Journal of Machine Learning Research 12 (2011) 1501-1536 Submitted 10/10; Revised 2/11; Published 5/11

Learning from Partial Labels

Timothee Cour TIMOTHEE.COUR@GMAIL.COM
NEC Laboratories America
10080 N Wolfe Rd # Sw3350
Cupertino, CA 95014, USA

Benjamin Sapp BENSAPP@CIS.UPENN.EDU
Ben Taskar TASKAR@SEAS.UPENN.EDU
Department of Computer and Information Science
University of Pennsylvania
3330 Walnut Street
Philadelphia, PA 19107, USA

Editor: Yoav Freund

Abstract

We address the problem of partially-labeled multiclass classification, where instead of a single label per instance, the algorithm is given a candidate set of labels, only one of which is correct. Our setting is motivated by a common scenario in many image and video collections, where only partial access to labels is available. The goal is to learn a classifier that can disambiguate the partially-labeled training instances, and generalize to unseen data. We define an intuitive property of the data distribution that sharply characterizes the ability to learn in this setting and show that effective learning is possible even when all the data is only partially labeled. Exploiting this property of the data, we propose a convex learning formulation based on minimization of a loss function appropriate for the partial label setting. We analyze the conditions under which our loss function is asymptotically consistent, as well as its generalization and transductive performance. We apply our framework to identifying faces culled from web news sources and to naming characters in TV series and movies; in particular, we annotated and experimented on a very large video data set and achieve 6% error for character naming on 16 episodes of the TV series *Lost*.

Keywords: weakly supervised learning, multiclass classification, convex learning, generalization bounds, names and faces

1. Introduction

We consider a weakly-supervised multiclass classification setting where each instance is partially labeled: instead of a single label per instance, the algorithm is given a candidate set of labels, only one of which is correct. A typical example arises in photographs containing several faces per image and a caption that only specifies who is in the picture but not which name matches which face. In this setting each face is ambiguously labeled with the set of names extracted from the caption, see Figure 1 (bottom). Photograph collections with captions have motivated much recent interest in weakly annotated images and videos (Duygulu et al., 2002; Barnard et al., 2003; Berg et al., 2004; Gallagher and Chen, 2007). Another motivating example is shown in Figure 1 (top), which shows a setting where we can obtain plentiful but weakly labeled data: videos and screenplays. Using a screenplay, we can tell who is in a given scene, but for every detected face in the scene, the person's

Figure 1: Two examples of partial labeling scenarios for naming faces. **Top:** using a screenplay, we can tell who is in a movie scene, but for every face in the corresponding images, the person's identity is ambiguous (green labels). **Bottom:** images in photograph collections and webpages are often tagged ambiguously with several potential names in the caption or nearby text. In both cases, our goal is to learn a model from ambiguously labeled examples so as to disambiguate the training labels and also generalize to unseen examples.

identity is ambiguous: each face is partially labeled with the set of characters appearing at some point in the scene (Satoh et al., 1999; Everingham et al., 2006; Ramanan et al., 2007). The goal in each case is to learn a person classifier that can not only disambiguate the labels of the training faces, but also generalize to unseen data. Learning accurate models for face and object recognition from such imprecisely annotated images and videos can improve the performance of many applications, including image retrieval and video summarization.

This partially labeled setting is situated between fully supervised and fully unsupervised learning, but is qualitatively different from the semi-supervised setting where both labeled and unlabeled data are available. There have been several papers that addressed this partially labeled (also called ambiguously labeled) problem. Many formulations use the expectation-maximization-like algorithms to estimate the model parameters and "fill-in" the labels (Côme et al., 2008; Ambroise et al., 2001; Vannoorenberghe and Smets, 2005; Jin and Ghahramani, 2002). Most methods involve either non-convex objectives or procedural, iterative reassignment schemes which come without any guarantees of achieving global optima of the objective or classification accuracy. To the best of our knowledge, there has not been theoretical analysis of conditions under which proposed approaches are guaranteed to learn accurate classifiers. The contributions of this paper are:

- We show theoretically that effective learning is possible under reasonable distributional assumptions even when all the data is partially labeled, leading to useful upper and lower bounds on the true error.

- We propose a convex learning formulation based on this analysis by extending general multiclass loss functions to handle partial labels.

- We apply our convex learning formulation to the task of identifying faces culled from web news sources, and to naming characters in TV series. We experiment on a large data set consisting of 100 hours of video, and in particular achieve 6% (resp. 13%) error for character naming across 8 (resp. 32) labels on 16 episodes of *Lost*, consistently outperforming several strong baselines.

- We contribute the *Annotated Faces on TV data set*, which contains about 3,000 cropped faces extracted from 8 episodes of the TV show *Lost* (one face per track). Each face is registered and annotated with a groundtruth label (there are 40 different characters). We also include a subset of those faces with the partial label set automatically extracted from the screenplay.

- We provide the *Convex Learning from Partial Labels Toolbox*, an open-source matlab and C++ implementation of our approach as well as the baseline approach discussed in the paper. The code includes scripts to illustrate the process on Faces in the Wild Data Set (Huang et al., 2007a) and our Annotated Faces on TV data set.

The paper is organized as follows.[1] We review related work and relevant learning scenarios in Section 2. We pose the partially labeled learning problem as minimization of an ambiguous loss in Section 3, and establish upper and lower bounds between the (unobserved) true loss and the (observed) ambiguous loss in terms of a critical distributional property we call ambiguity degree. We propose the novel *Convex Learning from Partial Labels* (CLPL) formulation in Section 4, and show it offers a tighter approximation to the ambiguous loss, compared to a straightforward formulation. We derive generalization bounds for the inductive setting, and in Section 5 also provide bounds for the transductive setting. In addition, we provide reasonable sufficient conditions that will guarantee a consistent labeling in a simple case. We show how to solve proposed CLPL optimization problems by reducing them to more standard supervised optimization problems in Section 6, and provide several concrete algorithms that can be adapted to our setting, such as support vector machines and boosting. We then proceed to a series of controlled experiments in Section 7, comparing CLPL to several baselines on different data sets. We also apply our framework to a naming task in TV series, where screenplay and closed captions provide ambiguous labels. The code and data used in the paper can be found at: `http://www.vision.grasp.upenn.edu/video`.

2. Related Work

We review here the related work for learning under several forms of weak supervision, as well concrete applications.

2.1 Weakly Supervised Learning

To put the partially-labeled learning problem into perspective, it is useful to lay out several related learning scenarios (see Figure 2), ranging from fully supervised (supervised and multi-label learning), to weakly-supervised (semi-supervised, multi-instance, partially-labeled), to unsupervised.

- In **semi-supervised** learning (Zhu and Goldberg, 2009; Chapelle et al., 2006), the learner has access to a set of labeled examples as well as a set of unlabeled examples.

1. A preliminary version of this work appeared in Cour et al. (2009). Sections 4.2 to 6 present new material, and Sections 7 and 8 contain additional experiments, data sets and comparisons.

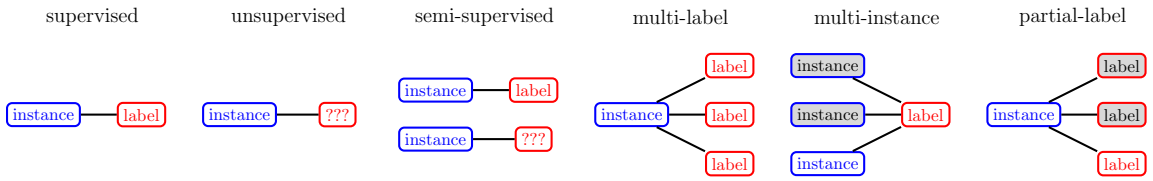

Figure 2: Range of supervision in classification. Training may be: **supervised** (a label is given for each instance), **unsupervised** (no label is given for any instance), **semi-supervised** (labels are given for some instances), **multi-label** (each instance can have multiple labels), **multi-instance** (a label is given for a group of instances where at least one instance in the group has the label), or **partially-labeled** (for each instance, several possible labels are given, only one of which is correct).

- In **multi-label** learning (Boutell et al., 2004; Tsoumakas et al., 2010), each example is assigned multiple labels, all of which can be true.

- In **multi-instance** learning (Dietterich et al., 1997; Andrews and Hofmann, 2004; Viola et al., 2006), examples are not individually labeled but grouped into sets which either contain at least one positive example, or only negative examples. A special case considers the easier scenario where **label proportions** in each bag are known (Kuck and de Freitas, 2005), allowing one to compute convergence bounds on the estimation error of the correct labels (Quadrianto et al., 2009).

- Finally, in our setting of **partially labeled** learning, also called ambiguously labeled learning, each example again is supplied with multiple labels, *only one of which is correct.* A formal definition is given in Section 3.

Clearly, these settings can be combined, for example with multi-instance multi-label learning (MIML) (Zhou and Zhang, 2007), where training instances are associated with not only multiple instances but also multiple labels. Another combination of interest appears in a recent paper building on our previous work (Cour et al., 2009) that addresses the case where sets of instances are ambiguously labeled with candidate labeling sets (Luo and Orabona, 2010).

2.2 Learning From Partially-labeled or Ambiguous Data

There have been several papers that addressed the ambiguous label problem. A number of these use the expectation-maximization algorithm (EM) to estimate the model parameters and the true label (Côme et al., 2008; Ambroise et al., 2001; Vannoorenberghe and Smets, 2005; Jin and Ghahramani, 2002). For example Jin and Ghahramani (2002) use an EM-like algorithm with a discriminative log-linear model to disambiguate correct labels from incorrect ones. Grandvalet and Bengio (2004) add a minimum entropy term to the set of possible label distributions, with a non-convex objective as in the case of (Jin and Ghahramani, 2002). Hullermeier and Beringer (2006) propose several non-parametric, instance-based algorithms for ambiguous learning based on greedy heuristics. These papers only report results on synthetically-created ambiguous labels for data sets such as the UCI repository. Also, the algorithms proposed rely on iterative non-convex learning.

2.3 Images and Captions

A related multi-class setting is common for images with captions: for example, a photograph of a beach with a palm tree and a boat, where object locations are not specified. Duygulu et al. (2002) and Barnard et al. (2003) show that such partial supervision can be sufficient to learn to identify the object locations. The key observation is that while text and images are separately ambiguous, jointly they complement each other. The text, for instance, does not mention obvious appearance properties, but the frequent co-occurrence of a word with a visual element could be an indication of association between the word and a region in the image. Of course, words in the text without correspondences in the image and parts of the image not described in the text are virtually inevitable. The problem of naming image regions can be posed as translation from one language to another. Barnard et al. (2003) address it using a multi-modal extension to mixture of latent Dirichlet allocations.

2.4 Names and Faces

The specific problem of naming faces in images and videos using text sources has been addressed in several works (Satoh et al., 1999; Berg et al., 2004; Gallagher and Chen, 2007; Everingham et al., 2006). There is a vast literature on fully supervised face recognition, which is out of the scope of this paper. Approaches relevant to ours include Berg et al. (2004), which aims at clustering face images obtained by detecting faces from images with captions. Since the name of the depicted people typically appears in the caption, the resulting set of images is ambiguously labeled if more than one name appears in the caption. Moreover, in some cases the correct name may not be included in the set of potential labels for a face. The problem can be solved by using unambiguous images to estimate discriminant coordinates for the entire data set. The images are clustered in this space and the process is iterated. Gallagher and Chen (2007) address the similar problem of retrieval from consumer photo collections, in which several people appear in each image which is labeled with their names. Instead of estimating a prior probability for each individual, the algorithm estimates a prior for groups using the ambiguous labels. Unlike Berg et al. (2004), the method of Gallagher and Chen (2007) does not handle erroneous names in the captions.

2.5 People in Video

In work on video, a wide range of cues has been used to automatically obtain supervised data, including: captions or transcripts (Everingham et al., 2006; Cour et al., 2008; Laptev et al., 2008), sound (Satoh et al., 1999) to obtain the transcript, or clustering based on clothing, face and hair color within scenes to group instances (Ramanan et al., 2007). Most of the methods involve either procedural, iterative reassignment schemes or non-convex optimization.

3. Formulation

In the standard supervised multiclass setting, we have labeled examples $S = \{(x_i, y_i)_{i=1}^m\}$ from an unknown distribution $P(X,Y)$ where $X \in \mathcal{X}$ is the input and $Y \in \{1,\ldots,L\}$ is the class label. In the partially supervised setting we investigate, instead of an unambiguous single label per instance we have a set of labels, one of which is the correct label for the instance. We will denote $\mathbf{y}_i = \{y_i\} \cup \mathbf{z}_i$ as the ambiguity set actually observed by the learning algorithm, where $\mathbf{z}_i \subseteq \{1,\ldots,L\} \setminus \{y_i\}$ is a set of additional labels, and y_i the latent groundtruth label which we would like to recover. Throughout the paper, we will use boldface to denote sets and uppercase to denote random variables

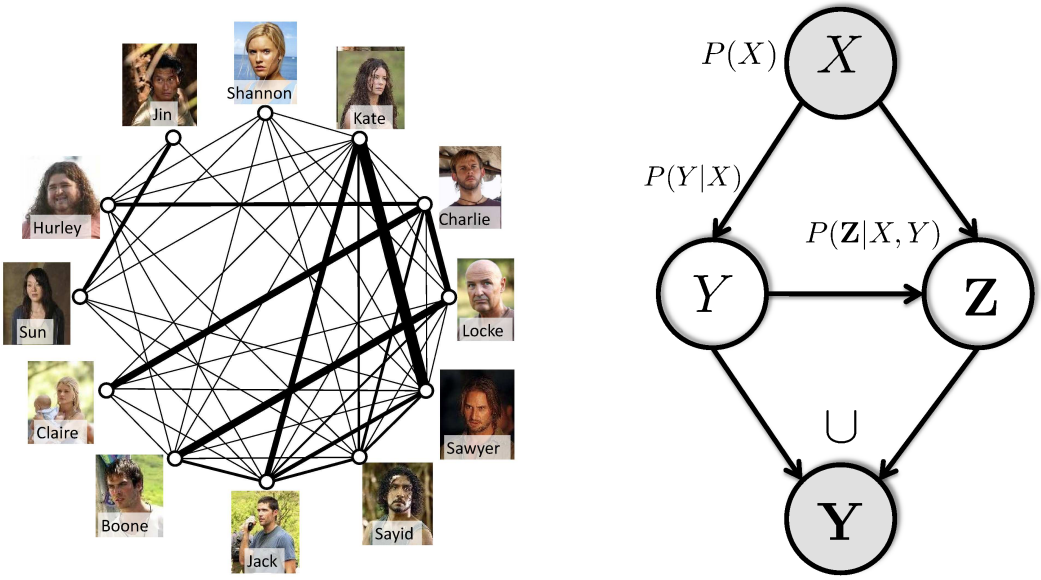

Figure 3: **Left**: Co-occurrence graph of the top characters across 16 episodes of *Lost*. Edge thickness corresponds to the co-occurrence frequency of characters. **Right**: The model of the data generation process: $(X,\mathbf{Y})$ are observed, $(Y,\mathbf{Z})$ are hidden, with $\mathbf{Y} = Y \cup \mathbf{Z}$.

with corresponding lowercase values of random variables. We suppose $X,Y,\mathbf{Z}$ are distributed according to an (unknown) distribution $P(X,Y,\mathbf{Z}) = P(X)P(Y \mid X)P(\mathbf{Z} \mid X,Y)$ (see Figure 3, right), of which we only observe samples of the form $S = \{(x_i,\mathbf{y}_i)_{i=1}^m\} = \{(x_i,\{y_i\} \cup \mathbf{z}_i)_{i=1}^m\}$. (In case X is continuous, $P(X)$ is a density with respect to some underlying measure μ on $\mathcal{X}$, but we will simply refer to the joint $P(X,Y,\mathbf{Z})$ as a distribution.) With the above definitions, $y_i \in \mathbf{y}_i, \mathbf{z}_i \subset \mathbf{y}_i, y_i \notin \mathbf{z}_i$ and $Y \in \mathbf{Y}, \mathbf{Z} \subset \mathbf{Y}, Y \notin \mathbf{Z}$.

Clearly, our setup generalizes the standard semi-supervised setting where some examples are labeled and some are unlabeled: an example is labeled when the corresponding ambiguity set $\mathbf{y}_i$ is a singleton, and unlabeled when $\mathbf{y}_i$ includes all the labels. However, we do not explicitly consider the semi-supervised setting this paper, and our analysis below provides essentially vacuous bounds for the semi-supervised case. Instead, we consider the middle-ground, where all examples are partially labeled as described in our motivating examples and analyze assumptions under which learning can be guaranteed to succeed.

In order to learn from ambiguous data, we must make some assumptions about the distribution $P(\mathbf{Z} \mid X,Y)$. Consider a very simple ambiguity pattern that makes accurate learning impossible: $L = 3$, $|\mathbf{z}_i| = 1$ and label 1 is present in every set $\mathbf{y}_i$, for all i. Then we cannot distinguish between the case where 1 is the true label of every example, and the case where it is not a label of any example. More generally, if two labels always co-occur when present in $\mathbf{y}$, we cannot tell them apart. In order to disallow this case, below we will make an assumption on the distribution $P(\mathbf{Z} \mid X,Y)$ that ensures some diversity in the ambiguity set. This assumption is often satisfied in practice. For example, consider our initial motivation of naming characters in TV shows, where the ambiguity set for any given detected face in a scene is given by the set of characters occurring at some point in that scene. In Figure 3 (left), we show the co-occurrence graph of characters in a season of the TV show *Lost*,

Symbol	Meaning
x, X	observed input value/variable: $x, X \in \mathcal{X}$
y, Y	hidden label value/variable: $y, Y \in \{1, \ldots, L\}$
$\mathbf{z}, \mathbf{Z}$	hidden additional label set/variable: $\mathbf{z}, \mathbf{Z} \subseteq \{1, \ldots, L\}$
$\mathbf{y}, \mathbf{Y}$	observed label set/variable: $\mathbf{y} = \{y\} \cup \mathbf{z}, \mathbf{Y} = \{Y\} \cup \mathbf{Z}$
$h(x), h(X)$	multiclass classifier mapping $h : \mathcal{X} \mapsto \{1, \ldots, L\}$
$\mathcal{L}(h(x), y), \mathcal{L}_A(h(x), \mathbf{y})$	standard and partial $0/1$ loss

Table 1: Summary of notation used.

where the thickness of the edges corresponds to the number of times characters share a scene. This suggests that for most characters, ambiguity sets are diverse and we can expect that the ambiguity degree is small. A more quantitative diagram will be given in Figure 11 (left).

Many formulations of fully-supervised multiclass learning have been proposed based on minimization of convex upper bounds on risk, usually, the expected $0/1$ loss (Zhang, 2004):

$$0/1 \textbf{ loss:} \quad \mathcal{L}(h(x), y) = \mathbb{1}(h(x) \neq y),$$

where $h(x) : \mathcal{X} \mapsto \{1, \ldots, L\}$ is a multiclass classifier.

We cannot evaluate the $0/1$ loss using our partially labeled training data. We define a surrogate loss which we can evaluate, and we call ambiguous or partial $0/1$ loss (where A stands for ambiguous):

$$\textbf{Partial } 0/1 \textbf{ loss:} \quad \mathcal{L}_A(h(x), \mathbf{y}) = \mathbb{1}(h(x) \notin \mathbf{y}).$$

3.1 Connection Between Partial and Standard $0/1$ Losses

An obvious observation is that the partial loss is an underestimate of the true loss. However, in the ambiguous learning setting we would like to minimize the true $0/1$ loss, with access only to the partial loss. Therefore we need a way to upper-bound the $0/1$ loss using the partial loss. We first introduce a measure of hardness of learning under ambiguous supervision, which we define as ambiguity degree ε of a distribution $P(X, Y, \mathbf{Z})$:

$$\textbf{Ambiguity degree:} \quad \varepsilon = \sup_{x,y,z:P(x,y)>0, z \in \{1,\ldots,L\}} P(z \in \mathbf{Z} \mid X = x, Y = y). \tag{1}$$

In words, ε corresponds to the maximum probability of an extra label z co-occurring with a true label y, over all labels and inputs. Let us consider several extreme cases: When $\varepsilon = 0$, $\mathbf{Z} = \emptyset$ with probability one, and we are back to the standard supervised learning case, with no ambiguity. When $\varepsilon = 1$, some extra label always co-occurs with a true label y on an example x and we cannot tell them apart: no learning is possible for this example. For a fixed ambiguity set size C (i.e., $P(|\mathbf{Z}| = C) = 1$), the smallest possible ambiguity degree is $\varepsilon = C/(L-1)$, achieved for the case where $P(\mathbf{Z} \mid X, Y)$ is uniform over subsets of size C, for which we have $P(z \in \mathbf{Z} \mid X, Y) = C/(L-1)$ for all $z \in \{1, \ldots, L\} \setminus \{y\}$. Intuitively, the best case scenario for ambiguous learning corresponds to a distribution with high conditional entropy for $P(\mathbf{Z} \mid X, Y)$.

The following proposition shows we can bound the (unobserved) $0/1$ loss by the (observed) partial loss, allowing us to approximately minimize the standard loss with access only to the partial one. The tightness of the approximation directly relates to the ambiguity degree.

Proposition 1 (Partial loss bound via ambiguity degree ε) *For any classifier h and distribution $P(X,Y,\mathbf{Z})$, with $\mathbf{Y} = X \cup \mathbf{Z}$ and ambiguity degree ε:*

$$\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})] \leq \mathbb{E}_P[\mathcal{L}(h(X),Y)] \leq \frac{1}{1-\varepsilon}\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})],$$

with the convention $1/0 = +\infty$. These bounds are tight, and for the second one, for any (rational) ε, we can find a number of labels L, a distribution P and classifier h such that equality holds.

Proof. All proofs appear in Appendix B. ■

3.2 Robustness to Outliers

One potential issue with Proposition 1 is that unlikely (outlier) pairs x,y (with vanishing $P(x,y)$) might force ε to be close to 1, making the bound very loose. We show we can refine the notion of ambiguity degree ε by excluding such pairs.

Definition 2 (ε,δ)***-ambiguous distribution.*** *A distribution $P(X,Y,\mathbf{Z})$ is (ε,δ)-ambiguous if there exists a subset G of the support of $P(X,Y)$, $G \subseteq \mathcal{X} \times \{1,\ldots,L\}$ with probability mass at least $1-\delta$, that is, $\int_{(x,y)\in G} P(X=x,Y=y)d\mu(x,y) \geq 1-\delta$, integrated with respect to the appropriate underlying measure μ on $\mathcal{X} \times \{1,\ldots,L\}$, for which*

$$\sup_{(x,y)\in G, z\in\{1,\ldots,L\}} P(z \in \mathbf{Z} \mid X=x, Y=y) \leq \varepsilon.$$

Note that in the extreme case $\varepsilon = 0$ corresponds to standard semi-supervised learning, where $1-\delta$-proportion of examples are unambiguously labeled, and δ are (potentially) fully unlabeled. Even though we can accommodate it, semi-supervised learning is not our focus in this paper and our bounds are not well suited for this case.

This definition allows us to bound the $0/1$ loss even in the case when some unlikely set of pairs x,y with probability $\leq \delta$ would make the ambiguity degree large. Suppose we mix an initial distribution with small ambiguity degree, with an outlier distribution with large overall ambiguity degree. The following proposition shows that the bound degrades only by an additive amount, which can be interpreted as a form of robustness to outliers.

Proposition 3 (Partial loss bound via (ε,δ)) *For any classifier h and (ε,δ)-ambiguous $P(\mathbf{Z} \mid X,Y)$,*

$$\mathbb{E}_P[\mathcal{L}(h(X),Y)] \leq \frac{1}{1-\varepsilon}\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})] + \delta.$$

A visualization of the bounds in Proposition 1 and Proposition 3 is shown in Figure 4.

3.3 Label-specific Recall Bounds

In the types of data from video experiments, we observe that certain subsets of labels are harder to disambiguate than others. We can further tighten our bounds between ambiguous loss and standard

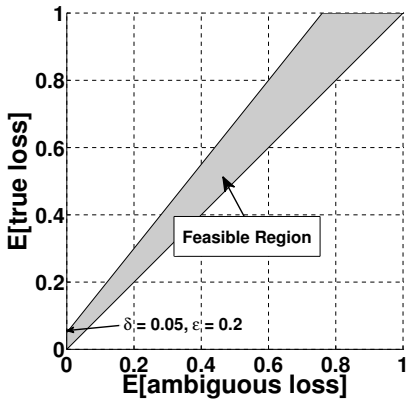

Figure 4: Feasible region for expected ambiguous and true loss, for $\varepsilon = 0.2, \delta = 0.05$.

$0/1$ loss if we consider label-specific information. We define the *label-specific ambiguity degree* ε_a of a distribution (with $a \in \{1,\ldots,L\}$) as:

$$\varepsilon_a = \sup_{x,z:P(X=x,Y=a)>0;z\in\{1,\ldots,L\}} P(z \in \mathbf{Z} \mid X = x, Y = a).$$

We can show a label-specific analog of Proposition 1:

Proposition 4 (Label-specific partial loss bound) *For any classifier h and distribution $P(X,Y,\mathbf{Z})$ with* label-specific ambiguity degree ε_a *,*

$$\mathbb{E}_P[\mathcal{L}(h(X),Y) \mid Y = a] \leq \frac{1}{1-\varepsilon_a}\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y}) \mid Y = a],$$

where we see that ε_a bounds per-class recall.

These bounds give a strong connection between ambiguous loss and real loss when ε is small. This assumption allows us to approximately minimize the expected real loss by minimizing (an upper bound on) the ambiguous loss, as we propose in the following section.

4. A Convex Learning Formulation

We have not assumed any specific form for our classifier $h(x)$ above. We now focus on a particular family of classifiers, which assigns a score $g_a(x)$ to each label a for a given input x and select the highest scoring label:

$$h(x) = \arg\max_{a\in 1..L} g_a(x).$$

We assume that ties are broken arbitrarily, for example, by selecting the label with smallest index a. We define the vector $g(x) = [g_1(x)\ldots g_L(x)]^\top$, with each component $g_a : \mathcal{X} \mapsto \mathbb{R}$ in a function class $\mathcal{G}$. Below, we use a multi-linear function class $\mathcal{G}$ by assuming a feature mapping $\mathbf{f}(x) : \mathcal{X} \mapsto \mathbb{R}^d$ from inputs to d real-valued features and let $g_a(x) = \mathbf{w}_a \cdot \mathbf{f}(x)$, where $\mathbf{w}_a \in \mathbb{R}^d$ is a weight vector for each class, bounded by some norm: $||\mathbf{w}_a||_p \leq B$ for $p = 1,2$.

We build our learning formulation on a simple and general multiclass scheme, frequently used for the fully supervised setting (Crammer and Singer, 2002; Rifkin and Klautau, 2004; Zhang, 2004; Tewari and Bartlett, 2005), that combines convex binary losses $\psi(\cdot) : \mathbb{R} \mapsto \mathbb{R}_+$ on individual

components of g to create a multiclass loss. For example, we can use hinge, exponential or logistic loss. In particular, we assume a type of one-against-all scheme for the supervised case:

$$\mathcal{L}_\psi(g(x),y) = \psi(g_y(x)) + \sum_{a \neq y} \psi(-g_a(x)).$$

A classifier $h_g(x)$ is selected by minimizing the empirical loss $\mathcal{L}_\psi$ on the sample $S = \{x_i, y_i\}_{i=1}^m$ (called empirical ψ-risk) over the function class $\mathcal{G}$:

$$\inf_{g \in \mathcal{G}} \mathbb{E}_S[\mathcal{L}_\psi(g(X),Y)] = \inf_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^m \mathcal{L}_\psi(g(x_i),y_i).$$

For the fully supervised case, under appropriate assumptions, this form of the multiclass loss is infinite-sample consistent. This means that a minimizer $\hat{g}$ of ψ-risk achieves optimal $0/1$ risk $\inf_g \mathbb{E}_S[\mathcal{L}_\psi(g(X),Y)] = \inf_g \mathbb{E}_P[\mathcal{L}(g(X),Y)]$ as the number of samples m grows to infinity, provided that the function class $\mathcal{G}$ grows appropriately fast with m to be able to approximate any function from $\mathcal{X}$ to $\mathbb{R}$ and $\psi(u)$ satisfies the following conditions: (1) $\psi(u)$ is convex, (2) bounded below, (3) differentiable and (4) $\psi(u) < \psi(-u)$ when $u > 0$ (Theorem 9 in Zhang (2004)). These conditions are satisfied, for example, for the exponential, logistic and squared hinge loss $\max(0, 1-u)^2$. Below, we construct a loss function for the partially labeled case and consider when the proposed loss is consistent.

4.1 Convex Loss for Partial Labels

In the partially labeled setting, instead of an unambiguous single label y per instance we have a set of labels Y, one of which is the correct label for the instance. We propose the following loss, which we call our *Convex Loss for Partial Labels* (CLPL):

$$\mathcal{L}_\psi(g(x),\mathbf{y}) = \psi\left(\frac{1}{|\mathbf{y}|}\sum_{a \in \mathbf{y}} g_a(x)\right) + \sum_{a \notin \mathbf{y}} \psi(-g_a(x)). \tag{2}$$

Note that if $\mathbf{y}$ is a singleton, the CLPL function reduces to the regular multiclass loss. Otherwise, CLPL will drive up the *average* of the scores of the labels in $\mathbf{y}$. If the score of the correct label is large enough, the other labels in the set do not need to be positive. This tendency alone does not guarantee that the correct label has the *highest* score. However, we show in Proposition 6 that $\mathcal{L}_\psi(g(x),\mathbf{y})$ upperbounds $\mathcal{L}_A(g(x),\mathbf{y})$ whenever $\psi(\cdot)$ is an upper bound on the $0/1$ loss.

Of course, minimizing an upperbound on the loss does not always lead to sensible algorithms. We show next that our loss function is consistent under certain assumptions and offers a tighter upperbound to the ambiguous loss compared to a more straightforward multi-label approach.

4.2 Consistency for Partial Labels

We derive conditions under which the minimizer of the CLPL in Equation 2 with partial labels achieves optimal $0/1$ risk: $\inf_{g \in \mathcal{G}} \mathbb{E}_S[\mathcal{L}_\psi(g(X),\mathbf{Y})] = \inf_{g \in \mathcal{G}} \mathbb{E}_P[\mathcal{L}(g(X),Y)]$ in the limit of infinite data and arbitrarily rich $\mathcal{G}$. Not surprisingly, our loss function is not consistent without making some additional assumptions on $P(\mathbf{Y} \mid X)$ beyond the assumptions for the fully supervised case. Note that the Bayes optimal classifier for $0/1$ loss satisfies the condition $h(x) \in \arg\max_a P(Y = a \mid X = x)$, and

may not be unique. First, we require that $\arg\max_a P(Y=a \mid X=x) = \arg\max_a P(a \in \mathbf{Y} \mid X=x)$, since otherwise $\arg\max_a P(Y=a \mid X=x)$ cannot be determined by any algorithm from partial labels $\mathbf{Y}$ without additional information even with an infinite amount of data. Second, we require a simple dominance condition as detailed below and provide a counterexample when this condition does not hold. The dominance relation defined formally below states that when a is the most (or one of the most) likely label given x according to $P(\mathbf{Y} \mid X=x)$ and b is not, $\mathbf{c} \cup \{a\}$ has higher (or equal) probability than $\mathbf{c} \cup \{b\}$ for any set of other labels $\mathbf{c}$.

Proposition 5 (Partial label consistency) *Suppose the following conditions hold:*

- $\psi(\cdot)$ *is differentiable, convex, lower-bounded and non-increasing, with* $\psi'(0) < 0$.
- *When* $P(X=x) > 0$, $\arg\max_{a'} P(Y=a' \mid X=x) = \arg\max_{a'} P(a' \in \mathbf{Y} \mid X=x)$.
- *The following dominance relation holds:* $\forall a \in \arg\max_{a'} P(a' \in \mathbf{Y} \mid X=x),\ \ \forall b \notin \arg\max_{a'} P(a' \in \mathbf{Y} \mid X=x),\ \ \forall \mathbf{c} \subset \{1,\ldots,L\} \backslash \{a,b\}$*:*

$$P(\mathbf{Y} = \mathbf{c} \cup \{a\} \mid X=x) \geq P(\mathbf{Y} = \mathbf{c} \cup \{b\} \mid X=x).$$

Then $\mathcal{L}_\psi(g(x), \mathbf{y})$ *is infinite-sample consistent:*

$$\inf_{g \in \mathcal{G}} \mathbb{E}_S[\mathcal{L}_\psi(g(X), \mathbf{Y})] = \inf_{g \in \mathcal{G}} \mathbb{E}_P[\mathcal{L}(g(X), Y)],$$

as $|S| = m \to \infty$ *and* $\mathcal{G} \to \mathbb{R}^L$ *. As a corollary, consistency is implied when ambiguity degree* $\varepsilon < 1$ *and* $P(Y \mid X)$ *is deterministic, that is,* $P(Y \mid X) = \mathbb{1}(Y = h(X))$ *for some* $h(\cdot)$.

If the dominance relation does not hold, we can find counter-examples where consistency fails. Consider a distribution with a single x with $P(x) > 0$, and let $L = 4$, $P(|\mathbf{Y}| = 2 \mid X = x) = 1$, ψ be the square-hinge loss, and $P(\mathbf{Y} \mid X = x)$ be such that:

		a			
	$250 \cdot P_{ab}$	1	2	3	4
b	1	0	29	44	**0**
	2	29	0	17	**26**
	3	44	17	0	9
	4	**0**	**26**	9	0
	$250 \cdot P_a$	73	72	70	35

Above, the abbreviations are $P_{ab} = P(\mathbf{Y} = \{a,b\} \mid X = x)$ and $P_a = \sum_b P_{ab}$, and the entries that do not satisfy the dominance relation are in bold. We can explicitly compute the minimizer of $\mathcal{L}_\psi$, which is $g = (\frac{1}{2}P_{ab} + \mathrm{diag}(2 - \frac{3}{2}P_a))^{-1}(3P_a - 2) \approx -\begin{bmatrix} 0.6572 & 0.6571 & 0.6736 & 0.8568 \end{bmatrix}$. It satisifes $\arg\max_a g_a = 2$ but $\arg\max_a \sum_b P_{ab} = 1$.

4.3 Comparison to Other Loss Functions

The "naive" partial loss, proposed by Jin and Ghahramani (2002), treats each example as having multiple correct labels, which implies the following loss function

$$\mathcal{L}_\psi^{naive}(g(x), \mathbf{y}) = \frac{1}{|\mathbf{y}|} \sum_{a \in \mathbf{y}} \psi(g_a(x)) + \sum_{a \notin \mathbf{y}} \psi(-g_a(x)). \tag{3}$$

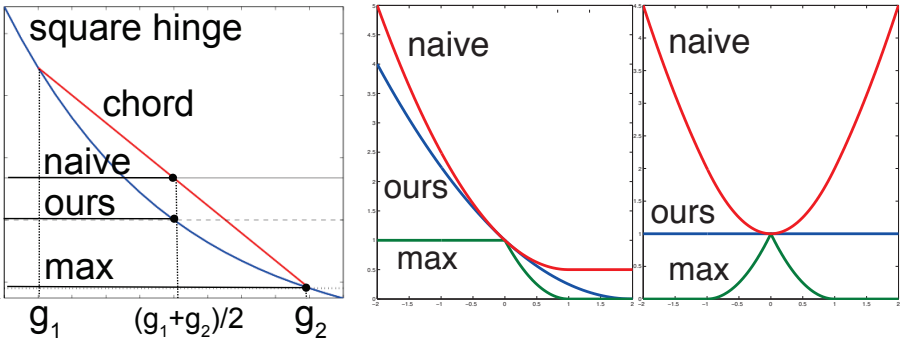

Figure 5: Our loss function in Equation 2 provides a tighter convex upperbound than the naive loss Equation 3 on the non-convex max-loss Equation 4. (**Left**) We show the square hinge ψ (blue) and a chord (red) touching two points g_1, g_2. The horizontal lines correspond to our loss $\psi(\frac{1}{2}(g_1+g_2))$ Equation 2, the max-loss $\psi(\max(g_1,g_2))$, and the naive loss $\frac{1}{2}(\psi(g_1)+\psi(g_2))$ (ignoring negative terms and assuming $\mathbf{y}=\{1,2\}$). (**Middle**) Corresponding losses as we vary $g_1 \in [-2,2]$ (with $g_2=0$). (**Right**) Same, with $g_2=-g_1$.

One reason we expect our loss function to outperform the naive approach is that we obtain a tighter convex upper bound on $\mathcal{L}_A$. Let us also define

$$\mathcal{L}_\psi^{max}(g(x),\mathbf{y}) = \psi\left(\max_{a\in\mathbf{y}} g_a(x)\right) + \sum_{a\notin\mathbf{y}} \psi(-g_a(x)), \tag{4}$$

which is not convex, but is in some sense closer to the desired true loss. The following inequalities are verified for common losses ψ such as square hinge loss, exponential loss, and log loss with proper scaling:

Proposition 6 (Comparison between partial losses) *Under the usual conditions that ψ is a convex, decreasing upper bound of the step function, the following inequalities hold:*

$$2\mathcal{L}_A \leq \mathcal{L}_\psi^{max} \leq \mathcal{L}_\psi \leq \mathcal{L}_\psi^{naive}.$$

The 2^{nd} and 3^{rd} bounds are tight, and the first one is tight provided $\psi(0)=1$ and $\lim_{+\infty}\psi=0$.

This shows that our CLPL $\mathcal{L}_\psi$ is a tighter approximation to $\mathcal{L}_A$ than $\mathcal{L}_\psi^{naive}$, as illustrated in Figure 5. To gain additional intuition as to why CLPL is better than the naive loss Equation 3: for an input x with ambiguous label set (a,b), CLPL only encourages the *average* of $g_a(x)$ and $g_b(x)$ to be large, allowing the correct score to be positive and the extraneous score to be negative (e.g., $g_a(x)=2, g_b(x)=-1$). In contrast, the naive model encourages both $g_a(x)$ and $g_b(x)$ to be large.

4.4 Generalization Bounds

To derive concrete generalization bounds on multiclass error for CLPL we define our function class for g. We assume a feature mapping $\mathbf{f}(x): \mathcal{X} \mapsto \mathbb{R}^d$ from inputs to d real-valued features and let $g_a(x) = \mathbf{w}_a \cdot \mathbf{f}(x)$, where $\mathbf{w}_a \in \mathbb{R}^d$ is a weight vector for each class, bounded by L_2 norm : $||\mathbf{w}_a||_2 \leq B$.

We use $\psi(u) = \max(0, 1-u)^p$ (for example hinge loss with $p = 1$, squared hinge loss with $p = 2$). The corresponding margin-based loss is defined via a truncated, rescaled version of ψ:

$$\psi_\gamma(u) = \begin{cases} 1 & \text{if } u \leq 0, \\ (1-u/\gamma)^p & \text{if } 0 < u \leq \gamma, \\ 0 & \text{if } u > \gamma. \end{cases}$$

Proposition 7 (Generalization bound) *For any integer m and any $\eta \in (0,1)$, with probability at least $1-\eta$ over samples $S = \{(x_i, \mathbf{y}_i)\}_{i=1}^m$, for every g in $\mathcal{G}$:*

$$\mathbb{E}_P[\mathcal{L}_A(g(X), \mathbf{Y})] \leq \mathbb{E}_S[\mathcal{L}_{\psi_\gamma}(g(X), \mathbf{Y})] + \frac{4pBL^{5/2}}{c\gamma}\sqrt{\frac{\mathbb{E}_S[||\mathbf{f}(X)||^2]}{m}} + L\sqrt{\frac{8\log(2/\eta)}{m}}.$$

where c is an absolute constant from Lemma 12 in the appendix, $\mathbb{E}_S$ is the sample average and L is the number of labels.

The proof in the appendix uses definition 11 for Rademacher and Gaussian complexity, Lemma 12, Theorem 13 and Theorem 14 from Bartlett and Mendelson (2002), reproduced in the appendix and adapted to our notations for completeness. Using Proposition 7 and Proposition 1, we can derive the following bounds on the true expected $0/1$ loss $\mathbb{E}_P[\mathcal{L}(g(X), Y)]$ from purely ambiguous data:

Proposition 8 (Generalization bounds on true loss) *For any distribution ε-ambiguous distribution P, integer m and $\eta \in (0,1)$, with probability at least $1-\eta$ over samples $S = \{(x_i, \mathbf{y}_i)\}_{i=1}^m$, for every $g \in \mathcal{G}$:*

$$\mathbb{E}_P[\mathcal{L}(g(X), Y)] \leq \frac{1}{1-\varepsilon}\left(\mathbb{E}_S[\mathcal{L}_{\psi_\gamma}(g(X), \mathbf{Y})] + \frac{4pBL^{5/2}}{c\gamma}\sqrt{\frac{\mathbb{E}_S[||\mathbf{f}(X)||^2]}{m}} + L\sqrt{\frac{8\log\frac{2}{\eta}}{m}}\right).$$

5. Transductive Analysis

We now turn to the analysis of our *Convex Loss for Partial Labels* (CLPL) in the transductive setting. We show guarantees on disambiguating the labels of instances under fairly reasonable assumptions.

Example 1 *Consider a data set S of two points, x, x', with label sets $\{1,2\}, \{1,3\}$, respectively and suppose that the total number of labels is 3. The objective function is given by:*

$$\psi(\frac{1}{2}(g_1(x) + g_2(x))) + \psi(-g_3(x)) + \psi(\frac{1}{2}(g_1(x') + g_3(x'))) + \psi(-g_2(x')).$$

Suppose the correct labels are $(1,1)$. It is clear that without further assumptions about x and x' we cannot assume that the minimizer of the loss above will predict the right label. However, if $\mathbf{f}(x)$ and $\mathbf{f}(x')$ are close, it should be intuitively clear that we should be able to deduce the label of the two examples is 1.

A natural question is under what conditions on the data will CLPL produce a labeling that is consistent with groundtruth. We provide an analysis under several assumptions.

5.1 Definitions

In the remainder of this section, we denote $y(x)$ (resp. $\mathbf{y}(x)$) as the true label (resp. ambiguous label set) of some $x \in S$, and $\mathbf{z}(x) = \mathbf{y}(x) \backslash \{y(x)\}$. $||\cdot||$ denotes an arbitrary norm, with $||\cdot||^*$ its dual norm. As above, ψ denotes a decreasing upper bound on the step function and g a classifier satisfying: $\forall a,\ ||\mathbf{w}_a||^* \leq 1$ (we can easily generalize the remaining propositions to the case where g_a is 1-Lipschitz and $\mathbf{f}$ is the identity). For $x \in S$ and $\eta > 0$, we define $B_\eta(x)$ as the set of neighbors of x that have the same label as x:

$$B_\eta(x) = \{x' \in S \backslash \{x\} : ||f(x') - f(x)|| < \eta, y(x') = y(x)\}.$$

Lemma 9 *Let $x \in S$. If $\mathcal{L}_\psi(g(x), \mathbf{y}(x)) \leq \psi(\eta/2)$ and $\forall a \in \mathbf{z}(x), \exists x' \in B_\eta(x)$ such that $g_a(x') \leq -\eta/2$, then g predicts the correct label for x.*

In other words, g predicts the correct label for x when its loss is sufficiently small, and for each of its ambiguous labels a, we can find a neighbor with same label whose score $g_a(x')$ is small enough. Note that this does not make any assumption on the *nearest* neighbors of x.

Corollary 10 *Let $x \in S$. Suppose $\exists q \geq 0$, $x_1 \ldots x_q \in B_\eta(x)$ such that $\cap_{i=0..q} \mathbf{z}(x_i) = \emptyset$, $\max_{i=0..q} \mathcal{L}_\psi(g(x_i), \mathbf{y}(x_i)) \leq \psi(\eta/2)$ (with $x_0 := x$). Then g predicts the correct label for x.*

In other words, g predicts the correct label for x if we can find a set of neighbors of the same label with small enough loss, and without any common extra label. This simple condition often arises in our experiments.

6. Algorithms

Our formulation is quite flexible and we can derive many alternative algorithms depending on the choice of the binary loss $\psi(u)$, the regularization, and the optimization method. We can minimize Equation 2 using off-the-shelf binary classification solvers. To do this, we rewrite the two types of terms in Equation 2 as linear combinations of $m \cdot L$ feature vectors. We stack the parameters and features into one vector as follows below, so that $g_a(x) = \mathbf{w}_a \cdot \mathbf{f}(x) = \mathbf{w} \cdot \mathbf{f}(x, a)$:

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_1 \\ \ldots \\ \mathbf{w}_L \end{pmatrix};\ \mathbf{f}(x, a) = \begin{pmatrix} \mathbb{1}(a = 1)\mathbf{f}(x) \\ \ldots \\ \mathbb{1}(a = L)\mathbf{f}(x) \end{pmatrix}.$$

We also define $\mathbf{f}(x, \mathbf{y})$ to be the average feature vector of the labels in the set $\mathbf{y}$:

$$\mathbf{f}(x, \mathbf{y}) = \frac{1}{|\mathbf{y}|} \sum_{a \in \mathbf{y}} \mathbf{f}(x, a).$$

With these definitions, we have:

$$\mathcal{L}_\psi(g(x), \mathbf{y}) = \psi(\mathbf{w} \cdot \mathbf{f}(x, \mathbf{y})) + \sum_{a \notin \mathbf{y}} \psi(-\mathbf{w} \cdot \mathbf{f}(x, a)).$$

Then to use a binary classification method to solve CLPL optimization, we simply transform the m partially labelled training examples $S = \{x_i, \mathbf{y}_i\}_{i=1}^m$ into m positive examples $S_+ = \{\mathbf{f}(x_i, \mathbf{y}_i)\}_{i=1}^m$

and $\sum_i L - |\mathbf{y}_i|$ negative examples $S_- = \{\mathbf{f}(x_i, a)\}_{i=1,a\notin\mathbf{y}_i}^{m}$. Note that the increase in dimension of the features by a factor of L does not significantly affect the running time of most methods since the vectors are sparse. We use the off-the-shelf implementation of binary SVM with squared hinge (Fan et al., 2008) in most of our experiments, where the objective is:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_i \max(0, 1 - \mathbf{w}\cdot\mathbf{f}(x_i,\mathbf{y}_i))^2 + C\sum_{i,a\notin\mathbf{y}_i} \max(0, 1 + \mathbf{w}\cdot\mathbf{f}(x_i,a))^2.$$

Using hinge loss and L_1 regularization lead to a linear programming formulation, and using L_1 with exponential loss leads naturally to a boosting algorithm. We present (and experiment with) a boosting variant of the algorithm, allowing efficient feature selection, as described in Appendix A. We can also consider the case where the regularization is L_2 and $\mathbf{f}(x) : \mathcal{X} \mapsto \mathbb{R}^d$ is a nonlinear mapping to a high, possibly infinite dimensional space using kernels. In that case, it is simple to show that

$$\mathbf{w} = \sum_i \alpha_i \mathbf{f}(x_i,\mathbf{y}_i) - \sum_{i,a\notin\mathbf{y}_i} \alpha_{i,a}\mathbf{f}(x_i,a),$$

for some set of non-negative α's, where α_i corresponds to the positive example $\mathbf{f}(x_i,\mathbf{y}_i)$, and $\alpha_{i,a}$ corresponds to the negative example $\mathbf{f}(x_i,a)$, for $a \notin \mathbf{y}_i$. Letting $K(x,x') = \mathbf{f}(x)\cdot\mathbf{f}(x')$ be the kernel function, note that $\mathbf{f}(x,a)\cdot\mathbf{f}(x',b) = \mathbb{1}(a=b)K(x,x')$. Hence, we have:

$$\mathbf{w}\cdot\mathbf{f}(x,b) = \sum_{i,a\in\mathbf{y}_i} \frac{\alpha_i}{|\mathbf{y}_i|}\mathbb{1}(a=b)K(x_i,x) - \sum_{i,a\notin\mathbf{y}_i} \alpha_{i,a}\mathbb{1}(a=b)K(x_i,x).$$

This transformation allows us to use kernels with standard off-the-shelf binary SVM implementations.

7. Controlled Partial Labeling Experiments

We first perform a series of controlled experiments to analyze our *Convex Learning from Partial Labels* (CLPL) framework on several data sets, including standard benchmarks from the **UCI repository** (Asuncion and Newman, 2007), a **speaker identification** task from audio extracted from movies, and a **face naming task** from Labeled Faces in the Wild (Huang et al., 2007b). In Section 8 we also consider the challenging task of **naming characters in TV shows** throughout an entire season. In each case the goal is to correctly label faces/speech segments/instances from examples that have multiple potential labels (transductive case), as well as learn a model that can generalize to other unlabeled examples (inductive case).

We analyze the effect on learning of the following factors: distribution of ambiguous labels, size of ambiguous bags, proportion of instances which contain an ambiguous bag, entropy of the ambiguity, distribution of true labels and number of distinct labels. We compare our CLPL approach against a number of **baselines**, including a generative model, a discriminative maximum-entropy model, a naive model, two K-nearest neighbor models, as well as models that ignore the ambiguous bags. We also propose and compare several variations on our cost function. We conclude with a comparative summary, analyzing our approach and the baselines according to several criteria: accuracy, applicability, space/time complexity and running time.

7.1 Baselines

In the experiments, we compare CLPL with the following baselines.

7.1.1 CHANCE BASELINE

We define the *chance* baseline as randomly guessing between the possible ambiguous labels only. Defining the (empirical) average ambiguous size to be $\mathbb{E}_S[|\mathbf{y}|] = \frac{1}{m}\sum_{i=1}^{m}|\mathbf{y}_i|$, then the expected error from the *chance* baseline is given by $\text{error}_{\text{chance}} = 1 - \frac{1}{\mathbb{E}_S[|\mathbf{y}|]}$.

7.1.2 NAIVE MODEL

We report results on an un-normalized version of the naive model introduced in Equation 3: $\sum_{a\in\mathbf{y}}\psi(g_a(x)) + \sum_{a\notin\mathbf{y}}\psi(-g_a(x))$, but both normalized and un-normalized versions produce very similar results. After training, we predict the label with the highest score (in the transductive setting): $\hat{y} = \arg\max_{a\in\mathbf{y}} g_a(x)$.

7.1.3 IBM MODEL 1

This generative model was originally proposed in Brown et al. (1993) for machine translation, but we can adapt it to the ambiguous label case. In our setting, the conditional probability of observing example $x \in \mathbb{R}^d$ given that its label is a is Gaussian: $x \sim N(\mu_a, \Sigma_a)$. We use the expectation-maximization (EM) algorithm to learn the parameters of the Gaussians (mean μ_a and diagonal covariance matrix $\Sigma_a = diag(\sigma_a)$ for each label).

7.1.4 DISCRIMINATIVE EM

We compare with the model proposed in Jin and Ghahramani (2002), which is a discriminative model with an EM procedure adapted for the ambiguous label setting. The authors minimize the KL divergence between a maximum entropy model P (estimated in the M-step) and a distribution over ambiguous labels $\hat{P}$ (estimated in the E-step):

$$J(\theta, \hat{P}) = \sum_i \sum_{a\in\mathbf{y}} \hat{P}(a \mid x_i) \log\left(\frac{\hat{P}(a \mid x_i)}{P(a \mid x_i, \theta)}\right).$$

7.1.5 K-NEAREST NEIGHBOR

Following Hullermeier and Beringer (2006), we adapt the k-Nearest Neighbor Classifier to the ambiguous label setting as follows:

$$knn(x) = \arg\max_{a\in\mathbf{y}} \sum_{i=1}^{k} w_i \mathbb{1}(a \in \mathbf{y}_i), \tag{5}$$

where x_i is the i^{th} nearest-neighbor of x using Euclidean distance, and w_i are a set of weights. We use two kNN baselines: **kNN** assumes uniform weights $w_i = 1$ (model used in Hullermeier and Beringer, 2006), and **weighted kNN** uses linearly decreasing weights $w_i = k - i + 1$. We use $k = 5$ and break ties randomly as in Hullermeier and Beringer (2006).

7.1.6 SUPERVISED MODELS

Finally we also consider two baselines that *ignore* the ambiguous label setting. The first one, denoted as **supervised model**, removes from Equation 3 the examples with $|\mathbf{y}| > 1$. The second model, denoted as **supervised kNN**, removes from Equation 5 the same examples.

7.2 Data Sets and Feature Description

We describe below the different data sets used to report our experiments. The experiments for automatic naming of characters in TV shows can be found in Section 8. A concise summary is given in Table 2.

Data Set	# instances (m)	# features (d)	# labels (L)	prediction task
UCI: dermatology	366	34	6	disease diagnostic
UCI: ecoli	336	8	8	site prediction
UCI: abalone	4177	8	29	age determination
FIW(10b)	500	50	10 (balanced)	face recognition
FIW(10)	1456	50	10	face recognition
FIW(100)	3011	50	100	face recognition
Lost audio	522	50	19	speaker id
TV+movies	10,000	50	100	face recognition

Table 2: Summary of data sets used in our experiments. The TV+movies experiments are treated in Section 8. Faces in the Wild (1) uses a balanced distribution of labels (first 50 images for the top 10 most frequent people).

7.2.1 UCI DATA SETS

We selected three biology related data sets from the publicly available UCI repository (Asuncion and Newman, 2007): dermatology, ecoli, abalone. As a preprocessing step, each feature was independently scaled to have zero mean and unit variance.

7.2.2 FACES IN THE WILD (FIW)

We experiment with different subsets of the publicly available Labeled Faces in the Wild (Huang et al., 2007a) data set. We use the images registered with funneling (Huang et al., 2007a), and crop out the central part corresponding to the approximate face location, which we resize to 60x90. We project the resulting grayscale patches (treated as 5400x1 vectors) onto a 50-dimensional subspace using PCA.[2] In Table 2, FIW(10b) extracts the first 50 images for each of the top 10 most frequent people (balanced label distribution); FIW(10) extracts *all* images for each of the top 10 most frequent people (heavily unbalanced label distribution, with 530 hits for George Bush and 53 hits for John Ashcroft); FIW(100) extracts up to 100 faces for each of the top 100 most frequent people (again, heavily unbalanced label distribution).

2. We kept the features simple by design; more sophisticated part-based registration and representation would further improve results, as we will see in Section 8.

7.2.3 SPEAKER IDENTIFICATION FROM AUDIO

We also investigate a speaker identification task based on audio in an uncontrolled environment. The audio is extracted from an episode of *Lost* (season 1, episode 5) and is initially completely unaligned. Compared to recorded conversation in a controlled environment, this task is more realistic and very challenging due to a number of factors: background noise, strong variability in tone of voice due to emotions, and people shouting or talking at the same time. We use the Hidden Markov Model Toolkit (HTK) (`http://htk.eng.cam.ac.uk/`) to compute forced alignment (Moreno et al., 1998; Sjölander, 2003), between the closed captions and the audio (given the rough initial estimates from closed caption time stamps, which are often overlapping and contain background noise). After alignment, our data set is composed of 522 utterances (each one corresponding to a closed caption line, with aligned audio and speaker id obtained from aligned screenplay), with 19 different speakers. For each speech segment (typically between 1 and 4 seconds) we extract standard voice processing audio features: pitch (Talkin, 1995), Mel-Frequency Cepstral Coefficients (MFCC) (Mermelstein, 1976), Linear predictive coding (LPC) (Proakis and Manolakis, 1996). This results in a total of 4,000 features, which we normalize to the range $[-1,1]$ and then project onto 50 dimensions using PCA.

7.3 Experimental Setup

For the **inductive experiments**, we split randomly in half the instances into (1) **ambiguously labeled training set**, and (2) **unlabeled testing set**. The ambiguous labels in the training set are generated randomly according to different noise models which we specify in each case. For each method and parameter setting, we report the **average test error rate** over **20 trials** after training the model on the ambiguous train set. We also report the corresponding **standard deviation** as an error bar in the plots. Note, in the inductive setting we consider the test set as unlabeled, thus the classifier votes among *all* possible labels:

$$a^* = h(x) = \arg \max_{a \in \{1..L\}} g_a(x).$$

For the **transductive experiments**, there is no test set; we report the error rate for disambiguating the ambiguous labels (also averaged over 20 trials corresponding to random settings of ambiguous labels). The main differences with the inductive setting are: (1) the model is trained on all instances and tested on the same instances; and (2) the classifier votes only among the ambiguous labels, which is easier:

$$a^* = h(x) = \arg \max_{a \in \mathbf{y}} g_a(x).$$

We compare our CLPL approach (denoted as **mean** in figures, due to the form of the loss) against the **baselines** presented in Section 7.1: Chance, Model 1, Discriminative EM model, k-Nearest Neighbor, weighted k-Nearest Neighbor, Naive model, supervised model, and supervised kNN. Note, in our experiments the Discriminative EM model was much slower to converge than all the other methods, and we only report the first series of experiments with this baseline.

Table 3 summarizes the different settings used in each experiment. We experiment with different noise models for ambiguous bags, parametrized by p, q, ε, see Figure 6. p represents the proportion of examples that are ambiguously labeled. q represents the number of *extra* labels for each ambiguous example. ε represents the degree of ambiguity (defined in 1) for each ambiguous

example.[3] We also vary the dimensionality by increasing the number of PCA components from 1 to 200, with half of extra labels added uniformly at random. In Figure 7, we vary the ambiguity size q for three different subsets of Faces in the Wild. We report results on additional data sets in Figure 8.

Experiment	fig	induct.	data set	parameter
# of ambiguous bags	6	yes	FIW(10b)	$p \in [0, 0.95], q = 2$
degree of ambiguity	6	yes	FIW(10b)	$p = 1, q = 1, \varepsilon \in [1/(L-1), 1]$
degree of ambiguity	6	**no**	FIW(10b)	$p = 1, q = 1, \varepsilon \in [1/(L-1), 1]$
dimension	6	yes	FIW(10b)	$p = 1, q = \frac{L-1}{2}, d \in [1, .., 200]$
ambiguity size	7	yes	FIW(10b)	$p = 1, q \in [0, 0.9(L-1)]$
ambiguity size	7	yes	FIW(10)	$p = 1, q \in [0, 0.9(L-1)]$
ambiguity size	7	yes	FIW(100)	$p = 1, q \in [0, 0.9(L-1)]$
ambiguity size	8	yes	*Lost* audio	$p = 1, q \in [0, 0.9(L-1)]$
ambiguity size	8	yes	ecoli	$p = 1, q \in [0, 0.9(L-1)]$
ambiguity size	8	yes	derma	$p = 1, q \in [0, 0.9(L-1)]$
ambiguity size	8	yes	abalone	$p = 1, q \in [0, 0.9(L-1)]$

Table 3: Summary of controlled experiments. We experiment with 3 different noise models for ambiguous bags, parametrized by p, q, ε. p represents the proportion of examples that are ambiguously labeled. q represents the number of *extra* labels for each ambiguous example (generated uniformly without replacement). ε represents the degree of ambiguity for each ambiguous example (see definition 1). L is the total number of labels. We also study the effects of data set choice, inductive vs transductive learning, and feature dimensionality.

7.3.1 EXPERIMENTS WITH A BOOSTING VERSION OF CLPL

We also experiment with a boosting version of our CLPL optimization, as presented in Appendix A. Results are shown in Figure 9, comparing our method with kNN and the naive method (also using boosting). Despite the change in learning algorithm and loss function, the trends remain the same.

7.4 Comparative Summary

We can draw several conclusions. Our proposed CLPL model **uniformly outperformed** all baselines in all but one experiment (UCI dermatology data set), where it ranked second closely behind Model 1. In particular CLPL always uniformly outperformed the naive model. The naive model ranks in second. As expected, increasing ambiguity size monotonically affects error rate. We also see that increasing ε significantly affects error, even though the ambiguity size is constant, consistent with our bounds in Section 3.3. We also note that the supervised models defined in Section 7.1.6 (which ignore the ambiguously labeled examples) consistently perform worse than their counterparts adapted for the ambiguous setting. For example, in Figure 6 (Top Left), a model trained with nearly all examples ambiguously labeled ("mean" curve), $p = 95\%$) performs as good as a model which uses 60% of *fully labeled* examples ("supervised" curve, $p = 40\%$). The same holds between the "kNN" curve at $p = 95\%$ and the "supervised kNN" curve at $p = 40\%$.

3. We first choose at random for each label a dominant co-occurring label which is sampled with probability ε; the rest of the labels are sampled uniformly with probability $(1-\varepsilon)/(L-2)$ (there is a single extra label per example).

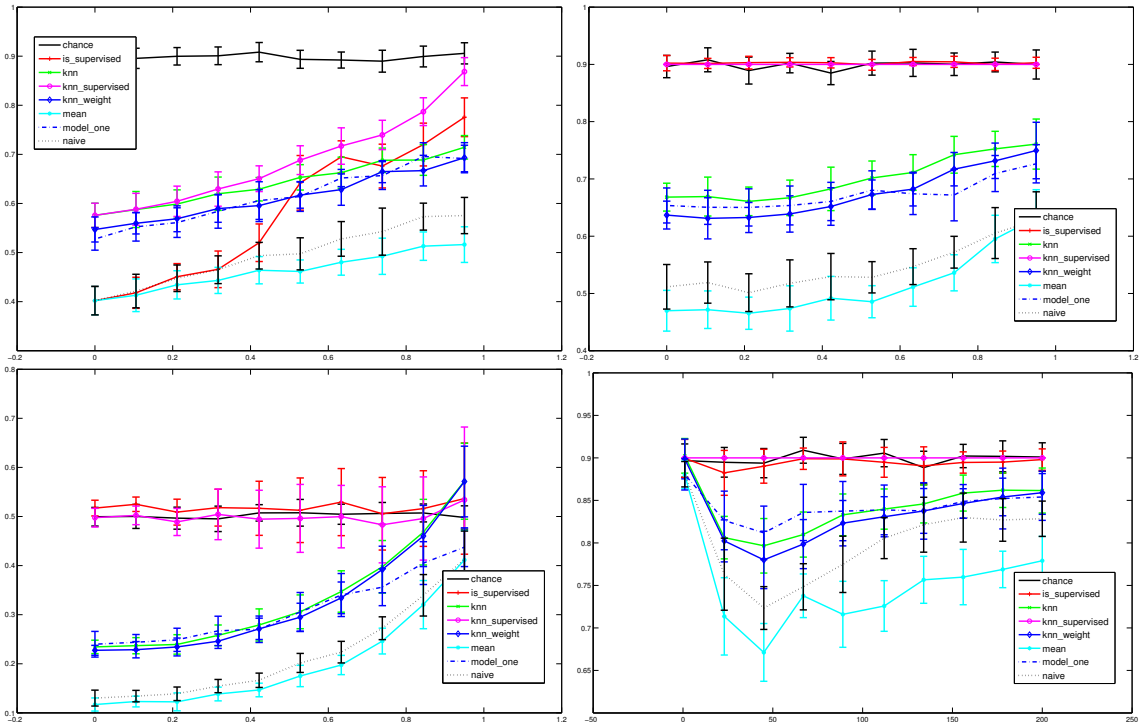

Figure 6: Results on Faces in the Wild in different settings, comparing our proposed CLPL (denoted as **mean**) to several baselines. In each case, we report the average error rate (y-axis) and standard deviation over 20 trials as in Figure 7. **(top left)** increasing proportion of ambiguous bags q, inductive setting. **(top right)** increasing ambiguity degree ε (Equation 1), inductive setting. **(bottom left)** increasing ambiguity degree ε (Equation 1), transductive setting. **(bottom right)** increasing dimensionality, inductive setting.

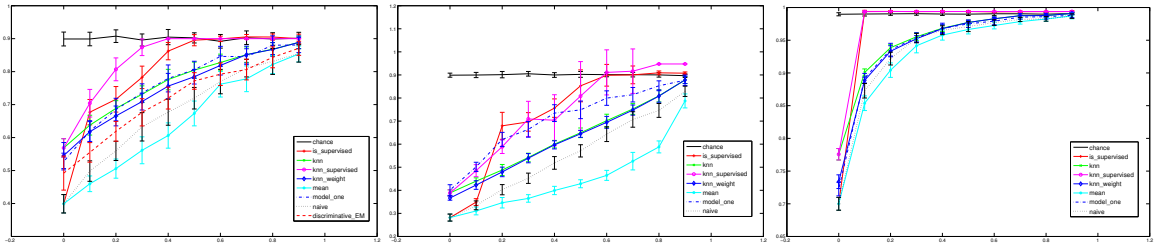

Figure 7: Additional results on Faces in the Wild, obtained by varying the ambiguity size q on the x-axis (inductive case). **Left:** balanced data set using 50 faces for each of the top 10 labels. **Middle:** unbalanced data set using all faces for each of the top 10 labels. **Right:** unbalanced data set using up to 100 faces for each of the top 100 labels.

7.4.1 COMPARISON WITH VARIANTS OF OUR APPROACH

In order to get some intuition on CLPL (Equation 2), which we refer to as the **mean** model in our experiments, we also compare with the following **sum** and **contrastive** alternatives:

$$\mathcal{L}_\psi^{\text{sum}}(g(x),\mathbf{y}) = \psi\left(\sum_{a\in\mathbf{y}} g_a(x)\right) + \sum_{a\notin\mathbf{y}} \psi(-g_a(x)), \tag{6}$$

$$\mathcal{L}_\psi^{\text{contrastive}}(g(x),\mathbf{y}) = \sum_{a'\notin\mathbf{y}} \psi\left(\frac{1}{|\mathbf{y}|}\sum_{a\in\mathbf{y}} g_a(x) - g_{a'}(x)\right). \tag{7}$$

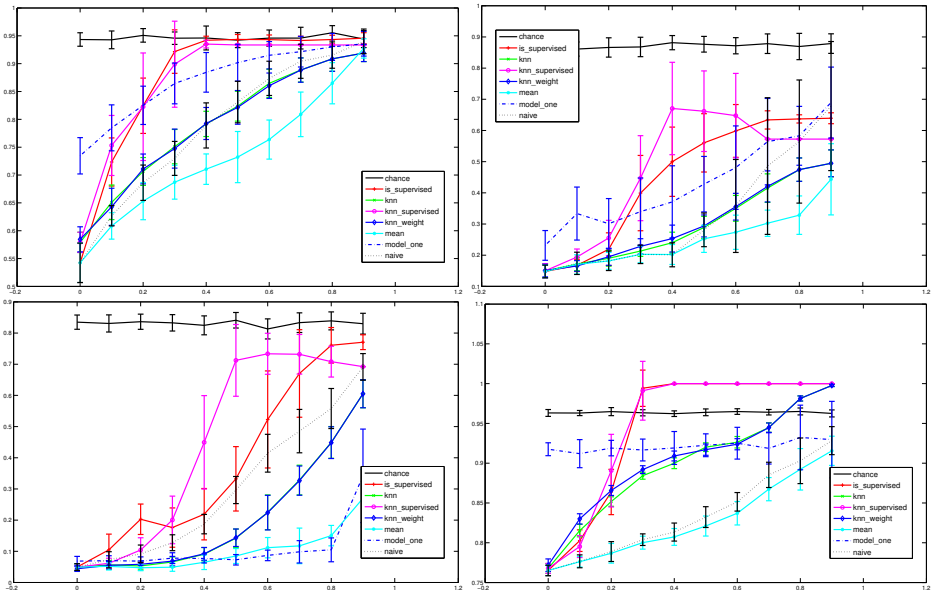

Figure 8: Inductive results on different data sets. In each case, we report the average error rate (y-axis) and standard deviation over 20 trials as in Figure 7. **Top Left:** speaker identification from *Lost* audio. **Top Right:** ecoli data set (UCI). **Bottom Left:** dermatology data set (UCI). **Bottom Right:** abalone data set (UCI).

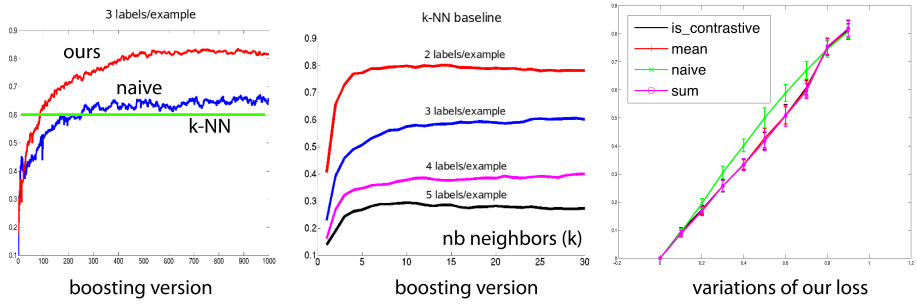

Figure 9: **Left:** We experiment with a boosting version of the ambiguous learning, and compare to a boosting version of the naive baseline (here with ambiguous bags of size 3). We plot *accuracy* vs number of boosting rounds. The green horizontal line corresponds to the best performance (across k) of the k-NN baseline. **Middle:** accuracy of k-NN baseline across k. **Right:** we compare CLPL (labeled **mean**) with two variants defined in Equation 6,Equation 7, along with the naive model (same setting as Figure 6, Top Left).

When $\psi(\cdot)$ is the hinge loss, the mean and sum model are very similar, but this is not the case for strictly convex binary losses. Figure 9 shows that variations on our cost function have *little effect* in the transductive setting. In the inductive setting, other experiments we performed show that the mean and sum version are still very similar, but the contrastive version is worse. In general it seems that models based on minimization of a convex loss function (naive and different versions of our model) usually outperform the other models.

Figure 10: Predictions on *Lost* and *C.S.I.*. Incorrect examples are: row 1, column 3 (truth: Boone); row 2, column 2 (truth: Jack).

8. Experiments with Partially Labeled Faces in Videos

We now return to our introductory motivating example, naming people in TV shows (Figure 1, right). Our goal is to identify characters given ambiguous labels derived from the screenplay. Our data consists of 100 hours of *Lost* and *C.S.I.*, from which we extract ambiguously labeled faces to learn models of common characters. We use the same features, learning algorithm and loss function as in Section 7.2.2. We also explore using additional person- and movie-specific constraints to improve performance. Sample results are shown in Figure 10.

8.1 Data Collection

We adopt the following filtering pipeline to extract face tracks, inspired by Everingham et al. (2006): **(1)** Run the off-the-shelf OpenCV face detector over all frames, searching over rotations and scales. **(2)** Run face part detectors[4] over the face candidates. **(3)** Perform a 2D rigid transform of the parts to a template. **(4)** Compute the score of a candidate face $s(x)$ as the sum of part detector scores plus rigid fit error, normalizing each to weight them equally, and filtering out faces with low score. **(5)** Assign faces to tracks by associating face detections within a shot using normalized cross-correlation in RGB space, and using dynamic programming to group them together into tracks. **(6)** Subsample face tracks to avoid repetitive examples. In the experiments reported here we use the best scoring face in each track, according to $s(x)$.

Concretely, for a particular episode, step (1) finds approximately 100,000 faces, step (4) keeps approximately 10,000 of those, and after subsampling tracks in step (6) we are left with 1000 face examples.

8.2 Ambiguous Label Selection

Screenplays for popular TV series and movies are readily available for free on the web. Given an alignment of the screenplay to frames, we have ambiguous labels for characters in each scene: the set of speakers mentioned at some point in the scene, as shown in Figure 1. Alignment of screenplay to video uses methods presented in Cour et al. (2008) and Everingham et al. (2006), linking closed captions to screenplay.

4. The detectors use boosted cascade classifiers of Haar features for the eyes, nose and mouth.

Lost (#labels, #episodes)	(8,16)	(16,8)†	(16,16)	(32,16)
Naive	14%	18.6%	16.5%	18.5%
ours (CLPL / "mean")	10%	12.6%	14%	17%
ours+constraints	**6%**	n/a	**11%**	**13%**

Table 4: Misclassification rates of different methods on TV show *Lost*. In comparison, for (16,16) the baseline performances are *knn*: 30%; *Model 1*: 44%; *chance*: 53%. †: This column contains results exactly reproducible from our publicly available reference implementation, which can be found at `http://vision.grasp.upenn.edu/video`. For simplicity, this public code does not include a version with extra constraints.

We use the ambiguous sets to select face tracks filtered through our pipeline. We prune scenes which contain characters other than the set we choose to focus on for experiments (top $\{8,16,32\}$ characters), or contain 4 or more characters. This leaves ambiguous bags of size 1, 2 or 3, with an average bag size of 2.13 for *Lost*, and 2.17 for *C.S.I.*.

8.3 Errors in Ambiguous Label Sets

In the TV episodes we considered, we observed that approximately 1% of ambiguous label sets were wrong, in that they didn't contain the ground truth label of the face track. This came from several reasons: presence of a non-english speaking character (Jin Kwon in *Lost*, who speaks Korean) whose dialogue is not transcribed in the closed captions; sudden occurence of an unknown, uncredited character on screen, and finally alignment problems due to large discrepencies between screenplay and closed captions. While this is not a major problem, it becomes so when we consider additional cues (mouth motion, gender) that restrict the ambiguous label set. We will see how we tackle this issue with a robust confidence measure for obtaining good precision recall curves in Section 8.5.

8.4 Results with the Basic System

Now that we have a set of instances (face tracks), feature descriptors for the face track and ambiguous label sets for each face track, we can apply the same method as described in the previous section. We use a transductive setting: we test our method on our ambiguously labeled training set.

The confusion matrix displaying the distribution of ambiguous labels for the top 16 characters in *Lost* is shown in Figure 11 (left). The confusion matrix of our predictions after applying our ambiguous learning algorithm is shown in Figure 11 (right). Our method had the most trouble disambiguating Ethan Rom from Claire Littleton (Ethan Rom only appears in 0.7% of the ambiguous bags, 3 times less then the second least common character) and Liam Pace from Charlie Pace (they are brothers and co-occur frequently, as can be seen in the top figure). The case of Sun Kwon and Jin Kwon is a bit special, as Jin does not speak English in the series and is almost never mentioned in the closed-captions, which creates alignment errors between screenplay and closed captions. These difficulties illustrate some of the interesting challenges in ambiguously labeled data sets. As we can see, the most difficult classes are the ones with which another class is strongly correlated in the ambiguous label confusion matrix. This is consistent with the theoretical bounds we obtained in Section 3.3, which establish a relation between the class specific error rate and class specific degree of ambiguity ε.

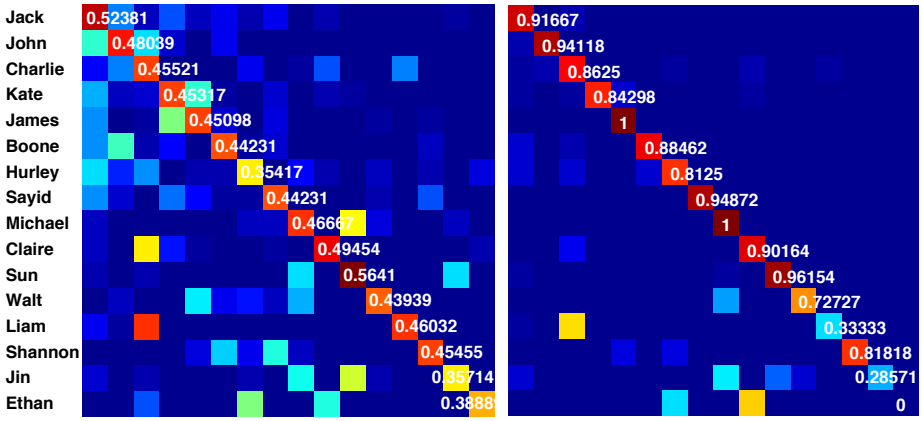

Figure 11: **Left**: Label distribution of top 16 characters in *Lost* (using the standard matlab color map). Element D_{ij} represents the proportion of times class i was seen with class j in the ambiguous bags, and $D\mathbf{1} = \mathbf{1}$. **Right**: Confusion matrix of predictions from Section 8.4. Element A_{ij} represents the proportion of times class i was classified as class j, and $A\mathbf{1} = \mathbf{1}$. Class priors for the most frequent, the median frequency, and the least frequent characters in *Lost* are Jack Shephard, 14%; Hugo Reyes, 6%; Liam Pace 1%.

Quantitative results are shown in Table 4. We measure error according to average 0-1 loss with respect to hand-labeled groundtruth labeled in 8 entire episodes of *Lost*. Our model outperforms all the baselines, and we will further improve results. We now compare several methods to obtain the best possible precision at a given recall, and propose a confidence measure to this end.

8.5 Improved Confidence Measure for Precision-recall Evaluation

We obtain a precision-recall curve using a refusal to predict scheme, as used by Everingham et al. (2006): we report the precision p for the r most confident predictions, varying $r \in [0,1]$. We compare several **confidence measures** based on the classifier scores $g(x)$ and propose a novel one that significantly improves precision-recall, see Figure 12 for results.

1. the **max** and **ratio** confidence measures (as used in Everingham et al., 2006) are defined as:

$$C_{\max}(g(x)) = \max_a g_a(x),$$
$$C_{\text{ratio}}(g(x)) = \max_a \frac{\exp(g_a(x))}{\sum_b \exp(g_b(x))}.$$

2. the **relative** score can be defined as the difference between the best and second best scores over all classifiers $(g_a)_{a\in\{1..L\}}$ (where $a^* = \arg\max_{a\in\{1..L\}} g_a(x)$):

$$C_{\text{rel}}(g(x)) = g_{a^*}(x) - \max_{a\in\{1..L\}-\{a^*\}} g_a(x).$$

3. we can define the **relative-constrained** score as an adaptation to the ambiguous setting; we only consider votes among ambiguous labels $\mathbf{y}$ (where $a^* = \arg\max_{a\in\mathbf{y}} g_a(x)$):

$$C_{\text{rel},\mathbf{y}}(g(x)) = g_{a^*}(x) - \max_{a\in\mathbf{y}-\{a^*\}} g_a(x).$$

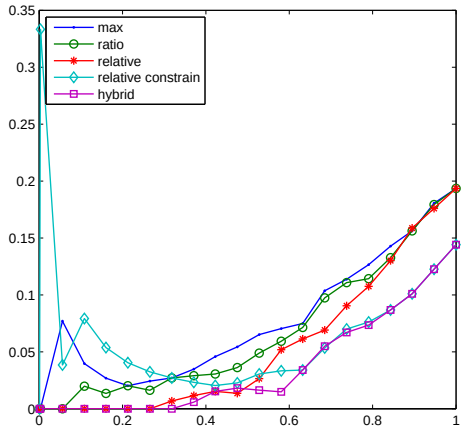

Figure 12: Improved **hybrid** confidence measure for precision-recall evaluation. x axis: recall; y axis: naming error rate for CLPL on 16 episodes of *Lost* (top 16 characters). **max** confidence score performs rather poorly as it ignores other labels. **relative** improves the high precision/low recall region by considering the margin instead. The **relative-constrain** improves the high-recall/low-precision region by only voting among the ambiguous bags, but it suffers in high-precision/low recall region because some ambiguous bags may be erroneous. Our **hybrid** confidence score gets the best of both worlds.

There are some problems with all of those choices, especially in the case where we have some errors in ambiguous label set ($a \notin Y$ for the true label a). This can occur for example if we restrict them with some heuristics to prune down the amount of ambiguity, such as the ones we consider in Section 8.6 (mouth motion cue, gender, etc). At **low recall**, we want maximum precision, therefore we cannot trust too much the heuristic used in relative-constrained confidence. At **high recall**, the errors in the classifier dominate the errors in ambiguous labels, and relative-constrained confidence gives better precision because of the restriction. We introduce a **hybrid** confidence measure that performs well for all recall levels r, interpolating between the two confidence measures:

$$h_r^a(x) = \begin{cases} g_a(x) & \text{if } a \in \mathbf{y}, \\ (1-r)g_a(x) + r \min_b g_b(x) & \text{else.} \end{cases}$$
$$C_r(g(x)) = C_{\text{rel}}(h_r(x)).$$

By design, in the limit $r \to 0$, $C_r(g(x)) \approx C_{\text{rel}}(g(x))$. In the limit $r \to 1$, $h_r^a(x)$ is small for $a \notin \mathbf{y}$ and so $C_r(g(x)) \approx C_{\text{rel},\mathbf{y}}(g(x))$.

8.6 Additional Cues

We investigate additional features to further improve the performance of our system: mouth motion, grouping constraints, gender. Final misclassification results are reported in Table 4.

8.6.1 MOUTH MOTION

We use a similar approach to Everingham et al. (2006) to detect mouth motion during dialog and adapt it to our ambiguous label setting.[5] For a face track x with ambiguous label set $\mathbf{y}$ and a temporally overlapping utterance from a speaker $a \in \{1..L\}$ (after aligning screenplay and closed captions), we restrict $\mathbf{y}$ as follows:

$$\mathbf{y} := \begin{cases} \{a\} & \text{if mouth motion,} \\ \mathbf{y} & \text{if refuse to predict or } \mathbf{y} = \{a\}, \\ \mathbf{y} - \{a\} & \text{if absence of mouth motion.} \end{cases}$$

8.6.2 GENDER CONSTRAINTS

We introduce a gender classifier to constrain the ambiguous labels based on predicted gender. The gender classifier is trained on a data set of registered male and female faces, by boosting a set of decision stumps computed on Haar wavelets. We use the average score over a face track output by the gender classifier. We assume known the gender of names mentioned in the screenplay (using automatically extracted cast list from IMDB). We use gender by filtering out the labels that do not match by gender the predicted gender of a face track, if the confidence exceeds a threshold (one for females and one for males are set on a validation data to achieve 90% precision for each direction of the gender prediction). Thus, we modify ambiguous label set $\mathbf{y}$ as:

$$\mathbf{y} := \begin{cases} \mathbf{y} & \text{if gender uncertain,} \\ \mathbf{y} - \{a : a \text{ is male}\} & \text{if gender predicts female,} \\ \mathbf{y} - \{a : a \text{ is female}\} & \text{if gender predicts male.} \end{cases}$$

8.6.3 GROUPING CONSTRAINTS

We propose a very simple must-not-link constraint, which states $y_i \neq y_j$ if face tracks x_i, x_j are in two consecutive shots (modeling alternation of shots, common in dialogs). This constraint is active only when a scene has 2 characters. Unlike the previous constraints, this constraint is incorporated as additional terms in our loss function, as in Yan. et al. (2006). We also propose *groundtruth* grouping constraints for comparison: $y_i = y_j$ for each pair of face tracks x_i, x_j of the same label, and that are separated by at most one shot.

8.7 Ablative Analysis

Figure 13 is an ablative analysis, showing error rate vs recall curves for different sets of cues. We see that the constraints provided by mouth motion help most, followed by gender and link constraints. The best setting (without using groundtruth) combines the former two cues. Also, we notice, once again, a significant performance improvement of our method over the naive method.

8.8 Qualitative Results and Video Demonstration

We show examples with predicted labels and corresponding accuracy, for various characters in *C.S.I.*, see Figure 14. Those results were obtained with the basic system of Section 8.4. Full-frame

5. Motion or absence of motion are detected with a low and high threshold on normalized cross-correlation around mouth regions in consecutive frames.

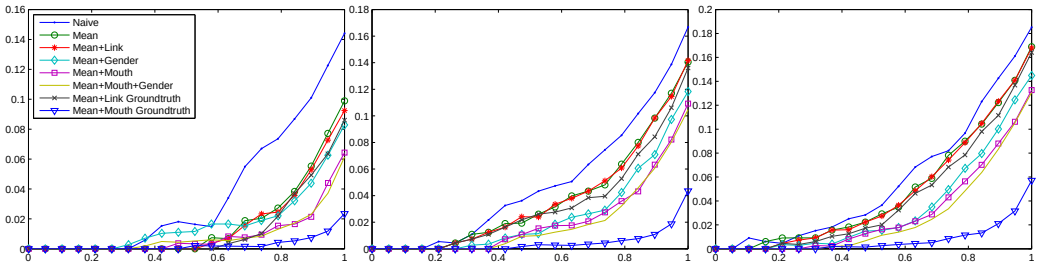

Figure 13: Ablative analysis. *x*-axis: recall; *y*-axis: error rate for character naming across 16 episodes of *Lost*, and the 8, 16, and 32 most common labels (respectively for the left, middle, right plots). We compare our method, **mean**, to the **Naive** model and show the effect of adding several cues to our system. **Link**: simple must-not-link constraints from shot alternation, **Gender**: gender cue for simplification of ambiguous bags; **Mouth**: mouth motion cue for detecting the speaker with synchronous mouth motion; we also consider the combination **Mouth+Gender**, as well as swapping in perfect components such as **Groundtruth link** constraints and **Groundtruth Mouth** motion.

Figure 14: **Left:** Examples classified as Catherine Willows in *C.S.I.* data set using our method (zoom-in for details). Results are sorted by classifier score, in column major format; this explains why most of the errors occur in the last columns. The precision is 85.3%. **Right:** Examples classified as Sara Sidle in *C.S.I.*. The precision is 78.3%.

detections for *Lost* and *C.S.I.* data sets can be seen in Figure 10. We also propagate the predicted labels of our model to all faces in the same face track throughout an episode. Video results of several episodes can be found at the following website `http://www.youtube.com/user/AmbiguousNaming`.

9. Conclusion

We have presented an effective learning approach for partially labeled data, where each instance is tagged with more than one label. Theoretically, under reasonable assumptions on the data distribu-

tion, one can show that our algorithm will produce an accurate classifier. We applied our method to two partially-supervised naming tasks: one on still images and one on video from TV series. We also compared to several strong competing algorithms on the same data sets and demonstrated that our algorithm achieves superior performance. We attribute the success of the approach to better modeling of the mutual exclusion between labels than the simple multi-label approach. Moreover, unlike recently published techniques that address similar ambiguously labeled problems, our method does not rely on heuristics and does not suffer from local optima of non-convex methods.

Acknowledgments

The authors would like to thank Jean Ponce and the anonymous reviewers for helpful suggestions. This work was supported by NSF grant 0803538 and a grant from Google. B. Taskar was partially supported by DARPA CSSG, the ONR Young Investigator Award N000141010746 and the Sloan Fellowship.

Appendix A. CLPL with Feature Selection Using Boosting

We derive Algorithm 1 by taking the second order Taylor expansion of the loss $\mathcal{L}_\psi(g(x),\mathbf{y})$, with $\psi(u)=\exp(-u)$. The updates of the algorithm are similar to a multiclass version of Gentleboost (Friedman et al., 2000), but keep a combined weight v_i for the positive example $\mathbf{f}(x_i,\mathbf{y}_i)$ and weights $v_{i,a}$ for the negative examples $\mathbf{f}(x_i,a), a\notin\mathbf{y}_i$.

Algorithm 1 Boosting for CLPL with exponential loss

Initialize weights: $v_i=1 \quad \forall i, \quad v_{i,a}=1 \quad \forall i, a\notin\mathbf{y}_i$
for $t=1\ldots T$ **do**
  for $a=1\ldots L$ **do**
    Fit the parameters of each weak classifier $u(x)$ to minimize the second-order Taylor approximation of the cost function with respect to the a^{th} classifier:

$$\frac{1}{2}\sum_i \left[v_i\cdot\mathbb{1}(a\in\mathbf{y}_i)(u(x_i)/|\mathbf{y}_i|-1)^2+v_{i,a}\cdot\mathbb{1}(a\notin\mathbf{y}_i)(u(x_i)+1)^2\right]+constant.$$

  end for
  Choose the combination of u,a with lowest residual error.
  Update $g_a(x)=g_a(x)+u(x)$
  for $i=1\ldots m$ **do**
    if $a\in\mathbf{y}_i$ **then**
      $v_i=v_i\cdot\exp(-u(x_i))$
    else
      $v_{i,a}=v_{i,a}\cdot\exp(u(x_i))$
    end if
  end for
  Normalize v to sum to 1.
end for

Appendix B. Proofs

Proof of Proposition 1 (Partial loss bound via ambiguity degree ε). The first inequality comes from the fact that $h(x) \notin \mathbf{y} \implies h(x) \neq y$. For the second inequality, fix an $x \in \mathcal{X}$ with $P(X = x) > 0$ and define $\mathbb{E}_P[\cdot \mid x]$ as the expectation with respect to $P(\mathbf{Y} \mid X = x)$.

$$\begin{aligned}\mathbb{E}_P[\mathcal{L}_A(h(x),\mathbf{Y})|x] &= P(h(x) \notin \mathbf{Y} \mid X = x) = P(h(x) \neq Y, h(x) \notin \mathbf{Z} \mid X = x) \\ &= \sum_{a \neq h(x)} P(Y = a \mid X = x)(1 - \underbrace{P(h(x) \in \mathbf{Z} \mid X = x, Y = a)}_{\leq \varepsilon \text{ by definition}}) \\ &\geq \sum_{a \neq h(x)} P(Y = a \mid X = x)(1-\varepsilon) = (1-\varepsilon)\mathbb{E}_P[\mathcal{L}(h(x),Y)|x].\end{aligned}$$

Hence, $\mathbb{E}_P[\mathcal{L}(h(x),Y)|x] \leq \frac{1}{1-\varepsilon}\mathbb{E}_P[\mathcal{L}_A(h(x),\mathbf{Y})|x]$ for any x. We conclude by taking expectation over x.The first inequality is tight: equality can be achieved, for example, when $P(y|x)$ is deterministic, and a perfect classifier h such that for all x, $h(x) = y$. The second inequality is also tight: for example consider the uniform case with a fixed ambiguity size $|\mathbf{z}| = C$ and for all $x, y, z \neq y$, $P(z \in \mathbf{z} \mid X = x, Y = y) = C/(L-1)$. In the proof above (second inequality), the only inequality becomes an equality. In fact, this also shows that for any (rational) ε, we can find a number of labels L, a distribution P and a classifer h such that there is equality. ■

Proof of Proposition 3 (Partial loss bound via (ε, δ)). We split up the expectation in two parts:

$$\begin{aligned}\mathbb{E}_P[\mathcal{L}(h(X),Y)] &= \mathbb{E}_P[\mathcal{L}(h(X),Y)|(X,Y) \in G](1-\delta) + \mathbb{E}_P[\mathcal{L}(h(X),Y)|(X,Y) \notin G]\delta \\ &\leq \mathbb{E}_P[\mathcal{L}(h(X),Y)|(X,Y) \in G](1-\delta) + \delta \\ &\leq \frac{1}{1-\varepsilon}\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})|(X,Y) \in G](1-\delta) + \delta.\end{aligned}$$

We applied Proposition 1 in the last step. Using a symmetric argument,

$$\begin{aligned}\mathbb{E}_P[\mathcal{L}_A(h(X),Y)] &= \mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})|(X,Y) \in G](1-\delta) + \mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})|(X,Y) \notin G]\delta \\ &\geq \mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})|(X,Y) \in G](1-\delta).\end{aligned}$$

Finally we obtain $\mathbb{E}_P[\mathcal{L}(h(X),Y)] \leq \frac{1}{1-\varepsilon}\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y})] + \delta$. ■

Proof of Proposition 4 (Label-specific partial loss bound). Fix $x \in \mathcal{X}$ such that $P(X = x) > 0$ and $P(Y = a|x) > 0$ and define $\mathbb{E}_P[\cdot \mid x, a]$ as the expectation w.r.t. $P(\mathbf{Z} \mid X = x, Y = a)$. We consider two cases:

a) if $h(x) = a$, $\quad \mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y}) \mid x, a] = P(h(x) \neq a, h(x) \notin \mathbf{y} \mid X = x, Y = a) = 0.$

b) if $h(x) \neq a$, $\quad \mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y}) \mid x, a] = P(h(x) \notin \mathbf{Z} \mid X = x, Y = a)$
$= 1 - P(h(x) \in \mathbf{Z} \mid X = x, Y = a) \geq 1 - \varepsilon_a.$

We conclude by taking expectation over x:

$$\begin{aligned}\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y}) \mid Y = a] &= P(h(X) = a|Y = a)\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y}) \mid h(X) = a, Y = a] \\ &\quad + P(h(X) \neq a|Y = a)\mathbb{E}_P[\mathcal{L}_A(h(X),\mathbf{Y}) \mid h(X) \neq a, Y = a] \\ &\geq 0 + P(h(X) \neq a \mid Y = a) \cdot (1 - \varepsilon_a) \\ &= (1 - \varepsilon_a) \cdot \mathbb{E}_P[\mathcal{L}(h(X),Y) \mid Y = a].\end{aligned}$$

■

Proof of Proposition 5 (Partial label consistency). We assume $g(x)$ is found by minimizing over an appropriately rich sequence of function classes (Tewari and Bartlett, 2005), in our case, as $m \to \infty$, $\mathcal{G} \to \mathbb{R}^L$. Hence we can focus on analysis for a fixed x (with $P(X = x) > 0$), writing $g_a = g_a(x)$, and for any set $\mathbf{c} \subseteq \{1, \ldots, L\}$, $g_{\mathbf{c}} = \sum_{a \in \mathbf{c}} g_a / |\mathbf{c}|$ and $P_{\mathbf{c}} = P(\mathbf{Y} = \mathbf{c} | X = x)$. We also write $P_a = P(a \in \mathbf{Y} | X = x)$ for any label a, and use shorthand $P_{\mathbf{c},a} = P_{\mathbf{c} \cup \{a\}}$ and $g_{\mathbf{c},a} = g_{\mathbf{c} \cup \{a\}}$. We have:

$$\mathcal{L}_\psi(g) = \sum_{\mathbf{c}} P_{\mathbf{c}} \cdot \left(\psi(g_{\mathbf{c}}) + \sum_{a \notin \mathbf{c}} \psi(-g_a) \right).$$

Note that the derivative $\psi'(\cdot)$ exists and is non-positive and non-decreasing by assumption and $\psi'(z) < 0$ for $z \leq 0$. The assumptions imply that $\psi(-\infty) \to \infty$, so assuming that $P_a < 1$, minimizers are upper-bounded: $g_a < \infty$. The case of $P_a = 0$ leads to $g_a \to -\infty$ and it can be ignored without loss of generality, so we can assume that optimal g is bounded for fixed p with $0 < P_a < 1$.

Taking the derivative of the loss with respect to g_a and setting to 0, we have the first order optimality conditions:

$$\frac{\partial \mathcal{L}_\psi(g)}{\partial g_a} = \sum_{\mathbf{c}: a \notin \mathbf{c}} \frac{P_{\mathbf{c},a} \psi'(g_{\mathbf{c},a})}{|\mathbf{c}| + 1} - (1 - P_a) \psi'(-g_a) = 0.$$

Now suppose (for contradiction) that at a minimizer g, $b \in \arg\max_{a'} g_{a'}$ but $P_a > P_b$ for some $a \in \arg\max_{a'} P_{a'}$. Subtracting the optimality conditions for a, b from each other, we get

$$\sum_{\mathbf{c}: a, b \notin \mathbf{c}} \frac{P_{\mathbf{c},a} \psi'(g_{\mathbf{c},a}) - P_{\mathbf{c},b} \psi'(g_{\mathbf{c},b})}{|\mathbf{c}| + 1} = (1 - P_a) \psi'(-g_a) - (1 - P_b) \psi'(-g_b).$$

Since $g_a \leq g_b$, $\psi'(g_{\mathbf{c},a}) \leq \psi'(g_{\mathbf{c},b})$ and $\psi'(-g_a) \geq \psi'(-g_b)$. Plugging in on both sides:

$$\sum_{\mathbf{c}: a, b \notin \mathbf{c}} \frac{(P_{\mathbf{c},a} - P_{\mathbf{c},b}) \psi'(g_{\mathbf{c},b})}{|\mathbf{c}| + 1} \geq (P_b - P_a) \psi'(-g_b).$$

By dominance assumption, $(P_{\mathbf{c},a} - P_{\mathbf{c},b}) \geq 0$ and since $(P_b - P_a) < 0$ and $\psi'(\cdot)$ is non-positive, the only possibility of the inequality holding is that $\psi'(-g_b) = 0$ (which implies $g_b > 0$) and $(P_{\mathbf{c},a} - P_{\mathbf{c},b}) \psi'(g_{\mathbf{c},a}) = 0$ for all $\mathbf{c}$. But $(P_b - P_a) < 0$ implies that there exists a subset $\mathbf{c}$ such that $(P_{\mathbf{c},a} - P_{\mathbf{c},b}) > 0$. Since $b \in \arg\max g$, $g_{\mathbf{c},b} \leq g_b$, so $g_{\mathbf{c},b} \leq 0$, hence $\psi'(g_{\mathbf{c},b}) < 0$, a contradiction.

When $P(y \mid x)$ is deterministic, let $P(y|x) = \mathbb{1}(y = a)$. Clearly, if $\varepsilon < 1$, then $a = \arg\max_{a'} P_{a'}$ and $P_a = 1 > P_{a'}, \forall a' \neq a$. Then the minimizer g satisfies either (1) $g_a \to \infty$ (this happens if $\psi'(\cdot) < 0$ for finite arguments) while $g_{a'}$ are finite because of $(1 - P_{a'}) \psi(-g_{a'})$ terms in the objective or (2) g is finite and the proof above applies since dominance holds: $P_{\mathbf{c},b} = 0$ if $a \notin \mathbf{c}$, so we can apply the theorem.

■

Proof of Proposition 6 (Comparison between partial losses). Let $a^* = \arg\max_{a \in 1..L} g_a(x)$. For the first inequality, if $a^* \in \mathbf{y}$, $\mathcal{L}_\psi^{max}(g(x),\mathbf{y}) \geq 0 = 2\mathcal{L}_A(g(x),\mathbf{y})$. Otherwise $a^* \notin \mathbf{y}$:

$$\begin{aligned}\mathcal{L}_\psi^{max}(g(x),\mathbf{y}) &\geq \psi(\max_{a\in\mathbf{y}} g_a(x)) + \psi(-g_{a^*}(x)) \geq \psi(g_{a^*}(x)) + \psi(-g_{a^*}(x)) \\ &\geq 2\psi\left(\frac{g_{a^*}(x) - g_{a^*}(x)}{2}\right) = 2\psi(0) \geq 2\mathcal{L}_A(g(x),\mathbf{y}).\end{aligned}$$

The second inequality comes from the fact that

$$\max_{a\in\mathbf{y}} g_a(x) \geq \frac{1}{|\mathbf{y}|}\sum_{a\in\mathbf{y}} g_a(x).$$

For the third inequality, we use the convexity of ψ:

$$\psi\left(\frac{1}{|\mathbf{y}|}\sum_{a\in\mathbf{y}} g_a(x)\right) \leq \frac{1}{|\mathbf{y}|}\sum_{a\in\mathbf{y}} \psi(g_a(x)).$$

For the tightness proof: When $g_a(x) = \textit{constant}$ over $a \in \mathbf{y}$, we have

$$\psi\left(\max_{a\in\mathbf{y}} g_a(x)\right) = \psi\left(\frac{1}{|\mathbf{y}|}\sum_{a\in\mathbf{y}} g_a(x)\right) = \frac{1}{|\mathbf{y}|}\sum_{a\in\mathbf{y}} \psi\left(g_a(x)\right),$$

implying $\mathcal{L}_\psi^{max}(g(x),\mathbf{y}) = \mathcal{L}_\psi(g(x),\mathbf{y}) = \mathcal{L}_\psi^{naive}(g(x),\mathbf{y})$.

As for the first inequality, we provide a sequence $g^{(n)}$ that verifies equality in the limit: let $g_a^{(n)}(x) = -1/n$ if $a \in \mathbf{y}$, $g_b^{(n)}(x) = 0$ for some $b \notin \mathbf{y}$, and $g_c^{(n)}(x) = -n$ for all $c \notin \mathbf{y}, c \neq b$. Then provided $\psi(0) = 1$ and $\lim_{u\to\infty}\psi(u) = 0$, we have $\lim_{n\to+\infty}\mathcal{L}_\psi^{max}(g^{(n)}(x),\mathbf{y}) = 2$ and for all n, $\mathcal{L}_A(g^{(n)}(x),\mathbf{y}) = 1$. ■

Proof of Proposition 7 (Generalization bounds). The proof uses Definition 11 for Rademacher and Gaussian complexity, Lemma 12, Theorem 13 and Theorem 14 from Bartlett and Mendelson (2002), reproduced below and adapted to our notations for completeness. We apply Theorem 13 with $\mathcal{L} := \frac{1}{L}\mathcal{L}_A$, $\phi := \frac{1}{L}\mathcal{L}_{\psi_\gamma}$:

$$\frac{1}{L}\mathbb{E}_P[\mathcal{L}_A(g(X),\mathbf{Y})] \leq \frac{1}{L}\mathbb{E}_S[\mathcal{L}_{\psi_\gamma}(g(X),\mathbf{Y})] + R_m(\bar{\phi}\circ\mathcal{G}) + \sqrt{\frac{8\log(2/\eta)}{m}}.$$

From Lemma 12, $R_m(\bar{\phi}\circ\mathcal{G}) \leq \frac{1}{c}G_m(\bar{\phi}\circ\mathcal{G})$. From Theorem 14, $G_m(\bar{\phi}\circ\mathcal{G}) \leq 2\lambda\sum_{a=1}^{L}\hat{G}_m(\mathcal{G}_a)$. Let (ν_i) be m independent standard normal random variables.

$$\begin{aligned}
\hat{G}_m(\mathcal{G}_a) &= \mathbb{E}_\nu\left[\sup_{g_a\in\mathcal{G}_a}\frac{2}{m}\sum_i \nu_i g_a(x_i) \mid S\right] = \frac{2}{m}\mathbb{E}_\nu\left[\sup_{||\mathbf{w}_a||\leq B}\mathbf{w}_a\cdot\sum_i\nu_i\mathbf{f}(x_i) \mid S\right] \\
&= \frac{2B}{m}\mathbb{E}_\nu\left[||\sum_i\nu_i\mathbf{f}(x_i)|| \mid S\right] = \frac{2B}{m}\mathbb{E}_\nu\left[\sqrt{\sum_{ij}\nu_i\nu_j\mathbf{f}(x_i)^T\mathbf{f}(x_j)} \mid S\right] \\
&\leq \frac{2B}{m}\sqrt{\mathbb{E}_\nu\left[\sum_{ij}\nu_i\nu_j\mathbf{f}(x_i)^T\mathbf{f}(x_j) \mid S\right]} = \frac{2B}{m}\sqrt{\sum_i\mathbb{E}_\nu\left[\nu_i^2||\mathbf{f}(x_i)||^2 \mid S\right]} \\
&= \frac{2B}{m}\sqrt{\sum_i||\mathbf{f}(x_i)||^2}.
\end{aligned}$$

Putting everything together, $R_m(\bar{\phi}\circ\mathcal{G}) \leq \frac{2\lambda L}{c}\hat{G}_m(\mathcal{G}_a) \leq \frac{4\lambda LB}{mc}\sqrt{\sum_i||\mathbf{f}(x_i)||^2}$ and:

$$\mathbb{E}_P[\mathcal{L}_A(g(X),Y)] \leq \mathbb{E}_S[\mathcal{L}_{\psi_\gamma}(g(X),\mathbf{Y})] + \frac{4\lambda BL^2}{mc}\sqrt{\sum_i||\mathbf{f}(x_i)||^2} + L\sqrt{\frac{8\log(2/\eta)}{m}}.$$

The Lipschitz constant from 14 can be computed as $\lambda := \frac{p}{\gamma}\sqrt{L}$, using the Lipschitz constant of the scalar function ψ_γ, which is $\frac{p}{\gamma}$, and the fact that $||g(x)||_1 \leq \sqrt{L}||g(x)||_2$. ■

Definition 11 (Definition 2 from Bartlett and Mendelson (2002)) *Let μ be a probability distribution on a set $\mathcal{X}$ and suppose that $S = \{x_i\}_{i=1}^m$ are independent samples sampled from μ. Let $\mathcal{G}$ be a class of functions $\mathcal{X}\to\mathbb{R}$. Define the random variables*

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{2}{m}\sum_i\sigma_i f(x_i) \mid S\right],$$
$$\hat{G}_m(\mathcal{F}) = \mathbb{E}_\nu\left[\sup_{f\in\mathcal{F}}\frac{2}{m}\sum_i\nu_i f(x_i) \mid S\right],$$

where (σ_i) are m independent uniform $\{\pm1\}$-valued random variables and (ν_i) are m independent standard normal random variables. Then the Rademacher (resp. Gaussian) complexity of $\mathcal{G}$ is $R_m(\mathcal{F}) = \mathbb{E}_S[\hat{R}_m(\mathcal{F})]$ (resp. $G_m(\mathcal{F}) = \mathbb{E}_S[\hat{F}_m(\mathcal{F})]$).

$R_m(\mathcal{F})$ and $G_m(\mathcal{F})$ quantify how much can a $f\in\mathcal{F}$ be correlated with a noise sequence of length m.

Lemma 12 (Lemma 4 from Bartlett and Mendelson (2002)) *There are absolute constants c and C such that for every class $\mathcal{G}$ and every integer m,*

$$cR_m(\mathcal{G}) \leq G_m(\mathcal{G}) \leq C\log mR_m(\mathcal{G}).$$

Theorem 13 (Theorem 8 from Bartlett and Mendelson (2002)) *Consider a loss function $\mathcal{L}: A \times \mathcal{Y} \mapsto [0,1]$ and a dominating cost function $\phi: A \times \mathcal{Y} \to [0,1]$, where A is an arbitrary output space. Let $\mathcal{G}$ be a class of functions mapping from $\mathcal{X}$ to A and let $S = \{(x_i, y_i)\}_{i=1}^m$ be independently selected according to the probability measure P. Define $\bar{\phi} \circ \mathcal{G} = \{(x,y) \mapsto \phi(g(x),y) - \phi(0,y) : g \in \mathcal{G}\}$. Then, for any integer m and any $\eta \in (0,1)$, with probability at least $1-\eta$ over samples of length m, $\forall g \in \mathcal{G}$:*

$$\mathbb{E}_P[\mathcal{L}(g(X),Y)] \leq \mathbb{E}_S \phi(g(X),y) + R_m(\bar{\phi} \circ \mathcal{G}) + \sqrt{\frac{8 \log(2/\eta)}{m}}.$$

Theorem 14 (Theorem 14 from Bartlett and Mendelson (2002)) *Let $A = \mathbb{R}^L$, and let $\mathcal{G}$ be a class of functions mapping $\mathcal{X}$ to A. Suppose that there are real-valued classes $\mathcal{G}_1, \ldots, \mathcal{G}_L$ such that $\mathcal{G}$ is a subset of their direct sum. Assume further that $\phi: A \times \mathcal{Y} \to \mathbb{R}$ is such that, for all $y \in \mathcal{Y}$, $\phi(\cdot, y)$ is a Lipschitz function (with respect to Euclidean distance on A) with constant λ which passes through the origin and is uniformly bounded. For $g \in \mathcal{G}$, define $\phi \circ g$ as the mapping $(x,y) \mapsto \phi(g(x),y)$. Then, for every integer m and every sample $S = \{(x_i, y_i)\}_{i=1}^m$,*

$$\hat{G}_m(\phi \circ \mathcal{G}) \leq 2\lambda \sum_{a=1}^{L} \hat{G}_m(\mathcal{G}_a),$$

where $\hat{G}_m(\phi \circ \mathcal{G})$ are the Gaussian averages of $\phi \circ \mathcal{G}$ with respect to the sample $\{(x_i, y_i)\}_{i=1}^m$ and $\hat{G}_m(\mathcal{G}_a)$ are the Gaussian averages of $\mathcal{G}_a$ with respect to the sample $\{x_i\}_{i=1}^m$.

Proof of Proposition 8 (Generalization bounds on true loss). This follows from Propositions 7 and 1. ■

Proof of Lemma 9. Let us write $\mathbf{z} = \mathbf{z}(x), y = y(x), \mathbf{y} = \mathbf{y}(x)$.

- Let $a \in \mathbf{z}$. By hypothesis, $\exists x' \in B_\eta(x) : g_a(x') \leq -\frac{\eta}{2}$. By definition of $B_\eta(x)$,

  $$g_a(x) = g_a(x') + \mathbf{w}_a \cdot (\mathbf{f}(x) - \mathbf{f}(x')) \leq g_a(x') + ||\mathbf{w}_a||^* \eta \leq g_a(x') + \eta \leq \frac{\eta}{2}.$$

  In fact, we also have $g_a(x) < \frac{\eta}{2}$, by considering two cases ($\mathbf{w}_a = 0$ or $\mathbf{w}_a \neq 0$) and using the fact that $||\mathbf{f}(x) - \mathbf{f}(x')|| < \eta$.

- Let $a \notin \mathbf{y}$. Since $\mathcal{L}_\psi(g(x), \mathbf{y}) \leq \psi(\eta/2)$ and each term is nonnegative, we have:

  $$\psi(-g_a(x)) \leq \psi(\frac{\eta}{2}) \implies g_a(x) \leq -\frac{\eta}{2}.$$

- Let $a = y$. $\mathcal{L}_\psi(g(x), \mathbf{y}) \leq \psi(\eta/2)$ also implies the following:

  $$\begin{aligned} & \psi\left(\tfrac{1}{|\mathbf{y}|} \textstyle\sum_{b \in \mathbf{y}} g_b(x)\right) && \leq \psi(\frac{\eta}{2}) \\ \implies \quad & \tfrac{1}{|\mathbf{y}|} \textstyle\sum_{b \in \mathbf{y}} g_b(x) && \geq \frac{\eta}{2} \\ \implies \quad & g_y(x) && \geq \frac{|\mathbf{y}|\eta}{2} - \sum_{b \in \mathbf{z}} g_b(x) \\ & && > \frac{|\mathbf{y}|\eta}{2} - \frac{|\mathbf{z}|\eta}{2} = \frac{\eta}{2}. \end{aligned}$$

Finally, $\forall a \neq y, g_a(x) < g_y(x)$ and g classifies x correctly. ■

Proof of corollary 10. Let $a \in \mathbf{z}(x)$, by the empty intersection hypothesis, $\exists i \geq 1 : a \notin \mathbf{z}(x_i)$ and since $y(x_i) = y(x)$ and $a \neq y(x)$ we also have $a \notin \mathbf{y}(x_i)$. Since $\mathcal{L}_\psi(g(x_i), \mathbf{y}(x_i) \leq \psi(\eta/2)$, we have $g_a(x_i) \leq -\frac{\eta}{2}$, as in the previous proof. We can apply Lemma 9 (with $x' = x_i$). ■

References

C. Ambroise, T. Denoeux, G. Govaert, and P. Smets. Learning from an imprecise teacher: Probabilistic and evidential approaches. In *Applied Stochastic Models and Data Analysis*, volume 1, pages 100–105, 2001.

S. Andrews and T. Hofmann. Multiple instance learning via disjunctive programming boosting. In *Advances in Neural Information Processing Systems*, 2004.

A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

K. Barnard, P. Duygulu, D.A. Forsyth, N. de Freitas, D.M. Blei, and M.I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

T.L. Berg, A.C. Berg, J.Edwards, M.Maire, R.White, Y.W. Teh, E.G. Learned-Miller, and D.A. Forsyth. Names and faces in the news. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 848–854, 2004.

M.R. Boutell, J. Luo, X. Shen, and C.M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

P. E. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2006.

E. Côme, L. Oukhellou, T. Denœux, and P. Aknin. Mixture model estimation with soft labels. *International Conference on Soft Methods in Probability and Statistics*, 2008.

T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *Proc. European Conference on Computer Vision*, 2008.

T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2002.

T.G. Dietterich, R.H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

P. Duygulu, K. Barnard, J.F.G. de Freitas, and D.A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. European Conference on Computer Vision*, pages 97–112, 2002.

M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy – automatic naming of characters in tv video. In *British Machine Vision Conference*, 2006.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.

A.C. Gallagher and T. Chen. Using group prior to identify people in consumer images. In *CVPR Workshop on Semantic Learning Applications in Multimedia*, 2007.

Y. Grandvalet and Y. Bengio. Learning from partial labels with minimum entropy. *Centre interuniversitaire de recherche en analyse des organisations (CIRANO)*, 2004.

G.B. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *Proc. International Conference on Computer Vision*, 2007a.

G.B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007b.

E. Hullermeier and J. Beringer. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5):419–439, 2006.

R. Jin and Z. Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems*, pages 897–904, 2002.

H. Kuck and N. de Freitas. Learning about individuals from group statistics. In *Uncertainty in Artificial Intelligence*, 2005.

I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

J. Luo and F. Orabona. Learning from candidate labeling sets. In *Advances in Neural Information Processing Systems*, 2010.

P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence*, pages 374–388, 1976.

P.J. Moreno, C. Joerg, J.M.V. Thong, and O. Glickman. A recursive algorithm for the forced alignment of very long audio segments. In *International Conference on Spoken Language Processing*, 1998.

J.G. Proakis and D.G. Manolakis. *Digital signal processing: principles, algorithms, and applications*. Prentice Hall, 1996.

N. Quadrianto, A.J. Smola, T.S. Caetano, and Q.V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009. ISSN 1532-4435.

D. Ramanan, S. Baker, and S. Kakade. Leveraging archival video for building face datasets. In *Proc. International Conference on Computer Vision*, 2007.

R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE MultiMedia*, 6(1):22–35, 1999.

K. Sjölander. An HMM-based system for automatic segmentation and alignment of speech. In *Fonetik*, pages 93–96, 2003.

D. Talkin. A robust algorithm for pitch tracking (RAPT). *Speech Coding and Synthesis*, pages 495–518, 1995.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. In *International Conference on Learning Theory*, volume 3559, pages 143–157, 2005.

G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*, pages 667–685, 2010.

P. Vannoorenberghe and P. Smets. Partially supervised learning by a credal EM approach. In *European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 956–967, 2005.

P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. *Advances in Neural Information Processing Systems*, 18:1417, 2006.

R. Yan., J. Zhang, J. Yang, and A.G. Hauptmann. A discriminative learning framework with pairwise constraints for video object classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):578–593, 2006.

T. Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004. ISSN 1533-7928.

Z.H. Zhou and M.L. Zhang. Multi-instance multi-label learning with application to scene classification. *Advances in Neural Information Processing Systems*, 19:1609, 2007.

X. Zhu and A.B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.