On Symmetric Losses for Learning from Corrupted Labels

Nontawat Charoenphakdee[1,2] Jongyeong Lee[1,2] Masashi Sugiyama[2,1]

1: The University of Tokyo 2: RIKEN AIP

Introduction

Learning from corrupted labels is possible, but…

Lu+ 2019: We need to know proportions of clean positive data in corrupted labeled data to optimize accuracy.

Problem: Proportions are unidentifiable from samples **(Scott+, 2013).**

Q: What can we learn without estimating proportions?

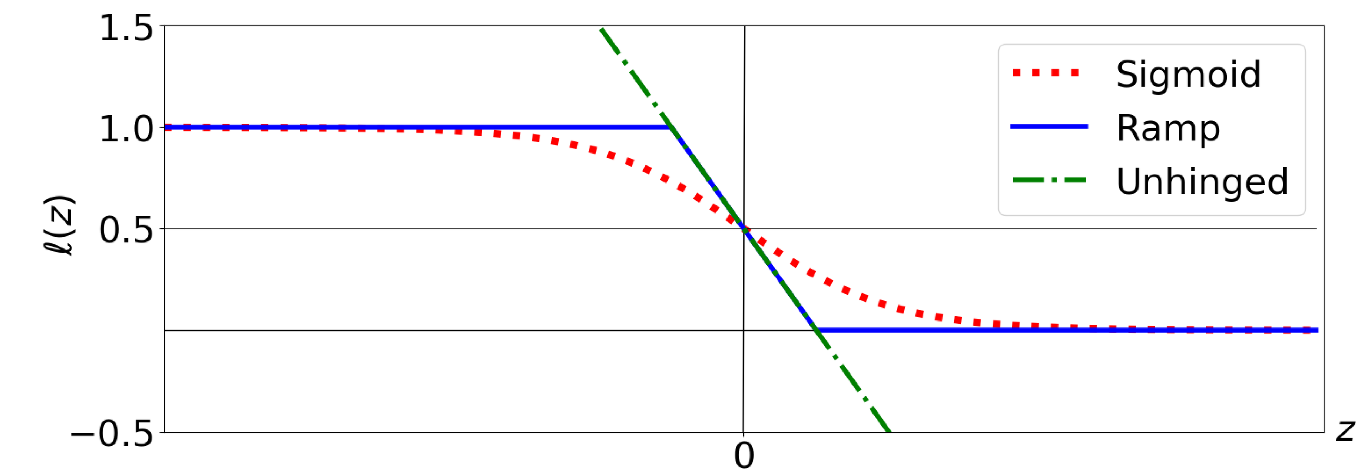
van Rooyen+ 2015: Optimizing Balanced error rate (BER) is effective with symmetric losses (no experiments).

Menon+ 2015: Optimizing BER or **Area under ROC curve (AUC)** is effective with many losses (experiments with squared loss).

Ours: using symmetric loss is preferable for both BER & AUC optimization theoretically and experimentally!

Symmetric Loss

$\ell(z) + \ell(-z) = \text{Constant}$ $\quad \ell: \mathbb{R} \to \mathbb{R}$



Learning from Corrupted Labels

(Scott+, 2013, Menon+, 2015, Lu+, 2019)

Given: Two sets of corrupted data

Positive: $X_{\mathrm{CP}} := \{\boldsymbol{x}_i^{\mathrm{CP}}\}_{i=1}^{n_{\mathrm{CP}}} \overset{\text{i.i.d.}}{\sim} \pi \mathrm{pos}(\boldsymbol{x}) + (1-\pi)\,\mathrm{neg}(\boldsymbol{x})$

Negative: $X_{\mathrm{CN}} := \{\boldsymbol{x}_j^{\mathrm{CN}}\}_{j=1}^{n_{\mathrm{CN}}} \overset{\text{i.i.d.}}{\sim} \pi' \mathrm{pos}(\boldsymbol{x}) + (1-\pi')\,\mathrm{neg}(\boldsymbol{x})$

$\mathrm{pos}(\boldsymbol{x})$: $p(\boldsymbol{x}|y=+1)$
$\mathrm{neg}(\boldsymbol{x})$: $p(\boldsymbol{x}|y=-1)$
$\pi, \pi' \in [0,1]$ and $\pi > \pi'$

Find: $g: \mathbb{R}^d \to \mathbb{R}$ that minimizes

AUC risk, i.e., bipartite ranking risk:

$$R_{\mathrm{AUC}}^{\ell_{0\text{-}1}}(g) = \mathbb{E}_{\mathrm{P}}[\mathbb{E}_{\mathrm{N}}[\ell_{0\text{-}1}(g(\boldsymbol{x}^{\mathrm{P}}) - g(\boldsymbol{x}^{\mathrm{N}}))]]$$

$\mathbb{E}_{\mathrm{P}}[\cdot]: \underset{\boldsymbol{x}\sim\mathrm{pos}(\boldsymbol{x})}{\mathbb{E}}[\cdot]$
$\mathbb{E}_{\mathrm{N}}[\cdot]: \underset{\boldsymbol{x}\sim\mathrm{neg}(\boldsymbol{x})}{\mathbb{E}}[\cdot]$

g **outputs higher values for positive data than negative data**

Similar results hold for **BER** in this paper and thus omitted for brevity.

AUC Maximization from corrupted labels

The following theorem relates **corrupted AUC risk** to **clean AUC risk**.

Theorem 1. *Let $\gamma^\ell(\boldsymbol{x}, \boldsymbol{x}') = \ell(f(\boldsymbol{x}', \boldsymbol{x})) + \ell(f(\boldsymbol{x}, \boldsymbol{x}'))$. Then $R^\ell_{\text{AUC-Corr}}(g)$ can be expressed as*

$$\underbrace{R^\ell_{\text{AUC-Corr}}(g)}_{\text{Corrupted risk}} = (\pi - \pi')\underbrace{R^\ell_{\mathrm{AUC}}(g)}_{\text{Clean risk}} + \underbrace{(\pi' - \pi\pi')\mathbb{E}_+[\mathbb{E}_-[\gamma^\ell(\boldsymbol{x}_+, \boldsymbol{x}_-)]]}_{\text{Excessive term}}$$
$$+ \underbrace{\frac{\pi\pi'}{2}\mathbb{E}_{+'}[\mathbb{E}_+[\gamma^\ell(\boldsymbol{x}_{+'}, \boldsymbol{x}_+)]] + \frac{(1-\pi)(1-\pi')}{2}\mathbb{E}_{-'}[\mathbb{E}_-[\gamma^\ell(\boldsymbol{x}_{-'}, \boldsymbol{x}_-)]]}_{\text{Excessive term}}.$$

$f(\boldsymbol{x}, \boldsymbol{x}') = g(\boldsymbol{x}) - g(\boldsymbol{x}')$

Minimizing the corrupted risk can be ineffective with excessive terms!! ☹

When $\gamma^\ell(\boldsymbol{x}, \boldsymbol{x}') = K$ which holds for symmetric losses, we have

$$\underbrace{R^\ell_{\text{AUC-Corr}}(g)}_{\text{Corrupted risk}} = (\pi - \pi')\underbrace{R^\ell_{\mathrm{AUC}}(g)}_{\text{Clean risk}} + \boxed{K\left(\frac{1-\pi+\pi'}{2}\right)}.$$

Excessive terms become constant!

With symmetric losses, excessive terms are constant and thus can be safely ignored. ☺

Properties of Symmetric Losses

Symmetric loss is useful but its theoretical understanding is **limited**...

Why? because nonnegative symmetric losses are **non-convex**.

Theory of convex losses cannot be applied. ☹ (du Plessis+, 2014, Ghosh+, 2015)

We prove several properties of symmetric losses, e.g.,

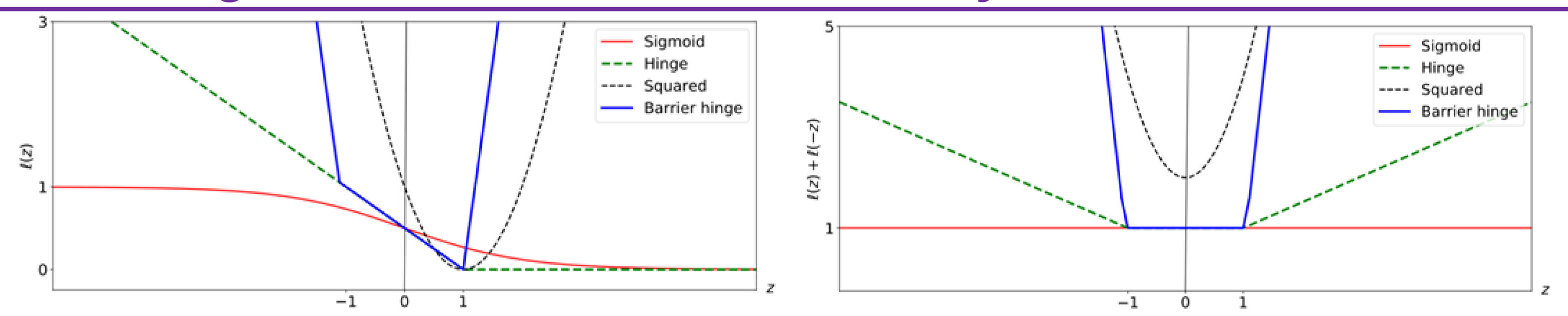
- **Necessary and sufficient condition** for **classification-calibration**
- **Inability** to estimate **class posterior probability**
- A **sufficient condition** for **AUC-consistency**

Well-known symmetric losses, e.g., sigmoid, ramp are classification-calibrated and AUC-consistent!

Barrier Hinge Loss

Q: Can we have a nonnegative convex loss that benefits from symmetric condition?

Barrier hinge loss: A convex loss that is symmetric in the middle region.



$\ell(z) = \max(-s(w+z) + w, \max(s(z-w), w-z))$

$s > 1$ **slope** of the non-symmetric region

$w > 0$ **width** of symmetric region.

Can be viewed as a soft-constrained unhinged loss $\ell(z) = 1 - z$ (van Rooyen+, 2015)

Experiments $\pi = 0.65, \pi' = 0.45$

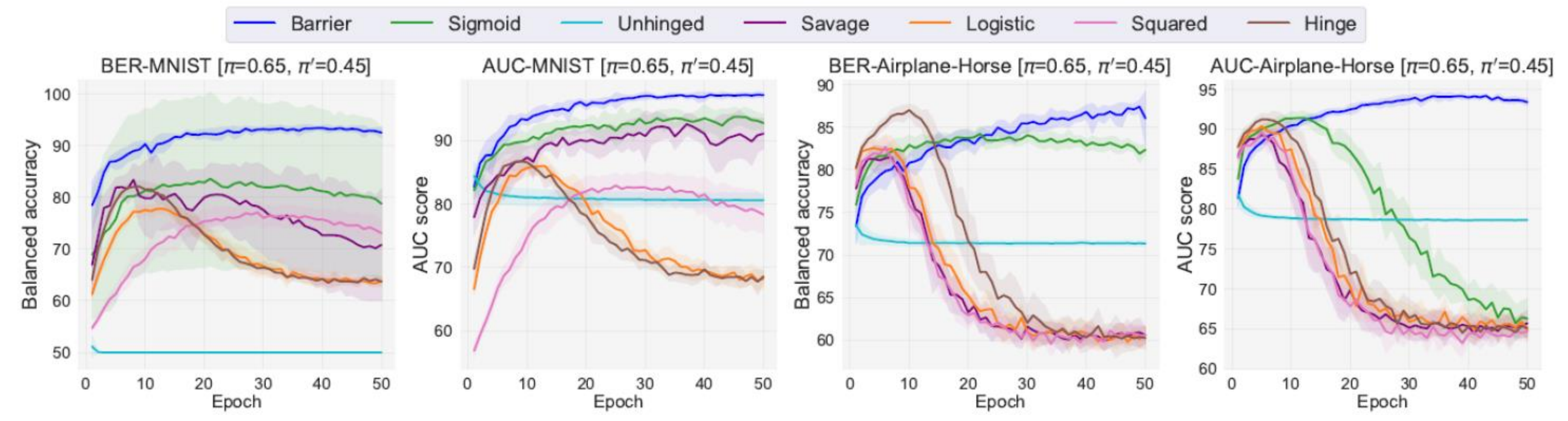
Q: Does the symmetric condition significantly help in practice?

Simple one-hidden layer neural networks (UCI datasets)

Table 2. Mean balanced accuracy (BAC=1-BER) and AUC score using multilayer perceptrons (rescaled to 0-100), where $\pi = 0.65$ and $\pi' = 0.45$. Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%. The experiments were conducted 20 times.

Dataset	Task	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	BAC	82.3(0.8)	**84.1 (0.6)**	80.9(0.6)	72.6(0.7)	74.7(0.7)	69.5(0.7)	73.6(0.6)
	AUC	86.8(0.7)	**90.9 (0.4)**	86.0(0.4)	79.2(0.8)	77.7(0.7)	73.6(0.8)	80.1(0.8)
waveform	BAC	**86.1 (0.4)**	**87.1 (0.6)**	85.4(0.6)	75.8(0.7)	78.3(0.7)	69.2(0.6)	73.2(0.6)
	AUC	**92.2 (0.4)**	**91.7 (0.6)**	**90.9 (0.6)**	82.3(0.7)	79.8(0.9)	75.1(0.7)	80.1(0.6)
twonorm	BAC	**96.2 (0.3)**	**96.7 (0.2)**	95.4(0.4)	80.2(0.5)	82.8(0.9)	71.6(0.7)	75.9(0.6)
	AUC	99.1(0.1)	**99.6 (0.0)**	98.0(0.2)	88.3(0.5)	83.9(0.7)	77.3(0.7)	82.7(0.5)
mushroom	BAC	**93.4 (0.8)**	91.1(0.9)	**94.4 (0.7)**	81.3(0.5)	84.5(1.0)	72.2(0.6)	79.5(0.8)
	AUC	**98.4 (0.2)**	97.2(0.4)	**97.8 (0.3)**	89.0(0.5)	82.2(0.6)	77.8(0.6)	88.1(0.7)

Convolutional neural networks (MNIST, CIFAR-10)



Architecture: Convolution neural networks with fully connected layer followed by dropout layer:

dim-Conv[18,5,1,0]-Max[2,2]-Conv[48,5,1,0]-Max[2,2]-800-400-1

Activation function: Rectifier linear units (ReLU)

Symmetric losses & barrier hinge loss are preferable!

*A negatively unbounded unhinged loss is observed to be less preferable when using complex models.

References

[1] Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In COLT, 2013.

[2] du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In NeurIPS, 2014

[3] Menon, A., van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class probability estimation. In ICML, 2015.

[4] van Rooyen, B., "Machine Learning via Transitions", PhD thesis (available online), 2015.

[5] Ghosh, A., Manwani, N., and Sastry, P. Making risk minimization tolerant to label noise. Neurocomputing, 2015.

[6] Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier from only unlabeled data. In ICLR, 2019