Multi-class Bounds for Majority Vote Classifiers in the Case of Label Noise

Vasilii Feofanov[1,2] **Emilie Devijver**[2] **Massih-Reza Amini**[2]

[1]Huawei Noah's Ark Lab [2] Univ. Grenoble Alpes

Introduction

In many applications, we do not access perfect labels (pseudo-labeling, distribution shift, noisy annotation). Due to this label noise, theoretical analysis is more intricate.

Contribution:

1. Relationship between the risk on the true and the noisy label.
2. Upper-bound for majority vote classifier's risk in this noisy scenario.

Problem Setup

Consider multi-class classification:

- Input $\mathcal{X} \in \mathbb{R}^d$ and output $\mathcal{Y} = \{1, \ldots, K\}$ spaces.
- Hypothesis space of classifiers $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$.
- R.V.: Input $\mathbf{X} \in \mathcal{X}$, true output $Y \in \mathcal{Y}$, noisy output $\hat{Y} \in \mathcal{Y}$.

Weighted majority vote classifier:

- $B_Q(\mathbf{x}) := \mathrm{argmax}_{c\in\mathcal{Y}}\, \mathbb{E}_{h\sim Q}\, \mathbb{I}(h(\mathbf{x}) = c)$,
- Margin: $m_Q(\mathbf{x}, y) := \mathbb{E}_{h\sim Q}\, \mathbb{I}(h(\mathbf{x}) = y) - \max_{c\in\mathcal{Y}\setminus y} \mathbb{E}_{h\sim Q}\, \mathbb{I}(h(\mathbf{x}) = c)$,

What we want: risk on true labels,

- $r(B_Q, \mathbf{x}) := \sum_{\mathcal{Y}\setminus\{B_Q(\mathbf{x})\}} P(Y = c|\mathbf{X} = \mathbf{x}),\ \ R(B_Q) := \mathbb{E}_{\mathbf{X}} r(B_Q, \mathbf{X})$,

What we have: risk on noisy labels,

- $\hat{r}(B_Q, \mathbf{x}) := \sum_{\mathcal{Y}\setminus\{B_Q(\mathbf{x})\}} P(\hat{Y} = c|\mathbf{X} = \mathbf{x}),\ \ \hat{R}(B_Q) := \mathbb{E}_{\mathbf{X}} \hat{r}(B_Q, \mathbf{X})$.

Labels Are Perfect $\Rightarrow$ C-Bound

Let $M_Q := m_Q(\mathbf{X}, Y)$ with its 1st and 2nd stat. moments $\mu_1^{M_Q}$ and $\mu_2^{M_Q}$, resp. Then, $\forall\, Q$ over $\mathcal{H}$, any density $f_{\mathbf{X}}$ over $\mathcal{X}$ and any distr. $P(Y|\mathbf{X})$ over $\mathcal{Y}$ s.t. $\mu_1^{M_Q} > 0$, we have:

$$R(B_Q) \leq 1 - \frac{\left(\mu_1^{M_Q}\right)^2}{\mu_2^{M_Q}}. \qquad \text{(CB)}$$

Minimization of C-Bound implies simultaneously:

- Maximization of the margin mean (**individual performance of members**),
- Minimization of the margin variance (**correlation between members**).

Want to Know More?

paper

What this paper is also about:

- Another multi-class bound for the transductive setting;
- Application to self-training: automatic threshold selection;
- Great results on tabular data, more robust to distribution shift than other policies (acc. to Odonnat et al, 2024).

Mislabeling Error Model

Guinea Pig **Hamster** **Orangutan**

Simplification: assume that $P(\mathbf{X}|Y, \hat{Y}) = P(\mathbf{X}|Y)$.

- Class-related mislabeling model:

$$\mathbf{P} = (p_{i,j})_{1\leq i,j\leq K} \text{ with } p_{i,j} := P(\hat{Y} = i|Y = j).$$

- Posterior transformation:

$$P(\hat{Y} = i|\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{K} p_{i,j} P(Y = j|\mathbf{X} = \mathbf{x}).$$

$$\begin{array}{cccc} & \text{Guinea Pig} & \text{Hamster} & \text{Orangutan} \end{array}$$
$$\mathbf{P} = \begin{pmatrix} 0.65 & 0.32 & 0.01 \\ 0.33 & 0.67 & 0.01 \\ 0.02 & 0.01 & 0.98 \end{pmatrix}$$

if $h(\mathbf{x}) =$ "Guinea Pig" $\Rightarrow$
$\alpha(\mathbf{x}) = 0.65$
$\delta(\mathbf{x}) = 0.65 - 0.32 = 0.33$

Connection b/w True and Noisy Risk

For all classifiers $h\colon \mathcal{X}\to\mathcal{Y}$, $\forall\, \mathbf{x} \in \mathcal{X}$, $\forall\, \lambda \geq 0$ such that $p_{i,i} > p_{i,j} - \lambda,\ \forall i, j \in \mathcal{Y}^2$, we have:

$$r(h, \mathbf{x}) \leq u(h, \mathbf{x}) := \frac{\hat{r}(h, \mathbf{x})}{\lambda + \delta(\mathbf{x})} - \frac{1 - \lambda - \alpha(\mathbf{x})}{\lambda + \delta(\mathbf{x})},$$

with

- $\alpha(\mathbf{x}) := p_{h(\mathbf{x}),h(\mathbf{x})}$,
- $\delta(\mathbf{x}) := p_{h(\mathbf{x}),h(\mathbf{x})} - \max_{j\in\mathcal{Y}\setminus\{h(\mathbf{x})\}} p_{h(\mathbf{x}),j}$.

Remarks:

- Equality when no mislabeling ($\alpha(\mathbf{x}) = \delta(\mathbf{x}) = 1$) and $\lambda = 0$;
- Holds also for $\mathbf{x}$-dependent mislabeling probs: $p_{i,j}^{\mathbf{x}} := P(\hat{Y} = i|Y = j, \mathbf{X} = \mathbf{x})$;
- Hyperparameter λ: can relax assumptions and prevent an arbitrarily large bound.

C-Bound with Imperfect Labels (CBIL)

Let $\hat{M}_Q := m_Q(\mathbf{X}, \hat{Y})$. Then, $\forall\, Q$ over $\mathcal{H}$, any density $f_{\mathbf{X}}$ over $\mathcal{X}$, all distr. $P(Y|\mathbf{X})$ and $P(\hat{Y}|\mathbf{X})$ over $\mathcal{Y}$, $\forall\, \lambda \geq 0$ such that $p_{i,i} > p_{i,j} - \lambda,\ \forall i, j \in \mathcal{Y}^2$, we have:

$$R(B_Q) \leq \psi_{\mathbf{P}} - \frac{\left(\mu_1^{\hat{M}_Q,\mathbf{P}}\right)^2}{\mu_2^{\hat{M}_Q,\mathbf{P}}}, \qquad \text{(CBIL)}$$

if $\mu_1^{\hat{M}_Q,\mathbf{P}} > 0$, where

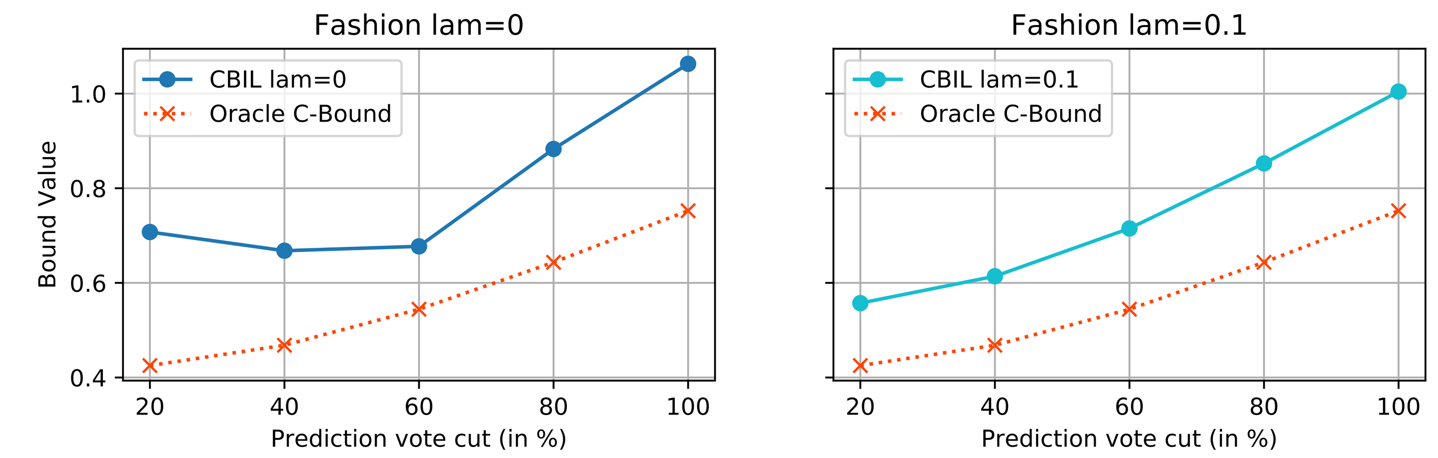
- $\psi_{\mathbf{P}} := \mathbb{E}_{\mathbf{X}}(\alpha(\mathbf{x}) + \lambda)/(\delta(\mathbf{x}) + \lambda)$,
- $\mu_1^{\hat{M}_Q,\mathbf{P}} := \int_{\mathbb{R}^{d+1}} m/(\delta(\mathbf{x}) + \lambda) f_{\hat{M}_Q,\mathbf{X}}(m, \mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}m$,
- $\mu_2^{\hat{M}_Q,\mathbf{P}} := \int_{\mathbb{R}^{d+1}} m^2/(\delta(\mathbf{x}) + \lambda) f_{\hat{M}_Q,\mathbf{X}}(m, \mathbf{x}) \mathrm{d}\mathbf{x}\mathrm{d}m$.

Remarks:

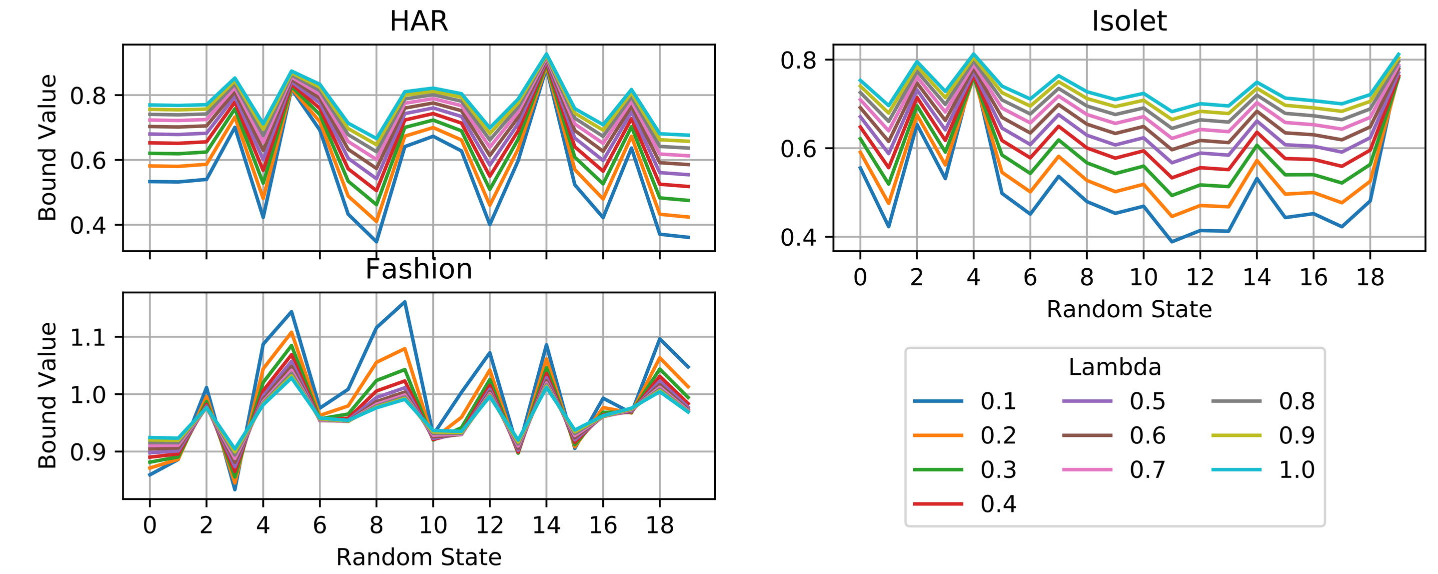
- "Weighted" moments: each margin is penalized by $(\delta(\mathbf{x}) + \lambda)$;
- Holds for any Q, so can be used as a criterion to optimize Q;
- When estimated from data, can be further bounded using the PAC-Bayesian theorem.

Benign Relaxation

When $\lambda > 0$, we relax the bound, but its value can be tighter!



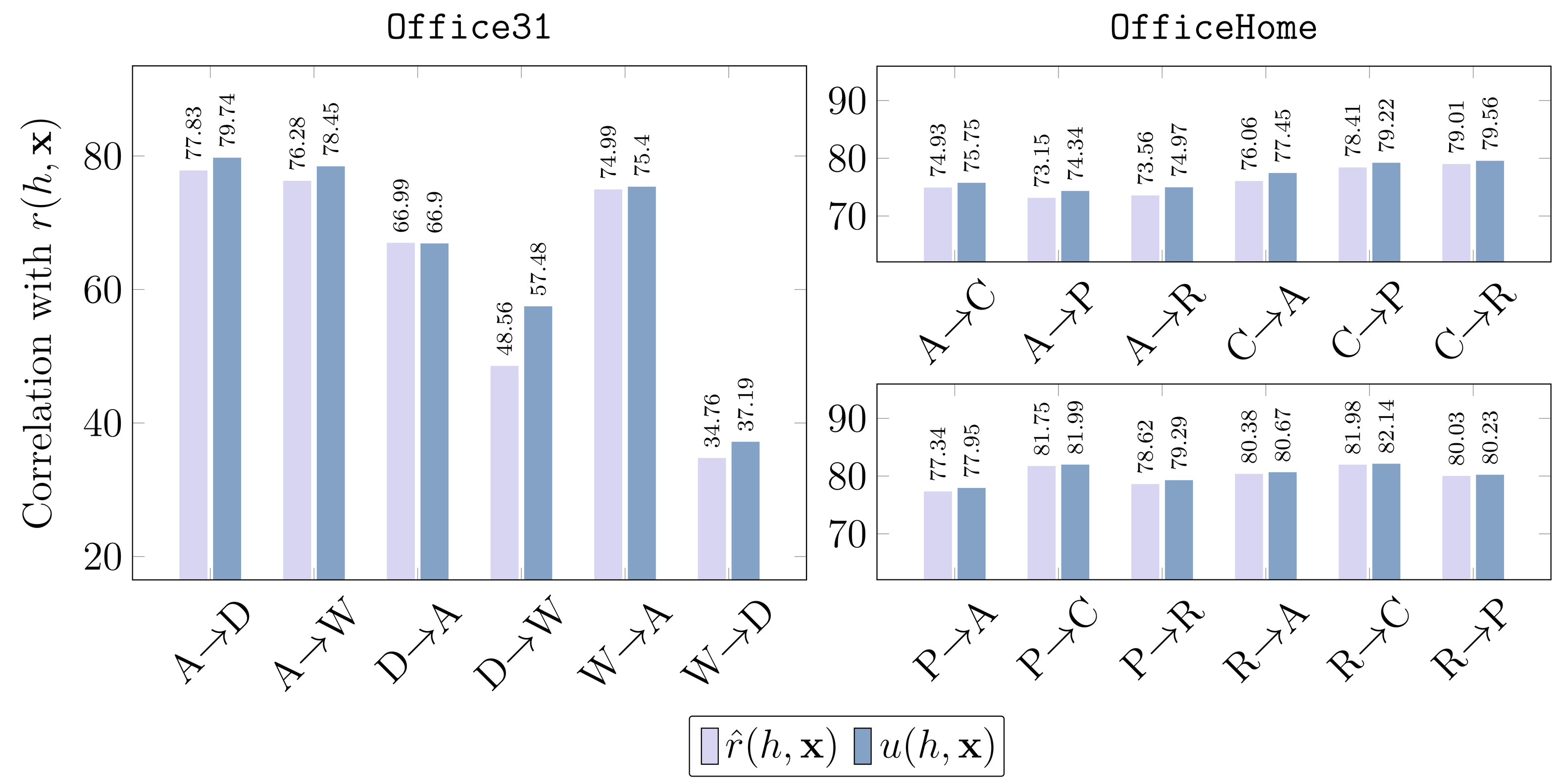
Higher values of λ makes behavior smoother and generally help to correlate better with the true risk.



Domain Shift Experiment

What people do: use logits to estimate accuracy on unlabeled data.
Problem: logits can be biased under distribution shift.
Experiment: Given h (ResNet-18, pre-trained on a source domain), compare correlations on a target domain b/w $r(h, \mathbf{x})$ and

- $\hat{r}(h, \mathbf{x})$ computed using softmax probs,
- $u(h, \mathbf{x})$ with oracle $\delta(\mathbf{x})$ and $\alpha(\mathbf{x})$.



Published in JMLR, presented at ICML 2024, Vienna, Austria

Correspondence to vasilii.feofanov@gmail.com