Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Theoretical Analysis of Weak-to-Strong Generalization

Qijie Zhu

Northwestern University

qijiezhu2029@u.northwestern.edu

10/18/2024

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Introduction

- Weakly-supervised learning allows practitioners to train models with possibly-incorrect, easy-to-obtain pseudolabels instead of accurate and expensive ground-truth labels.
- **Example** Classify documents based on: text x_i has positive or negative sentiment? Weakly-supervised learning enables models to learn from simple rules like if 'incredible' $\in x_i$, it is positive, else it is negative. Let $\mathcal{X}$ be the space of text documents and $\mathcal{Y} = \{-1, 1\}$. The weak label $\tilde{y}$ can be defined as:

$$\tilde{y}(x) = \begin{cases} +1, & \text{if 'incredible'} \in x, \\ -1, & \text{if 'horrible'} \in x, \\ \varnothing, & \text{otherwise.} \end{cases}$$

- **Fact** The student model often outperforms its "teacher".

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Introduction

- In the example, pseudolabels (teacher model) gives the pseudolabels $\tilde{y}(x)$ for $x \in \mathcal{X}$.
- Note that
  - it may make errors. i.e. $\exists x \in \mathcal{X}$ s.t. $\tilde{y}(x) \neq y(x)$
  - it may not cover every points. i.e. $\exists x \in \mathcal{X}$, s.t. $\tilde{y}(x) = \varnothing$
- it seems that a powerful enough classifier should exactly fit the pseudolabeler on the covered data and have trivial performance on the uncovered data. However, this is not what happens in practice.
- **Two outcomes:** (a) Pseudolabel correction: The performance of the model exceeds the performance of the pseudolabels used to train it; and (b) Coverage expansion: The model performs well even on the portion of example space X that is not covered by pseudolabels. These empirical outcomes are key to the success of weak supervision.

Symbols and Definitions

- x: A random variable with distribution $\mathcal{D}$.
- $\mathcal{X}$: Input space (assumed to be discrete but possibly large, e.g., $\mathbb{R}^d$).
- $y : \mathcal{X} \to \mathcal{Y} = \{1, \ldots, k\}$: Ground-truth label function.
- $\tilde{y} : \mathcal{X} \to \mathcal{Y} \cup \{\varnothing\}$: Pseudolabeler assigning each x a label in $\mathcal{Y}$ or abstention symbol $\varnothing$.
- $S = \{x \mid \tilde{y}(x) \neq \varnothing\}$: Covered set, where pseudolabels are assigned.
- $T = \{x \mid \tilde{y}(x) = \varnothing\} = \mathcal{X} \setminus S$: Uncovered set, without pseudolabels.
- $\{\mathcal{X}_i\}$: Partition of $\mathcal{X}$, where within each $\mathcal{X}_i$, the ground-truth label is constant.
- $S_i = S \cap \mathcal{X}_i$: Covered subset of $\mathcal{X}_i$.
- $T_i = T \cap \mathcal{X}_i$: Uncovered subset of $\mathcal{X}_i$.
- $S_i^{\text{good}} = \{x \in S_i \mid \tilde{y}(x) = y(x)\}$: Correctly-pseudolabeled examples.
- $S_i^{\text{bad}} = S_i \setminus S_i^{\text{good}}$: Incorrectly-pseudolabeled examples.
- $\alpha_i := \mathbb{P}(S_i^{\text{bad}} \mid S_i)$: Error rate of $\tilde{y}$ on S_i, assumed to satisfy $0 < \alpha_i < \frac{1}{2}$.

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Problem Setup

- **Goals:** For two classifiers $f, g : \mathcal{X} \to \mathcal{Y}$ and a set $U \subset \mathcal{X}$, the probability of disagreement is:

$$\mathrm{err}(f, g \mid U) = \mathbb{P}_{x \sim \mathcal{D}}(f(x) \neq g(x) \mid x \in U),$$

  We focus on minimizing: $\arg\min_{f \in \mathcal{F}} \mathrm{err}(f, \tilde{y} \mid S)$. The ultimate goal is to upper bound the error $\mathrm{err}(f, y \mid \mathcal{X})$, i.e., the error of a classifier f on the *true labels* over the entire space $\mathcal{X}$.
- **Key Challenges:**
  1. **Training with $\tilde{y}$:** Classifier is trained using $\tilde{y}$, and $\tilde{y}$ may contain arbitrary errors not captured by common noise models (e.g., class-conditional noise).
  2. **Performance on Entire $\mathcal{X}$:** Focus is on f's performance over $\mathcal{X}$, though training samples come from $S \subset \mathcal{X}$.
- **Fact** We can give an upper bound for $\mathrm{err}(f, y \mid \mathcal{X})$, by posing some assumption on the structure of the dataset and the property of classifiers.

Presentation

A. Uthor

Introduction

Notation

Problem Setup

Shortcomings of Existing Bounds

Definitions

Theorem 4.1 and 4.2

Revised Definitions

Theorem 4.3

Experiments

Text Classification Example: Let $\mathcal{X}$ be the space of text documents and $\mathcal{Y} = \{-1, 1\}$. The weak label $\tilde{y}$ is defined as:

$$\tilde{y}(x) = \begin{cases} +1, & \text{if 'incredible'} \in x, \\ -1, & \text{if 'horrible'} \in x, \\ \varnothing, & \text{otherwise.} \end{cases}$$

Assumptions:

- 'Incredible' and 'horrible' never co-occur, so $\tilde{y}$ is well-defined.
- $\mathcal{F}$ is the class of bag-of-words classifiers.

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Proposition 3.1: Suppose the label marginals for the above example satisfy $\mathbb{P}(y = Y) = \frac{1}{2}$ for $Y \in \{-1, 1\}$, and assume that the weak label error rates $\alpha_{-1} = \alpha_1 = \alpha$, and that the weak labels cover each class equally often: $\mathbb{P}(\tilde{y} = \varnothing | y = Y) = \mathbb{P}(\tilde{y} = \varnothing)$. Let $\tilde{f} = \min_{f \in \mathcal{F}} \text{err}(f, \tilde{y}|S)$ be the classifier minimizing the weak label error on the covered set. Then the bound from Fu et al. simplifies (in our notation) to:

$$\text{err}(\tilde{f}, y) \leq \mathbb{P}(S) \cdot 4\alpha(1 - \alpha) + \mathbb{P}(T).$$

Interpretation: The error bound has two terms:

- The first term accounts for error on S, which has weight $4\alpha(1 - \alpha)$. Note that $4\alpha(1 - \alpha) > \alpha, \alpha < \frac{3}{4}$, so it doesn't allow for pseudolabel correction. $\text{err}(\tilde{f}, y|B) < 4\alpha(1 - \alpha)$
- The second term accounts for error on T. Worse than random guess. $\text{err}(\tilde{f}, y|T) \leq 1$

Presentation

A. Uthor

Introduction

Notation

Problem Setup

Shortcomings of Existing Bounds

Definitions

Theorem 4.1 and 4.2

Revised Definitions

Theorem 4.3

Experiments

Definition 1 (Neighborhood)

Let $\mathcal{N}$ be a neighborhood function that maps each point x to a set of points $\mathcal{N}(x) \subseteq \mathcal{X}$ that we call the neighborhood of x. We assume $\mathcal{N}$ satisfies $x \in \mathcal{N}(x') \iff x' \in \mathcal{N}(x)$, i.e., the neighborhoods are symmetric. We extend $\mathcal{N}$ to a function of sets as $\mathcal{N}(A) = \cup_{x \in A} \mathcal{N}(x)$.
Examples: $\mathcal{N}(x) = \{x' : \|\phi(x) - \phi(x')\| \leq r\}$ for some representation $\phi : \mathcal{X} \to \mathbb{R}^d$, or for text inputs, the set of fluent paraphrases of x.

Definition 2 (η-robust)

For any classifier f and point x, define $r(f, x) = \mathbb{P}(f(x') \neq f(x) \mid x' \in \mathcal{N}(x))$ as the probability f assigns different labels to x and a random neighbor x' of x. A classifier $f : \mathcal{X} \to \mathcal{Y}$ is said to be η-robust at a point x if $r(f, x) \leq \eta$. Define $R_\eta(f) = \{x : r(f, x) \leq \eta\}$ as the set of η-robust points for f
If $\eta = 0$, f is η-robust at x if and only if f is adversarially robust over $\mathcal{N}(x)$, thus generalizing adversarial robustness. By Markov's inequality, any classifier f with:

$$\mathbb{E}_{x,x' \sim D|\mathcal{N}(x)}[\mathbb{P}(f(x') \neq f(x))] \leq \gamma$$

is η-robust on a set of probability at least $1 - \gamma/\eta$.

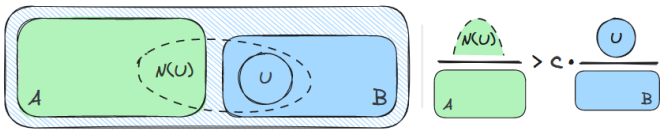
Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Definition 3 (Expansion)

Fix sets $A, B \subset \mathcal{X}$. We say the distribution $\mathbb{P}_x$ satisfies (c, q)-expansion on (A, B) if for all subsets $U \subset B$, $\mathbb{P}(U|B) > q$ and $\mathbb{P}(\mathcal{N}(U)|A) > c\mathbb{P}(U|B)$.
Fix sets $A, B \subset \mathcal{X}$ and suppose M is a collection *of* subsets of B. Then we say M satisfies (c, q)-expansion on (A, B) *if all sets* $U \in \mathcal{M}$ with $\mathbb{P}(U|B) > q$ satisfy $\mathbb{P}(\mathcal{N}(U)|A) > c\mathbb{P}(U|B)$.

Definition 4 (Adversarial Robustness)

A classifier f is said to be adversarially robust at a point x if for all $x' \in \mathcal{N}(x)$, $f(x) = f(x')$. This means that small perturbations of the input do not change the output label, i.e., the classifier's decision is consistent within the neighborhood of x.



Graphical representation of relative expansion for sets A and B.

Presentation

A. Uthor

Introduction

Notation

Problem Setup

Shortcomings of Existing Bounds

Definitions

Theorem 4.1 and 4.2

Revised Definitions

Theorem 4.3

Experiments

Definition 5 (Expanding Families)

We first define the families $\mathcal{M}$ and $\mathcal{M}'$ of sets that must expand according to Definition 4. Let $\mathcal{F}$ be the hypothesis class of the strong model and for each $f \in \mathcal{F}$, define:

$$R(f) = R_0(f) = \{x : r(f, x) = 0\}$$

to be the set of adversarially robust points for f. For $B \subset \mathcal{X}$ and $f \in \mathcal{F}$, define:

$$U(B, f) = \{x \in B \cap R(f) : f(x) \neq y(x)\}$$

as the set of robust points in B where f makes a mistake on the true label y. Now define $\mathcal{M}(B, \mathcal{F})$ to be the class of these robust mistake sets: $\mathcal{M}(B, \mathcal{F}) = \{U(B, f) : f \in \mathcal{F}\}$
Similarly, $\mathcal{M}'(B, \mathcal{F}) = \{(B \setminus U(B, f)) \cap R(f) : f \in \mathcal{F}\}$

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Theorem 4.1 (Pseudolabel Correction)

Suppose $\mathcal{M}'(S_i^{\text{good}}, \mathcal{F})$ satisfies (c, q)-expansion on the sets $(S_i^{\text{bad}}, S_i^{\text{good}})$ for $q < \frac{3}{4}(1 - 2\alpha)$. Consider an arbitrary classifier $f \in \mathcal{F}$ such that:

$$\mathbb{P}(f(x) \neq \tilde{y}(x) \text{ or } f \text{ not robust at } x \mid S_i) \leq \frac{1 - \alpha + 3\alpha c}{4}.$$

Then the true error of f on S_i satisfies:

$$\text{err}(f, y \mid S_i) \leq \frac{2\alpha_i}{1 - 2\alpha_i} \mathbb{P}(\overline{R(f)} \mid S_i) + \text{err}(f, \tilde{y} \mid S_i) + \alpha_i \left(1 - \frac{3}{2}c\right).$$

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Explanation 1: (c, q)-Expansion

- We assume the collection of sets $\mathcal{M}'(S_i^{\text{good}}, \mathcal{F})$ satisfies (c, q)-expansion on the sets $(S_i^{\text{bad}}, S_i^{\text{good}})$.
- i.e. $\forall U \in \mathcal{M}'(S_i^{\text{good}}, \mathcal{F})$, if $\mathbb{P}(U|S_i^{\text{good}}) > q$, then $\mathbb{P}(\mathcal{N}(U)|S_i^{\text{bad}}) > c\mathbb{P}(U|S_i^{\text{good}})$

Explanation 2: Robustness and Weak Labels

- We consider an arbitrary classifier $f \in \mathcal{F}$ such that $\mathbb{P}(f(x) \neq \tilde{y}(x)$ or f not robust at $x \mid S_i)$: here f not being robust means $x \in \overline{R(f)}$

Presentation

A. Uthor

Introduction

Notation

Problem Setup

Shortcomings of Existing Bounds

Definitions

Theorem 4.1 and 4.2

Revised Definitions

Theorem 4.3

Experiments

The Result of Theorem 4.1

$$\text{err}(f, y \mid S_i) \leq \frac{2\alpha_i}{1-2\alpha_i}\mathbb{P}(\overline{R(f)} \mid S_i) + \text{err}(f, \tilde{y} \mid S_i) + \alpha_i \left(1 - \frac{3}{2}c\right).$$

- We have the trivial error bound $\text{err}(f, y \mid S_i) \leq err(f, \tilde{y}|S_i) + \alpha_i$
- The upper bound in Theorem 4.1 is much tighter, when c is large, $\mathbb{P}(\overline{R(f)} \mid S_i)$ and $\text{err}(f, \tilde{y} \mid S_i)$ are small

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

- Let $M_i = \{x \in S_i : f(x) \neq y(x)\}$ be the set of mistakes of f on the true labels in S_i. Similarly, let $D_i = \{x \in S_i : f(x) \neq \tilde{y}(x)\}$ be the set of mistakes of f on the weak labels in S_i. Define $U_i = S_i \setminus M_i$ and let $V_i = R(f) \cap U_i \cap S_i^{\text{good}}$.
- Note that $V_i \in \mathcal{M}'(S_i^{\text{good}}, \mathcal{F})$. We can try to show that V_i is large enough to expand, i.e. to prove $\mathbb{P}(V_i \mid S_i^{\text{good}}) > q$. It follows $P(N(U)|S_i^{\text{good}}) > c\mathbb{P}(V_i \mid S_i^{\text{good}})$
- We define $\mathcal{N}'(A) \subset \mathcal{N}(A)$ as the subset of points reachable from A by a good edge $f(x) = f(x')$. We can show $\mathcal{N}'(A) \subset U_i$, and we can prove

$$\mathbb{P}(U_i \mid S_i^{\text{bad}}) \geq c\mathbb{P}(V_i \mid S_i^{\text{good}}).$$

- This inequality thus guarantees that there must be some points that f gets correct and y gets incorrect. It also gives a quantitative lower bound on how much probability is assigned to these points. What remains is deriving the final bound.

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Theorem 4.2 (Error Bound for Uncovered Points)

Suppose $\mathcal{M}(T_i, \mathcal{F})$ satisfies (c, q)-expansion on (S_i^{good}, T_i), and $\mathcal{M}'(T_i, \mathcal{F})$ satisfies (c, q)-expansion on (S_i^{bad}, T_i). Consider an arbitrary classifier $f \in \mathcal{F}$ that fits the weak labels well on S_i and is fairly robust on T_i:

$$\text{err}(f, \tilde{y} \mid S_i) + \mathbb{P}(\overline{R(f)} \mid T_i) < c(1 - q - \alpha_i).$$

Then the true error of f on T_i satisfies:

$$\text{err}(f, y \mid T_i) \leq \left(1 + \frac{\alpha_i}{1 - 2\alpha_i}\right) \mathbb{P}(\overline{R(f)} \mid T_i) + \max\left(q, \frac{\text{err}(f, \tilde{y} \mid S_i) - c\alpha_i}{c(1 - 2\alpha_i)}\right).$$

Presentation

A. Uthor

Introduction

Notation

Problem Setup

Shortcomings of Existing Bounds

Definitions

Theorem 4.1 and 4.2

Revised Definitions

Theorem 4.3

Experiments

The Result of Theorem 4.2

$$\mathrm{err}(f, y \mid T_i) \leq \left(1 + \frac{\alpha_i}{1 - 2\alpha_i}\right) \mathbb{P}(\overline{R(f)} \mid T_i) + \max\left(q, \frac{\mathrm{err}(f, \tilde{y} \mid S_i) - c\alpha_i}{c(1 - 2\alpha_i)}\right).$$

- f must fit the weak labels well on S_i, so $\mathrm{err}(f, \tilde{y} \mid S_i)$ is small.
- f must be adversarially robust at most points on T_i, so that $\mathbb{P}(\overline{R(f)} \mid T_i)$ is small.
- c have the same influence on the upper bound.

Presentation
A. Uthor
Introduction
Notation
Problem Setup
Shortcomings of Existing Bounds
Definitions
Theorem 4.1 and 4.2
Revised Definitions
Theorem 4.3
Experiments

Definition 6 (Example graph)

Let $G = (X, E)$ be a graph with one node for each element of X (we assume X is a possibly very large, but finite, set), and connect two nodes (x, x') if $x \in \mathcal{N}(x')$ or equivalently, if $x' \in \mathcal{N}(x)$, with an edge weight of $w(x, x') := \mathbb{P}(x)\mathbb{P}(x')\mathbf{1}[x \in \mathcal{N}(x')]$.

Definition 7 (η-robust neighborhood size)

Let $A, U \subset \mathcal{X}$. The size of the η-robust neighborhood of U in A is:

$$P_{1-\eta}(U, A) := \min_{V \subset A} \{\mathbb{P}(v \mid A) : w(V, U) \geq (1 - \eta) w(\mathcal{N}(U), U)\}.$$

$P_{1-\eta}(U, A)$ is the probability of the "smallest" subset of A that still captures at least a $1 - \eta$ fraction of the edge weight incident on U. When $\eta = 0$, we have $P_1(U, A) = \mathbb{P}(\mathcal{N}(U) \mid A)$.

Definition 8 (Robust expansion)

Fix sets $A, B \subset \mathcal{X}$ and suppose $\mathcal{M}$ is a collection of subsets of B. $\mathcal{M}$ satisfies (c, q, η)-robust expansion on (A, B) if for all $U \in \mathcal{M}$ with $\mathbb{P}(U \mid B) > q$, $P_{1-\eta}(U, A) \geq c\mathbb{P}(U \mid B)$. This recovers Definition 3 when $\eta = 0$.

Presentation

A. Uthor

Introduction

Notation

Problem Setup

Shortcomings of Existing Bounds

Definitions

Theorem 4.1 and 4.2

Revised Definitions

Theorem 4.3

Experiments

Theorem 4.3 (Informal)

Theorems 4.1 and 4.2 hold exactly with (c, q, η)-expansion instead of (c, q)-expansion and $R_\eta(f)$ instead of $R(f)$.

Model	i	(S_i^{bad}, S_i^{good}) exp.	α_i	$\text{err}(f, \tilde{y}\|S_i)$	Bound val	$\text{err}(f, y\|S_i)$
SentenceBERT	0	0.848	0.11	0.12	0.05	0.04
	1	0.497	0.33	0.29	0.37	0.35

Table 2: Measured expansion values and error bounds for the uncovered sets T_i. Expansion values for the families $\mathcal{M}(S_i^{bad}, \mathcal{F})$ on (S_i^{good}, S_i^{bad}) and $\mathcal{M}'(S_i^{good}, \mathcal{F})$ on (S_i^{bad}, S_i^{good}), are measured using the heuristic described in Section 5. The detection of *both* types of expansion (expansion from T_i to S_i^{good} *and* to S_i^{bad}) gives evidence for the extra structure we described in Section 4.1 and justifies our use of Theorem 4.2, which uses this structure, instead of Theorem B.3, which only uses expansion from T_i to S_i^{good} and gives a looser bound. Worst-case value of the error bound in Theorem 4.2 (specifically, the tighter version B.2, which allows for different amounts of expansion between S_i^{good}/T_i and S_i^{bad}/T_i), computed using the smallest expansion values and largest weak errors $\text{err}(f, \tilde{y}|S_i)$ from the 5 training runs. The $\text{err}(f, \tilde{y}|S_i)$ and α_i values used in the bound computation are identical to the values in Table 1. Unlike in Table 1, where the "baseline" for pseudolabel correction effects is to have error bounds strictly better than α_i, for coverage expansion, the more relevant comparison is against random/arbitrary guessing. The actual worst-case error of the student on each T_i is shown as $\text{err}(f, \tilde{y}|T_i)$. As suggested by our bound values, the errors on each T_i are non-trivial (much better than random or arbitrary guessing).

Model	i	(S_i^{good}, T_i)	(S_i^{bad}, T_i)	$\text{err}(f, \tilde{y}\|S_i)$	Bd. val	$\text{err}(f, y\|T_i)$
SentenceBERT	0	0.16	0.98	0.12	0.37	0.16
	1	0.75	0.55	0.29	0.33	0.29

Measured expansion and error bounds for the covered and uncovered sets S_i and T_i. The results are detailed in Tables 1 and 2, demonstrating the error bounds and expansion values for models like SentenceBERT.