Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels

Yu-Ting Chou, Gang Niu, Hsuan-Tien Lin, Masashi Sugiyama

ICML 2020 work done during Chou’s internship at RIKEN AIP, Japan; resulting M.S. thesis of Chou won the 2020 thesis award of TAAI

October 30, 2021, SSC, Kaohsiung, Taiwan
Supervised Learning
(Slide Modified from My ML Foundations MOOC)

unknown target function \( f : \mathcal{X} \rightarrow \mathcal{Y} \)

training examples \( \mathcal{D} : (x_1, y_1), \ldots, (x_N, y_N) \)

learning algorithm \( \mathcal{A} \)

final hypothesis \( g \approx f \)

hypothesis set \( \mathcal{H} \)

supervised learning:
every input vector \( x_n \) with
its (possibly expensive) label \( y_n \),
Weakly-supervised: Learning without True Labels $y_n$

- positive-unlabeled: some of true $y_n = +1$ revealed
- complementary: ‘not label’ $\bar{y}_n$ instead of true $y_n$
- noisy: noisy label $y'_n$ instead of true $y_n$

**weakly-supervised:** a realistic and hot research direction to reduce labeling burden

[EN08] Learning classifiers from only positive and unlabeled data, KDD’08.
[Ish+17] Learning from complementary labels, NeurIPS’17.
Introduction

Motivation

**popular weakly-supervised models [DNS15; Ish+19; Pat+17]**

- derive **Unbiased Risk Estimators (URE)** as new loss
- theoretically, nice properties (unbiased, consistent, etc.) [Ish+17]
- practically, sometimes **bad performance** (overfitting)

**our contributions: on Learning w/ Complementary Labels (LCL)**

- analysis: **identify weakness** of URE framework
- algorithm: propose an **improved framework**
- experiment: demonstrate **stronger performance**

next: introduction to LCL

---

Motivation behind LCL

complementary label $\bar{y}_n$ instead of true $y_n$

Figure 1 of [Yu+18]

complementary label: easier/cheaper to obtain for some applications
Fruit Labeling Task (Image from AI Cup in 2020)

hard: true label
• orange ?
• mango ?

cherry
banana

easy: complementary label
• orange
• mango
• cherry
• banana

complementary: less labeling

cost/expertise required
Comparison

Ordinary (Supervised) Learning

Training: \( \{(x_n, y_n = \text{mango})\} \rightarrow \text{classifier} \)

Complementary Learning

Training: \( \{(x_n, \overline{y}_n = \text{banana})\} \rightarrow \text{classifier} \)

Testing goal: \( \text{classifier} (\ ) \rightarrow \text{cherry} \)

Ordinary versus complementary: same goal via different training data
Learning with Complementary Labels Setup

**Given**

$N$ examples (input $x_n$, complementary label $\overline{y}_n) \in \mathcal{X} \times \{1, 2, \ldots K\}$ in data set $\mathcal{D}$ such that $\overline{y}_n \neq y_n$ for some hidden ordinary label $y_n \in \{1, 2, \ldots K\}$.

**Goal**

a multi-class classifier $g(x)$ that **closely predicts** (0/1 error) the ordinary label $y$ associated with some **unseen** inputs $x$

LCL model design: connecting **complementary & ordinary**
Unbiased Risk Estimation for LCL

Ordinary Learning

- empirical risk minimization (ERM) on training data

\[
\text{risk: } \mathbb{E}_{(x,y)}[\ell(y, g(x))] \quad \text{empirical risk: } \mathbb{E}_{(x_n,y_n) \in \mathcal{D}}[\ell(y_n, g(x_n))]
\]

- loss \( \ell \): usually surrogate of 0/1 error

LCL [Ish+19]

- rewrite the loss \( \ell \) to \( \overline{\ell} \), such that

\[
\text{unbiased risk estimator: } \mathbb{E}_{(x,\overline{y})}[\overline{\ell}(\overline{y}, g(x))] = \mathbb{E}_{(x,y)}[\ell(y, g(x))]
\]

- LCL by minimizing URE

URE: pioneer models for LCL
Example of URE

Cross Entropy Loss

for \( g(x) = \text{argmax}_{k \in \{1, 2, ..., K\}} p(k \mid x) \),

- \( \ell_{CE} \): derived by maximum likelihood as surrogate of 0/1 risk:
  \[
  R(g; \ell_{CE}) = \mathbb{E}_{(x,y)} \left( - \log(p(y \mid x)) \right)
  \]

Complementary Learning [Ish+19]

URE:

\[
\bar{R}(g; \ell) = \mathbb{E}_{(x,\bar{y})} \left[ (K - 1) \log(p(\bar{y} \mid x)) - \sum_{k=1}^{K} \log(p(k \mid x)) \right]
\]

under uniform \( \bar{y} \) assumption

ERM with URE: \( \min_{p} \bar{R} \) with \( \mathbb{E} \) taken on \( \mathcal{D} \)
Problems of URE

URE overfits on single label

\[ \ell = - \log(p(y \mid x)) \]

\[ \bar{\ell} = (K - 1) \log(p(\bar{y} \mid x)) - \sum_{k=1}^{K} \log(p(k \mid x)) \]

ordinary risk and URE very different

- \( \ell > 0 \) → ordinary risk non-negative
- small \( p(\bar{y} \mid x) \) (often) → possibly very negative \( \bar{\ell} \)

empirical URE can be negative: observing some but not all \( \bar{y} \)

- negative empirical URE drags minimization towards overfitting

practical remedy: [Ish+19]

NN-URE: constrain empirical URE to be non-negative

how can we avoid negative empirical URE?
Proposed Framework

Minimize Complementary 0/1

- Recall the goal: minimize 0-1 loss, not $\ell$
- The unbiased estimator of $R_{01}$

$$
\overline{R}_{01} : \mathbb{E}_y[\ell_{01}(\overline{y}, g(x))] = \ell_{01}(y, g(x))
$$

- We denote $\overline{\ell}_{01}$ as the complementary 0-1 loss:

$$
\overline{\ell}_{01}(\overline{y}, g(x)) = [\overline{y} = g(x)]
$$

Surrogate Complementary Loss (SCL)

- Surrogate loss to optimize $\overline{\ell}_{01}$
- Unify previous work as surrogates of $\overline{\ell}_{01}$ [Yu+18; Kim+19]

---

[Yu+18] Learning with biased complementary labels, ECCV’18.

Negative Risk Avoided

Unbiased Risk Estimator (URE)

URE loss $\bar{\ell}_{CE}$ [Ish+19] from cross-entropy $\ell_{CE}$,

$$
\bar{\ell}_{CE}(\hat{y}, g(x)) = (K - 1) \log(p(\hat{y} \mid x)) - \sum_{j=1}^{K} \log(p(j \mid x))
$$

can go negative.

Surrogate Complementary Loss (SCL)

a surrogate of $\bar{\ell}_{01}$ [Kim+19]

$$
\phi_{NL}(\hat{y}, g(x)) = -\log(1 - p(\hat{y} \mid x)))
$$

remains non-negative.
Illustrative Difference between URE and SCE

**URE:** Ripple effect of errors

- Theoretical motivation [Ish+17]
- Estimation step (E) amplifies approximation error (A) in $\ell$

**SCL:** ‘Directly’ minimize complementary likelihood

- Non-negative loss $\phi$
- Practically prevents ripple effect
Proposed Framework

Classification Accuracy

Methods

1. Unbiased risk estimator (URE) [Ish+19]
2. Non-negative correction methods on URE (NN) [Ish+19]
3. Surrogate complementary loss (SCL)

Table: URE and NN are based on $\bar{\ell}$ rewritten from cross-entropy loss, while SCL is based on exponential loss $\phi_{\text{EXP}}(\bar{y}, g(x)) = \exp(p_{\bar{y}})$.

<table>
<thead>
<tr>
<th>Data set + Model</th>
<th>URE</th>
<th>NN</th>
<th>SCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST + Linear</td>
<td>0.850</td>
<td>0.818</td>
<td>0.902</td>
</tr>
<tr>
<td>MNIST + MLP</td>
<td>0.801</td>
<td>0.867</td>
<td>0.925</td>
</tr>
<tr>
<td>CIFAR10 + ResNet</td>
<td>0.109</td>
<td>0.308</td>
<td>0.492</td>
</tr>
<tr>
<td>CIFAR10 + DenseNet</td>
<td>0.291</td>
<td>0.338</td>
<td>0.544</td>
</tr>
</tbody>
</table>
Gradient Analysis

**Gradient Direction of URE**

- Very diverse directions on each \( \bar{y} \)
- Low correlation to the target \( \ell_{01} \)

\[
\nabla \ell(y, g(x))
\]

\[
\nabla \bar{\ell}(\bar{y}, g(x))
\]

**Figure**: Illustration of URE

**Gradient Direction of SCL**

- Targets to minimum likelihood objective
- High correlation to the target \( \bar{\ell}_{01} \)
Gradient Estimation Error

Bias-Variance Decomposition

\[
\text{MSE} = \mathbb{E}[(f - c)^2] = \mathbb{E}[(f - h)^2] + \mathbb{E}[(h - c)^2]
\]

Gradient Estimation

1. Ordinary gradient \( f = \nabla \ell(y, g(x)) \)
2. Complementary gradient \( c = \nabla \ell(\bar{y}, g(x)) \)
3. Expected complementary gradient \( h \)
Bias-Variance Tradeoff

Findings

- SCL reduces variance by introducing small bias (towards $\bar{y}$)

<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>Variance</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>URE</td>
<td>0</td>
<td>Big</td>
<td>Big</td>
</tr>
<tr>
<td>SCL</td>
<td>Small</td>
<td>Small</td>
<td>Small</td>
</tr>
</tbody>
</table>

Chou et al. Learning with Complementary Labels
**Conclusion**

**Explain Overfitting of URE**
- Unbiased methods only do well in expectation
- Single fixed complementary label cause overfitting

**Surrogate Complementary Loss (SCL)**
- Minimum likelihood principle
- Avoids negative risk issue

**Experiment Results**
- SCL significantly outperforms other methods
- Introduce small bias for lower gradient variance
minimize $\ell_{0/1}$—hypothesis that least matches complementary data:

is this **minimum likelihood** principle well-justified? **Not yet.**

bias-variance decomposition of gradient based on **empirical findings**:

is there a theoretical guarantee to play with the trade-off? **Not yet.**

current results based on **uniform** complementary labels:

do we understand the assumptions to make LCL ‘learnable’? **Not yet.**

Thank you!
References


