

On Symmetric Losses for Learning from Corrupted Labels

Nontawat Charoenphakdee^{1,2}, Jongyeong Lee^{1,2}
and Masashi Sugiyama^{2,1}

The University of Tokyo¹

RIKEN Center for Advanced Intelligence Project (AIP)²



東京大学
THE UNIVERSITY OF TOKYO



Supervised learning

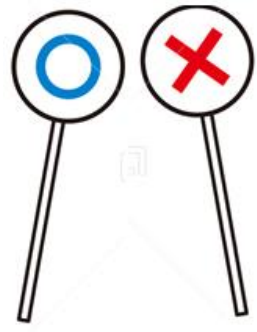
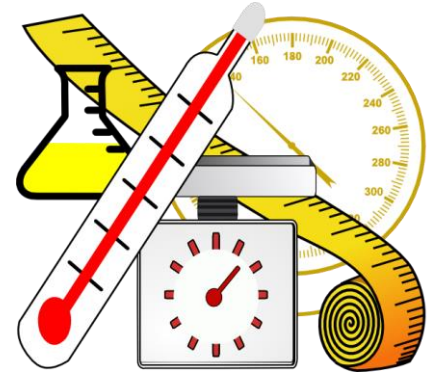
Learn from input-output pairs



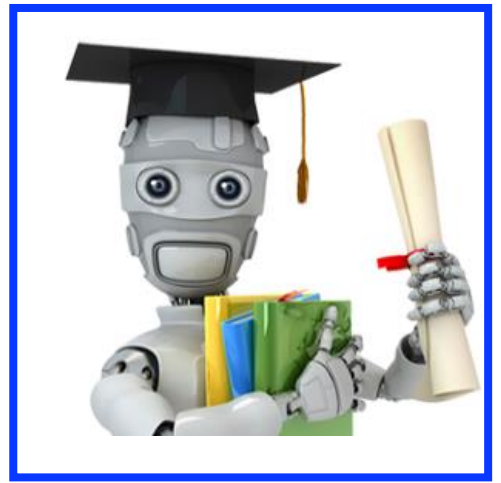
Predict output of **unseen input** accurately

Data collection

Features (Input) Labels (Output)



Prediction function

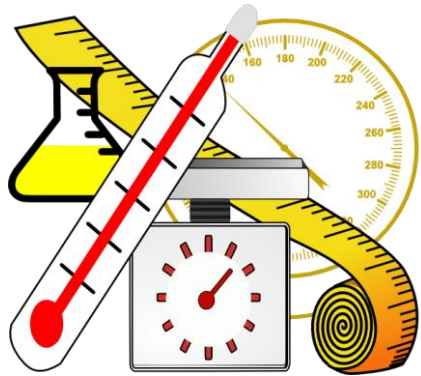


Learning from corrupted labels

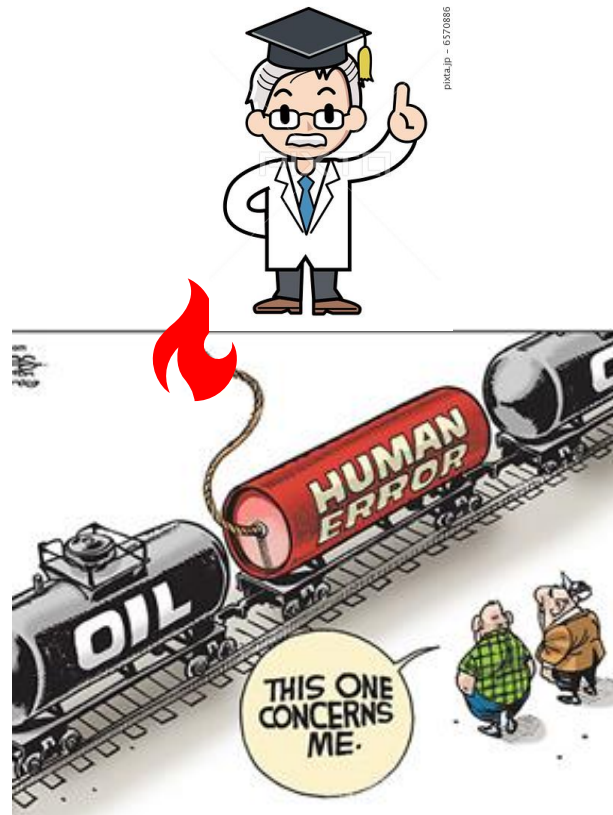
Data collection

Prediction function

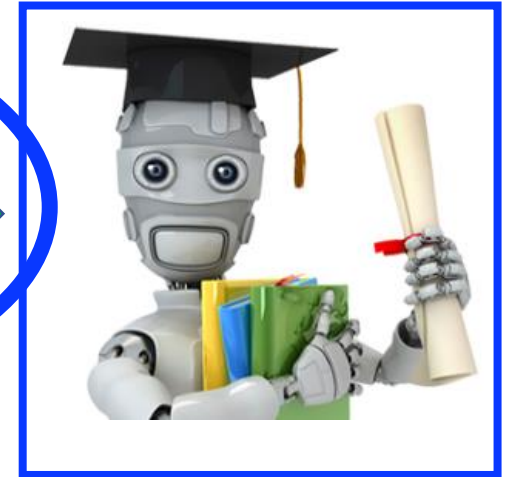
Feature collection



Labeling process



Our goal
Noise-robust ML



Examples:

- Expert labelers (human error)
- Crowdsourcing (non-expert error)

Contents

- Background and related work
- The importance of symmetric losses
- Theoretical properties of symmetric losses
- Barrier hinge loss
- Experiments

Warmup: Binary classification

- **Given:** input-output pairs:

$$\{\mathbf{x}_i, y_i\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} p(\mathbf{x}, y)$$

- **Goal:** minimize **expected error:**

$$R^{\ell_{0-1}}(g) = \mathbb{E}_{(\mathbf{x}, y) \sim p(\mathbf{x}, y)} [\ell_{0-1}(yg(\mathbf{x}))]$$

- **No access to distribution:** minimize **empirical error** (Vapnik, 1998):

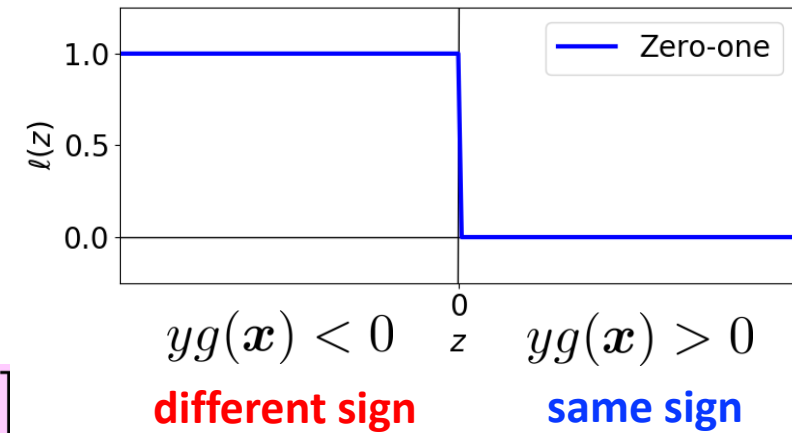
$$\hat{R}^{\ell_{0-1}}(g) = \frac{1}{n} \sum_{i=1}^n \ell_{0-1}(y_i g(\mathbf{x}_i))$$

$y \in \{-1, 1\}$: Label

$g: \mathbb{R}^d \rightarrow \mathbb{R}$: Prediction function

$\mathbf{x} \in \mathbb{R}^d$: Feature vector

$\ell: \mathbb{R} \rightarrow \mathbb{R}$: Margin loss function

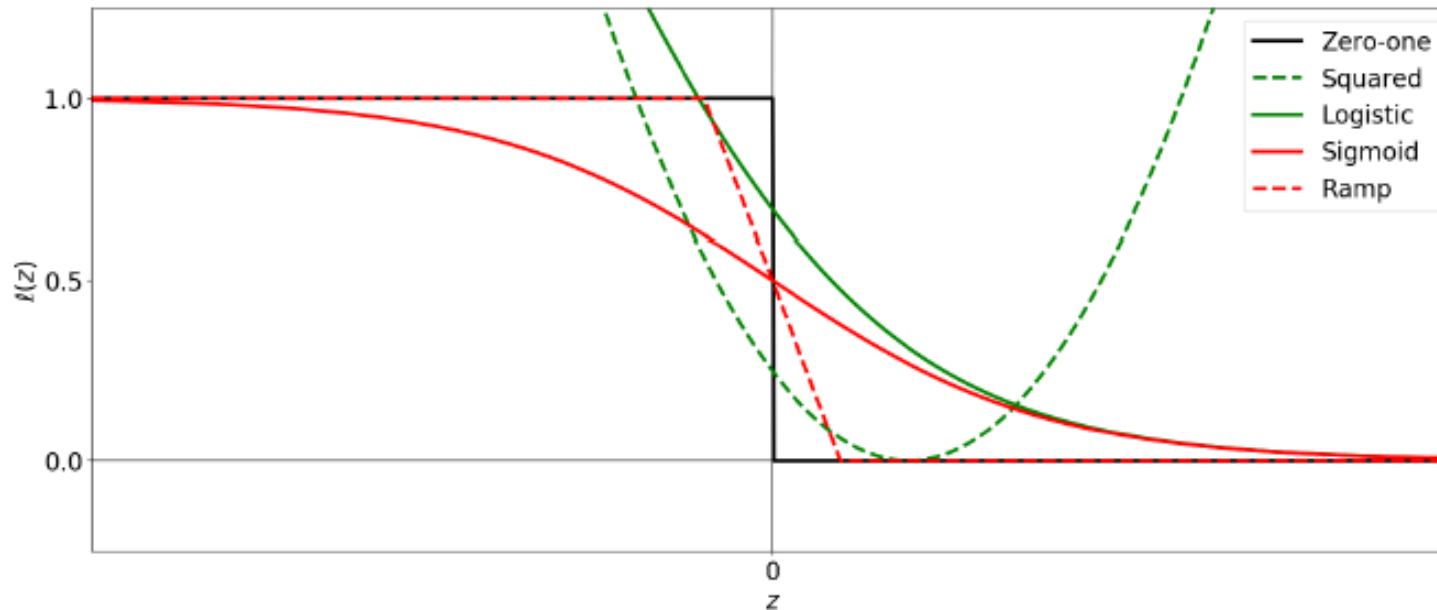


Surrogate losses

Minimizing 0-1 loss directly is **difficult**.

- Discontinuous and not differentiable (Ben-david+, 2003, Feldman+, 2012)

In practice, we minimize a **surrogate loss** (Zhang, 2004, Bartlett+, 2006).



$z = yg(\mathbf{x})$: Margin

$y \in \{-1, 1\}$: Label

$g: \mathbb{R}^d \rightarrow \mathbb{R}$: Prediction function

$\mathbf{x} \in \mathbb{R}^d$: Feature vector

Learning from corrupted labels

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

Given: Two sets of corrupted data:

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

Negative: $X_{\text{CN}} := \{\mathbf{x}_i^{\text{CN}}\}_{i=1}^{n_{\text{CN}}} \stackrel{\text{i.i.d.}}{\sim} \pi' \text{pos}(\mathbf{x}) + (1 - \pi') \text{neg}(\mathbf{x})$

Class priors

Clean: $\pi = 1, \pi' = 0$

Positive-unlabeled: $\pi = 1, \pi' < 1$ (du Plessis+, 2014)

$$\begin{aligned} \pi, \pi' &\in [0, 1] \\ \text{pos}(\mathbf{x}) &: p(\mathbf{x}|y = 1) \\ \text{neg}(\mathbf{x}) &: p(\mathbf{x}|y = -1) \end{aligned}$$

This setting covers many weakly-supervised settings (Lu+, 2019).

Issue on class priors

Given: Two sets of corrupted data:

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

Negative: $X_{\text{CN}} := \{\mathbf{x}_i^{\text{CN}}\}_{i=1}^{n_{\text{CN}}} \stackrel{\text{i.i.d.}}{\sim} \pi' \text{pos}(\mathbf{x}) + (1 - \pi') \text{neg}(\mathbf{x})$

Assumption: $\pi > \pi'$

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

How to learn **without estimating** π, π' ?

Related work:

Class priors are needed! (Lu+, 2019)

Classification error:

$$R^{\ell_{0-1}}(g) = \mathbb{E}_{(\mathbf{x}, y) \sim p(\mathbf{x}, y)} [\ell_{0-1}(yg(\mathbf{x}))]$$

$$\mathbb{E}_{\mathbf{P}}[\cdot] : \mathbb{E}_{\mathbf{x} \sim \text{pos}(\mathbf{x})} [\cdot]$$

$$\mathbb{E}_{\mathbf{N}}[\cdot] : \mathbb{E}_{\mathbf{x} \sim \text{neg}(\mathbf{x})} [\cdot]$$

Class priors are not needed! (Menon+, 2015)

Balanced error rate (BER):

$$R_{\text{Bal}}^{\ell_{0-1}}(g) = \frac{1}{2} \mathbb{E}_{\mathbf{P}} [\ell_{0-1}(g(\mathbf{x}^{\mathbf{P}}))] + \frac{1}{2} \mathbb{E}_{\mathbf{N}} [\ell_{0-1}(-g(\mathbf{x}^{\mathbf{N}}))]$$

Area under the receiver operating characteristic curve (AUC) risk:

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

Related work: BER and AUC optimization

Menon+, 2015: we can treat *corrupted data as if they were clean*.

The proof relies on a **property of 0-1 loss**.

Squared loss was used in experiments.

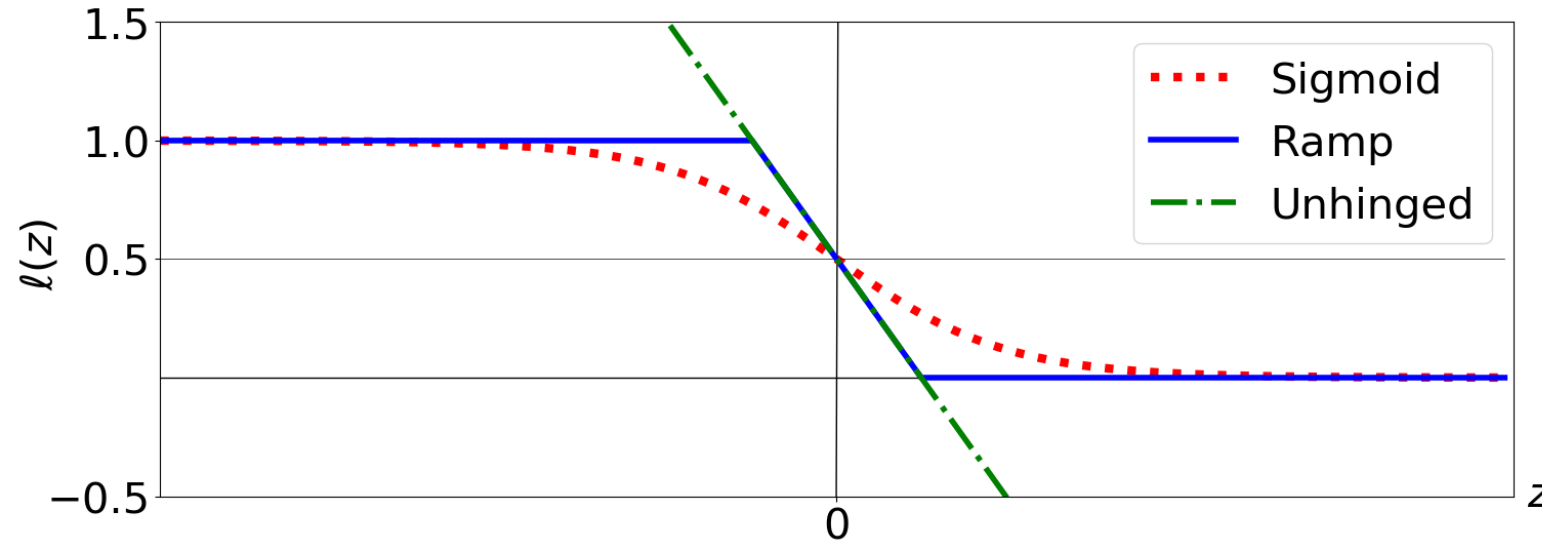
van Rooyen+, 2015: **symmetric losses** are also useful for **BER** minimization (no experiments).

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

Contents

- Background and related work
- **The importance of symmetric losses**
- Theoretical properties of symmetric losses
- Barrier hinge loss
- Experiments

Symmetric losses $\ell(z) + \ell(-z) = \text{Constant}$



Applications:

Risk estimator simplification in weakly-supervised learning

(du Plessis+, 2014, Kiryo+, 2017, Lu+, 2018)

Robustness under symmetric noise (label flip with a fixed probability)

(Ghosh+, 2015, van Rooyen+, 2015)

AUC maximization

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

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

$$R_{\text{AUC-Corr}}^\ell(g) = (\pi - \pi') R_{\text{AUC}}^\ell(g) + \underbrace{(\pi' - \pi \pi') \mathbb{E}_+ [\mathbb{E}_- [\gamma^\ell(\mathbf{x}_+, \mathbf{x}_-)]]}_{\text{Excessive term}}$$

$$+ \underbrace{\frac{\pi \pi'}{2} \mathbb{E}_{+'} [\mathbb{E}_+ [\gamma^\ell(\mathbf{x}_{+'}, \mathbf{x}_+)]]}_{\text{Excessive term}} + \underbrace{\frac{(1 - \pi)(1 - \pi')}{2} \mathbb{E}_{-' } [\mathbb{E}_- [\gamma^\ell(\mathbf{x}_{-'}, \mathbf{x}_-)]]}_{\text{Excessive term}}.$$

Symmetric losses: $\ell(z) + \ell(-z) = K$

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

$$R_{\text{AUC-Corr}}^\ell(g) = (\pi - \pi') R_{\text{AUC}}^\ell(g) + K \left(\frac{1 - \pi + \pi'}{2} \right).$$

Excessive terms become constant!

Excessive terms can be safely ignored with symmetric losses 😊

BER minimization

Theorem 3. Let $\gamma^\ell(\mathbf{x}) = \ell(g(\mathbf{x})) + \ell(-g(\mathbf{x}))$, $R_{\text{Bal-Corr}}^\ell(g)$ can be expressed as

$$R_{\text{Bal-Corr}}^\ell(g) = \underbrace{(\pi - \pi')R_{\text{Bal}}^\ell(g)}_{\text{Clean risk}} + \underbrace{\frac{\pi' \mathbb{E}_+[\gamma^\ell(\mathbf{x})] + (1 - \pi) \mathbb{E}_-[\gamma^\ell(\mathbf{x})]}{2}}_{\text{Excessive term}}.$$

Corrupted risk

Symmetric losses: $\ell(z) + \ell(-z) = K$

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

$$R_{\text{Bal-Corr}}^\ell(g) = (\pi - \pi')R_{\text{Bal}}^\ell(g) + K \left(\frac{1 - \pi + \pi'}{2} \right).$$

Coincides with **van Rooyen 2015+**

Excessive term becomes constant!

Excessive terms can be safely ignored with symmetric losses 😊

Contents

- Background and related work
- The importance of symmetric losses
- **Theoretical properties of symmetric losses**
- Barrier hinge loss
- Experiments

Theoretical properties of symmetric losses

Nonnegative symmetric losses are **non-convex**.

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

We provide a better understanding of symmetric losses: 😊

- **Necessary and sufficient condition** for **classification-calibration**
- **Excess risk bound** in binary classification
- **Inability** to estimate **class posterior probability**
- A **sufficient condition** for **AUC-consistency**
 - Covers many symmetric losses, e.g., sigmoid, ramp.

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

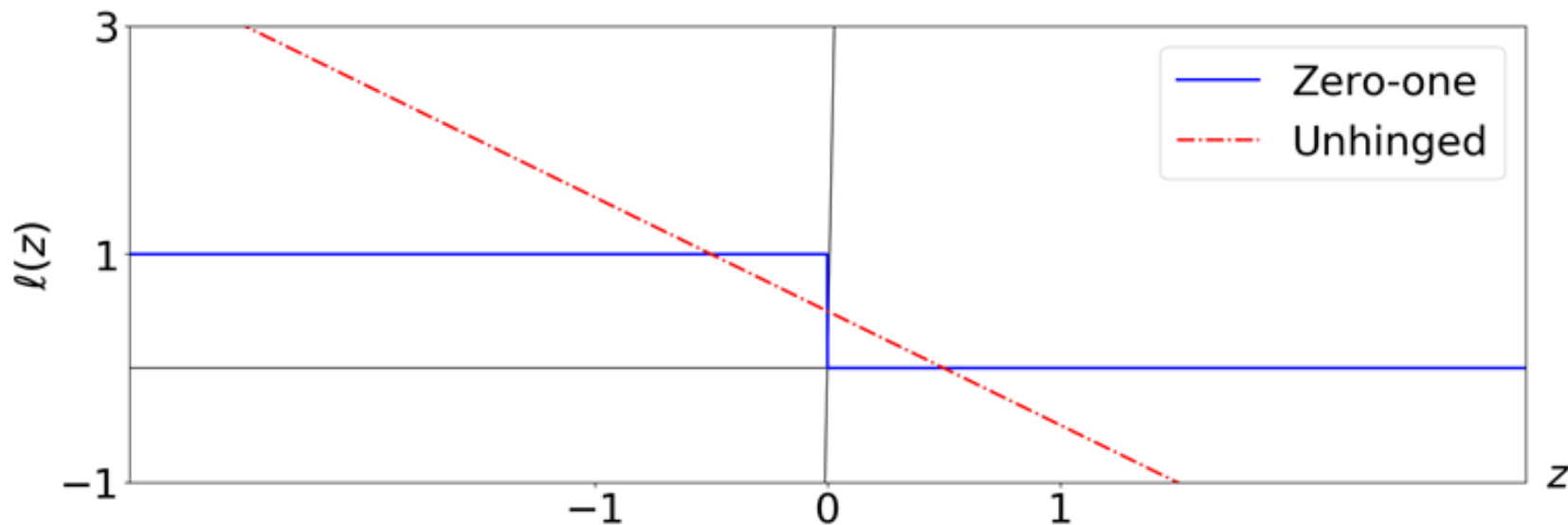
Contents

- Background and related work
- The importance of symmetric losses
- Theoretical properties of symmetric losses
- **Barrier hinge loss**
- Experiments

Convex symmetric losses?

By sacrificing nonnegativity:

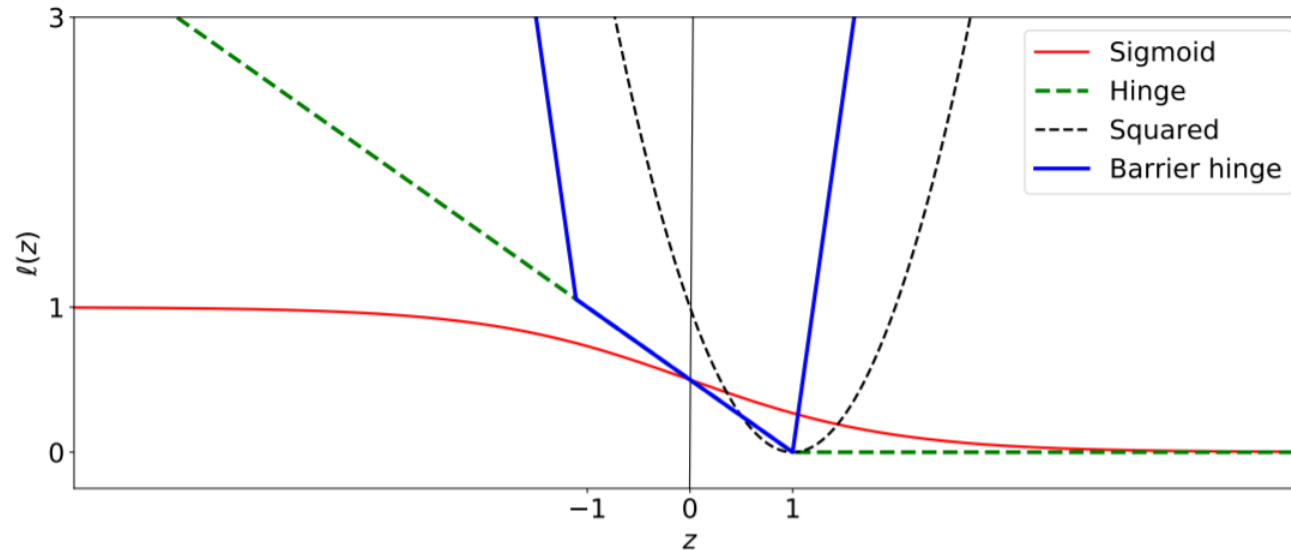
only unhinged loss is convex and symmetric (van Rooyen+, 2015).



This loss has been considered (although robustness was not discussed).

(Devroye+, 1996, Schoelkopf+, 2002, Shawe-Taylor+, 2004, Sriperumbudur+, 2009, Reid+, 2011)

Barrier hinge loss



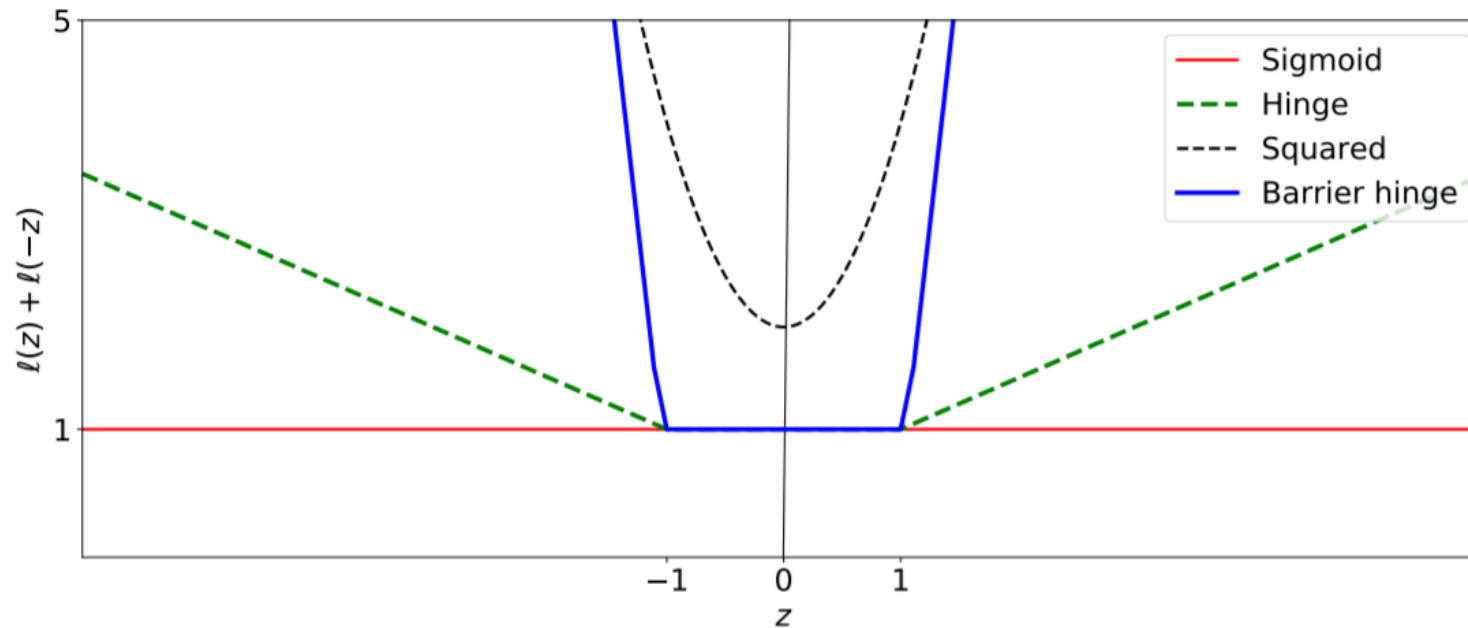
$$l(z) = \max(-s(w + z) + w, \max(s(z - w), w - z))$$

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

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

High penalty if **misclassify** or output is **outside symmetric region**.

Symmetry of barrier hinge loss



Satisfies symmetric property in an interval.

If output range is restricted in a symmetric region:

unhinged, hinge, barrier are equivalent.

Contents

- Background and related work
- The importance of symmetric losses
- Theoretical properties of symmetric losses
- Barrier hinge loss
- **Experiments**

Experiments: BER/AUC optimization from corrupted labels

To empirically answer the following questions:

1. Does the symmetric condition significantly help?
2. Do we need a loss to be symmetric everywhere?
3. Does the negative unboundedness degrade the practical performance?

We conducted the following experiments: Fix the models, vary the loss functions

Losses: **Barrier [s=200, w=50]**, **Unhinged**, **Sigmoid**, Logistic, Hinge, Squared, Savage

Experiment 1:

MLPs on UCI/LIBSVM datasets.

Experiment 2:

CNNs on more difficult datasets (MNIST, CIFAR-10).

Experiments: BER/AUC optimization from corrupted labels

For UCI datasets:

Multilayered perceptrons (MLPs) with one hidden layer: [d-500-1]

Activation function: Rectifier Linear Units (ReLU) (Nair+, 2010)

MNIST and CIFAR-10:

Convolutional neural networks (CNNs):

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

ReLU after fully connected layer follows by dropout layer (Srivastava+, 2010)

MNIST: Odd numbers vs Even numbers

CIFAR: One class vs Airplane (follows Ishida+, 2017)

Conv[18, 5, 1, 0]: 18 channels, 5 x 5 convolutions, stride 1, padding 0

Max[2,2]: max pooling with kernel size 2 and stride 2

Experiment 1: MLPs on UCI/LIBSVM datasets

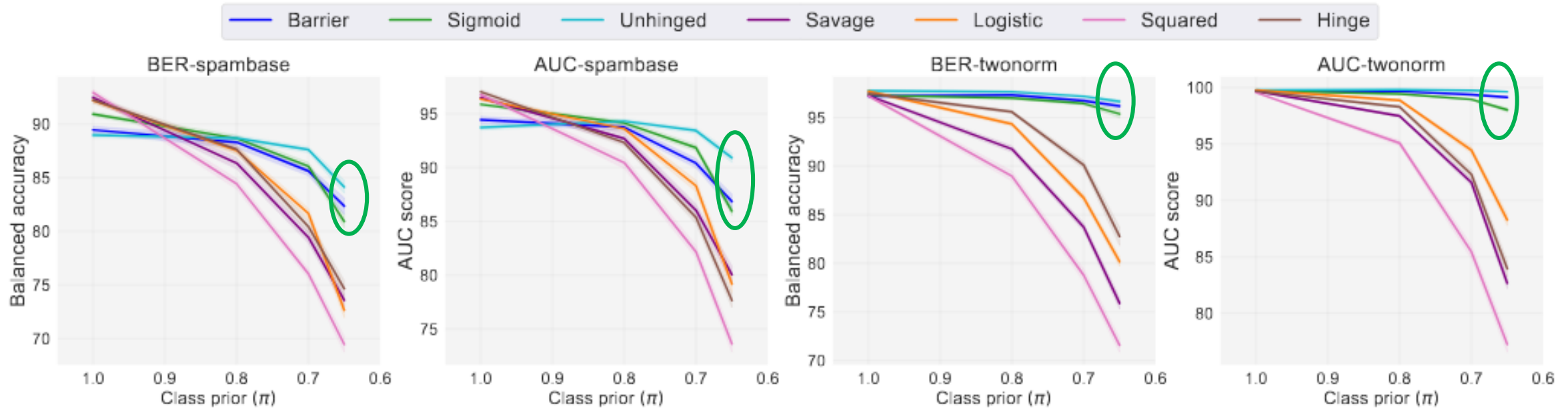


Figure 4: Mean balanced accuracy (1-BER) and AUC score using multilayer perceptrons (rescaled to 0-100) with varying noise rates ($\pi = 1.0, \pi' = 0.0$), ($\pi = 0.8, \pi' = 0.3$), ($\pi = 0.7, \pi' = 0.4$), ($\pi = 0.65, \pi' = 0.45$). The experiments were conducted 20 times.

The higher the better.

Dataset information and more experiments and can be found in our paper.

Experiment 1: MLPs on UCI/LIBSVM datasets

Symmetric losses and barrier hinge loss are preferable!

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) |

The higher the better.

Experiment 2: CNNs on MNIST/CIFAR-10

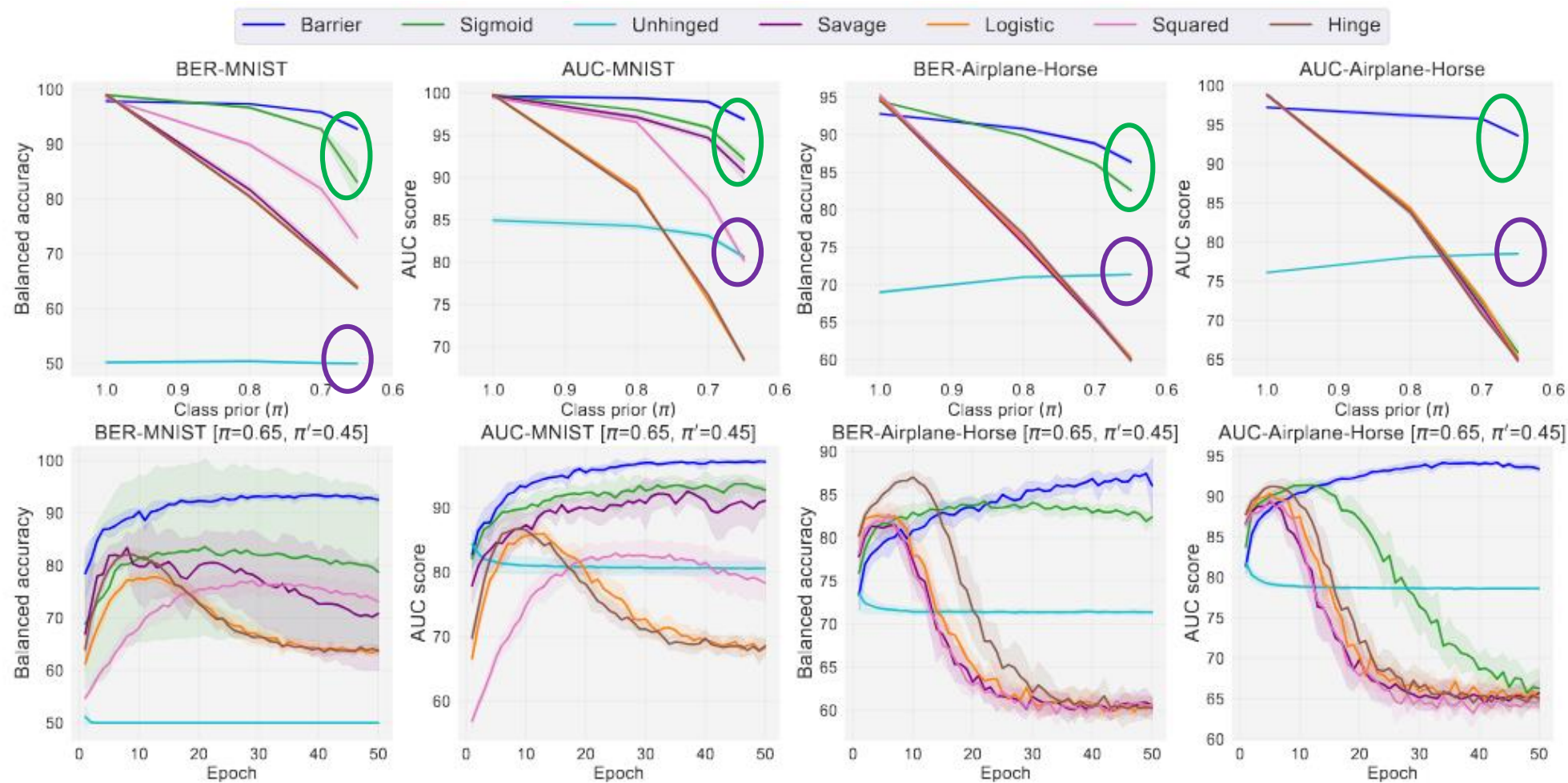
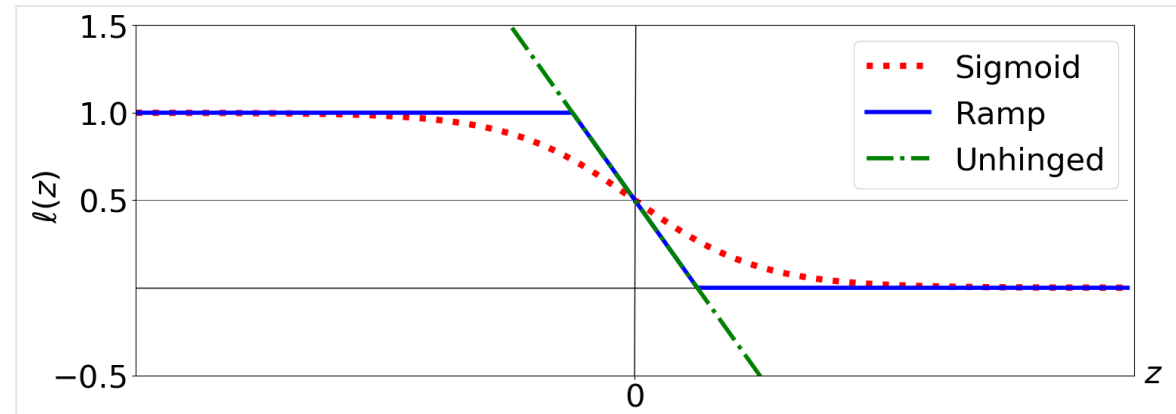


Figure 5: Mean balanced accuracy (1-BER) and AUC score using convolutional neural networks (rescaled to 0-100). (Top) the varying noise rates ranged from $(\pi = 1.0, \pi' = 0.0)$, $(\pi = 0.8, \pi' = 0.3)$, $(\pi = 0.7, \pi' = 0.4)$, $(\pi = 0.65, \pi' = 0.45)$. (Bottom) the noise rate is $\pi = 0.65$ and $\pi' = 0.45$. The experiments were conducted 10 times.

Conclusion



We showed that **symmetric loss is preferable under corrupted labels** for:

- Area under the receiver operating characteristic curve (**AUC**) maximization
- Balanced error rate (**BER**) minimization

We provided **general theoretical properties** for symmetric losses:

- Classification-calibration, excess risk bound, AUC-consistency
- Inability of estimating the class posterior probability

We proposed a **barrier hinge loss**:

- As a proof of concept of the importance of symmetric condition
- **Symmetric only in an interval** but benefits greatly from symmetric condition
- Significantly outperformed all losses in BER/AUC optimization using CNNs