A. Proofs

We provide the proofs in this section.

A.1. Proof of Theorem 1

Proof. Recall that the AUC risk is:

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

Corrupted AUC risk where X_{CP} is assigned to be positive and X_{CN} as negative:

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

where

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

$R^{\ell}_{\text{AUC-Corr}}(g)$ can be rewritten as follows:

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

Let

$$\begin{aligned}
A &= \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\ell(f(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}}))]], \\
B &= \mathbb{E}_{\mathrm{N}}[\mathbb{E}_{\mathrm{P}}[\ell(f(\boldsymbol{x}_{\mathrm{N}}, \boldsymbol{x}_{\mathrm{P}}))]], \\
C &= \mathbb{E}_{\mathrm{P}}[\mathbb{E}_{\mathrm{N}}[\ell(f(\boldsymbol{x}_{\mathrm{P}}, \boldsymbol{x}_{\mathrm{N}}))]] = R^{\ell}_{\mathrm{AUC}}(g), \\
D &= \mathbb{E}_{\mathrm{N}'}[\mathbb{E}_{\mathrm{N}}[\ell(f(\boldsymbol{x}_{\mathrm{N}'}, \boldsymbol{x}_{\mathrm{N}}))]], \\
\gamma^{\ell} &= \mathbb{E}_{\mathrm{P}}[\mathbb{E}_{\mathrm{N}}[\ell(f(\boldsymbol{x}_{\mathrm{P}}, \boldsymbol{x}_{\mathrm{N}})) + \ell(f(\boldsymbol{x}_{\mathrm{N}}, \boldsymbol{x}_{\mathrm{P}}))]] = B + C, \\
\gamma^{\ell}(\boldsymbol{x}, \boldsymbol{x}') &= \ell(f(\boldsymbol{x}, \boldsymbol{x}')) + \ell(f(\boldsymbol{x}', \boldsymbol{x})).
\end{aligned}$$

First, we show that $A = \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\ell(f(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}}))]] = \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\frac{\gamma^{\ell}(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}})}{2}]]$:

$$\begin{aligned}
\mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\ell(f(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}}))]] &= \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\mathbb{1}_{\boldsymbol{x}_{\mathrm{P}'} = \boldsymbol{x}_{\mathrm{P}}}\ell(0) + \mathbb{1}_{\boldsymbol{x}_{\mathrm{P}'} \neq \boldsymbol{x}_{\mathrm{P}}}\ell(f(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}}))]] \\
&= \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\mathbb{1}_{\boldsymbol{x}_{\mathrm{P}'} = \boldsymbol{x}_{\mathrm{P}}}\ell(0)]] + \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\mathbb{1}_{\boldsymbol{x}_{\mathrm{P}'} \neq \boldsymbol{x}_{\mathrm{P}}}\ell(f(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}}))]] \\
&= 0 + \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[1 \times \ell(f(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}}))]] \\
&= \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\frac{\ell(f(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}})) + \ell(f(\boldsymbol{x}_{\mathrm{P}}, \boldsymbol{x}_{\mathrm{P}'}))}{2}]] \\
&= \mathbb{E}_{\mathrm{P}'}[\mathbb{E}_{\mathrm{P}}[\frac{\gamma^{\ell}(\boldsymbol{x}_{\mathrm{P}'}, \boldsymbol{x}_{\mathrm{P}})}{2}]].
\end{aligned}$$

D can also be rewritten in a same manner so it is omitted for brevity.

$$D = \mathbb{E}_{\mathrm{N}'}[\mathbb{E}_{\mathrm{N}}[\ell(f(\boldsymbol{x}_{\mathrm{N}'}, \boldsymbol{x}_{\mathrm{N}}))]] = \mathbb{E}_{\mathrm{N}'}[\mathbb{E}_{\mathrm{N}}[\frac{\gamma^{\ell}(\boldsymbol{x}_{\mathrm{N}'}, \boldsymbol{x}_{\mathrm{N}})}{2}]].$$

Then, we get the following result:

$$\begin{aligned}
R^{\ell}_{\text{AUC-Corr}}(g) &= \pi\pi' A + (1-\pi)\pi' B + \pi(1-\pi')C + (1-\pi)(1-\pi')D \\
&= \pi\pi' A + (1-\pi)\pi'(\gamma^{\ell} - C) + \pi(1-\pi')C + (1-\pi)(1-\pi')D \\
&= \pi\pi' A + (\pi' - \pi\pi')\gamma^{\ell} + (\pi - \pi')C + (1-\pi)(1-\pi')D \\
&\quad = (\pi - \pi')R^{\ell}_{\text{AUC}}(g) + (1-\pi)\pi'\mathbb{E}_{\text{P}}[\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x}_{\text{P}}, \boldsymbol{x}_{\text{N}})]] \\
&\qquad + \pi\pi'\mathbb{E}_{\text{P}'}[\mathbb{E}_{\text{P}}[\frac{\gamma^{\ell}(\boldsymbol{x}_{\text{P}'}, \boldsymbol{x}_{\text{P}})}{2}]] + (1-\pi)(1-\pi')\mathbb{E}_{\text{N}'}[\mathbb{E}_{\text{N}}[\frac{\gamma^{\ell}(\boldsymbol{x}_{\text{N}'}, \boldsymbol{x}_{\text{N}})}{2}]] \\
&\quad = (\pi - \pi')R^{\ell}_{\text{AUC}}(g) + (\pi' - \pi\pi')\mathbb{E}_{\text{P}}[\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x}_{\text{P}}, \boldsymbol{x}_{\text{N}})]] \\
&\qquad + \frac{\pi\pi'}{2}\mathbb{E}_{\text{P}'}[\mathbb{E}_{\text{P}}[\gamma^{\ell}(\boldsymbol{x}_{\text{P}'}, \boldsymbol{x}_{\text{P}})]] + \frac{(1-\pi)(1-\pi')}{2}\mathbb{E}_{\text{N}'}[\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x}_{\text{N}'}, \boldsymbol{x}_{\text{N}})]].
\end{aligned}$$

Therefore, minimizing $R^{\ell}_{\text{AUC-Corr}}(g)$ does not imply minimizing $R^{\ell}_{\text{AUC}}(g)$ unless $\ell(f(\boldsymbol{x}, \boldsymbol{x}')) + \ell(f(\boldsymbol{x}', \boldsymbol{x}))$ is a constant. □

A.2. Proof of 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

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

Proof. Recall that the balanced risk is:

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

Balanced corrupted risk where X_{CP} is assigned to be positive and X_{CN} as negative:

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

where

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

$R^{\ell}_{\text{BER-Corr}}(g)$ can be rewritten as follows:

$$\begin{aligned}
2R^{\ell}_{\text{BER-Corr}}(g) &= \pi\mathbb{E}_{\text{P}}[\ell(g(\boldsymbol{x}))] + (1-\pi)\mathbb{E}_{\text{N}}[\ell(g(\boldsymbol{x}))] \\
&\quad + \pi'\mathbb{E}_{\text{P}}[\ell(-g(\boldsymbol{x}))] + (1-\pi')\mathbb{E}_{\text{N}}[\ell(-g(\boldsymbol{x}))] \\
&= \pi\mathbb{E}_{\text{P}}[\ell(g(\boldsymbol{x}))] + (1-\pi)\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x}) - \ell(-g(\boldsymbol{x}))] \\
&\quad + \pi'\mathbb{E}_{\text{P}}[\gamma^{\ell}(\boldsymbol{x}) - \ell(g(\boldsymbol{x}))] + (1-\pi')\mathbb{E}_{\text{N}}[\ell(-g(\boldsymbol{x}))] \\
&= \pi\mathbb{E}_{\text{P}}[\ell(g(\boldsymbol{x}))] + (1-\pi)\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x})] - (1-\pi)\mathbb{E}_{\text{N}}[\ell(-g(\boldsymbol{x}))] \\
&\quad + \pi'\mathbb{E}_{\text{P}}[\gamma^{\ell}(\boldsymbol{x})] - \pi'\mathbb{E}_{\text{P}}[\ell(g(\boldsymbol{x}))] + (1-\pi')\mathbb{E}_{\text{N}}[\ell(-g(\boldsymbol{x}))] \\
&= \pi\mathbb{E}_{\text{P}}[\ell(g(\boldsymbol{x}))] - \pi'\mathbb{E}_{\text{N}}[\ell(-g(\boldsymbol{x}))] + \pi\mathbb{E}_{\text{N}}[\ell(-g(\boldsymbol{x}))] \\
&\quad - \pi'\mathbb{E}_{\text{P}}[\ell(g(\boldsymbol{x}))] + (1-\pi)\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x})] + \pi'\mathbb{E}_{\text{P}}[\gamma^{\ell}(\boldsymbol{x})] \\
&= (\pi - \pi')[\mathbb{E}_{\text{P}}[\ell(g(\boldsymbol{x}))] + \mathbb{E}_{\text{N}}[\ell(-g(\boldsymbol{x}))]] + \pi'\mathbb{E}_{\text{P}}[\gamma^{\ell}(\boldsymbol{x})] + (1-\pi)\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x})] \\
&= 2(\pi - \pi')R^{\ell}_{\text{BER}}(g) + \pi'\mathbb{E}_{\text{P}}[\gamma^{\ell}(\boldsymbol{x})] + (1-\pi)\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x})] \\
R^{\ell}_{\text{BER-Corr}}(g) &= (\pi - \pi')R^{\ell}_{\text{BER}}(g) + \frac{\pi'\mathbb{E}_{\text{P}}[\gamma^{\ell}(\boldsymbol{x})] + (1-\pi)\mathbb{E}_{\text{N}}[\gamma^{\ell}(\boldsymbol{x})]}{2}.
\end{aligned}$$

□

A.3. Conditional risk for binary classification

By making use of the symmetric property, i.e., $\ell(z) + \ell(-z) = K$, a pointwise conditional risk can be rewritten such that there is only one term depending on α as follows for a fixed $\boldsymbol{x}$:

$$\begin{aligned} C_\eta^\ell(\alpha) &= \eta\ell(\alpha) + (1-\eta)\ell(-\alpha) \\ &= \eta\ell(\alpha) + (1-\eta)(K - \ell(\alpha)) \\ &= (1-\eta)K + (2\eta - 1)\ell(\alpha), \end{aligned}$$

where $\eta = p(y = 1|\boldsymbol{x})$. It can be observed that $\ell(-z)$ can be expressed by $K - \ell(z)$. The symmetric property makes analysis simpler because $\ell(-z)$ can be rewritten as $\ell(z)$and the following general properties can be obtained by only rely on the symmetric property.

A.4. Proof of Theorem 5

Proof. Let $H(\eta) = \inf_{\alpha\in\mathbb{R}} C_\eta^\ell(\alpha)$ and $H^-(\eta) = \inf_{\alpha:\alpha(2\eta-1)\leq 0} C_\eta^\ell(\alpha)$.

First, consider the ψ-transform from the definition 2 of Bartlett et al. (2006). Consider $\ell\colon \mathbb{R} \to [0,\infty)$, function $\psi : [0,1] \to [0,\infty)$ by $\psi = \tilde{\psi}^{**}$, where

$$\tilde{\psi}(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right),$$

$g^{**}\colon [0,1] \to \mathbb{R}$ is the Fenchel-Legendre biconjugate of $g : [0,1] \to \mathbb{R}$ characterized by

$$\operatorname{epi} g^{**} = \overline{co}\operatorname{epi} g.$$

It is known that $\psi = \tilde{\psi}$ if and only if $\tilde{\psi}$ is convex. For more details, please refer to Bartlett et al. (2006).

Next, we use the following statements in Lemma 5 from Bartlett et al. (2006) which can be interpreted that, ℓ is classification-calibrated if and only if $\psi(\theta) > 0$ for all $\theta \in (0,1]$. Based on this statement, we prove the sufficient and necessary condition for symmetric losses to be classification-calibrated by showing that $\psi(\theta) > 0$ for all $\theta \in (0,1]$ if and only if $\inf_{\alpha>0} \ell(\alpha) < \inf_{\alpha\leq 0} \ell(\alpha)$.

Using the conditional risk of symmetric losses in the previous section, H and H^- can be written as

$$\begin{aligned} H(\eta) &= \inf_{\alpha\in\mathbb{R}} C_\eta^\ell(\alpha) \\ &= (1-\eta)K + \inf_{\alpha\in\mathbb{R}} (2\eta-1)\ell(\alpha), \\ H^-(\eta) &= \inf_{\alpha:\alpha(2\eta-1)\leq 0} C_\eta^\ell(\alpha) \\ &= (1-\eta)K + \inf_{\alpha:\alpha(2\eta-1)\leq 0} (2\eta-1)\ell(\alpha). \end{aligned}$$

Let $\tilde{\psi}(\theta) = H^-(\frac{1+\theta}{2}) - H(\frac{1+\theta}{2})$ where $\theta \in (0,1]$,

$$\begin{aligned} \tilde{\psi}(\theta) &= H^-(\frac{1+\theta}{2}) - H(\frac{1+\theta}{2}) \\ &= \inf_{\alpha:\alpha\theta\leq 0} \theta\ell(\alpha) - \inf_{\alpha\in\mathbb{R}} \theta\ell(\alpha) \\ &= \theta[\inf_{\alpha\leq 0} \ell(\alpha) - \inf_{\alpha\in\mathbb{R}} \ell(\alpha)]. \end{aligned}$$

Let C $= \inf_{\alpha\leq 0} \ell(\alpha) - \inf_{\alpha\in\mathbb{R}} \ell(\alpha)$ is a constant depends on the function.

$$\tilde{\psi}(\theta) = C\theta.$$

Here, $\tilde{\psi}(\theta)$ is linear and therefore convex. As a result, $\psi = \tilde{\psi}$. Based on Lemma 5 of Bartlett et al. (2006). ℓ is classification-callibrated if and only if $\psi(\theta) > 0$ for all $\theta \in (0, 1]$. In this case, θ is positive therefore, any symmetric loss function is classification-calibrated if and only if $C > 0$.

$$\inf_{\alpha \leq 0} \ell(\alpha) - \inf_{\alpha \in \mathbb{R}} \ell(\alpha) > 0$$
$$\inf_{\alpha \in \mathbb{R}} \ell(\alpha) < \inf_{\alpha \leq 0} \ell(\alpha)$$
$$\inf_{\alpha > 0} \ell(\alpha) < \inf_{\alpha \leq 0} \ell(\alpha).$$

Therefore, a symmetric loss ℓ is classification-calibrated if and only if $\inf_{\alpha > 0} \ell(\alpha) < \inf_{\alpha \leq 0} \ell(\alpha)$. □

A.5. Proof of Theorem 7

Proof. Once $\psi(\theta) = [\inf_{\alpha \leq 0} \ell(\alpha) - \inf_{\alpha > 0} \ell(\alpha)]\theta$ is obtained in the previous proof of classification-calibration for a symmetric loss. It is straightforward to obtain an excess risk bound based on Bartlett et al. (2006):

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

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

A.6. Proof of Theorem 8

Proof. Consider a conditional risk minimizer of a symmetric loss ℓ

$$\begin{aligned} f^*_{\ell}(\boldsymbol{x}) &= \arg\min_{\alpha \in \mathbb{R}} C^{\ell}_{\eta(\boldsymbol{x})}(\alpha) \\ &= \arg\min_{\alpha \in \mathbb{R}} (1 - \eta(\boldsymbol{x}))K + (2\eta(\boldsymbol{x}) - 1)\ell(\alpha). \end{aligned}$$

The constants can be ignored as it does not depend on α. Let us consider two cases of $\eta > \frac{1}{2}$ and $\eta < \frac{1}{2}$:

Case 1: $\eta > \frac{1}{2}$

$$\begin{aligned} f^*_{\ell}(\boldsymbol{x}) &= (1 - \eta(\boldsymbol{x}))K + \arg\min_{\alpha \in \mathbb{R}} (2\eta(\boldsymbol{x}) - 1)\ell(\alpha) \\ &= \arg\min_{\alpha \in \mathbb{R}} \ell(\alpha). \end{aligned}$$

Case 2: $\eta < \frac{1}{2}$

$$\begin{aligned} f^*_{\ell}(\boldsymbol{x}) &= (1 - \eta(\boldsymbol{x}))K + \arg\min_{\alpha \in \mathbb{R}} (2\eta(\boldsymbol{x}) - 1)\ell(\alpha) \\ &= \arg\max_{\alpha \in \mathbb{R}} \ell(\alpha). \end{aligned}$$

Suppose there are many α to satisfy the conditions. Due to the symmetric condition, We can express the following relations.

$$\arg\min_{\alpha \in \mathbb{R}} \ell(\alpha) = -\arg\max_{\alpha \in \mathbb{R}} \ell(\alpha),$$

where $-\arg\max_{\alpha \in \mathbb{R}} \ell(\alpha)$ means a set such that each element in the set $\arg\max_{\alpha \in \mathbb{R}} \ell(\alpha)$ is multiplied by -1. As a result, $f^*_{\ell}(\boldsymbol{x})$ can be simply written as follows:

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

where $\mathrm{M} \in \arg\min_{\alpha \in \mathbb{R}} \ell(\alpha)$. This result shows that the conditional risk minimizer of a symmetric loss can be expressed as the bayes classifier scaled by a constant. In the case of functions such that it is classification-calibrated and argmin cannot be obtained, $M \to \infty$. □

A.7. Introduction of AUC-consistency

In AUC maximization, we want to find the function g that minimizes the following risk:

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

Gao & Zhou (2015) showed that the Bayes optimal functions can be expressed as follows:

$$\begin{aligned}\mathcal{B} &= \{g : R^{\ell_{0\text{-}1}}_{\mathrm{AUC}}(g) = R^{\ell_{0\text{-}1}*}_{\mathrm{AUC}}(g)\} \\ &= \{g : (g(\boldsymbol{x}) - g(\boldsymbol{x}'))(\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}')) > 0 \text{ if } \eta(\boldsymbol{x}) \neq \eta(\boldsymbol{x}')\}\end{aligned}$$

Unlike classification-calibration, the Bayes optimal functions for AUC maximization depend on the *pairwise* class probability, i.e., the class probabilities for two data points are compared. The optimal function g is a function such that the sign of $g(\boldsymbol{x}) - g(\boldsymbol{x}')$ matches the sign of $\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}')$. Therefore, one solution of g is the class probability itself. Because when $g(\boldsymbol{x}) = \eta(\boldsymbol{x})$ for all $\boldsymbol{x}$, then $g(\boldsymbol{x}) - g(\boldsymbol{x}') = \eta(\boldsymbol{x}) - \eta(\boldsymbol{x}')$ which is exactly the same value as the function we want to match the sign with. As a result, it is arguable that the bipartite ranking problem based on the AUC score is easier than the class conditional probability estimation problem in the sense that the problem is solved if we have an access to $\eta(\boldsymbol{x})$. However, we only need to find a function g such that $\mathrm{sign}(g(\boldsymbol{x}) - g(\boldsymbol{x}')) = \mathrm{sign}(\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))$. *AUC-consistency* property can be treated as the minimum requirement of a loss function to be suitable for bipartite ranking (Gao & Zhou, 2015).

A.8. Proof of Lemma 9

A proof is based on a necessary of the notion of calibration in Gao & Zhou (2015), which we call AUC-calibration to avoid confusion in this paper. According to Gao & Zhou (2015), AUC-calibration is a necessary condition for AUC-consistency. Here, we prove that a symmetric loss is AUC-calibrated if and only if a symmetric loss is classification-calibrated.

Proof. For a symmetric loss ℓ, we can rewrite a pairwise conditional risk term in the infimum as follows:

$$\begin{aligned}\eta(1-\eta')\ell(\alpha) + \eta'(1-\eta)\ell(-\alpha) &= \eta(1-\eta')\ell(\alpha) + \eta'(1-\eta)(K - \ell(\alpha)) \\ &= \eta(1-\eta')\ell(\alpha) + \eta' K(1-\eta) - \eta'\ell(\alpha) + \eta\eta'\ell(\alpha) \\ &= (\eta-\eta')\ell(\alpha) + \eta' K(1-\eta).\end{aligned}$$

$$\begin{aligned}H^{-}(\eta,\eta') &> H(\eta,\eta') \\ H^{-}(\eta,\eta') - H(\eta,\eta') &> 0 \\ \frac{1}{2\pi(1-\pi)}[\inf_{\alpha:\alpha(\eta-\eta')\leq 0}(\eta-\eta')\ell(\alpha) - \inf_{\alpha\in\mathbb{R}}(\eta-\eta')\ell(\alpha)] &> 0 \\ \inf_{\alpha:\alpha(\eta-\eta')\leq 0}(\eta-\eta')\ell(\alpha) - \inf_{\alpha\in\mathbb{R}}(\eta-\eta')\ell(\alpha) &> 0\end{aligned}$$

Case 1: $\eta - \eta' > 0$

$$\begin{aligned}(\eta-\eta')[\inf_{\alpha:\alpha(\eta-\eta')\leq 0}\ell(\alpha) - \inf_{\alpha\in\mathbb{R}}\ell(\alpha)] &> 0 \\ \inf_{\alpha:\alpha\leq 0}\ell(\alpha) - \inf_{\alpha\in\mathbb{R}}\ell(\alpha) &> 0.\end{aligned}$$

Case 2: $\eta - \eta' < 0$

$$\begin{aligned}(\eta-\eta')[\sup_{\alpha:\alpha(\eta-\eta')\leq 0}\ell(\alpha) - \sup_{\alpha\in\mathbb{R}}\ell(\alpha)] &> 0 \\ \sup_{\alpha:\alpha(\eta-\eta')\leq 0}\ell(\alpha) - \sup_{\alpha\in\mathbb{R}}\ell(\alpha) &< 0 \\ \sup_{\alpha:\alpha\geq 0}\ell(\alpha) - \sup_{\alpha\in\mathbb{R}}\ell(\alpha) &< 0.\end{aligned}$$

The two inequalities are equivalent which proved in Section 4.9.6. Therefore, a symmetric loss must satisfy $\inf_{\alpha>0} \ell(\alpha) < \inf_{\alpha\leq 0} \ell(\alpha)$ to be AUC-calibrated. This is equivalent to classification-calibration condition for a symmetric loss. Next, it is known that AUC-calibration is a necessary condition for AUC-consistent (Gao & Zhou, 2015), therefore, a symmetric loss that is not classification-calibrated must not satisfy this condition, and thus not AUC-consistent.

This elucidates that classification-calibration is a necessary condition for a symmetric loss to be AUC-consistent. □

A.9. Proof of Proposition 10

Proof. Consider a pairwise conditional risk:

$$\begin{aligned}\eta(1-\eta')\ell(\alpha)+\eta'(1-\eta)\ell(-\alpha) &= \eta(1-\eta')\ell(\alpha)+\eta'(1-\eta)(K-\ell(\alpha))\\ &= \eta(1-\eta')\ell(\alpha)+\eta' K(1-\eta)-\eta'\ell(\alpha)+\eta\eta'\ell(\alpha)\\ &= (\eta-\eta')\ell(\alpha)+\eta' K(1-\eta). \end{aligned} \tag{3}$$

Then, let us consider a symmetric loss ℓ_{EX} such that $\ell_{\mathrm{EX}}(1) = 0$, $\ell_{\mathrm{EX}}(-1) = 1$, and 0.5 otherwise. It is straightforward to see that it is a symmetric loss where $\ell_{\mathrm{EX}}(z) + \ell_{\mathrm{EX}}(-z) = 1$. We are going to show that this loss is classification-calibrated but AUC-consistent. Moreover, we can see that $\inf_{\alpha>0} \ell_{\mathrm{EX}}(\alpha) < \inf_{\alpha\leq 0} \ell_{\mathrm{EX}}(\alpha)$. Therefore, ℓ_{EX} is classification-calibrated based on the previous theorem on a necessary and sufficient condition of a symmetric loss to be classification-calibrated.

Next, let us consider a uniform discrete distribution D_{U} that contains 3 possible supports $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\}$. Moreover, let $\eta(\boldsymbol{x}_1) = 1$, $\eta(\boldsymbol{x}_2) = 0.5$, $\eta(\boldsymbol{x}_3) = 0$.

Here, we prove Proposition 10 by a counterexample that the minimizer of the AUC risk with respect to ℓ_{EX} resulted in a function that behaves differently the Bayes-optimal solution of AUC maximization of a function that has a strictly monotonic relationship with the class probability $\eta(\boldsymbol{x})$ (Menon & Williamson, 2016), and therefore AUC-inconsistent.

Consider the following pairwise risk:

$$\begin{aligned} R^{\mathrm{pair}}_{\ell_{\mathrm{EX}}}(g) &= \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2} [\eta(\boldsymbol{x})(1-\eta(\boldsymbol{x}'))\ell_{\mathrm{EX}}(g(\boldsymbol{x})-g(\boldsymbol{x}'))\\ &\quad + \eta(\boldsymbol{x}')(1-\eta(\boldsymbol{x})\ell_{\mathrm{EX}}(g(\boldsymbol{x}')-g(\boldsymbol{x}))]\\ &= \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2} [(\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell_{\mathrm{EX}}(g(\boldsymbol{x})-g(\boldsymbol{x}'))+\eta' K(1-\eta)] \end{aligned}$$

Since we are only interested in the minimizer of the risk, let us ignore the constant term and rewrite the risk pair as follows:

$$R^{\mathrm{pair}}_{\ell_{\mathrm{EX}}}(g) = C_0 + C_1 \sum_{i=1}^{3}\sum_{j\neq i} \ell_{\mathrm{EX}}(g(\boldsymbol{x}_i)-g(\boldsymbol{x}_j)),$$

where C_0 and C_1 are some constants.

Let us consider the following g_1, g_2, g_3, g_4: $\mathbb{R}^d \to \mathbb{R}$,

$$\begin{aligned} g_1(\boldsymbol{x}_1) &= g_1(\boldsymbol{x}_2)+1 = g_1(\boldsymbol{x}_3)+1,\\ g_2(\boldsymbol{x}_1) &= g_2(\boldsymbol{x}_2)+1 = g_2(\boldsymbol{x}_3)+2,\\ g_3(\boldsymbol{x}_1) &= g_3(\boldsymbol{x}_2) = g_3(\boldsymbol{x}_3),\\ g_4(\boldsymbol{x}_1) &= g_4(\boldsymbol{x}_2) = g_4(\boldsymbol{x}_3)+1. \end{aligned}$$

Then, a function g that minimizes the risk $R^{\mathrm{pair}}_{\ell_{\mathrm{EX}}}(g)$ is the one that minimizes $\sum_{i=1}^{3}\sum_{j\neq i}\ell_{\mathrm{EX}}(g(\boldsymbol{x}_i)-g(\boldsymbol{x}_j)) = \sum_{i=1}^{3}\sum_{j\neq i}\ell_{\mathrm{EX}}(g(\boldsymbol{x}_i,\boldsymbol{x}_j))$. More precisely, there are six pairs to consider as can be observed in the following table.

Table 3. The illustrations of the values for each pair in the uniform discrete distribution supports.

Pair	$\eta_i - \eta_j$	$\ell_{\mathrm{EX}}(g_1(\boldsymbol{x}_i, \boldsymbol{x}_j))$	$\ell_{\mathrm{EX}}(g_2(\boldsymbol{x}_i, \boldsymbol{x}_j))$	$\ell_{\mathrm{EX}}(g_3(\boldsymbol{x}_i, \boldsymbol{x}_j))$	$\ell_{\mathrm{EX}}(g_4(\boldsymbol{x}_i, \boldsymbol{x}_j))$
$\eta_1 - \eta_2$	0.5	0	0	0.5	0.5
$\eta_1 - \eta_3$	1	0	0.5	0.5	0
$\eta_2 - \eta_1$	-0.5	1	1	0.5	0.5
$\eta_2 - \eta_3$	0.5	0.5	0	0.5	0
$\eta_3 - \eta_1$	-1	1	0.5	0.5	1
$\eta_3 - \eta_2$	-0.5	0.5	1	0.5	1

We can rank the score of each g by taking a weighted sum of column "$\eta_i - \eta_j$" in Table 3 to the column of the loss function of a function g. For example, for g_1, the score is $0.5 \times 0 + 1 \times 0 + (-0.5) \times 1 + 0.5 \times 0.5 + (-1) \times 1 + (-0.5) \times 0.5 = -1.5$. Note that the lower sum the better since we are interested in the minimizer.

The function g_2 is a function that is optimal with respect to the pairwise risk with respect to the zero-one loss, i.e., has a strictly monotonic relationship with the class probability $\eta(\boldsymbol{x})$. However, the score of g_2 is -1 which is worse than g_1 and g_4. In this scenario, g_1 and g_4 minimize the risk in this distribution which contradicts to the optimal solution of AUC optimization.

Note that g_1 and g_4 are the global minimizer of the risk, not only among g_1, g_2, g_3, g_4. Since ℓ_{EX} returns the same value of all input *except two points* which are 1 and -1, the minimizer of the risk is the one that the loss function returns 1 for the lowest weight, i.e., for $\eta_i - \eta_j = -1$ and $\eta_i - \eta_j = -0.5$.

Intuitively, to fill in the blanks for all pairs, once we pick where the loss will return 1 for two pairs, all other pairs will be fixed. For other terms, they will cancel each other out and therefore the variable term minimum pairwise risk in the distribution D with respect to the loss ℓ_{EX} is -1.5, which includes the one that is not the Bayes-optimal solution and the one that conforms to the Bayes-optimal solution is not included.

Thus, we conclude that ℓ_{EX}, which is a classification-calibrated symmetric loss is AUC-inconsistent. This suggests the gap between classification calibration and AUC-consistency for a symmetric loss. $\square$

A.10. Proof of Theorem 11

Proof. Recall the Bayes optimal functions for AUC-optimization Gao & Zhou (2015) :

$$\begin{aligned}\mathcal{B} &= \{g : R^{\ell_{0\text{-}1}}_{\mathrm{AUC}}(g) = R^{\ell_{0\text{-}1}*}_{\mathrm{AUC}}(g)\} \\ &= \{g : (g(\boldsymbol{x}) - g(\boldsymbol{x}'))(\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}')) > 0 \text{ if } \eta(\boldsymbol{x}) \neq \eta(\boldsymbol{x}')\}.\end{aligned}$$

Here, we consider ℓ as a non-increasing loss $\ell\colon \mathbb{R} \to \mathbb{R}$ such that $\ell(z) + \ell(-z)$ is a constant and $\ell'(0) < 0$.

Let us write

$$\begin{aligned}R^{\mathrm{pair}}_{\ell}(g) &= \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}' \sim D_X^2} [\eta(\boldsymbol{x})(1 - \eta(\boldsymbol{x}'))\ell(g(\boldsymbol{x}) - g(\boldsymbol{x}')) \\ &\quad + \eta(\boldsymbol{x}')(1 - \eta(\boldsymbol{x}))\ell(g(\boldsymbol{x}') - g(\boldsymbol{x}))] \\ &= \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}' \sim D_X^2} [(\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))\ell(g(\boldsymbol{x}) - g(\boldsymbol{x}')) + \eta' K(1 - \eta)].\end{aligned}$$

Next, we show that the minimizer of the AUC risk of ℓ, has a strictly monotonic relationship with the class probability $\eta(\boldsymbol{x})$. More precisely, we will prove the following inequality:

$$\inf_{g \notin \mathcal{B}} R^{\mathrm{pair}}_{\ell}(g) > \inf_{g} R^{\mathrm{pair}}_{\ell}(g). \tag{4}$$

We will prove by contradiction. First, let us assume that there is a function g_{B} that is not strictly monotonic to the class probability $\eta(\boldsymbol{x})$ but is a minimizer of the AUC risk R^{pair}_{ℓ}. Then, we prove that it is impossible since there always exists a function that can further minimize the AUC risk AUC risk R_ℓ^{pair}. Note that the key idea of the proof is similar to that of Gao & Zhou (2015) except the fact that a loss is not convex and we can make use of the symmetric property.

First, similarly to the proof of the previous proposition, by making use of symmetric property, let C_0, C_1, C_2, C_3 be some constants, we obtain the following

$$\begin{aligned} R_\ell^{\mathrm{pair}}(g) &= \frac{1}{2\pi(1-\pi)} \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2}[(\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))+\eta' K(1-\eta)]. \\ &= C_0 \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2}[(\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))] + C_1. \\ &= C_0 \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2,\eta(\boldsymbol{x})>\eta(\boldsymbol{x}')}[(\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\left(\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))-\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))\right)] + C_1. \\ &= C_0 \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2,\eta(\boldsymbol{x})>\eta(\boldsymbol{x}')}[(\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\left(2\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))-K\right)] + C_1. \\ &= C_2 \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2,\eta(\boldsymbol{x})>\eta(\boldsymbol{x}')}[(\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))] + C_3. \end{aligned} \tag{5}$$

The key advantage for the symmetric loss is that there is only one term that involves a loss for each pair $\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))$, this helps us handle the conditional risk easier similarly to the binary classification scenario.

Next, we will show that for any $g_{\mathrm{B}} \notin \mathcal{B}$ there exists a better function g_{G} such that

$$R_\ell^{\mathrm{pair}}(g_{\mathrm{B}}) > R_\ell^{\mathrm{pair}}(g_{\mathrm{G}}). \tag{6}$$

By ignoring constants, the term that a function g can minimize the risk for a symmetric loss is

$$R_\ell^{\mathrm{comp}}(g) = \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{x}'\sim D_X^2,\eta(\boldsymbol{x})>\eta(\boldsymbol{x}')}[(\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell(g(\boldsymbol{x})-g(\boldsymbol{x}'))]$$

To show that (6) holds, it suffices to show that

$$R_\ell^{\mathrm{comp}}(g_{\mathrm{B}}) > R_\ell^{\mathrm{comp}}(g_{\mathrm{G}}) \tag{7}$$

Then, we know that there exists $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, which is a pair such that $g_{\mathrm{B}}(\boldsymbol{x}_1) \leq g_{\mathrm{B}}(\boldsymbol{x}_2)$, but $\eta(\boldsymbol{x}_1) > \eta(\boldsymbol{x}_2)$. Let $\delta = |g_{\mathrm{B}}(\boldsymbol{x}_1) - g_{\mathrm{B}}(\boldsymbol{x}_2)| + \epsilon$, where $\epsilon > 0$.

Let us construct g_{G} as follows.

$$\begin{aligned} g_{\mathrm{G}}(\boldsymbol{x}) &= g_{\mathrm{B}}(\boldsymbol{x}) - \delta, \text{ if } \eta(\boldsymbol{x}) \leq \eta(\boldsymbol{x}_1) \\ g_{\mathrm{G}}(\boldsymbol{x}) &= g_{\mathrm{B}}(\boldsymbol{x}) + \delta, \text{ if } \eta(\boldsymbol{x}) > \eta(\boldsymbol{x}_1) \end{aligned}$$

Since $\eta(\boldsymbol{x}_1) > \eta(\boldsymbol{x}_2)$, $g_{\mathrm{B}}(\boldsymbol{x}_1) - g_{\mathrm{B}}(\boldsymbol{x}_2) \leq 0$, $g_{\mathrm{G}}(\boldsymbol{x}_1) - g_{\mathrm{G}}(\boldsymbol{x}_2) > 0$, and ℓ is non-increasing and $\ell'(0) < 0$, it is straightforward to see that

$$(\eta(\boldsymbol{x}_1)-\eta(\boldsymbol{x}_2))\ell(g_{\mathrm{B}}(\boldsymbol{x}_1)-g_{\mathrm{B}}(\boldsymbol{x}_2)) > (\eta(\boldsymbol{x}_1)-\eta(\boldsymbol{x}_2))\ell(g_{\mathrm{G}}(\boldsymbol{x}_1)-g_{\mathrm{G}}(\boldsymbol{x}_2)). \tag{8}$$

Next, we show that modifications of other pairs from the construction of g_{G} will not further increase the R_ℓ^{comp} with respect to $R_\ell^{\mathrm{comp}}(g_{\mathrm{B}})$. There are three following cases to consider.

Case 1: $A_1 = \{\boldsymbol{x} \text{ such that } \eta(\boldsymbol{x}) > \eta(\boldsymbol{x}_1)\}$. Since all $\boldsymbol{x} \in A_1$ are modified equally, i.e., $g_{\mathrm{G}}(\boldsymbol{x}) = g_{\mathrm{B}}(\boldsymbol{x}) + \delta$. For all $\boldsymbol{x}, \boldsymbol{x}' \in A_1$

$$\begin{aligned} (\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell(g_{\mathrm{B}}(\boldsymbol{x})-g_{\mathrm{B}}(\boldsymbol{x}')) &= (\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell((g_{\mathrm{B}}(\boldsymbol{x})+\delta)-(g_{\mathrm{B}}(\boldsymbol{x}')+\delta)) \\ &= (\eta(\boldsymbol{x})-\eta(\boldsymbol{x}'))\ell(g_{\mathrm{G}}(\boldsymbol{x})-g_{\mathrm{G}}(\boldsymbol{x}')) \end{aligned}$$

Case 2: $A_2 = \{\boldsymbol{x} \text{ such that } \eta(\boldsymbol{x}) \leq \eta(\boldsymbol{x}_1)\}$. Since all $\boldsymbol{x} \in A_2$ are modified equally, i.e., $g_{\mathrm{G}}(\boldsymbol{x}) = g_{\mathrm{B}}(\boldsymbol{x}) - \delta$. For all $\boldsymbol{x}, \boldsymbol{x}' \in A_2$

$$\begin{aligned}(\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))\ell(g_{\mathrm{B}}(\boldsymbol{x}) - g_{\mathrm{B}}(\boldsymbol{x}')) &= (\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))\ell((g_{\mathrm{B}}(\boldsymbol{x}) - \delta) - (g_{\mathrm{B}}(\boldsymbol{x}') - \delta)) \\ &= (\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))\ell(g_{\mathrm{G}}(\boldsymbol{x}) - g_{\mathrm{G}}(\boldsymbol{x}'))\end{aligned}$$

Case 3: For all $\boldsymbol{x} \in A_1$ and $\boldsymbol{x}' \in A_2$. Since ℓ is a non-increasing function and $\delta > 0$.

$$\begin{aligned}(\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))\ell(g_{\mathrm{B}}(\boldsymbol{x}) - g_{\mathrm{B}}(\boldsymbol{x}')) &\geq (\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))\ell(g_{\mathrm{B}}(\boldsymbol{x}) + \delta - g_{\mathrm{B}}(\boldsymbol{x}') + \delta) \\ &= (\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}'))\ell(g_{\mathrm{B}}(\boldsymbol{x}) - g_{\mathrm{B}}(\boldsymbol{x}') + 2\delta)\end{aligned}$$

Therefore, with the strict inequality (8) and other pairs will not further increase the risk higher than a bad function as shown in the analysis of three cases, we show that (7) must hold, and therefore (6) and (4) hold. As a result, it is impossible that $\inf_{g \notin \mathcal{B}} R_\ell^{\mathrm{pair}}(g) = \inf_g R_\ell^{\mathrm{pair}}(g)$ since we can always find a better function g_{G} compared with a function $g_{\mathrm{B}} \notin \mathcal{B}$.

Thus, we conclude that (4) holds. Once we show that (4) holds, we can directly use the results from the proof of Theorem 2 in Gao & Zhou (2015) without modification to show that ℓ is AUC-consistent. □

Note that we can further relax the condition $\ell'(0) < 0$, we only have to make sure a loss is not a constant function. Nevertheless, we prove this condition for $\ell'(0) < 0$ since this is not difficult to satisfy in practice and covers many surrogate losses in the literature to the best of our knowledge.

B. Details of Implementation and Datasets

B.1. Experiments on UCI and LIBSVM Datasets

We used nine datasets, namely *spambase*, *phoneme*, *phishing*, *phishing*, *waveform*, *susy*, *w8a*, *adult*, *twonorm*, *mushroom*. We used the one hiddent layer multilayer perceptron as a model ($d - 500 - 1$). We used 500 corrupted positive data, 500 corrupted negative data, and balanced 500 test data. The corruption for the training data can be done 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 used rectifier linear units (ReLU) (Nair & Hinton, 2010). Learning rate was set to 0.001, batch size was 500, and the number of epoch was 100. We ran 20 trials for each experiment and reported the mean values and standard error. 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).

B.2. Experiments on MNIST and CIFAR-10

MNIST: The MNIST dataset contains 60,000 gray-scale training images and 10,000 test images from digits 0 to 9. In this experiment which consider the binary classification, we used even and odd digits as positive and negative classes respectively. To make sure same data were not used as both positive and negative class, we sampled 15,000 images for each class. For instance, when noise rate is ($\pi = 0.7, \pi' = 0.4$), positive class consists of 10,500 even digits images and 4,500 odd digits images and negative class consists of 6,000 even digits images and 9,500 odd digits images respectively. The model used for MNIST was convolutional neural networks which is same architecture of Ishida et al. (2018): d-Conv[18,5,1,0]-Max[2,2]-Conv[48,5,1,0]-Max[2,2]-800-400-1, where Conv[18, 5, 1, 0] means 18 channels of 5×5 convolutions with stride 1 and padding 0, and Max[2,2] means max pooling with kernel size 2 and stride 2. We used rectifier linear units (ReLU) (Nair & Hinton, 2010) as activation function after fully connected layer followed by dropout layer (Srivastava et al., 2014) in the first two fully connected layer.

CIFAR-10: The CIFAR-10 dataset contains natural RGB images from 10 classes with 5,000 training images and 1,000 test images per class. Following Ishida et al. (2018), we set a class 'airplane' as the positive class and set one of other classes as negative class in order to construct binary classification problem. Thus, we conducted experiments on 9 pairs of airplane vs others. To make sure same data were not used as both positive and negative class, we sampled 4,540 images for each class. Note that we have a few data differently from MNIST, 4,540 is the highest number we can sure that same data were not duplicated. Same architecture of CNNs was used for experiment of CIFAR-10.

C. Additional Experimental Results

In this section, we show the experimental results on additional datasets from the main body.

C.1. BER Optimization Using UCI and LIBSVM Datasets

Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%. The experiments were conducted 20 times.

Table 4. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 1.0$ and $\pi' = 0$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	89.4(0.3)	89.0(0.3)	90.9(0.2)	92.2(0.2)	**92.2 (0.3)**	**92.9 (0.3)**	**92.5 (0.2)**
phoneme	5	75.2(0.4)	76.4(0.4)	78.9(0.4)	**82.0 (0.4)**	**82.5 (0.5)**	**82.1 (0.3)**	**82.5 (0.4)**
phishing	30	91.1(0.4)	87.5(0.3)	92.3(0.2)	**93.0 (0.2)**	**92.7 (0.2)**	**92.5 (0.3)**	**92.7 (0.2)**
waveform	21	86.7(0.4)	86.2(0.2)	89.8(0.3)	**91.2 (0.3)**	**91.3 (0.3)**	**90.7 (0.2)**	**90.8 (0.3)**
susy	18	71.3(0.4)	71.3(0.6)	74.1(0.5)	**77.0 (0.5)**	**77.5 (0.4)**	**77.2 (0.3)**	**77.1 (0.3)**
w8a	300	87.8(0.3)	83.6(0.4)	**89.6 (0.3)**	**89.8 (0.3)**	88.2(0.3)	**90.2 (0.3)**	**89.7 (0.3)**
adult	104	78.8(0.4)	79.2(0.3)	78.7(0.4)	**80.6 (0.5)**	79.6(0.4)	79.6(0.4)	**80.8 (0.4)**
twonorm	20	97.2(0.1)	**97.7 (0.1)**	97.3(0.2)	**97.7 (0.1)**	**97.5 (0.2)**	97.2(0.1)	97.2(0.2)
mushroom	98	98.3(0.2)	91.0(0.5)	**99.8 (0.0)**	**99.9 (0.1)**	**99.8 (0.1)**	**99.9 (0.0)**	**99.9 (0.1)**

Table 5. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 0.8$ and $\pi' = 0.3$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	**88.3 (0.5)**	**88.7 (0.3)**	**88.7 (0.3)**	87.5(0.4)	87.6(0.4)	84.4(0.5)	86.3(0.5)
phoneme	5	75.0(0.5)	75.7(0.4)	76.9(0.5)	**79.3 (0.5)**	79.0(0.4)	**79.7 (0.4)**	**80.2 (0.5)**
phishing	30	89.9(0.4)	86.1(0.4)	**91.5 (0.3)**	89.7(0.3)	90.5(0.3)	85.7(0.4)	88.5(0.5)
waveform	21	87.4(0.4)	86.8(0.3)	**88.7 (0.4)**	87.6(0.4)	**88.6 (0.3)**	84.4(0.5)	87.4(0.4)
susy	18	71.1(0.4)	71.2(0.5)	**73.6 (0.4)**	**73.1 (0.4)**	**74.1 (0.6)**	71.8(0.6)	**73.2 (0.5)**
w8a	300	**85.8 (0.5)**	84.0(0.5)	81.2(0.4)	76.5(0.5)	73.2(0.7)	74.1(0.5)	78.1(0.4)
adult	104	**77.9 (0.4)**	**78.1 (0.5)**	**77.4 (0.4)**	75.2(0.6)	73.7(0.5)	70.8(0.5)	74.6(0.6)
twonorm	20	**97.3 (0.2)**	**97.6 (0.1)**	97.0(0.2)	94.3(0.2)	95.6(0.2)	89.0(0.5)	91.8(0.3)
mushroom	98	97.9(0.3)	94.8(0.6)	**99.1 (0.2)**	97.5(0.2)	**98.9 (0.1)**	93.6(0.3)	97.7(0.2)

Table 6. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 0.7$ and $\pi' = 0.4$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	85.6(0.4)	**87.6 (0.3)**	86.1(0.4)	81.7(0.5)	80.4(0.6)	76.1(0.5)	79.4(0.5)
phoneme	5	**75.8 (0.3)**	75.5(0.6)	**76.8 (0.7)**	**76.9 (0.6)**	**76.1 (0.6)**	**76.6 (0.8)**	**76.2 (0.7)**
phishing	30	**87.9 (0.7)**	86.0(0.5)	**89.2 (0.5)**	84.1(0.5)	84.4(0.6)	77.5(0.5)	82.2(0.6)
waveform	21	86.6(0.3)	86.6(0.5)	**88.3 (0.4)**	82.4(0.4)	84.6(0.5)	76.0(0.6)	79.4(0.6)
susy	18	70.2(0.5)	**70.6 (0.7)**	**71.3 (0.4)**	68.3(0.8)	68.4(0.5)	66.9(0.5)	67.8(0.5)
w8a	300	77.7(0.7)	**80.4 (0.6)**	71.2(0.6)	68.0(0.5)	65.9(0.6)	65.7(0.6)	68.4(0.8)
adult	104	**75.9 (0.4)**	**76.9 (0.6)**	75.3(0.5)	69.4(0.5)	69.0(0.6)	63.2(0.6)	67.4(0.5)
twonorm	20	96.7(0.2)	**97.2 (0.1)**	96.4(0.2)	86.7(0.4)	90.1(0.4)	78.8(0.6)	83.7(0.4)
mushroom	98	**96.8 (0.5)**	92.2(0.9)	**96.6 (0.5)**	90.8(0.5)	95.1(0.6)	79.5(0.6)	90.2(0.4)

Table 7. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 0.65$ and $\pi' = 0.45$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	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)
phoneme	5	**74.5 (0.8)**	**73.4 (0.9)**	**74.5 (0.6)**	**73.4 (0.8)**	**73.8 (1.1)**	71.3(0.9)	71.0(0.7)
phishing	30	**86.2 (0.4)**	82.8(0.7)	84.9(0.7)	77.7(0.6)	78.8(0.9)	69.1(0.8)	73.3(0.7)
waveform	21	**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)
susy	18	**68.3 (0.6)**	**68.9 (0.8)**	**66.9 (0.9)**	64.8(0.8)	65.1(0.8)	61.7(0.7)	64.6(0.7)
w8a	300	71.3(0.8)	**73.1 (0.5)**	65.1(0.7)	62.4(0.7)	61.1(0.6)	60.6(0.5)	62.3(0.6)
adult	104	73.2(0.7)	**74.7 (0.6)**	69.9(1.0)	64.8(0.8)	64.2(1.0)	59.1(0.6)	63.2(0.8)
twonorm	20	**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)
mushroom	98	**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)

C.2. AUC Optimization Using UCI and LIBSVM Datasets

Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%. The experiments were conducted 20 times.

Table 8. Mean AUC score and standard error for AUC maximization from corrupted labels, where $\pi = 1.0$ and $\pi' = 0.0$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	94.4(0.3)	93.7(0.2)	95.9(0.1)	96.4(0.2)	**97.0 (0.2)**	**96.8 (0.2)**	96.5(0.2)
phoneme	5	81.8(0.5)	82.3(0.4)	84.2(0.3)	**87.4 (0.3)**	**88.1 (0.4)**	**87.3 (0.3)**	**87.9 (0.4)**
phishing	30	97.3(0.1)	93.9(0.2)	97.6(0.1)	**97.9 (0.1)**	**97.9 (0.1)**	**97.7 (0.1)**	**97.8 (0.1)**
waveform	21	95.3(0.2)	90.3(0.4)	96.0(0.2)	**96.3 (0.2)**	**96.8 (0.1)**	96.1(0.2)	**96.6 (0.1)**
susy	18	81.3(0.3)	78.1(0.6)	83.1(0.5)	**84.7 (0.4)**	**85.5 (0.4)**	**85.0 (0.4)**	84.5(0.3)
w8a	300	96.5(0.2)	94.5(0.2)	**96.9 (0.2)**	**96.8 (0.1)**	96.7(0.1)	**96.7 (0.2)**	**97.1 (0.1)**
adult	104	86.1(0.3)	87.6(0.2)	87.4(0.3)	**88.6 (0.4)**	**88.3 (0.3)**	87.6(0.3)	**88.8 (0.3)**
twonorm	20	99.7(0.0)	**99.8 (0.0)**	99.7(0.0)	**99.8 (0.0)**	99.7(0.0)	99.6(0.0)	99.7(0.0)
mushroom	98	**99.9 (0.0)**	99.6(0.1)	**100.0 (0.0)**	**100.0 (0.0)**	**100.0 (0.0)**	**100.0 (0.0)**	**99.9 (0.1)**

Table 9. Mean AUC score and standard error for AUC maximization from corrupted labels, where $\pi = 0.8$ and $\pi' = 0.3$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	**93.8 (0.3)**	**94.3 (0.2)**	**94.1 (0.3)**	93.6(0.3)	92.3(0.3)	90.5(0.5)	92.7(0.5)
phoneme	5	81.0(0.5)	81.7(0.4)	82.1(0.5)	**85.3 (0.4)**	**85.1 (0.2)**	**85.6 (0.3)**	**85.7 (0.4)**
phishing	30	**96.8 (0.1)**	93.7(0.3)	**96.8 (0.2)**	96.2(0.2)	95.1(0.2)	92.9(0.3)	95.4(0.2)
waveform	21	**94.7 (0.2)**	91.3(0.3)	**95.1 (0.3)**	94.1(0.3)	93.8(0.2)	91.5(0.4)	94.1(0.3)
susy	18	80.0(0.5)	77.9(0.5)	**81.3 (0.4)**	**81.1 (0.4)**	**81.7 (0.5)**	79.0(0.6)	**80.8 (0.5)**
w8a	300	91.3(0.5)	**92.9 (0.2)**	90.8(0.3)	87.4(0.4)	83.2(0.6)	82.9(0.6)	88.7(0.4)
adult	104	85.3(0.3)	**86.1 (0.4)**	85.1(0.4)	82.2(0.5)	78.3(0.6)	77.4(0.5)	81.9(0.5)
twonorm	20	99.7(0.0)	**99.8 (0.0)**	99.4(0.0)	98.9(0.1)	98.3(0.1)	95.1(0.2)	97.5(0.1)
mushroom	98	**99.8 (0.1)**	99.3(0.1)	**99.7 (0.1)**	99.2(0.2)	98.6(0.2)	97.9(0.2)	99.6(0.1)

Table 10. Mean AUC score and standard error for AUC maximization from corrupted labels, where $\pi = 0.7$ and $\pi' = 0.4$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	90.4(0.4)	**93.4 (0.3)**	91.8(0.3)	88.3(0.5)	85.4(0.6)	82.2(0.6)	86.0(0.5)
phoneme	5	**81.0 (0.4)**	**81.1 (0.5)**	**82.2 (0.6)**	**82.2 (0.6)**	**81.8 (0.5)**	**82.2 (0.6)**	**81.9 (0.6)**
phishing	30	**95.9 (0.3)**	93.0(0.5)	94.9(0.4)	91.7(0.4)	88.1(0.5)	83.7(0.5)	90.2(0.5)
waveform	21	**93.5 (0.3)**	91.5(0.5)	**94.1 (0.2)**	90.1(0.4)	88.6(0.6)	82.4(0.8)	86.5(0.5)
susy	18	**77.9 (0.6)**	77.4(0.6)	**78.8 (0.5)**	75.6(1.0)	74.6(0.6)	73.2(0.6)	74.2(0.7)
w8a	300	79.1(0.7)	**89.5 (0.5)**	79.4(0.6)	75.6(0.4)	72.3(0.8)	71.5(0.6)	76.4(0.8)
adult	104	82.1(0.4)	**84.6 (0.4)**	81.7(0.5)	75.5(0.5)	72.6(0.6)	68.2(0.8)	73.4(0.6)
twonorm	20	99.4(0.1)	**99.7 (0.0)**	98.9(0.1)	94.5(0.3)	92.3(0.5)	85.4(0.6)	91.6(0.3)
mushroom	98	**99.6 (0.1)**	98.9(0.1)	98.8(0.2)	96.6(0.3)	92.6(0.5)	86.7(0.5)	96.7(0.3)

Table 11. Mean AUC score and standard error for AUC maximization from corrupted labels, where $\pi = 0.65$ and $\pi' = 0.45$.

Dataset	Dim.	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
spambase	57	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)
phoneme	5	**80.2 (0.6)**	**79.2 (0.9)**	78.4(0.8)	78.2(0.8)	77.8(0.8)	76.2(0.8)	76.2(0.7)
phishing	30	**94.7 (0.3)**	90.2(0.8)	91.1(0.6)	85.0(0.6)	82.0(0.8)	73.8(0.9)	80.3(0.8)
waveform	21	**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)
susy	18	**73.6 (0.8)**	**75.3 (0.8)**	72.5(1.0)	70.9(1.0)	69.9(1.0)	66.2(0.8)	69.9(0.9)
w8a	300	70.9(0.8)	**81.7 (0.8)**	71.3(0.9)	68.4(0.7)	66.8(0.8)	65.5(0.6)	68.3(0.6)
adult	104	79.0(0.7)	**81.2 (0.7)**	75.3(1.1)	69.6(0.8)	66.8(1.0)	62.3(0.8)	68.0(1.0)
twonorm	20	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	98	**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)

C.3. BER Minimization Using MNIST Dataset

Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%. The experiments were conducted 10 times.

Table 12. Mean balanced accuracy and standard error for BER minimization from corrupted labels with varying noises.

Dataset	(π, π')	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
MNIST	$(1.0, 0.0)$	97.8(0.0)	50.2(0.1)	99.0(0.0)	**99.1 (0.0)**	99.0(0.0)	98.4(0.0)	99.0(0.0)
	$(0.8, 0.3)$	**97.3 (0.0)**	50.4(0.2)	96.7(0.1)	80.5(0.2)	80.4(0.2)	89.9(0.4)	81.7(0.8)
	$(0.7, 0.4)$	**95.8 (0.2)**	50.0(0.0)	92.7(0.3)	69.6(0.3)	69.5(0.2)	81.8(1.2)	70.2(0.9)
	$(0.65, 0.45)$	**92.8 (0.3)**	50.0(0.0)	83.1(3.7)	64.0(0.2)	63.7(0.3)	73.0(1.3)	63.9(0.1)

C.4. AUC Maximization Using MNIST Dataset

Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%. The experiments were conducted 10 times.

Table 13. Mean AUC score and standard error for AUC maximization from corrupted labels with varying noises.

Dataset	(π, π')	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
MNIST	$(1.0, 0.0)$	$99.6(0.0)$	$85.0(0.5)$	$99.8(0.0)$	**99.8 (0.0)**	$99.8(0.0)$	$99.5(0.0)$	$99.7(0.0)$
	$(0.8, 0.3)$	**99.4 (0.0)**	$84.3(0.4)$	$98.0(0.1)$	$88.5(0.3)$	$88.2(0.2)$	$96.6(0.2)$	$97.2(0.4)$
	$(0.7, 0.4)$	**99.0 (0.0)**	$83.1(0.4)$	$95.9(0.2)$	$75.5(0.4)$	$76.1(0.3)$	$87.5(0.6)$	$94.7(0.4)$
	$(0.65, 0.45)$	**96.9 (0.2)**	$80.6(0.3)$	$92.2(0.7)$	$68.6(0.4)$	$68.5(0.4)$	$80.2(0.5)$	$90.6(1.0)$

C.5. BER Minimization Using CIFAR-10 Dataset

Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%. The experiments were conducted 10 times.

Table 14. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 1.0$ $\pi' = 0.0$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	$87.0(0.4)$	$69.5(0.2)$	$93.0(0.2)$	$93.6(0.1)$	$93.4(0.1)$	**94.3 (0.1)**	$93.4(0.1)$
bird	$84.0(0.2)$	$64.9(0.1)$	$88.2(0.1)$	**88.6 (0.2)**	**88.7 (0.2)**	**88.9 (0.1)**	$88.6(0.1)$
car	$88.5(0.1)$	$69.8(0.1)$	$91.8(0.1)$	$92.5(0.2)$	$92.6(0.1)$	**93.1 (0.1)**	$92.8(0.1)$
deer	$89.6(0.1)$	$71.6(0.2)$	$93.3(0.1)$	$93.7(0.2)$	$93.8(0.0)$	**94.1 (0.1)**	**94.0 (0.1)**
dog	$91.6(0.1)$	$67.6(0.2)$	$93.8(0.1)$	$94.1(0.2)$	$94.2(0.1)$	**94.9 (0.1)**	$94.4(0.1)$
frog	$93.3(0.1)$	$73.8(0.1)$	$95.6(0.1)$	**96.2 (0.1)**	**96.0 (0.1)**	**96.1 (0.1)**	$96.0(0.1)$
horse	$92.8(0.1)$	$69.0(0.2)$	$94.5(0.1)$	$94.9(0.1)$	$94.6(0.1)$	**95.3 (0.1)**	$94.9(0.1)$
ship	$80.9(0.5)$	$64.4(0.1)$	$87.5(0.3)$	$89.1(0.2)$	$89.1(0.2)$	**89.6 (0.1)**	$89.3(0.1)$
truck	$87.5(0.2)$	$69.4(0.1)$	$90.6(0.2)$	$91.2(0.2)$	$91.1(0.2)$	**91.6 (0.1)**	$91.1(0.2)$

Table 15. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 0.8$ $\pi' = 0.3$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	$86.6(0.2)$	$70.2(0.2)$	**88.5 (0.2)**	$74.3(0.2)$	$74.1(0.5)$	$74.2(0.3)$	$73.6(0.3)$
bird	$82.3(0.3)$	$66.7(0.2)$	**83.3 (0.3)**	$72.4(0.4)$	$72.3(0.4)$	$71.6(0.3)$	$71.3(0.4)$
car	$87.3(0.1)$	$71.2(0.1)$	**87.8 (0.1)**	$73.3(0.2)$	$74.4(0.4)$	$73.9(0.3)$	$73.5(0.4)$
deer	**88.5 (0.2)**	$72.9(0.2)$	**88.9 (0.1)**	$74.3(0.4)$	$75.3(0.5)$	$74.6(0.4)$	$74.2(0.3)$
dog	$90.0(0.1)$	$68.4(0.1)$	**90.6 (0.2)**	$75.4(0.4)$	$76.6(0.4)$	$75.9(0.2)$	$74.4(0.5)$
frog	$92.8(0.1)$	$76.2(0.2)$	**93.1 (0.1)**	$76.0(0.3)$	$78.2(0.6)$	$77.9(0.3)$	$76.5(0.5)$
horse	**90.8 (0.3)**	$71.0(0.1)$	$89.8(0.2)$	$76.0(0.3)$	$76.8(0.4)$	$76.3(0.3)$	$75.5(0.3)$
ship	$77.1(0.3)$	$65.7(0.1)$	**80.0 (0.2)**	$70.1(0.2)$	$69.7(0.2)$	$69.8(0.3)$	$69.8(0.3)$
truck	**86.3 (0.1)**	$70.0(0.1)$	**86.3 (0.3)**	$73.9(0.3)$	$73.8(0.5)$	$74.6(0.3)$	$73.7(0.4)$

Table 16. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 0.7$ $\pi' = 0.4$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	**85.4 (0.2)**	70.0(0.3)	83.8(0.3)	65.8(0.5)	65.6(0.4)	65.6(0.3)	64.1(0.4)
bird	**81.7 (0.2)**	66.9(0.1)	80.7(0.3)	63.1(0.4)	63.6(0.5)	63.6(0.3)	62.8(0.4)
car	**86.7 (0.2)**	71.4(0.1)	84.3(0.2)	64.8(0.5)	64.4(0.4)	64.5(0.2)	64.8(0.4)
deer	**87.5 (0.1)**	74.0(0.1)	84.3(0.2)	64.0(0.5)	63.8(0.6)	64.5(0.2)	64.1(0.5)
dog	**88.9 (0.2)**	68.4(0.1)	87.2(0.2)	65.8(0.7)	64.5(0.5)	65.2(0.3)	64.9(0.4)
frog	**92.3 (0.1)**	77.0(0.2)	90.9(0.2)	65.6(0.7)	66.0(0.5)	66.6(0.4)	67.0(0.4)
horse	**88.8 (0.2)**	71.2(0.2)	86.1(0.3)	65.7(0.6)	65.6(0.4)	66.0(0.4)	65.5(0.3)
ship	**74.9 (0.2)**	65.5(0.0)	**74.7 (0.2)**	62.1(0.4)	61.4(0.5)	62.2(0.4)	62.4(0.2)
truck	**84.7 (0.2)**	70.3(0.1)	82.4(0.3)	63.7(0.3)	64.4(0.6)	64.5(0.4)	64.3(0.5)

Table 17. Mean balanced accuracy and standard error for BER minimization from corrupted labels, where $\pi = 0.65$ $\pi' = 0.45$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	**84.0 (0.3)**	70.8(0.2)	79.7(0.3)	59.8(0.6)	59.1(0.5)	60.1(0.3)	60.4(0.3)
bird	**81.5 (0.2)**	67.6(0.2)	77.5(0.4)	58.5(0.4)	59.4(0.3)	58.4(0.4)	58.0(0.5)
car	**85.6 (0.1)**	71.8(0.1)	81.5(0.4)	60.6(0.4)	59.7(0.4)	59.8(0.3)	60.1(0.2)
deer	**86.2 (0.2)**	74.6(0.2)	80.3(0.5)	58.3(0.4)	58.9(0.5)	59.0(0.4)	58.5(0.4)
dog	**87.2 (0.4)**	68.6(0.2)	83.1(0.2)	59.7(0.2)	59.8(0.6)	59.9(0.4)	59.3(0.4)
frog	**91.0 (0.2)**	78.2(0.1)	88.6(0.3)	60.4(0.5)	61.0(0.4)	60.9(0.3)	61.6(0.5)
horse	**86.4 (0.4)**	71.4(0.1)	82.6(0.3)	60.3(0.4)	60.0(0.4)	60.0(0.4)	60.0(0.3)
ship	**71.7 (0.6)**	65.9(0.1)	68.9(0.4)	58.2(0.3)	58.4(0.3)	57.2(0.3)	58.1(0.3)
truck	**82.4 (0.2)**	70.6(0.1)	78.6(0.4)	60.0(0.4)	59.1(0.4)	60.0(0.2)	59.5(0.4)

C.6. AUC Maximization Using CIFAR-10 Dataset

Outperforming methods are highlighted in boldface using one-sided t-test with the significance level 5%. The experiments were conducted 10 times.

Table 18. Mean AUC score and standard error for AUC maximization from corrupted labels, $\pi = 1.0$ $\pi' = 0.0$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	95.8(0.1)	75.2(0.1)	**98.4 (0.0)**	**98.4 (0.0)**	**98.3 (0.0)**	**98.3 (0.1)**	98.2(0.0)
bird	91.5(0.1)	71.7(0.0)	95.0(0.1)	**95.2 (0.0)**	**95.2 (0.0)**	**95.2 (0.1)**	**95.1 (0.1)**
car	94.7(0.1)	76.5(0.0)	97.5(0.1)	97.6(0.0)	**97.6 (0.1)**	**97.7 (0.0)**	**97.7 (0.0)**
deer	95.3(0.1)	79.5(0.1)	**98.3 (0.1)**	**98.3 (0.1)**	**98.4 (0.0)**	**98.3 (0.1)**	**98.3 (0.1)**
dog	96.5(0.1)	74.4(0.2)	**98.5 (0.0)**	**98.5 (0.0)**	**98.5 (0.0)**	**98.5 (0.0)**	**98.5 (0.1)**
frog	97.2(0.1)	81.5(0.0)	99.1(0.0)	99.0(0.0)	**99.1 (0.0)**	99.0(0.0)	99.0(0.0)
horse	97.2(0.1)	76.1(0.0)	**98.9 (0.0)**	**98.9 (0.0)**	**98.9 (0.0)**	98.7(0.0)	**98.8 (0.0)**
ship	92.4(0.1)	70.7(0.1)	95.5(0.1)	**95.7 (0.1)**	95.5(0.1)	**95.5 (0.1)**	**95.6 (0.1)**
truck	94.2(0.1)	75.7(0.1)	**97.2 (0.0)**	**97.1 (0.0)**	**97.1 (0.1)**	96.9(0.1)	97.1(0.0)

Table 19. Mean AUC score and standard error for AUC maximization from corrupted labels, $\pi = 0.8$ $\pi' = 0.3$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	**94.8 (0.1)**	75.7(0.1)	83.5(0.5)	82.0(0.4)	81.4(0.2)	82.1(0.3)	82.1(0.3)
bird	**90.6 (0.1)**	73.3(0.0)	79.8(0.3)	79.2(0.3)	78.7(0.3)	79.1(0.2)	79.4(0.3)
car	**93.4 (0.4)**	78.4(0.0)	82.7(0.5)	82.1(0.3)	81.1(0.3)	82.0(0.4)	80.8(0.4)
deer	**94.6 (0.1)**	81.2(0.0)	83.6(0.6)	81.5(0.5)	81.6(0.3)	82.1(0.4)	82.3(0.3)
dog	**95.6 (0.1)**	76.2(0.1)	82.8(0.4)	83.5(0.4)	82.8(0.4)	83.3(0.3)	83.1(0.3)
frog	**96.9 (0.1)**	83.7(0.0)	85.2(0.4)	85.2(0.4)	84.4(0.2)	84.6(0.4)	84.3(0.4)
horse	**96.2 (0.4)**	78.1(0.1)	84.0(0.5)	84.3(0.3)	83.9(0.5)	83.9(0.3)	83.9(0.4)
ship	**89.0 (0.1)**	71.9(0.1)	78.3(0.4)	76.8(0.3)	77.4(0.3)	77.0(0.4)	77.2(0.3)
truck	**93.7 (0.1)**	76.6(0.1)	81.4(0.7)	81.7(0.3)	81.1(0.2)	81.0(0.4)	81.9(0.3)

Table 20. Mean AUC score and standard error for AUC maximization from corrupted labels, $\pi = 0.7$ $\pi' = 0.4$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	**93.2 (0.1)**	76.0(0.1)	72.3(0.8)	71.1(0.6)	70.3(0.3)	71.0(0.5)	70.7(0.4)
bird	**90.0 (0.2)**	73.8(0.0)	68.7(0.8)	68.2(0.3)	68.3(0.5)	67.0(0.4)	67.9(0.5)
car	**93.4 (0.2)**	78.5(0.0)	70.5(0.4)	70.2(0.5)	69.2(0.5)	70.0(0.5)	69.8(0.2)
deer	**93.3 (0.2)**	81.6(0.1)	69.3(0.6)	69.5(0.7)	69.5(0.4)	69.3(0.3)	69.9(0.5)
dog	**94.9 (0.1)**	76.4(0.1)	70.9(0.8)	71.8(0.4)	70.9(0.4)	70.9(0.4)	71.5(0.3)
frog	**96.7 (0.1)**	84.8(0.0)	73.1(0.7)	73.4(0.4)	72.9(0.6)	72.3(0.4)	72.7(0.4)
horse	**95.8 (0.1)**	78.4(0.1)	72.3(0.7)	72.6(0.4)	70.8(0.3)	71.2(0.5)	71.7(0.4)
ship	**84.5 (0.4)**	71.6(0.1)	69.8(0.4)	67.2(0.4)	66.4(0.3)	66.9(0.4)	67.4(0.3)
truck	**92.1 (0.1)**	76.8(0.1)	71.3(0.7)	70.1(0.4)	69.2(0.5)	69.8(0.4)	70.3(0.3)

Table 21. Mean AUC score and standard error for AUC maximization from corrupted labels, $\pi = 0.65$ $\pi' = 0.45$

Dataset	Barrier	Unhinged	Sigmoid	Logistic	Hinge	Squared	Savage
automobile	**91.3 (0.3)**	76.5(0.1)	64.1(0.4)	64.7(0.3)	63.9(0.3)	64.0(0.6)	64.0(0.5)
bird	**88.5 (0.1)**	74.4(0.0)	63.3(0.6)	62.0(0.5)	61.4(0.4)	61.6(0.3)	62.0(0.3)
car	**92.9 (0.2)**	78.9(0.1)	65.9(0.9)	63.7(0.6)	63.6(0.4)	63.9(0.3)	64.4(0.4)
deer	**92.3 (0.1)**	82.6(0.1)	64.3(0.8)	62.3(0.6)	62.8(0.5)	63.3(0.3)	62.4(0.5)
dog	**93.2 (0.2)**	77.3(0.1)	64.1(0.7)	63.4(0.6)	63.6(0.4)	63.5(0.4)	64.1(0.3)
frog	**96.4 (0.1)**	85.8(0.0)	67.2(0.6)	66.4(0.4)	65.9(0.4)	65.8(0.5)	65.2(0.6)
horse	**93.6 (0.2)**	78.5(0.1)	65.9(0.9)	65.3(0.4)	65.0(0.3)	65.0(0.5)	64.9(0.3)
ship	**77.8 (0.4)**	72.0(0.1)	62.8(0.3)	61.8(0.5)	60.9(0.4)	61.3(0.2)	60.9(0.3)
truck	**89.8 (0.2)**	77.1(0.0)	63.8(0.3)	63.5(0.5)	63.2(0.6)	63.2(0.3)	63.1(0.5)

C.7. Additional Figures for CIFAR-10

Similarly to the main part of the paper, we provide figures for additional eight pairs of CIFAR-10.

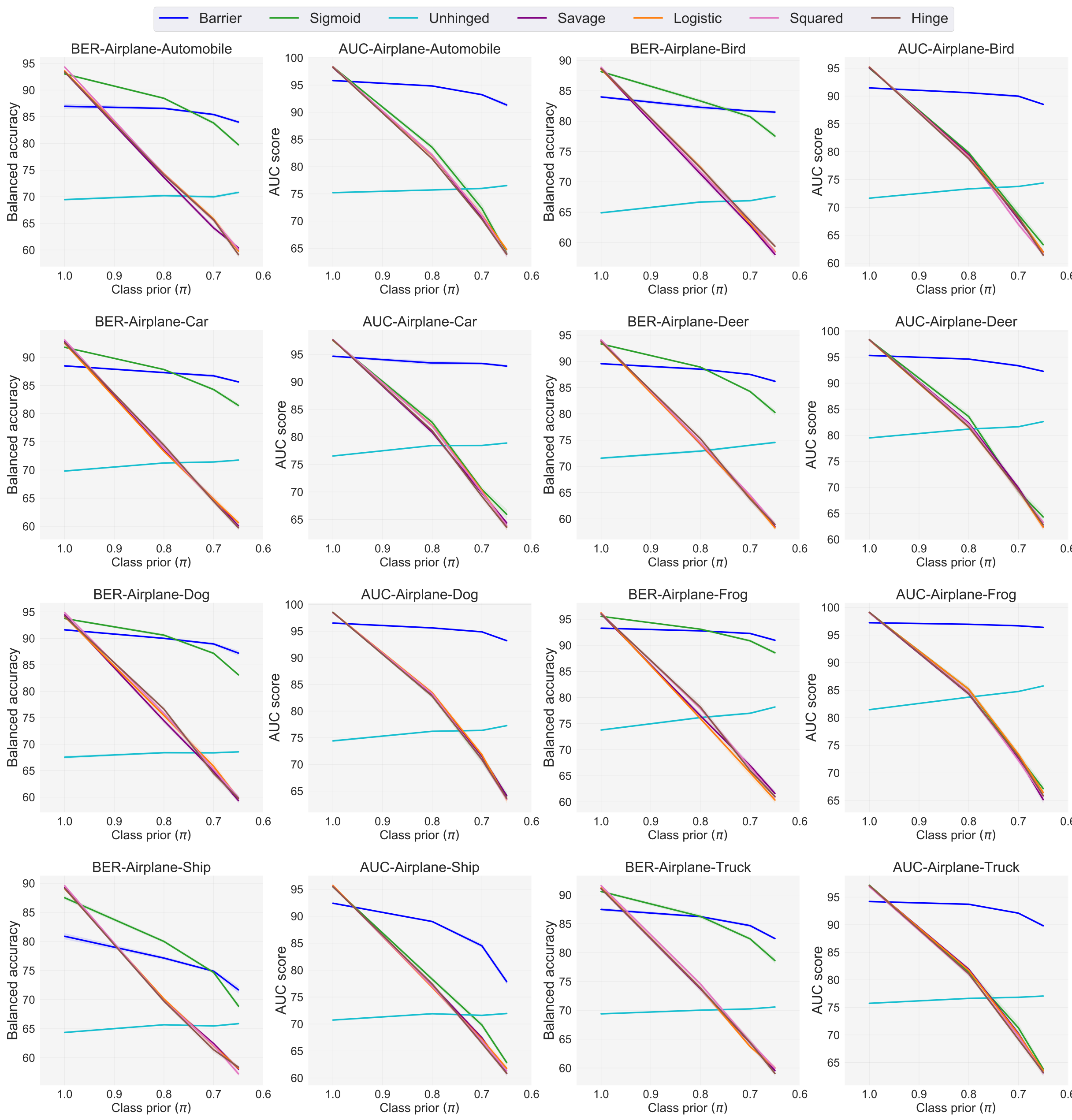

Figure 6. Mean balanced accuracy (1-BER) and AUC score using convolutional neural networks (rescaled to 0-100). The noise rate is ranged from $(\pi = 1.0, \pi' = 0.0)$, $(\pi = 0.8, \pi' = 0.3)$, $(\pi = 0.7, \pi' = 0.4)$, $(\pi = 0.65, \pi' = 0.45)$.

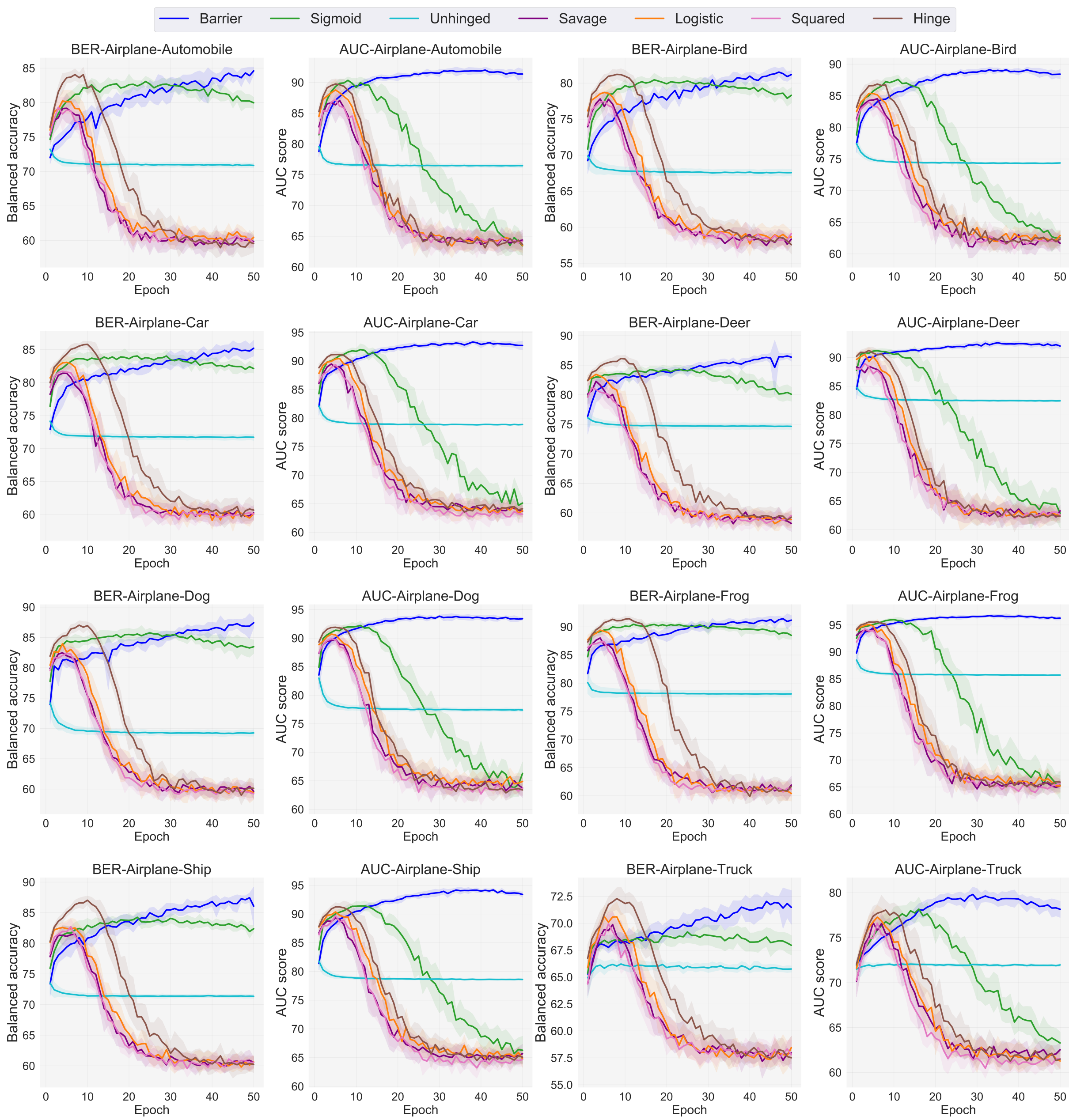

Figure 7. Mean balanced accuracy (1-BER) and AUC score using convolutional neural networks (rescaled to 0-100). The noise rate is $\pi = 0.65$ and $\pi' = 0.45$. The experiments were conducted 10 times.