langle \eta, \Psi(s)\rangle$, because $\nabla\langle \eta, \Psi^{s^\flat}\rangle(\tilde{s} - s^\flat) = \eta\nabla\Psi(s^\flat)(\tilde{s} - s^\flat) = \delta'(a-b)(\tilde{s}_i - \tilde{s}_j) < 0$. Therefore, $s^*_i > s^*_j$, as desired. □

However, it is untrue in that we can exhibit a convex, differentiable, symmetric loss which is top-1 calibrated but not top-2, 3, 4, ... calibrated. It is shown below:

$$\begin{aligned}\Psi^{CD}(s, y) &= \log(1 + \exp(-s_y))\\ &+ \sum_{i\neq y}\left(s_i - \frac{1}{M-1}\sum_{j\neq y} s_j\right)^2 + \sum_{i\neq y} s_i^2 \qquad (4)\end{aligned}$$

To show Ψ^{CD} is not calibrated for $1 < k < M$, we run gradient descent on $L_{\psi^{CD}}(s, \eta) = \langle \eta, \Psi^{CD}(s)\rangle$ with $\eta = [0.01, 0.02, 0.03, 0.04, 0.9]$ and reach an optimum of $[0.0114226, 0.011404, 0.011385, 0.011365, 0.470880]$. While the most probable class got the highest score, the scores of the others are reversed relative to probability. To see why this happens intuitively, the presence of the logistic loss makes s_5 at optimum much higher than the others, since η_5 is by far the largest. Now for $y \neq 5$, the best way to decrease the loss $\psi(s, y)$ is to increase s_j, $j \neq y$, because the mean is being blown up by s_5, and s_y is deliberately excluded from the mean differences.

Theorem 6.2. *Ψ^{CD} is top-1 calibrated.*

Proof. Consider $s \in \mathbb{R}^M$ and WLOG suppose $s \geq 0$, and s_1 is the maximum entry. We have

$$\begin{aligned}\psi(s, 2) - \psi(s, 1) \geq\ &(M-1)(\mathrm{Var}(X) + \mathbb{E}[X^2])\\ &- (M-1)(\mathrm{Var}(Y) - \mathbb{E}[Y^2])\end{aligned}$$

Where X is uniform over $\{s_1, s_3, s_4, \ldots, s_M\}$ and Y is uniform over $\{s_2, s_3, s_4, \ldots, s_M\}$. For the remainder of the proof, we let $n := M - 1$ for brevity. Showing $\psi(s, 2) - \psi(s, 1) > 0$ may be done by showing

$$\begin{aligned}&\mathrm{Var}(X) + \mathbb{E}[X^2] - \mathrm{Var}(Y) - \mathbb{E}[Y^2]\\ &\quad = 2(\mathbb{E}[X^2] - \mathbb{E}[Y^2]) - (\mathbb{E}[X]^2 - \mathbb{E}[Y]^2) > 0.\end{aligned}$$

Letting $m = \sum_{i=2}^{M} s_i$, we have

$$\begin{aligned}
&2(\mathbb{E}[X^2] - \mathbb{E}[Y^2]) - (\mathbb{E}[X]^2 - \mathbb{E}[Y]^2) \\
&= \frac{2(s_1^2 - s_2^2)}{n} - \frac{1}{n^2}[(m + s_1 - s_2)^2 - m^2)] \\
&= \frac{2(s_1^2 - s_2^2)}{n} - \frac{2m(s_1 - s_2) + (s_1 - s_2)^2}{n^2} \\
&= \frac{2(s_1^2 - s_2^2)}{n} - \frac{2(\frac{m}{n} + \frac{s_1 - s_2}{n})(s_1 - s_2)}{2n}
\end{aligned}$$

To complete the proof we just need to show that $\frac{m}{n} + \frac{s_1 - s_2}{n} < s_1 + s_2$. This is equivalent to showing that $m < (n-1)s_1 + (n+1)s_2$. But this is true; $\sum_{i=3}^{M} s_i \leq (M-2)s_1 + ns_2$, as each $s_i < s_1$ and $s_2 \geq 0$. This proves that $\psi(s, 2) - \psi(s, 1) > 0$ if $s_2 < s_1$. □

7. Synthetic Data Experiments

Here we describe experiments comparing an assortment of top-k surrogate loss functions on synthetic data, to see how their behavior compares with reference to the theory. One synthetic experiment empirically showcases the inconsistency of $\psi_1, \psi_2, \psi_3, \psi_4$ and consistency of ψ_5. A second and third experiment flesh out the behavior of the losses in different regimes. we also employ the classic cross entropy loss Ent, and the following truncated cross entropy losses:

$$\begin{aligned}
\mathrm{Ent}_{\mathrm{Tr}_1}(s, y) &= -\ln g(s)_y \\
\mathrm{Ent}_{\mathrm{Tr}_2}(s, y) &= -\ln g(s)_y + \sum_{i=1}^{M} g(s)_i - 1
\end{aligned}$$

with $g(s)_j = \frac{\exp(s_j)}{\exp(s_j) + \sum_{i=k}^{M-1} \exp(s_{\setminus j)_{[i]}}}$. $\mathrm{Ent}_{\mathrm{Tr}_1}$ was proposed in Lapin et al. (2016), and we propose $\mathrm{Ent}_{\mathrm{Tr}_2}$ by restoring the terms dropped from the Bregman Divergence by $\mathrm{Ent}_{\mathrm{Tr}_1}$. Since g is inverse order preserving, by Theorem 3.1 in fact $\mathrm{Ent}_{\mathrm{Tr}_2}$ is top-k calibrated for every k.

We use Pytorch to implement each loss and use them to train on synthetic data. A machine with an Intel Core i7 8th-gen CPU with 16GB of RAM was used.

The first synthetic data experiment we conduct highlights the consistency/inconsistency of the top-k hinge losses. By Proposition 4.2, if the $k+1$ least likely classes altogether have a probability of occurring greater than $\frac{k}{k+1}$, the predictions made by ψ_2, ψ_3, ψ_4 equal a constant vector, and by Theorem 4.1, ψ_1 will assign a value of $c+1$ to the $k-1$ most probable classes and c to the rest. This behavior is inconsistent. On the other hand, ψ_5, which is top-k consistent, will still assign values of $c+1$ to the k most probable classes, and c to the rest.

We construct training data which matches the above setting. The data contains 68 data points with each input data point equal to the zero vector in $\mathbb{R}^2$. Each class in $\{1, 2\}$ is assigned to 10 data points, and each class in $\{3, 4, 5, 6, 7, 8\}$ is assigned to 8 data points. We set $k = 2$ so that $\sum_{i=k+1}^{M} \eta_{[i]} = \frac{48}{68} > \frac{2}{3}$, as described in Proposition 4.2. We train our neural architecture on the data using batch gradient descent, setting the loss of the last layer to be each of $\{\psi_1, \ldots, \psi_5\}$ with $k = 2$. For each classifier obtained, we evaluate the top-2 error on the training set. This is repeated for 100 trials to ensure the robustness of our results.

One may surmise that even if the theoretical minimizers for a loss are not top-k Bayes optimal, they may be effective in practice due to the optimization process. For example, the learned classifier for ψ_2 could output a vector close to 0, but with the first two entries minutely greater than the rest. Interestingly, this is not the case: the returned classifiers for ψ_2, ψ_3, ψ_4 essentially pick randomly amongst the 8 possible classes. The classifier returned by ψ_1 chooses one of $\{0, 1\}$, and randomly picks from the rest of the classes. Finally, the classifier returned by ψ_5 returns the Bayes decision rule, $\{0, 1\}$. These results closely align with the theoretical optima of these losses.

We report average top-2 accuracy over the 100 trials in Table 2. For reference, predicting $\{0, 1\}$ yields a top-2 accuracy of $\frac{20}{68} = 0.294$, predicting one of them gives $\frac{18}{68} = 0.265$, and predicting none of them gives $\frac{16}{68} = 0.235$. Examples of score vectors returned by each loss are in the Appendix. We note that the neural net trained with ψ_5 predicts $\{0, 1\}$ every trial.

Table 2. Results for Top-2 accuracy on the synthetic dataset demonstrating consistency/inconsistency of hinge-like losses. Averaged over 100 trials.

	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
Top-2:	0.2671	0.2515	0.2500	0.2468	0.2941

To investigate a more interesting and realistic example, we also conduct the following synthetic experiment. Given an input N, we randomly sample from a d dimensional Gaussian until we find N vectors which are all at least $c\sqrt{d}$ apart from each other in ℓ_2 distance. Then, we assume there are M classes, where M is a parameter. For each class, we randomly select K of the N means, and then generate a random probability distribution over the K means. Then, we sample L points from the class, by randomly picking a mean according to the probability distribution and sampling from a Gaussian centered there. This models a situation where labels have overlapping distributions.

We set $d = 2, c = 2, K = 5, L = 40$ and vary N in $\{10, 50, 100\}$ to generate the training set. We generate a test set using the same Gaussians and classes with $l = 7$. Results are shown in Table 3, averaged over 10 trials of

generating the data followed by training and evaluation of classifiers on the test set. We optimize with Adam for 500 epochs, using a learning rate of 0.1 and full batch.

Usually, cross entropy dominates other losses in performance. However, in this experiment, due to the overlapping nature of the label distributions, and the function class being restricted to linear predictors, cross entropy actually does notably worse than certain losses which particularly perform well in this scenario – ψ_1, ψ_5, and $\mathrm{Ent}_{\mathrm{Tr}_1}$. This can be viewed as an empirical validation of our results on the linear-restricted inconsistency of cross entropy and consistency of ψ_1 and ψ_5. Furthermore, it light of our discussion of the relationship between convexity and calibration, it is interesting that specifically the nonconvex losses do well in this scenario.

Another interesting phenomenon we observe is that ψ_5 is in a sense robust to its setting of k. While the performance of ψ_1 and $\mathrm{Ent}_{\mathrm{Tr}_1}$ degrade noticeably for top-5 accuracy in the $N = 100$ case, the performance of ψ_5 stays about the same. This is in keeping with ψ_5 being more lenient, not caring as much as long as the top-k error is 0.

Table 3. Results of the second synthetic experiment. Superscript on loss function denotes which k is taken in the loss. We try out both $k = 5$ and $k = 4$ for the $N = 100$ case.

	$N = 10$		$N = 50$	
	Top-5	Acc	Top-5	Acc
Ent	0.699	**0.212**	0.755	**0.267**
ψ_1^5	**0.737**	0.120	0.869	0.134
ψ_2^5	0.639	0.189	0.734	0.245
ψ_3^5	0.649	0.191	0.741	0.241
ψ_4^5	0.651	0.185	0.740	0.205
ψ_5^5	0.726	0.117	**0.880**	0.149
$\mathrm{Ent}^5_{\mathrm{Tr}_1}$	0.711	0.125	0.879	0.118
$\mathrm{Ent}^5_{\mathrm{Tr}_2}$	0.636	0.169	0.656	0.196
	$N = 100$, $k = 5$		$N = 100$, $k = 4$	
	Top-5	Acc	Top-5	Acc
Ent	0.763	**0.242**	0.761	**0.224**
ψ_1^k	**0.896**	0.131	0.834	0.144
ψ_2^k	0.734	0.236	0.721	0.236
ψ_3^k	0.711	0.214	0.722	0.219
ψ_4^k	0.744	0.210	0.720	0.201
ψ_5^k	0.884	0.124	**0.868**	0.123
$\mathrm{Ent}^k_{\mathrm{Tr}_1}$	0.892	0.111	0.857	0.136
$\mathrm{Ent}^k_{\mathrm{Tr}_2}$	0.686	0.169	0.726	0.221

We also model more separated probability distributions. We generate N means as described earlier. For each mean, we sample kl points from the Gaussian centered at the vector with covariance matrix $I \in \mathbb{R}^{d\times d}$. Each set of kl points is divided into k classes of l points each. The top-k error is necessary to achieve 0 error because each Gaussian center spawns k classes that are indistinguishable from each other.

We set $d = 5, c = 2, k = 5, l = 20$ and vary N in $\{10, 50, 100\}$ to generate the training set. We generate a test set using the same Gaussians and classes with $l = 7$. Results are shown in Table 4, averaged over 10 trials of generating the data followed by training and evaluation of classifiers on the test set.

We find that while on this more conventional dataset, Ent dominates, the newly proposed ψ_4, ψ_5 do the best among the other losses.

Table 4. Third set of synthetic experiments, each value averaged over 10 trials. N is the number of Gaussian centers. Superscript on top-k losses indicates the value of k for that loss. Top-5 is top-5 accuracy=$1 - \mathrm{err}_5$, Acc. is accuracy, Δ_1 = test loss - test top-5 error. N/A means not computed due to numerical instability.

	$N = 10$		$N = 50$		$N = 100$	
	Top-5	Acc.	Top-5	Acc.	Top-5	Acc.
Ent	0.932	**0.196**	**0.914**	0.180	**0.888**	**0.183**
ψ_1^5	0.844	0.146	0.720	0.132	0.613	0.126
ψ_2^5	0.918	0.187	0.784	0.179	0.651	0.162
ψ_3^5	0.924	0.192	0.784	0.180	0.640	0.160
ψ_4^5	**0.933**	0.186	0.812	**0.181**	0.661	0.157
ψ_5^5	0.874	0.179	0.801	0.172	0.695	0.146
$\mathrm{Ent}^5_{\mathrm{Tr}_1}$	0.803	0.129	0.815	0.153	0.649	0.127
$\mathrm{Ent}^5_{\mathrm{Tr}_2}$	0.802	0.177	N/A	N/A	N/A	N/A

8. Conclusion

We laid out a theoretical framework for the consistency of surrogate losses used in top-k classification, by defining top-k preserving-ness and top-k calibration.

Our subsequent results on the calibration of losses possessing a form involving Bregman divergences and on the inconsistency of various hinge losses, in constrast to the consistency of a new one we propose, chart some of the consistency landscape of top-k surrogate losses.

We further develop the theory of top-k consistency by exploring a practically relevant extension: consistency restricted to a particular function class. Furthermore, we analyze the relationship of convexity to top-k calibration. With hinge losses, convexity seems antithetical to top-k calibration, and when differentiability is added, top-k calibrated losses are nice, up to a certain limit that is demonstrated via an interesting counterexample.

Future directions include investigating which losses generalize well in the context of top-k classification, as this is the natural and practical progression of the inherent infinite sam-

ple assumption of consistency, and determining consistency when restricted to deep learning function classes.

References

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. Towards good practice in large-scale learning for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36:507–520, 2012.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity , classification , and risk bounds. 2003a.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Large margin classifiers: Convex loss, low noise, and convergence rates. In *NIPS*, 2003b.

Ben-David, S., Eiron, N., and Long, P. M. On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.*, 66(3):496–514, May 2003. ISSN 0022-0000. doi: 10.1016/S0022-0000(03)00038-2. URL http://dx.doi.org/10.1016/S0022-0000(03)00038-2.

Berrada, L., Zisserman, A., and Kumar, M. P. Smooth loss functions for deep top-k classification. *CoRR*, abs/1802.07595, 2018.

Calauzènes, C., Usunier, N., and Gallinari, P. Calibration and regret bounds for order-preserving surrogate losses in learning to rank. *Machine Learning*, 93(2-3):227–260, November 2013. doi: 10.1007/s10994-013-5382-3. URL https://hal.archives-ouvertes.fr/hal-00834230.

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

Gao, W. and Zhou, Z.-H. On the consistency of multi-label learning. *Artif. Intell.*, 199-200(1):22–44, June 2013. ISSN 0004-3702. doi: 10.1016/j.artint.2013.03.001. URL http://dx.doi.org/10.1016/j.artint.2013.03.001.

Koyejo, O. O., Natarajan, N., Ravikumar, P., and Dhillon, I. S. Consistent multilabel classification. In *NIPS*, 2015.

Lapin, M., Hein, M., and Schiele, B. Top-k multiclass svm. In *NIPS*, 2015.

Lapin, M., Hein, M., and Schiele, B. Loss functions for top-k error: Analysis and insights. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1468–1477, 2016.

Lapin, M., Hein, M., and Schiele, B. Analysis and optimization of loss functions for multiclass, top-k, and multilabel classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1533–1554, 2018.

Long, P. and Servedio, R. Consistency versus realizable h-consistency for multiclass classification. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 801–809, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL http://proceedings.mlr.press/v28/long13.html.

Narasimhan, H., Ramaswamy, H., Saha, A., and Agarwal, S. Consistent multiclass algorithms for complex performance measures. In *International Conference on Machine Learning*, pp. 2398–2407, 2015.

Ravikumar, P., Tewari, A., and Yang, E. On ndcg consistency of listwise ranking methods. In *AISTATS*, 2011.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Tewari, A. and Bartlett, P. L. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2005.

Usunier, N., Buffoni, D., and Gallinari, P. Ranking with ordered weighted pairwise classification. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1057–1064, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553509. URL http://doi.acm.org/10.1145/1553374.1553509.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

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

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. 2004b.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:1452–1464, 2018.