On Symmetric Losses for Learning from Corrupted Labels

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

Abstract

This paper aims to provide a better understanding of a symmetric loss. First, we emphasize that using a symmetric loss is advantageous in the balanced error rate (BER) minimization and area under the receiver operating characteristic curve (AUC) maximization from corrupted labels. Second, we prove general theoretical properties of symmetric losses, including a classification-calibration condition, excess risk bound, conditional risk minimizer, and AUC-consistency condition. Third, since all nonnegative symmetric losses are non-convex, we propose a convex barrier hinge loss that benefits significantly from the symmetric condition, although it is not symmetric everywhere. Finally, we conduct experiments to validate the relevance of the symmetric condition.

1. Introduction

In the real-world, it is unrealistic to expect that clean fully-supervised data can always be obtained. Weakly-supervised learning is a learning paradigm to mitigate this problem (Zhou, 2017). For example, labelers are not necessarily experts or even human experts can make mistakes. Learning under noisy labels is an example of weakly-supervised learning that relaxes the assumption that labels are always accurate (Aslam & Decatur, 1996; Biggio et al., 2011; Cesa-Bianchi et al., 1999; Natarajan et al., 2013). Other examples of weakly-supervised learning are learning from positive and unlabeled data (du Plessis et al., 2015; 2014; Kiryo et al., 2017), learning from pairwise similarity and unlabeled data (Bao et al., 2018), and learning from complementary labels (Ishida et al., 2017).

A loss function that satisfies a symmetric condition has demonstrated its usefulness in weakly-supervised learning, e.g., one can use a symmetric loss to simplify a risk estimator in learning from positive-unlabeled data (du Plessis et al., 2014). This simplification allows the use of a cost-sensitive learning library to implement the risk estimator directly. Not limited to the simplification of the risk estimator, symmetric losses are known to be robust in the symmetric label noise scenarios (Manwani & Sastry, 2013; Ghosh et al., 2015). However, the symmetric label noise assumption is restrictive and may not be practical since it assumes that a label of each pattern may flip independently with the same probability.

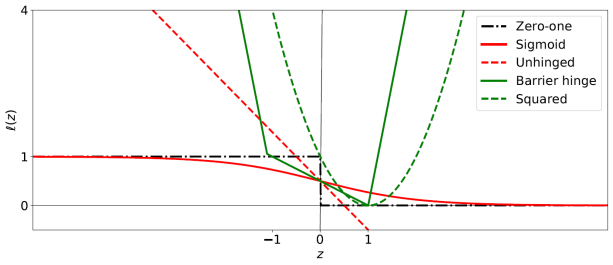

Figure 1. Examples of losses used in this paper. The zero-one loss, sigmoid loss, and unhinged loss are symmetric, i.e., $\ell(z) + \ell(-z)$ is a constant. The barrier hinge loss is our proposed loss.

This paper elucidates the robustness of symmetric losses in a more general noise framework called the mutually contaminated distributions or corrupted labels framework (Scott et al., 2013). Many weakly-supervised learning problems can be formulated in the corrupted labels framework (Natarajan et al., 2013; Lu et al., 2018). Therefore, the robustness of learning from corrupted labels is highly desirable for many real-world applications.

Although it has been shown by Menon et al. (2015) that BER and AUC optimization from corrupted labels can be optimized without knowing the noise information, we point out that the use of non-symmetric losses may degrade the performance and therefore using a symmetric losses is preferable. Our experiments show that symmetric losses significantly outperformed many well-known non-symmetric losses when the given labels are corrupted. Furthermore, we provide a better understanding of symmetric losses by elucidating several general theoretical properties of symmetric losses, including a classification-calibration condition, excess risk bound, conditional risk minimizer, and

[1]Department of Computer Science, The University of Tokyo, Tokyo, Japan [2]RIKEN Center of Artificial Intelligence Project, Tokyo, Japan. Correspondence to: Nontawat Charoenphakdee <nontawat@ms.k.u-tokyo.ac.jp>, Jongyeong Lee <lee@ms.k.u-tokyo.ac.jp>, Masashi Sugiyama <sugi@k.u-tokyo.ac.jp>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

AUC-consistency. We show that many well-known symmetric losses are suitable for both classification and bipartite ranking problems. We also discuss the negative result of symmetric losses, which is the inability to recover the class probability given the risk minimizer. This suggests a limitation to use such symmetric losses for a task that requires a prediction confidence such as learning with a reject option (Chow, 1970; Yuan & Wegkamp, 2010).

Unfortunately, it is known that a nonnegative symmetric loss must be non-convex (du Plessis et al., 2014; Ghosh et al., 2015). Van Rooyen et al. (2015a) proposed an unhinged loss, which is convex, symmetric but negatively unbounded. In this paper, we propose a barrier hinge loss which is convex, nonnegative, and satisfies the symmetric condition in a subset of the domain space, not everywhere.

2. Preliminaries

In this section, we review the notation and related work of symmetric losses and learning from corrupted labels.

2.1. Notation

Let $\boldsymbol{x} \in \mathbb{R}^d$ be a d-dimensional real-valued pattern, $y \in \{-1, +1\}$ denote a class label which can only be either positive or negative, and $g\colon \mathbb{R}^d \to \mathbb{R}$ denote a prediction function. In binary classification, we use $\operatorname{sign}(g(\boldsymbol{x}))$ to determine the predicted label of a prediction function, where $\operatorname{sign}(g(\boldsymbol{x})) = 1$ if $g(\boldsymbol{x}) > 0$, -1 if $g(\boldsymbol{x}) < 0$, and 0 otherwise. $\mathbb{E}_{\mathrm{P}}[\cdot]$ and $\mathbb{E}_{\mathrm{N}}[\cdot]$ denote the expectations of $\boldsymbol{x}$ over $p(\boldsymbol{x}|y=1)$ and $p(\boldsymbol{x}|y=-1)$, respectively. $\eta(\boldsymbol{x})$ indicates the class probability $p(y=1|\boldsymbol{x})$ of a pattern $\boldsymbol{x}$. In this paper, we consider a margin loss $\ell\colon \mathbb{R} \to \mathbb{R}$ that takes only one argument, which is typically $yg(\boldsymbol{x})$. Table 1 shows examples of margin losses.

2.2. Symmetric Losses

Note that the notion of a *symmetric loss* can be ambiguous since there are many definitions of symmetric loss (see Natarajan et al. (2013); Reid & Williamson (2010) for other definitions). In this paper, we consider a *symmetric* loss from the perspective that it is a margin loss, $\ell\colon \mathbb{R} \to \mathbb{R}$ that satisfies the symmetric condition, i.e., $\ell(z) + \ell(-z) = K$, where K is a constant. Examples of such losses are the zero-one loss, unhinged loss and sigmoid loss which are described in Figure 1.

The advantage of using a symmetric loss was investigated in the symmetric label noise scenario (Manwani & Sastry, 2013; Ghosh et al., 2015; Van Rooyen et al., 2015a). The results from Long & Servedio (2010) suggested that convex losses are non-robust in this scenario and this motivated the use of a robust non-convex loss in the symmetric label noise scenario. Ghosh et al. (2015) proved that the symmetric condition is sufficient for a loss to be robust in this scenario. Van Rooyen et al. (2015a) later proposed an unhinged loss, which is the only possible convex loss to be symmetric, but it needs to be negatively unbounded. The negative unboundedness is not a common property for a loss function, which avoids the condition in Long & Servedio (2010) to achieve the robustness in the symmetric label noise scenario. Another notable extension of a symmetric condition is the extension to a multiclass setting (Ghosh et al., 2017).

This paper considers a noise framework called mutually contaminated distributions or corrupted labels framework (Scott et al., 2013), where the symmetric label noise is a special case of the corrupted labels framework (Menon et al., 2015). Then, we discuss a problem of non-symmetric losses in this scenario and emphasize that advantage of symmetric losses.

2.3. Learning from Corrupted Labels

In the corrupted labels scenario, we are given two sets of data drawn from the corrupted positive and corrupted negative marginal distributions respectively as follows:

$$\{\boldsymbol{x}_i\}_{i=1}^{n_{\mathrm{CP}}} \overset{\text{i.i.d.}}{\sim} \pi p(\boldsymbol{x}|y=1) + (1-\pi)p(\boldsymbol{x}|y=-1),$$
$$\{\boldsymbol{x}_j\}_{j=1}^{n_{\mathrm{CN}}} \overset{\text{i.i.d.}}{\sim} \pi' p(\boldsymbol{x}|y=1) + (1-\pi')p(\boldsymbol{x}|y=-1),$$

where n_{CP} denotes the number of corrupted positive patterns and π is the class prior $p(y=1)$ for the corrupted positive distribution, i.e, a proportion of clean positive data in the corrupted positive data. n_{CN} and π' are defined similarly for the corrupted negative data. We denote $X_{\mathrm{CP}} := \{\boldsymbol{x}_i\}_{i=1}^{n_{\mathrm{CP}}}$ as a corrupted positive sample and $X_{\mathrm{CN}} := \{\boldsymbol{x}_j\}_{j=1}^{n_{\mathrm{CN}}}$ as a corrupted negative sample. $p(\boldsymbol{x}|y)$ denotes the class conditional density. In this setting, $\pi \neq \pi'$ but the class condition probabilities $p(\boldsymbol{x}|y)$ are identical for both sets. Clean data implies $\pi = 1, \pi' = 0$. The class prior in this case can also be interpreted as the noise rate (Menon et al., 2015), where $(1-\pi)$ is the noise rate for positive data and π' is the noise rate for negative data. We assume $\pi > \pi'$ for simplicity. Otherwise, labels from the classifier must be flipped.

Menon et al. (2015) first showed that BER and AUC optimization from corrupted labels yield the same minimizer as minimizing from the clean labels. However, in this paper, we take a closer look of this problem and point out that the use of surrogate losses may yield different minimizers and degrade the performance. Another notable work in this corrupted labels setting is the classification from two sets of unlabeled data (Lu et al., 2018). They proposed an unbiased risk estimator for the classification error metric in this setting. BER is a special case of the classification error metric where the class prior is balanced. Nevertheless, their unbiased risk estimator requires the knowledge of the class priors of the two training distributions and the test distribution. This paper only focuses on BER and AUC optimization and does not require any class prior information.

3. The Importance of Symmetric Losses in BER and AUC Optimization

In this section, we show that using a symmetric loss is preferable for BER and AUC optimization from corrupted labels without class prior estimation. BER and AUC are popular metrics for imbalanced data classification (Cheng et al., 2002; Guyon et al., 2005). Furthermore, AUC is also known as an evaluation metric for bipartite ranking (Narasimhan & Agarwal, 2013; Menon & Williamson, 2016). In the corrupted labels framework, the class prior estimation problem is known to be a bottleneck in this framework since it is an unidentifiable problem unless a restrictive condition is applied (Blanchard et al., 2010; Scott, 2015). Thus, being able to minimize BER and AUC without estimating class priors is a great advantage in practice.

Related work: Menon et al. (2015) proved that for the zero-one loss, the clean and corrupted BER/AUC risks have the same minimizer. However, it remains unclear whether the same result holds for any surrogate losses. Later, Van Rooyen et al. (2015b) generalized the result of BER minimization in Menon et al. (2015) from the zero-one loss to any symmetric losses. In this paper, we analyze both BER and AUC optimization from corrupted labels by first proving the relationship between the clean surrogate risk and corrupted surrogate risk for *any* surrogate losses. Our results indicate that using a non-symmetric loss may not yield the same minimizer for the clean and corrupted risks since it may suffer from excessive terms (see Sections 3.1 and 3.2). Then, we clarify that similarly to BER minimization that was proven by Van Rooyen et al. (2015b), using a symmetric loss is also advantageous for AUC maximization. We are also the first to provide the experimental results for validating the advantage of symmetric losses for BER and AUC optimization from corrupted labels in practice.

3.1. Area under the Receiver Operating Characteristic Curve (AUC) Maximization

In AUC maximization, we consider the following AUC risk (Narasimhan & Agarwal, 2013):

$$R^{\ell}_{\mathrm{AUC}}(g) = \mathbb{E}_{\mathrm{P}}[\mathbb{E}_{\mathrm{N}}[\ell(f(\boldsymbol{x}_{\mathrm{P}}, \boldsymbol{x}_{\mathrm{N}}))]], \tag{1}$$

where $f(\boldsymbol{x}, \boldsymbol{x}') = g(\boldsymbol{x}) - g(\boldsymbol{x}')$. The expected AUC score is $1 - R^{\ell_{0\text{-}1}}_{\mathrm{AUC}}(g)$. Therefore, we can maximize the AUC score by minimizing the AUC risk. Since we do not have access to clean data, let us consider a corrupted AUC risk with a surrogate loss ℓ that treats X_{CP} as being positive and X_{CN} as being negative:

$$R^{\ell}_{\text{AUC-Corr}}(g) = \mathbb{E}_{\mathrm{CP}}[\mathbb{E}_{\mathrm{CN}}[\ell(f(\boldsymbol{x}_{\mathrm{CP}}, \boldsymbol{x}_{\mathrm{CN}}))]].$$

The following theorem shows that by using a symmetric loss, the minimizers of $R^{\ell}_{\text{AUC-Corr}}(g)$ and $R^{\ell}_{\mathrm{AUC}}(g)$ are identical (its proof is given in Appendix).

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*

$$\begin{aligned} R^{\ell}_{\text{AUC-Corr}}(g) = (\pi - \pi') R^{\ell}_{\mathrm{AUC}}(g) \\ + \underbrace{(1-\pi)\pi' \mathbb{E}_{\mathrm{P}}[\mathbb{E}_{\mathrm{N}}[\gamma^{\ell}(\boldsymbol{x}_{\mathrm{P}}, \boldsymbol{x}_{\mathrm{N}})]]}_{\text{Excessive term}} \\ + \underbrace{\frac{\pi\pi'}{2} \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\gamma^{\ell}(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}})]]}_{\text{Excessive term}} \\ + \underbrace{\frac{(1-\pi)(1-\pi')}{2} \mathbb{E}_{\mathrm{N}'}[\mathbb{E}_{\mathrm{N}}[\gamma^{\ell}(\boldsymbol{x}_{\mathrm{N}'}, \boldsymbol{x}_{\mathrm{N}})]]}_{\text{Excessive term}}. \end{aligned}$$

Corollary 2. *Let ℓ be a symmetric loss such that $\ell(z) + \ell(-z) = K$, where K is a constant. $R^{\ell}_{\text{AUC-Corr}}(g)$ can be expressed as*

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

Corollary 2 can be obtained simply by substituting $\gamma^{\ell}(\boldsymbol{x}, \boldsymbol{x}')$ with K. This suggests that the excessive term becomes a constant when using a symmetric loss and guarantees that the minimizers of $R^{\ell}_{\text{AUC-Corr}}(g)$ and $R^{\ell}_{\mathrm{AUC}}(g)$ are identical. On the other hand, if a loss is non-symmetric, then the excessive terms are not constants and the minimizers of both risks may differ. A special case of this setting where $\pi = 1$ has been studied by Sakai et al. (2018). They showed that a convex surrogate loss can be applied but π' needs to be estimated in order to cancel the excessive term. By using a symmetric loss, the class prior estimation is not required and the given positive patterns can also be corrupted. More generally, our results indicate that using a symmetric loss for AUC maximization from corrupted labels yields the same minimizer as clean labels and can be applied to various weakly-supervised learning settings (Natarajan et al., 2013; Niu et al., 2016; Bao et al., 2018; Lu et al., 2018).

3.2. Balanced Error Rate (BER) Minimization

Consider the following misclassification risk:

$$R^{\ell}_{\mathrm{BER}}(g) = \frac{1}{2}\left[\mathbb{E}_{\mathrm{P}}\left[\ell(g(\boldsymbol{x}))\right] + \mathbb{E}_{\mathrm{N}}\left[\ell(-g(\boldsymbol{x}))\right]\right].$$

The BER minimization problem is equivalent to minimizing $R^{\ell_{0\text{-}1}}_{\mathrm{BER}}(g)$. i.e., the classification risk with the zero-one loss when the class prior of the test distribution is balanced.

Let us define

$$R^{\ell}_{\text{BER-Corr}}(g) = \frac{1}{2}\left[R^{\ell}_{\mathrm{CP}}(g) + R^{\ell}_{\mathrm{CN}}(g)\right],$$

where

$$\begin{aligned} R^{\ell}_{\mathrm{CP}}(g) &= \pi \mathbb{E}_{\mathrm{P}}[\ell(g(\boldsymbol{x}))] + (1-\pi)\mathbb{E}_{\mathrm{N}}[\ell(g(\boldsymbol{x}))], \\ R^{\ell}_{\mathrm{CN}}(g) &= \pi' \mathbb{E}_{\mathrm{P}}[\ell(-g(\boldsymbol{x}))] + (1-\pi')\mathbb{E}_{\mathrm{N}}[\ell(-g(\boldsymbol{x}))]. \end{aligned}$$

Then, we state the following theorem (its proof is given in Appendix).

Theorem 3. *Let* $\gamma^\ell(\boldsymbol{x}) = \ell(g(\boldsymbol{x})) + \ell(-g(\boldsymbol{x}))$, $R^\ell_{\text{BER-Corr}}(g)$ *can be expressed as*

$$R^\ell_{\text{BER-Corr}}(g) = (\pi - \pi')R^\ell_{\text{BER}}(g) + \underbrace{\frac{\pi'\mathbb{E}_{\text{P}}[\gamma^\ell(\boldsymbol{x})] + (1-\pi)\mathbb{E}_{\text{N}}[\gamma^\ell(\boldsymbol{x})]}{2}}_{\text{Excessive term}}.$$

By observing an excessive term, we can directly obtain the following corollary, which coincides with the existing result by Van Rooyen et al. (2015b).

Corollary 4 (Van Rooyen et al. (2015b))**.** *Let* ℓ *be a symmetric loss such that* $\ell(z) + \ell(-z) = K$*, where* K *is a constant.* $R^\ell_{\text{Bal-Corr}}(g)$ *can be expressed as*

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

Similarly to Corollary 2, if a loss ℓ is symmetric, then the excessive term is a constant and the minimizers of $R^\ell_{\text{Bal-Corr}}(g)$ and $R^\ell_{\text{Bal}}(g)$ are guaranteed to be identical.

4. Theoretical Properties of Symmetric Losses

In this section, we investigate general theoretical properties of symmetric losses. Since all nonnegative symmetric losses are non-convex, many convenient conditions that assume a loss function is convex cannot be applied (Zhang, 2004; Bartlett et al., 2006; Gao & Zhou, 2015; Niu et al., 2016). Nevertheless, thanks to the symmetric condition, we show that it is possible to derive general theoretical properties of a symmetric loss.

4.1. Classification-calibration

The main motivation to use a surrogate loss in binary classification is that the zero-one loss is discontinuous and therefore difficult to optimize (Ben-David et al., 2003; Feldman et al., 2012). A natural question is what kind of surrogate losses can be used instead of the zero-one loss. This problem has been studied extensively in binary classification (Zhang, 2004; Bartlett et al., 2006). Classification-calibration is known to be a minimal requirement of a loss function for the binary classification task (see Bartlett et al. (2006) for more details on classification-calibration).

We derive the following theorem that establishes a necessary and sufficient condition for a symmetric loss to be classification-calibrated (its proof is given in Appendix).

Theorem 5. *A symmetric loss* $\ell\colon \mathbb{R} \to \mathbb{R}$ *such that* $\ell(z) + \ell(-z)$ *is a constant is classification-calibrated if and only if* $\inf_{\alpha>0} \ell(\alpha) < \inf_{\alpha\leq 0} \ell(\alpha)$.

The following corollary is straightforward from the theorem above, but we emphasize it since it covers many surrogate symmetric losses, e.g., the sigmoid, ramp, and unhinged losses.

Corollary 6. *A non-increasing loss* $\ell\colon \mathbb{R} \to \mathbb{R}$ *such that* $\ell(z) + \ell(-z)$ *is a constant and* $\ell'(0) < 0$*, is classification-calibrated.*

Based on Theorem 5, by simply checking the condition whether $\inf_{\alpha>0} \ell(\alpha) < \inf_{\alpha\leq 0} \ell(\alpha)$ is necessary and sufficient to determine if a symmetric loss is classification-calibrated. Note that Corollary 6 is a sufficient condition that covers many symmetric losses such as the ramp loss and sigmoid loss. In general, the differentiability at zero of a symmetric loss is not required to verify the classification-calibrated condition unlike convex losses (Bartlett et al., 2006). Note that some specific symmetric losses such as the ramp loss and sigmoid loss were proven to be classification-calibrated (Bartlett et al., 2006; Niu et al., 2016). This paper provides a necessary and sufficient condition for all symmetric losses.

4.2. Excess Risk Bound

The excess risk bound provides a relationship between the excess risk of minimizing the misclassification risk with respect to the zero-one loss and the surrogate loss. It is known that an excess risk bound of a loss ℓ exists if and only if ℓ is classification-calibrated (Bartlett et al., 2006).

Consider the standard binary misclassification risk:

$$R^\ell(g) = \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim D}\left[\ell(yg(\boldsymbol{x}))\right]. \tag{2}$$

The following theorem indicates an excess risk bound for any classification-calibrated symmetric loss (its proof is given in Appendix).

Theorem 7. *An excess risk bound of a classification-calibrated symmetric loss* $\ell\colon \mathbb{R} \to \mathbb{R}$ *such that* $\ell(z) + \ell(-z)$ *is a constant can be expressed as*

$$R^{\ell_{0\text{-}1}}(g) - R^{\ell_{0\text{-}1}*} \leq \frac{R^\ell(g) - R^{\ell*}}{\inf_{\alpha\leq 0} \ell(\alpha) - \inf_{\alpha>0} \ell(\alpha)},$$

where $R^{\ell*} = \inf_g R^\ell(g)$ *and* $R^{\ell_{0\text{-}1}*} = \inf_g R^{\ell_{0\text{-}1}}(g)$.

The result suggests that the excess risk bound of any classification-calibrated symmetric loss is controlled only by the difference of the infima $\inf_{\alpha>0} \ell(\alpha) - \inf_{\alpha\leq 0} \ell(\alpha)$. Intuitively, the excess risk bound tells us that if the prediction function g minimizes the surrogate risk $R^\ell(g) = R^{\ell*}$, then the prediction function g must also minimize the misclassification risk $R^{\ell_{0\text{-}1}}(g) = R^{\ell_{0\text{-}1}*}$.

Table 1. Loss functions and their properties including the convexity, symmetricity, capability of recovering $\eta(\boldsymbol{x})$, and their conditional risk minimizers $f^{\ell*}(\boldsymbol{x})$. Although the conditional risk minimizers of each loss function are different, the sign of each minimizer $\mathrm{sign}(f^{\ell*}(\boldsymbol{x}))$ matches each other, which agrees with the Bayes-optimal classifier. The savage loss is proposed by Masnadi-Shirazi & Vasconcelos (2009). The minimizer $f^{\ell*}(\boldsymbol{x})$ of the ramp, sigmoid, and unhinged losses are unique if the prediction output is in $[-1, 1]$.

Loss name	$\ell(z)$	$f^{\ell*}(\boldsymbol{x})$	Convex	Symmetric	Recover $\eta(\boldsymbol{x})$
Zero-one	$-0.5\mathrm{sign}(z) + 0.5$	$\mathrm{sign}(\eta(\boldsymbol{x}) - 0.5)$	$\times$	✓	$\times$
Squared	$(1-z)^2$	$2\eta(\boldsymbol{x}) - 1$	✓	$\times$	✓
Hinge	$\max(0, 1-z)$	$\mathrm{sign}(\eta(\boldsymbol{x}) - 0.5)$	✓	$\times$	$\times$
Logistic	$\log(1+\exp(-z))$	$\log\left(\frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}\right)$	✓	$\times$	✓
Savage	$\left[(1+\exp(2z))^2\right]^{-1}$	$0.5\log\left(\frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}\right)$	$\times$	$\times$	✓
Ramp	$\max(0, \min(1, 0.5 - 0.5z))$	$\mathrm{sign}(\eta(\boldsymbol{x}) - 0.5)$	$\times$	✓	$\times$
Sigmoid	$[1+\exp(z)]^{-1}$	$\mathrm{sign}(\eta(\boldsymbol{x}) - 0.5)$	$\times$	✓	$\times$
Unhinged	$1 - z$	$\mathrm{sign}(\eta(\boldsymbol{x}) - 0.5)$	✓	✓	$\times$

4.3. Inability to Recover the Class Probability $\eta(\boldsymbol{x})$

We investigate the form of the conditional risk minimizer of a symmetric loss. The conditional risk minimizer is useful to know the behavior of a prediction function learned from minimizing such a surrogate loss. For example, we can recover a class probability $\eta(\boldsymbol{x})$ from a prediction function if a loss ℓ is a proper composite loss (Buja et al., 2005; Reid & Williamson, 2010). The mapping function to recover a class probability $\eta(\boldsymbol{x})$ depends on the conditional risk minimizer. For example, one can recover the class probability $\eta(\boldsymbol{x})$ of the squared loss by the relationship $\eta(\boldsymbol{x}) = \frac{f^{\ell_{\mathrm{sq}}*}(\boldsymbol{x})+1}{2}$. Table 1 shows the examples of classification-calibrated losses and their conditional risk minimizers.

Our following theorem states that the conditional risk minimizer of any classification-calibrated symmetric loss can be expressed as a scaled Bayes-optimal classifier (its proof is given in Appendix).

Theorem 8. *Let ℓ be a symmetric loss $\ell\colon \mathbb{R} \to \mathbb{R}$ such that $\ell(z) + \ell(-z)$ is a constant and classification-calibrated, if the minimum of ℓ exists and $M \in \arg\min_{\alpha\in\mathbb{R}} \ell(\alpha)$. Then, the condition risk minimizer of ℓ can be expressed as follows:*

$$f^{\ell*}(\boldsymbol{x}) = M\,\mathrm{sign}(\eta(\boldsymbol{x}) - \frac{1}{2}),$$

where $\eta(\boldsymbol{x}) = p(y = 1|\boldsymbol{x})$.

When a symmetric loss is classification-calibrated but the minimum does not exist, $M \to \infty$. Note that the minimizer of a symmetric loss does not need to be unique as there might exist many points that give the minimum value.

By observing the conditional risk minimizer in Theorem 8, it is obvious that the class probability $\eta(\boldsymbol{x})$ cannot be recovered from the conditional risk minimizer since it knows only whether $\eta(\boldsymbol{x}) > \frac{1}{2}$. This similar property has been observed and well-studied for the hinge loss $\ell_{\mathrm{hinge}}(z) = \max(0, 1-z)$, where its minimizer is the Bayes-optimal classifier $\mathrm{sign}(\eta(\boldsymbol{x}) - \frac{1}{2})$, which suggests that the hinge loss is not suitable for class probability estimation (Bartlett & Tewari, 2007; Buja et al., 2005; Reid & Williamson, 2010).

4.4. AUC-consistency

AUC-consistency is similar to classification-calibration but from the perspective of AUC maximization (Gao & Zhou, 2015), i.e., minimizing the pairwise conditional risk for AUC maximization instead of the pointwise conditional risk in bainry classification. The Bayes-optimal solution of AUC maximization is a function that has a strictly monotonic relationship with the class probability $\eta(\boldsymbol{x})$, which is a consequence of the Neyman-Pearson lemma (Menon & Williamson, 2016).

Our following lemma states that classification-calibration is necessary for a symmetric loss to be AUC-consistent (its proof is given in Appendix).

Lemma 9. *An AUC consistent symmetric loss $\ell\colon \mathbb{R} \to \mathbb{R}$ such that $\ell(z) + \ell(-z)$ is a constant, is classification-calibrated.*

Next, an interesting question is whether all classification-calibrated symmetric losses are AUC-consistency. We prove by giving a counterexample that unfortunately this is not the case (its proof is given in Appendix).

Proposition 10. *Classification-calibration is necessary yet insufficient for a symmetric loss $\ell\colon \mathbb{R} \to \mathbb{R}$ such that $\ell(z) + \ell(-z)$ to be AUC-consistent.*

Proposition 10 illustrates that there is a gap between classification-calibration and AUC-consistency for a symmetric loss. This gives rise to an important question whether well-known symmetric losses are AUC-consistent. We elucidate the positive result by establishing a sufficient condition for a symmetric loss to be AUC-consistent, which covers almost all existing surrogate symmetric losses to the best of our knowledge (its proof is given in Appendix).

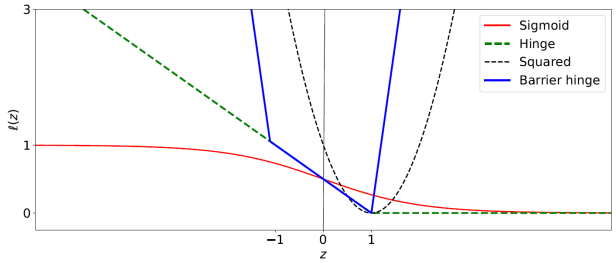

Figure 2. The barrier hinge loss scaled by 0.5 with $b = 10, r = 1$: $\ell(z) = 0.5\max(-10(1+z)+1, \max(10(z-1), 1-z))$, the hinge loss: $\ell(z) = \max(0, 0.5 - 0.5z)$ and the sigmoid loss: $\ell(z) = [1+\exp(z)]^{-1}$. The symmetric property holds for the barrier hinge loss for $z \in [-1, 1]$.

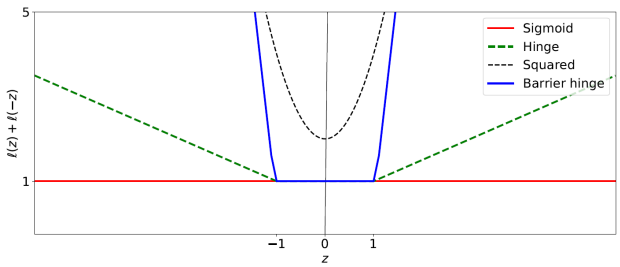

Figure 3. The plot of $\ell(z) + \ell(-z)$ of classification-calibrated losses. Only the sigmoid loss is symmetric. The hinge loss and barrier hinge loss satisfy the symmetric condition in $z \in [-1, 1]$.

Theorem 11. *A non-increasing loss $\ell\colon \mathbb{R} \to \mathbb{R}$ such that $\ell(z)+\ell(-z)$ is a constant and $\ell'(0) < 0$, is AUC-consistent.*

With Corollary 6 and Theorem 11, we show that a non-increasing symmetric loss that $\ell'(0) < 0$ is sufficient to be both classification-calibrated and AUC-consistent. Such conditions are not difficult to satisfy in practice. In fact, most surrogate symmetric losses that we are aware of satisfy this condition. Thus, the choice of symmetric losses is highly flexible for both the classification and bipartite ranking problems.

5. Barrier Hinge Loss

In this section, we propose a convex loss that benefits from the symmetric condition although it is not symmetric everywhere. Note that it is impossible to have a nonnegative symmetric loss (du Plessis et al., 2014; Ghosh et al., 2015). Our main idea to compensate this problem is to construct a loss that does not have to satisfy the symmetric condition everywhere, i.e ., $\ell(z) + \ell(-z)$ is a constant for every $z \in \mathbb{R}$. In this case, it is possible to find a classification-calibrated convex loss function that satisfies the symmetric condition only for an interval in $\mathbb{R}$. For example, the hinge loss satisfies the symmetric condition for $z \in [-1, 1]$. Nevertheless, the symmetric condition does not hold for z when $z \notin [-1, 1]$ and might suffer from the excessive term. Motivated by this observation, we propose a *barrier hinge loss*, which is a loss that satisfies a symmetric condition not everywhere and gives a large penalty when z is outside of the interval that is symmetric regardless of the correctness of the prediction.

Definition 12. A barrier hinge loss is defined as

$$\ell(z) = \max(-b(r+z)+r, \max(b(z-r), r-z)),$$

where $b > 1$ and $r > 0$.

Figure 2 shows a scaled barrier hinge loss with a specific parameter. Since a barrier hinge loss is convex, it is simple to verify that it is classification-calibrated since the derivative of the barrier hinge loss at zero is negative (Bartlett et al., 2006). Intuitively, barrier hinge losses are designed to give a very high penalty when z is in the non-symmetric area. As a result, a prediction function which is learned from a barrier hinge loss has an incentive to give a prediction value inside the symmetric area. The parameter r determines the width of the region that satisfies the symmetric property while the parameter b determines the slope of the penalty when z is in the non-symmetric area (b is expected to be a large value). In the experiment section, we show that our barrier hinge loss benefits from the symmetric condition and more robust than other non-symmetric losses. For fairness, we fix $b = 200$ and $r = 50$ for all datasets in the experiment section. Hence, one can further tune the parameters b and r to achieve a more preferable performance.

It is important to note that if we restrict the output of a loss to be in a symmetric region, e.g., $g(\boldsymbol{x}) \in [-1, 1]$ and $r \geq 1$, using the barrier hinge loss, unhinged loss, or standard hinge loss, are equivalent. Thus, the barrier hinge loss can also be viewed as a soft-constrained version of the unhinged loss.

6. Experimental Results

In this section, we present experimental results of BER and AUC optimization from corrupted labels. We used the balanced accuracy (1-BER) to evaluate the performance of BER minimization and the AUC score for AUC maximization. We also rescaled the score to be from 0 to 100. Note that higher balanced accuracy and AUC score are better. Training data were corrupted manually by simply mixing positive and negative data according to the class prior of the corrupted positive and corrupted negative data, i.e., π and π'. We compare the following loss functions: the squared loss, logistic loss, exponential loss, hinge loss, savage loss, sigmoid loss, unhinged loss, and barrier loss. Note that the class prior information is not given to the classifier. Moreover, only the sigmoid loss and unhinged loss are symmetric while our proposed barrier loss is not symmetric everywhere but is designed to benefit from the symmetric condition. One might suspect that the improve-

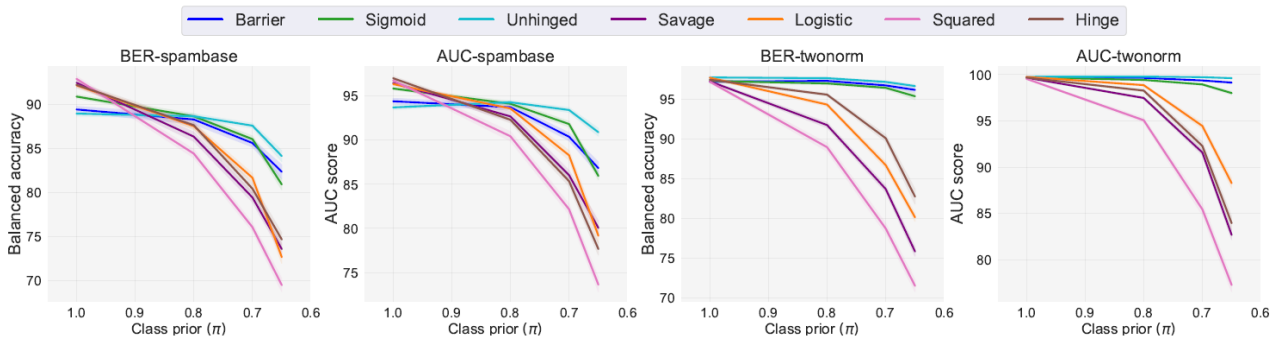

Figure 4. Mean balanced accuracy (1-BER) and AUC score using multilayer perceptrons (rescaled to 0-100) with varying noise rates $(\pi = 1.0, \pi' = 0.0), (\pi = 0.8, \pi' = 0.3), (\pi = 0.7, \pi' = 0.4), (\pi = 0.65, \pi' = 0.45)$. The experiments were conducted 20 times.

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)

ment of the performance comes from the fact that these symmetric losses are bounded from above and therefore more robust against noise. To emphasize the importance of the symmetric property, we also compare the performance with the savage loss, a loss function which is bounded and has demonstrated its robustness against outliers in classification (Masnadi-Shirazi & Vasconcelos, 2009). We also found that the double hinge loss (du Plessis et al., 2015) performed similarly to the hinge loss and thus we omit the results.

We design the experiments to answer the following three questions. First, does the symmetric condition helps significantly in BER and AUC optimization from corrupted labels? Second, do we need a loss to be symmetric everywhere to benefit from the robustness of symmetric losses? Third, does the negative unboundedness of the unhinged loss degrade the practical performance?

6.1. Experiments on UCI and LIBSVM Datasets

In this experiment, we used the one hidden layer multilayer perceptron $d - 500 - 1$ as a model. We used datasets from the UCI machine learning repository (Lichman et al., 2013) and LIBSVM (Chang & Lin, 2011). Training data consists of 500 corrupted positive data, 500 corrupted negative data, and balanced 500 clean test data. More details on the implementation, datasets, and full experimental results using more datasets can be found in Appendix. The objective functions of the neural networks were optimized using AMSGRAD (Reddi et al., 2018). The experiment code was implemented with Chainer (Tokui et al., 2015).

Figure 4 shows the performance of BER and AUC optimization with varying noise rates $(\pi = 1.0, \pi' = 0.0), (\pi = 0.8, \pi' = 0.3), (\pi = 0.7, \pi' = 0.4), (\pi = 0.65, \pi' = 0.45)$. Table 2 also shows the results where labels are highly corrupted ($\pi = 0.65$ and $\pi' = 0.45$). Although the savage loss is a bounded loss, its performance is not desirable when the labels are corrupted. It can be observed that when the data is clean ($\pi = 1.0$ and $\pi' = 0.0$), the performance of all losses are not significantly different. However, as the noise rate increases, the sigmoid loss, unhinged loss, and barrier loss significantly outperform other losses in this experiment. This suggests that only using a bounded loss is not sufficient to perform BER minimization from corrupted labels effectively. Therefore, the experimental results support our hypothesis that using symmetric losses can be preferable in the BER minimization problem from corrupted labels.

In this experiment, the unhinged loss performs well although it is negatively unbounded. This positive result of the un-

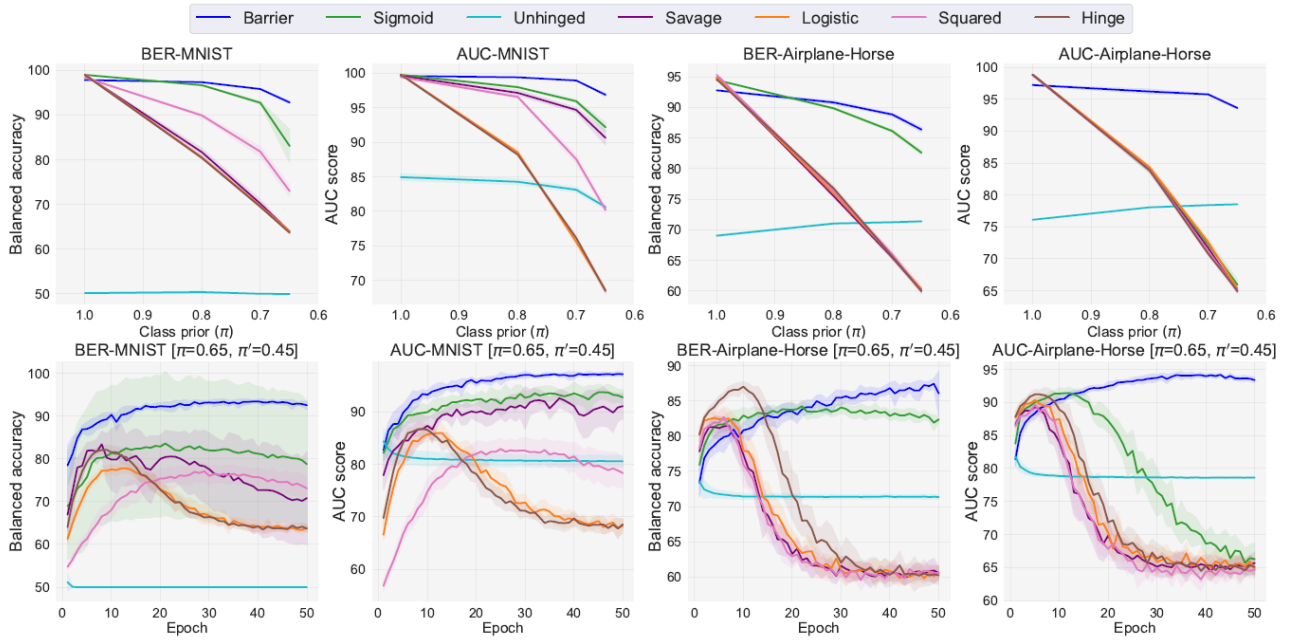

Figure 5. Mean balanced accuracy (1-BER) and AUC score using convolutional neural networks (rescaled to 0-100). (Top) the varying noise rates ranged from$(\pi = 1.0, \pi' = 0.0), (\pi = 0.8, \pi' = 0.3), (\pi = 0.7, \pi' = 0.4), (\pi = 0.65, \pi' = 0.45)$. (Bottom) the noise rate is $\pi = 0.65$ and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

hinged loss agrees with Van Rooyen et al. (2015a), where they used a linear-in-input model. However, our next experiment shows that the performance of the unhinged loss is less desirable when deeper neural networks are applied.

6.2. Experiments on MNIST and CIFAR-10

In this experiment, we used MNIST (LeCun, 1998) (Odd vs Even) and CIFAR-10 (Airplane vs Horse) (Krizhevsky & Hinton, 2009) as the datasets. We used the convolutional neural networks as the models for all losses. Full experimental results including the experiments on additional eight pairs of CIFAR10 and the implementation details can be found in Appendix. The objective functions were optimized using AMSGRAD (Reddi et al., 2018). The experiment code was implemented with PyTorch (Paszke et al., 2017).

Figure 5 (top) shows the performance on BER and AUC optimization with varying noise rates similarly to the previous experiment. It is observed that the unhinged loss failed miserably in BER minimization, although it outperformed other baselines when the labels are highly corrupted in CIFAR-10 (Airplane vs Horse). Our proposed barrier hinge loss is observed to be advantageous in this experiment.

Figure 5 (bottom) shows the performance on BER and AUC optimization from highly corrupted labels $(\pi = 0.65, \pi' = 0.45)$ as the training epoch increases. The unhinged loss is observed to converge very quickly but its performance is marginal. The performance of the barrier hinge loss is preferable and does not degrade as the number of epoch increases. For the sigmoid loss, it is observed that the performance also degraded for the AUC maximization in CIFAR-10 as the epoch increases although it degraded slower than other losses that do not benefit from the symmetric condition.

In summary, our experimental results support that the symmetric condition significantly contributes to improving the performance on BER and AUC optimization from corrupted labels. Our barrier hinge loss, which is not symmetric everywhere, also demonstrated its robustness in this experiment. Finally, the unhinged loss is observed to perform poorly when complex models such as the convolutional neural networks are applied for which the potential reason can be the negative unboundedness of the unhinged loss.

7. Conclusion

We analyze a class of symmetric losses. We showed that the symmetric condition of a loss contributes to the robustness of the BER and AUC optimization from corrupted labels. Moreover, we proved the general theoretical results to provide a better understanding of symmetric losses. We also proposed a convex barrier hinge loss that is not symmetric everywhere but benefits greatly from the symmetric condition. The experimental results showed the advantage of using a symmetric loss for the BER and AUC optimization from corrupted labels and also illustrated the problem when a loss is negatively unbounded, such as the unhinged loss.

Acknowledgement

We thank Han Bao and Zhenghang Cui for helpful discussion. We also thank anonymous reviewers for providing insightful comments. NC was supported by MEXT scholarship and MS was supported by JST CREST JPMJCR18A2.

References

Aslam, J. A. and Decatur, S. E. On the sample complexity of noise-tolerant learning. *Information Processing Letters*, 57(4):189–195, 1996.

Bao, H., Niu, G., and Sugiyama, M. Classification from pairwise similarity and unlabeled data. In *ICML*, pp. 452–461, 2018.

Bartlett, P. L. and Tewari, A. Sparseness vs estimating conditional probabilities: Some asymptotic results. *JMLR*, 8: 775–790, 2007.

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. *JASA*, 101(473):138–156, 2006.

Ben-David, S., Eiron, N., and Long, P. M. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003.

Biggio, B., Nelson, B., and Laskov, P. Support vector machines under adversarial label noise. In *ACML*, pp. 97–112, 2011.

Blanchard, G., Lee, G., and Scott, C. Semi-supervised novelty detection. *JMLR*, 11:2973–3009, 2010.

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft*, 2005.

Cesa-Bianchi, N., Dichterman, E., Fischer, P., Shamir, E., and Simon, H. U. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM*, 46(5): 684–719, 1999.

Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.-A., Morishita, S., Page, D., and Sese, J. KDD Cup 2001 report. *ACM SIGKDD Explorations Newsletter*, 3(2):47–64, 2002.

Chow, C. K. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970.

du Plessis, M. C., Niu, G., and Sugiyama, M. Analysis of learning from positive and unlabeled data. In *NeurIPS*, pp. 703–711, 2014.

du Plessis, M. C., Niu, G., and Sugiyama, M. Convex formulation for learning from positive and unlabeled data. In *ICML*, pp. 1386–1394, 2015.

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

Gao, W. and Zhou, Z.-H. On the consistency of auc pairwise optimization. In *IJCAI*, pp. 939–945, 2015.

Ghosh, A., Manwani, N., and Sastry, P. Making risk minimization tolerant to label noise. *Neurocomputing*, 160: 93–107, 2015.

Ghosh, A., Kumar, H., and Sastry, P. Robust loss functions under label noise for deep neural networks. In *AAAI*, pp. 1919–1925, 2017.

Guyon, I., Gunn, S., Ben-Hur, A., and Dror, G. Result analysis of the NIPS 2003 feature selection challenge. In *NeurIPS*, pp. 545–552, 2005.

Ishida, T., Niu, G., Hu, W., and Sugiyama, M. Learning from complementary labels. In *NeurIPS*, pp. 5639–5649, 2017.

Ishida, T., Niu, G., and Sugiyama, M. Binary classification from positive-confidence data. In *NeurIPS*, pp. 5917–5928, 2018.

Kiryo, R., Niu, G., du Plessis, M. C., and Sugiyama, M. Positive-unlabeled learning with non-negative risk estimator. In *NeurIPS*, pp. 1674–1684, 2017.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lichman, M. et al. UCI machine learning repository, 2013.

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. *Machine learning*, 78(3):287–304, 2010.

Lu, N., Niu, G., Menon, A. K., and Sugiyama, M. On the minimal supervision for training any binary classifier from only unlabeled data. *arXiv preprint arXiv:1808.10585*, 2018.

Manwani, N. and Sastry, P. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3): 1146–1151, 2013.

Masnadi-Shirazi, H. and Vasconcelos, N. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *NeurIPS*, pp. 1049–1056, 2009.

Menon, A., Van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *ICML*, pp. 125–134, 2015.

Menon, A. K. and Williamson, R. C. Bipartite ranking: a risk-theoretic perspective. *JMLR*, 17(1):6766–6867, 2016.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, pp. 807–814, 2010.

Narasimhan, H. and Agarwal, S. On the relationship between binary classification, bipartite ranking, and binary class probability estimation. In *NeurIPS*, pp. 2913–2921, 2013.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with noisy labels. In *NeurIPS*, pp. 1196–1204, 2013.

Niu, G., du Plessis, M. C., Sakai, T., Ma, Y., and Sugiyama, M. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *NeurIPS*, pp. 1199–1207, 2016.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of Adam and beyond. In *ICLR*, 2018.

Reid, M. D. and Williamson, R. C. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

Sakai, T., Niu, G., and Sugiyama, M. Semi-supervised auc optimization based on positive-unlabeled learning. *Machine Learning*, 107(4):767–794, 2018.

Scott, C. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *AISTATS*, pp. 838–846, 2015.

Scott, C., Blanchard, G., and Handy, G. Classification with asymmetric label noise: Consistency and maximal denoising. In *COLT*, pp. 489–511, 2013.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958, 2014.

Tokui, S., Oono, K., Hido, S., and Clayton, J. Chainer: a next-generation open source framework for deep learning. In *NeurIPS Workshop*, volume 5, 2015.

Van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In *NeurIPS*, pp. 10–18, 2015a.

Van Rooyen, B., Menon, A. K., and Williamson, R. C. An average classification algorithm. *arXiv preprint arXiv:1506.01520*, 2015b.

Yuan, M. and Wegkamp, M. Classification methods with reject option based on convex risk minimization. *JMLR*, 11:111–130, 2010.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pp. 56–85, 2004.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.