Bayes-Optimal Balanced Error Rate and Area Under the Receiver Operating Characteristic Curve: Estimation and Evaluation of Estimators

Ryota Ushio[1,2] **Takashi Ishida**[2,1] **Masashi Sugiyama**[2,1]

[1] Graduate School of Frontier Sciences, The University of Tokyo, Tokyo, Japan
[2] RIKEN AIP, Tokyo, Japan
{ushio@ms.,ishida@,sugi@}k.u-tokyo.ac.jp

Abstract

A fundamental quantity in machine learning is the *optimal performance* achievable by any model on a given task. Estimating this quantity allows us to distinguish the irreducible part of the error from a deficiency of the model, telling us how much room for improvement remains. Recent work has shown that the Bayes error, or equivalently the optimal accuracy, can be estimated from soft labels in binary classification. However, accuracy is often a poor summary of performance in settings with severe class imbalance or noisy annotations, where metrics such as the balanced error rate (BER) and the area under the receiver operating characteristic curve (AUC) are more appropriate. We address this gap with two complementary contributions. **(i) Estimation.** We propose soft-label-based estimators for the optimal BER and AUC. We first consider the clean setting in which true soft labels and the class prior are known, and then extend the estimators to a more realistic setting in which the class prior is unknown and the observed soft labels are *corrupted* by an unknown order-preserving transformation, possibly followed by additive noise. In the latter setting, we approximately recover the clean soft labels via isotonic regression with auxiliary hard labels and derive finite-sample error bounds for the resulting plug-in estimators. **(ii) Evaluation.** Since the optimum is unobservable on real datasets, evaluating any such estimator is itself nontrivial. We extend the FeeBee framework, originally proposed for evaluating Bayes-error estimators, to the optimal BER and AUC. The resulting procedure provides practical evaluation scores without requiring knowledge of the optimum, and applies to any estimator of the optimal BER or AUC, not only our proposed ones. Experiments on synthetic and real-world datasets validate both the estimators and the evaluation procedure.

1 Introduction

An integral part of machine learning workflows is evaluating models using performance metrics. Of natural interest, then, is the *Bayes-optimal performance*, i.e., the best value achievable by any model on a given task. This quantity represents the irreducible component of the error and helps us separate a deficiency of the current model from the inherent difficulty of the task. In practice, when the gap between current and optimal performance is small, it tells us that not much room is left for improvement and that further training of the current model or training of larger models is unlikely to help. This is particularly important in the era of large-scale machine learning models, which often require a huge amount of computational resources to train and have a significant environmental impact (Strubell et al., 2020; Luccioni et al., 2023). Moreover, recent work suggests that estimation of the Bayes-optimal performance can be useful for detecting data contamination (Ishida et al., 2026).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We focus on binary classification in this paper. Given a binary classifier $h : \mathcal{X} \to \{0, 1\}$, the most commonly used metric to measure its performance is the *error rate* $\mathrm{Err}(h) \coloneqq \mathbb{P}(h(x) \neq y)$ or the *accuracy* $1 - \mathrm{Err}(h)$, where (x, y) is drawn from an unknown joint distribution $\mathbb{P}$ over the space $\mathcal{X}$ of input features and the set $\{0, 1\}$ of class labels. When the class posterior, described by $\eta(x) \coloneqq \mathbb{P}(y = 1 \mid x)$, is genuinely uncertain, even a perfect predictor cannot avoid misclassification due to the stochasticity of the labels. The lowest achievable error rate $\mathrm{Err}^* \coloneqq \inf_{h:\mathcal{X}\to\{0,1\}} \mathrm{Err}(h)$, also known as the *Bayes error*, therefore quantifies the inherent label uncertainty and is determined by the underlying data distribution.

Estimation of the Bayes error has a long history in the literature. While most existing methods use input-label pairs $\{(x_i, y_i)\}_{i=1}^n$ (Fukunaga & Hostetler, 1975; Devijver, 1985; Berisha et al., 2016; Moon et al., 2018; Noshad et al., 2019; Theisen et al., 2021), a recent line of work suggests another approach that estimates the Bayes error from *soft labels* $\eta_i = \eta(x_i)$, $i = 1, \ldots, n$ (Ishida et al., 2023; Jeong et al., 2023; Ushio et al., 2026). A soft label η_i represents class membership as a probability value in $[0, 1]$ unlike a binary *hard label* $y_i \in \{0, 1\}$. This soft-label-based approach does not need access to the input features x_i, which makes it suitable for privacy-sensitive applications and also allows it to bypass the curse of dimensionality. Also, it does not make the distributional assumptions often imposed by the methods based on input-label pairs, such as lower/upper bounds on the probability density function or Lipschitzness of the density. Early work typically assumes that the true soft labels η_i or their high-quality estimates are available, e.g., by averaging many independent annotations per instance (Ishida et al., 2023; Jeong et al., 2024), which might be unrealistic in practice. More recent work (Ushio et al., 2026) relaxed this assumption by allowing the soft labels to be corrupted by unknown order-preserving transformations.

However, the error rate is a poor summary of performance under class imbalance or label noise, which are both common in practice, and other metrics such as the *balanced error rate (BER)* and the *area under the receiver operating characteristic curve (AUC)* are often preferred in such settings (Ling & Li, 1998; Gu et al., 2009; Menon et al., 2013, 2015; van Rooyen et al., 2015; Menon et al., 2018; Charoenphakdee et al., 2019). Concretely, the BER of a classifier $h : \mathcal{X} \to \{0, 1\}$ is the average of the in-class error rates, i.e., $\mathrm{BER}(h) \coloneqq (\mathrm{FPR}(h) + \mathrm{FNR}(h))/2$, where $\mathrm{FPR}(h) \coloneqq \mathbb{P}(h(x) = 1 \mid y = 0)$ and $\mathrm{FNR}(h) \coloneqq \mathbb{P}(h(x) = 0 \mid y = 1)$ are the false positive and false negative rates (Menon et al., 2015). The AUC of a scoring function $g : \mathcal{X} \to \mathbb{R}$ (which can be used to produce a classifier by thresholding the score) is the frequency that a random positive instance $x_+ \sim \mathbb{P}(x \mid y = 1)$ is ranked above a random negative instance $x_- \sim \mathbb{P}(x \mid y = 0)$, with ties counted as $1/2$ (Clémençon et al., 2008; Menon et al., 2015): $\mathrm{AUC}(g) \coloneqq \mathbb{P}(g(x_+) > g(x_-)) + \frac{1}{2}\mathbb{P}(g(x_+) = g(x_-))$. Under severe class imbalance, a trivial classifier that always predicts the majority class can achieve a near-zero error rate while learning nothing from data, and hence the error rate is not helpful. BER remains informative even under class imbalance by averaging the per-class error rates (Menon et al., 2013). AUC is also a preferred metric under class imbalance, as it summarizes the trade-off between true- and false-positive rates over all decision thresholds (Ling & Li, 1998; Fawcett, 2006; Gu et al., 2009). Moreover, BER and AUC are popular metrics under label noise due to their robustness; e.g., their optimizers remain unchanged under noisy labels (Menon et al., 2015; van Rooyen et al., 2015).

Our contribution Despite the practical importance of BER and AUC, the estimation of their Bayes-optimal values, i.e., $\mathrm{BER}^* \coloneqq \inf_{h:\mathcal{X}\to\{0,1\}} \mathrm{BER}(h)$ and $\mathrm{AUC}^* \coloneqq \sup_{g:\mathcal{X}\to\mathbb{R}} \mathrm{AUC}(g)$, remains underexplored. The first part of our contribution fills this gap by proposing soft-label-based estimators of BER^* and AUC^*. We start with the clean setting similar to the work by Ishida et al. (2023) for the Bayes error, in which the true soft labels η_i are available (Section 2). Then, inspired by Ushio et al. (2026), we extend the estimators to a more realistic setting where the soft labels are only observed through some unknown order-preserving corruption, possibly followed by additive noise (Section 3). Unlike the Bayes error, the optimal BER and AUC depend on the class prior $\theta \coloneqq \mathbb{P}(y = 1)$ $(0 < \theta < 1)$. We first assume that θ is known for simplicity and then handle the unknown-prior case.

A separate question is how to evaluate an estimator of BER^* or AUC^* on *real datasets*. This is challenging because, for real-world datasets, the underlying data distribution is unknown and does not reveal the ground-truth optimal values. For the Bayes error, Renggli et al. (2021) proposed *FeeBee*, which sidesteps this issue with a simple idea of injecting label noise. In the second part of our contribution, we extend this idea to the optimal BER and AUC (Section 4). Specifically, we

first derive a simple closed-form relationship between the original optima and their noise-injected counterparts. Then, we propose scores for evaluating estimators of BER^* and AUC^* based on the derived relationship. The procedure can be applied to any estimator of BER^* or AUC^*, not only the ones we propose. Our proposed method allows us to evaluate Bayes-optimal BER/AUC estimators on real-world datasets without knowledge of the true optima.

2 Estimation from clean soft labels

In this section, we first show that the optimal BER and AUC can be expressed as expectations of certain functions of the class posterior $\eta(x)$ (Lemmas 2.1 and 2.2). This reveals that, for each of BER and AUC, there are two different natural choices of unbiased estimators of the optimum. We discuss data-dependent criteria (or *discriminants*) for choosing between them in Section 2.1. Next, in Section 2.2, we present an efficient $O(n \log n)$ time algorithm for computing the proposed optimal AUC estimator and AUC discriminant, which would otherwise take $\Theta(n^2)$ time. Finally in Section 2.5, we discuss the case where the class prior θ is unknown.

Lemma 2.1. *The optimal BER is given by* $\mathrm{BER}^* = \mathbb{E}\left[\phi^{(1)}_{\mathrm{BER}}(\eta(x))\right] = \mathbb{E}\left[\phi^{(2)}_{\mathrm{BER}}(\eta(x))\right]$*, where* $\phi^{(1)}_{\mathrm{BER}}(z) \coloneqq \frac{1}{2}\min\left\{\frac{z}{\theta}, \frac{1-z}{1-\theta}\right\}$ *and* $\phi^{(2)}_{\mathrm{BER}}(z) \coloneqq 1 - \frac{1}{2}\max\left\{\frac{z}{\theta}, \frac{1-z}{1-\theta}\right\}$.

Lemma 2.2. *Let* x' *be an i.i.d. copy of* x*. Then, the optimal AUC is given by* $\mathrm{AUC}^* = \mathbb{E}\left[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right] = \mathbb{E}\left[\phi^{(2)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right]$*, where* $\phi^{(1)}_{\mathrm{AUC}}(z_1, z_2) \coloneqq 1 - \frac{1}{2\theta(1-\theta)}\min\{z_1(1-z_2), z_2(1-z_1)\}$ *and* $\phi^{(2)}_{\mathrm{AUC}}(z_1, z_2) \coloneqq \frac{1}{2\theta(1-\theta)}\max\{z_1(1-z_2), z_2(1-z_1)\}$.

Remark 2.3. Clémençon et al. (2008, Example 1) mentioned the former expression of $\mathrm{AUC}^* = \mathbb{E}[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))]$ but without a proof. We state the full proof for the sake of completeness (Appendix A.1).

2.1 Unbiased estimators: Min formula vs. max formula

In Lemma 2.1, we saw that the best possible BER can be expressed in two different ways: $\mathrm{BER}^* = \mathbb{E}\left[\phi^{(1)}_{\mathrm{BER}}(\eta(x))\right] = \mathbb{E}\left[\phi^{(2)}_{\mathrm{BER}}(\eta(x))\right]$. This naturally leads to two different unbiased estimators of BER^*: $\widehat{\mathrm{BER}}^*_1 \coloneqq \frac{1}{n}\sum_{i=1}^n \phi^{(1)}_{\mathrm{BER}}(\eta_i)$ and $\widehat{\mathrm{BER}}^*_2 \coloneqq \frac{1}{n}\sum_{i=1}^n \phi^{(2)}_{\mathrm{BER}}(\eta_i)$. We call $\widehat{\mathrm{BER}}^*_1$ the *min formula* for BER because $\phi^{(1)}_{\mathrm{BER}}$ uses the $\min$ operation, and $\widehat{\mathrm{BER}}^*_2$ the *max formula* because $\phi^{(2)}_{\mathrm{BER}}$ uses the $\max$ operation. Similarly, Lemma 2.2 suggests two different unbiased estimators of the optimal AUC, namely the min formula $\widehat{\mathrm{AUC}}^*_1 \coloneqq \frac{2}{n(n-1)}\sum_{i<j} \phi^{(1)}_{\mathrm{AUC}}(\eta_i, \eta_j)$ and the max formula $\widehat{\mathrm{AUC}}^*_2 \coloneqq \frac{2}{n(n-1)}\sum_{i<j} \phi^{(2)}_{\mathrm{AUC}}(\eta_i, \eta_j)$, where $\sum_{i<j}$ denotes the summation over all pairs of indices $\big\{(i,j) \in \{1,\dots,n\}^2 \mid i < j\big\}$.[1]

Thus, a natural question is: *Which estimator, the min formula or the max formula, is better?* Two unbiased estimators are usually compared in terms of their variances; the one with a smaller variance is considered better (Lehmann & Casella, 1998). Therefore, we want to be able to guess which estimator has a smaller variance for a given data distribution.

To this end, for each of BER and AUC, we define a *discriminant* that indicates which estimator, the min formula or the max formula, has a smaller variance. The discriminant for BER is defined as $\Delta_{\mathrm{BER}} \coloneqq (1-2\theta)\,\mathbb{E}\left[(\eta(x)-\theta)\,|\eta(x)-\theta|\right]$. For AUC, we define $\Delta_{\mathrm{AUC}} \coloneqq (1-2\theta)\,\mathbb{E}\left[(\eta(x)-\theta)\,|\eta(x)-\eta(x')|\right]$, where x' is an i.i.d. copy of x. The following theorem suggests that the sign of each discriminant can be used to choose the better estimator.

Theorem 2.4.

(i) For all $n \in \mathbb{N}$*, we have* $\mathrm{Var}\left[\widehat{\mathrm{BER}}^*_1\right] \leq \mathrm{Var}\left[\widehat{\mathrm{BER}}^*_2\right]$ *if and only if* $\Delta_{\mathrm{BER}} \geq 0$.

[1] Naively, $\widehat{\mathrm{AUC}}^*_1$ and $\widehat{\mathrm{AUC}}^*_2$ seem to require $\Theta(n^2)$ time to compute due to this double summation. However, we will show that they can be computed in $O(n \log n)$ time later in Section 2.2.

(ii) Assume that the random variable $\eta(x)$ is not a.s. constant. Then, we have $\mathrm{Var}\left[\widehat{\mathrm{AUC}_1^*}\right] < \mathrm{Var}\left[\widehat{\mathrm{AUC}_2^*}\right]$ *for all* $n \geq 2$ *if and only if* $\Delta_{\mathrm{AUC}} \geq 0$.

In practice, the discriminants can be unbiasedly estimated from soft labels as $\widehat{\Delta}_{\mathrm{BER}} := \frac{1-2\theta}{n}\sum_{i=1}^{n}(\eta_i - \theta)\,|\eta_i - \theta|$ and $\widehat{\Delta}_{\mathrm{AUC}} := \frac{1-2\theta}{n(n-1)}\sum_{i=1}^{n}\sum_{j:\,j\neq i}(\eta_i - \theta)\,|\eta_i - \eta_j|$,[2] and we can choose between the min formula and the max formula based on the signs of these estimates, i.e., $\mathrm{sign}(\widehat{\Delta}_{\mathrm{BER}})$ and $\mathrm{sign}(\widehat{\Delta}_{\mathrm{AUC}})$.[3]

Theoretically, some sufficient conditions can be shown for $\Delta_{\mathrm{BER}} \geq 0$ and $\Delta_{\mathrm{AUC}} \geq 0$, i.e., the min formulas to be better estimators. An obvious sufficient condition is that the two classes are balanced, i.e., $\theta = \frac{1}{2}$. Another condition is that the Bayes error Err^* is small enough, or in other words, the inherent label uncertainty of the data distribution is sufficiently low. See Proposition A.8 for the formal statement and proof.

In the rest of this paper, our theoretical results are stated mainly for the min formula estimators unless explicitly stated otherwise, and we denote them just by $\widehat{\mathrm{BER}^*}$ and $\widehat{\mathrm{AUC}^*}$ for brevity. However, similar results also hold for the max formula estimators.

2.2 Near-linear time algorithm for computing $\widehat{\mathrm{AUC}^*}$

The definitions of $\widehat{\mathrm{AUC}_1^*}$, $\widehat{\mathrm{AUC}_2^*}$ and $\widehat{\Delta}_{\mathrm{AUC}}$ involve summation over all pairs of indices (i, j), so a naive algorithm would take $\Theta(n^2)$ time to compute each of them. This can be a significant bottleneck, or even prohibitive, for large-scale datasets. This computational challenge is common for many *U-statistics* (Hoeffding, 1948) of degree 2 or higher. $\widehat{\mathrm{AUC}_1^*}$, $\widehat{\mathrm{AUC}_2^*}$ and $\widehat{\Delta}_{\mathrm{AUC}}$ are examples of degree-two U-statistics and are no exception. The literature has addressed this challenge typically by using Monte-Carlo approximation of original U-statistics, which are called *incomplete U-statistics* (Blom, 1976; Clémençon et al., 2016; Lee, 2019). This approach reduces the number of terms to be added up and thus the computational cost, but it inevitably introduces additional variance.

Fortunately, however, it turns out that $\widehat{\mathrm{AUC}_1^*}$, $\widehat{\mathrm{AUC}_2^*}$ and $\widehat{\Delta}_{\mathrm{AUC}}$ have special structures that enable efficient computation without relying on approximation. For example, the exact $\widehat{\mathrm{AUC}_1^*}$ can be computed in $O(n \log n)$ time in the worst case by Algorithm 2.1; the proof is given in Theorem A.13. We also provide $O(n \log n)$-time algorithms for exactly computing $\widehat{\mathrm{AUC}_2^*}$ (Algorithm A.2) and $\widehat{\Delta}_{\mathrm{AUC}}$ (Algorithm A.1) in the appendices.

Algorithm 2.1: An efficient algorithm for computing $\widehat{\mathrm{AUC}_1^*}$

input Soft labels $\eta_1, \ldots, \eta_n$, class prior θ
  $\eta_{(1)}, \ldots, \eta_{(n)} \leftarrow$ sort $\eta_1, \ldots, \eta_n$ in ascending order
  $\mathrm{SUM} \leftarrow 0$
  $\mathrm{PREFIX} \leftarrow 0$
  for $j = 2$ **to** n **do**
    $\mathrm{PREFIX} \leftarrow \mathrm{PREFIX} + \eta_{(j-1)}$
    $\mathrm{SUM} \leftarrow \mathrm{SUM} + (1 - \eta_{(j)}) \cdot \mathrm{PREFIX}$
  end for
output $1 - \frac{\mathrm{SUM}}{n(n-1)\theta(1-\theta)}$

2.3 Clipping the AUC estimator

Another minor note about the estimation of AUC^*: although the true value AUC^* lies between 0.5 and 1 for any data distribution, it is possible that the estimates $\widehat{\mathrm{AUC}_1^*}$ and $\widehat{\mathrm{AUC}_2^*}$ can fall outside the valid range $[0.5, 1]$ in rare events.[4] To address this issue, we can simply clip the raw estimates to the valid range $[0.5, 1]$. We denote the clipped estimators by $\widetilde{\mathrm{AUC}_k^*} := \mathrm{clip}_{0.5}^{1}(\widehat{\mathrm{AUC}_k^*}) = \min\{1, \max\{0.5, \widehat{\mathrm{AUC}_k^*}\}\}$, $k = 1, 2$. Here, $\mathrm{clip}_a^b(z) = \min\{b, \max\{a, z\}\}$ represents the operation that clips its argument to the interval $[a, b]$. It is straightforward to verify that each clipped estimator is always more accurate than the unclipped counterpart, i.e., $|\widetilde{\mathrm{AUC}_k^*} - \mathrm{AUC}^*| \leq |\widehat{\mathrm{AUC}_k^*} - \mathrm{AUC}^*|$.

[2] Despite the double summation, $\widehat{\Delta}_{\mathrm{AUC}}$ can be computed in $O(n \log n)$ time as shown in Algorithm A.1 in the appendix.

[3] It is also possible to construct Wald-type statistical hypothesis tests on the signs of Δ_{BER} and Δ_{AUC} based on the asymptotic normality of $\widehat{\Delta}_{\mathrm{BER}}$ and $\widehat{\Delta}_{\mathrm{AUC}}$; see Appendix A.2.1 for details about the AUC case.

[4] For example, imagine that we draw a tiny sample of size $n = 2$, which happens to be $\{\eta_1, \eta_2\} = \{\frac{1}{2}, \frac{1}{2}\}$. Then, we have $\widehat{\mathrm{AUC}_1^*} = 1 - \frac{1}{8\theta(1-\theta)}$, which is less than 0.5 unless $\theta = 0.5$.

Again, we present our theoretical analysis mainly for the min formula estimator $\widetilde{\mathrm{AUC}_1^*}$ and denote it just by $\widetilde{\mathrm{AUC}^*}$ for brevity, but the arguments for $\widetilde{\mathrm{AUC}_2^*}$ are similar. Note that the BER estimators do not require clipping as both the true value BER^* and the estimates $\widehat{\mathrm{BER}_1^*}$, $\widehat{\mathrm{BER}_2^*}$ lie in $[0, 0.5]$.

2.4 Estimation error bounds

Our estimators of the optimal BER and AUC enjoy the following finite-sample bounds on their estimation errors.

Theorem 2.5. *For any $\delta > 0$, with probability at least $1-\delta$, we have $|\widehat{\mathrm{BER}^*} - \mathrm{BER}^*| \leq \sqrt{\frac{\log(2/\delta)}{8n}}$.*

Theorem 2.6. *For any $\delta > 0$, with probability at least $1-\delta$, we have $|\widetilde{\mathrm{AUC}^*} - \mathrm{AUC}^*| \leq \sqrt{\frac{\log(2/\delta)}{\theta(1-\theta)n}}$.*

Remark 2.7. Theorem 2.5 simply follows from the Hoeffding bound. Theorem 2.6 needs a more careful analysis: while we can use basic tools like Hoeffding's inequality (11) for U-statistics or the bounded difference inequality, such tools lead to a bound of stochastic order $O_p(\frac{1}{\theta(1-\theta)\sqrt{n}})$. The constant factor $\frac{1}{\theta(1-\theta)}$ blows up under severe class imbalance, i.e., when θ is close to 0 or 1. We instead employ the Bernstein-type inequality (13) for U-statistics to improve the θ-dependence to $\frac{1}{\sqrt{\theta(1-\theta)}}$, resulting in the bound in Theorem 2.6. See Appendix A.4 for details.

2.5 Unknown class prior

So far, we have assumed that the class prior $\theta = \mathbb{P}(y=1)$ is known. When the class prior is unknown, we can still estimate it from the soft labels simply by the sample mean $\bar{\eta} \coloneqq \frac{1}{n}\sum_{i=1}^n \eta_i$[5] since we have $\mathbb{E}[\eta_i] = \theta$. However, using $\bar{\eta}$ directly in place of θ in the definitions of $\widehat{\mathrm{BER}^*}$ and $\widetilde{\mathrm{AUC}^*}$ may cause numerical instability or even zero division since both estimators involve division by θ and $1-\theta$. A simple solution to this problem is to use a clipped version of $\bar{\eta}$: $\widehat{\theta}_\varepsilon \coloneqq \mathrm{clip}_\varepsilon^{1-\varepsilon}(\bar{\eta}) = \min\{1-\varepsilon, \max\{\varepsilon, \bar{\eta}\}\}$, where $\varepsilon \in (0, \frac{1}{2})$ is some clipping threshold.

With a *fixed* threshold ε, however, we must choose ε carefully so that $0 < \varepsilon < \min\{\theta, 1-\theta\}$; otherwise, our clipped estimator $\widehat{\theta}_\varepsilon$ and the resulting plug-in estimators for BER^* and AUC^* will be statistically inconsistent and suffer from bias that never vanishes as $n \to \infty$. Such a choice of ε requires certain prior knowledge of θ. We can bypass this issue by using a *vanishing* threshold τ_n that depends on the sample size n and tends to zero as n increases. For our theoretical analysis, we set the threshold as follows.

Assumption 2.8. The clipping threshold is set to $\tau_n = \frac{c}{n}$ for some constant $0 < c < \frac{1}{2}$.

The error of the resulting class prior estimator $\widehat{\theta}_{\tau_n}$ is bounded as follows.

Lemma 2.9. *Under Assumption 2.8, for any $\delta > 0$ and $n \geq \frac{(2\log(2/\delta)+3c)^2}{18\theta(1-\theta)\log(2/\delta)}$, with probability at least $1-\delta$, we have $|\widehat{\theta}_{\tau_n} - \theta| \leq \sqrt{\frac{8\theta(1-\theta)\log(2/\delta)}{n}}$.*

We can now estimate BER^* and AUC^* by plugging in $\widehat{\theta}_{\tau_n}$ in place of θ in the definitions of $\widehat{\mathrm{BER}^*}$ and $\widetilde{\mathrm{AUC}^*}$. Let us abuse the notation a bit and denote these plug-in estimators by $\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n})$ and $\widetilde{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n})$, respectively. It can be shown that they are statistically consistent. More specifically, the following finite-sample error bounds hold; see Theorems A.22 and A.23 for the proofs.

Theorem 2.10. *Under Assumption 2.8, for any $\delta > 0$, each of the following holds with probability at least $1-\delta$: $\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \leq \frac{17}{8}\sqrt{\frac{2\log(4/\delta)}{\theta(1-\theta)n}}$ and $\left|\widetilde{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \leq 10\sqrt{\frac{\log(4/\delta)}{\theta(1-\theta)n}}$.*

Also, these estimators are asymptotically unbiased; see Propositions A.24 and A.25.

[5] Jeong et al. (2024) proposed to estimate the class prior by $\frac{1}{n}\sum_{i=1}^n \mathbb{1}[\eta_i \geq 0.5]$ in their discussion of the soft-label-based estimation of the FPR and FNR of the Bayes optimal classifier (i.e., the classifier $h^*(x) = \mathbb{1}[\eta(x) \geq 1/2]$ that achieves the Bayes error rate Err^*), where $\mathbb{1}[\cdot]$ denotes the indicator function. However, it is easy to construct examples where their estimator becomes biased and statistically inconsistent (see Example A.19).

3 Estimation from corrupted soft labels

In practice, clean soft labels $\eta_i = \mathbb{P}(y = 1 \mid x = x_i)$ may be unavailable. If the data generation process allows repeated sampling of class labels for the same input x_i, it is still possible to approximate $\eta_i = \mathbb{E}[y \mid x = x_i]$ by the average of the sampled labels, in which case the resulting plug-in Bayes error estimator has been shown to be asymptotically unbiased (Ishida et al., 2023; Ushio et al., 2026).

However, it is often difficult to obtain more than one class label per input due to various reasons. For example, the construction of image classification datasets, e.g., CIFAR-10 (Krizhevsky, 2009), typically proceeds as follows: first, images are retrieved from the web by searching with keywords associated with each class name; then, human annotators filter out images that are irrelevant to the class; finally, the images are downsampled to a fixed resolution. As such, each class label is associated with the original high-resolution image before downsampling. Although it is possible to collect multiple class labels for each low-resolution image after downsampling as in the CIFAR-10H dataset (Peterson et al., 2019), the downsampling process makes the class posterior distribution much more uncertain than that of the original high-resolution image, distorting the soft label estimates.

A natural way to model such situations is to assume that we have access to a single class label $y_i \in \{0, 1\}$ plus a *corrupted soft label* $\xi_i = f(\eta_i) \in [0, 1]$, instead of the clean soft label η_i, for each input x_i (Ushio et al., 2026). Here, the corrupted soft label ξ_i is the clean soft label η_i skewed by some unknown increasing transformation $f : [0, 1] \to [0, 1]$. In the CIFAR-10/10H example above, ξ_i is the class posterior of the low-resolution image, which can be obtained by averaging the labels from CIFAR-10H, and y_i is the class label from CIFAR-10.

Here, we consider the problem of estimating the optimal BER and AUC from a dataset of n pairs $(\xi_1, y_1), \ldots, (\xi_n, y_n)$. We first describe an algorithm based on isotonic regression (Ayer et al., 1955) to recover the clean soft labels from the corrupted ones. Next, we discuss the estimation of the class prior. Then, we present finite-sample error bounds for the resulting plug-in estimators. Finally, we extend these bounds to a noisy setting where the observed soft labels may not preserve the ordering of the clean soft labels.

Recovery of soft labels To approximately recover the clean soft labels $\eta_1, \ldots, \eta_n$, we adopt an approach based on isotonic regression, which was originally proposed by Ushio et al. (2026) in the context of Bayes error estimation. Isotonic regression is one of the most widely used methods for calibrating classifier outputs to produce reliable probability estimates (Zadrozny & Elkan, 2002), but here we use it for calibrating the corrupted soft labels. Then the calibrated soft labels can be used as estimates of the clean soft labels.

A technical description of the algorithm is as follows. We first sort the soft labels $\xi_1, \ldots, \xi_n$ in ascending order, and let (j) denote the index of the j-th smallest of them, i.e., $\xi_{(1)} \leq \cdots \leq \xi_{(n)}$. Then we use the pool adjacent violator algorithm (PAVA; Ayer et al., 1955) to find the monotonic sequence $\widehat{\eta}_{(1)} \leq \cdots \leq \widehat{\eta}_{(n)}$ that best fits the corresponding class labels $y_{(1)}, \ldots, y_{(n)}$ in the least squares sense: $\min_{\widehat{\eta}_{(1)} \leq \cdots \leq \widehat{\eta}_{(n)}} \frac{1}{n} \sum_{i=1}^{n} \left(y_{(i)} - \widehat{\eta}_{(i)}\right)^2$. Each resulting $\widehat{\eta}_i$ is expected to approximate the true clean soft label η_i. See also Figure 1. Assuming standard sorting algorithms with worst-case time complexity of $O(n \log n)$, this algorithm runs in $O(n \log n)$ time in total, since the PAVA part takes only $O(n)$ time (Best & Chakravarti, 1990; Busing, 2022).

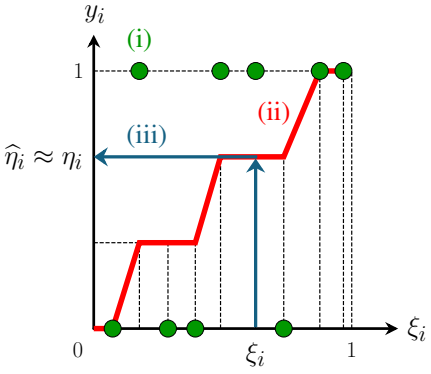


Figure 1: Illustration of the use of isotonic regression to recover clean soft labels from pairs $(\xi_1, y_1), \ldots, (\xi_n, y_n)$ of corrupted soft labels and class labels. (i) Plot the points (ξ_i, y_i). (ii) Find the non-decreasing function that best fits the points. (iii) Use the fitted function values as estimates of the clean soft labels. Roughly speaking, the fitted function is expected to approximate the inverse f^{-1} of the unknown skew function f.

Estimation of class prior As the BER and AUC estimators introduced in Section 2 involve the unknown class prior θ, it needs to be estimated from data. Note that Ushio et al. (2026) did not discuss this problem since their Bayes error estimator did not depend on the class prior.

Since $\theta = \mathbb{E}[y] = \mathbb{E}[\eta(x)]$, two natural candidate estimators of θ are $\frac{1}{n}\sum_i y_i$ and $\frac{1}{n}\sum_i \widehat{\eta}_i$. Interestingly, however, it turns out that they are actually no different from each other. Indeed, we have $\widehat{\boldsymbol{\eta}} + t\mathbf{1} \in \mathcal{M}_n$ for any $t \in \mathbb{R}$.[6] So by the optimality of $\widehat{\boldsymbol{\eta}}$, $t \mapsto \frac{1}{2}\|\boldsymbol{y} - (\widehat{\boldsymbol{\eta}} + t\mathbf{1})\|_2^2$ achieves its minimum at $t = 0$. Therefore, we have $\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{y} - (\widehat{\boldsymbol{\eta}} + t\mathbf{1})\|_2^2\Big|_{t=0} = \widehat{\boldsymbol{\eta}}^\top\mathbf{1} - \boldsymbol{y}^\top\mathbf{1} = 0$, which implies $\frac{1}{n}\sum_i y_i = \frac{1}{n}\sum_i \widehat{\eta}_i$. Similarly to Section 2.5, we use a clipped version of this estimator: $\tilde{\theta}_{\tau_n} \coloneqq \mathrm{clip}_{\tau_n}^{1-\tau_n}(\frac{1}{n}\sum_i y_i) = \mathrm{clip}_{\tau_n}^{1-\tau_n}(\frac{1}{n}\sum_i \widehat{\eta}_i)$, where the clipping threshold τ_n is set according to Assumption 2.8. This estimator $\tilde{\theta}_{\tau_n}$ converges to θ at rate $\frac{1}{\sqrt{n}}$ (see Lemma B.1).

Estimation error bounds Now we can estimate BER^* and AUC^* by plugging in $\widehat{\boldsymbol{\eta}}$ and $\tilde{\theta}_{\tau_n}$ in place of $\boldsymbol{\eta}$ and θ, respectively, in the definitions of $\widehat{\mathrm{BER}^*}$ and $\widetilde{\mathrm{AUC}^*}$. Let us denote these plug-in estimators by $\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n})$ and $\widetilde{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n})$. They enjoy the following finite-sample error bounds, where $\lesssim$ represents an inequality up to some constant not depending on n, θ or δ.

Theorem 3.1. *Under Assumption 2.8, for any $\delta \in (0,1)$, each of the following holds with probability at least $1-\delta$:* $\left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \lesssim \frac{1}{\min\{\theta, 1-\theta\}}\left(\frac{1}{n^{1/3}} + \sqrt{\frac{\log(5/\delta)}{n}}\right)$ *and* $\left|\widetilde{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \lesssim \frac{1}{\theta(1-\theta)}\left(\frac{1}{n^{1/3}} + \sqrt{\frac{\log(5/\delta)}{n}}\right)$.

Noisy corruption model Theorem 3.1 can be extended to the case where the corruption involves additive noise, i.e., $\xi_i = f(\eta_i) + \varepsilon_i$. Formally, we consider the following corruption model.

Assumption 3.2 (Noisy corruption model)**.** Corrupted soft labels $\xi_i \in [0,1]$ are generated as $\xi_i = f(\eta_i) + \varepsilon_i$. Here, the skew function $f : [0,1] \to [0,1]$ is differentiable with $f' \geq \gamma$ for some constant $\gamma > 0$ and satisfies $f(0) = 0$ and $f(1) = 1$. Also, given $\eta_1, \ldots, \eta_n$, ε_i is a zero-mean random variable with variance at most σ^2. We further assume that $\varepsilon_1, \ldots, \varepsilon_n, y_1, \ldots, y_n$ are conditionally independent given $\eta_1, \ldots, \eta_n$.

The following counterpart of Theorem 3.1 holds in this noisy setting. This result is similar to Theorem 3 of Ushio et al. (2026).

Theorem 3.3. *Under Assumption 2.8 and Assumption 3.2, for any $\delta \in (0,1)$, each of the following holds with probability at least $1-\delta$:*

$$\left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \lesssim \frac{1}{\min\{\theta, 1-\theta\}}\left(\frac{\sigma}{\gamma} + \frac{1}{n^{1/3}} + \sqrt{\frac{\log(5/\delta)}{n}}\right),$$

$$\left|\widetilde{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \lesssim \frac{1}{\theta(1-\theta)}\left(\frac{\sigma}{\gamma} + \frac{1}{n^{1/3}} + \sqrt{\frac{\log(5/\delta)}{n}}\right).$$

A simple example of this model is when the corrupted soft labels are obtained by averaging multiple hard labels drawn from a corrupted Bernoulli distribution. That is, for each $i = 1, \ldots, n$, we independently draw m hard labels $y_i^{(1)}, \ldots, y_i^{(m)}$ from a Bernoulli distribution whose mean is $f(\eta_i)$ rather than η_i. We then obtain the noisy corrupted soft label $\xi_i = \frac{1}{m}\sum_{j=1}^m y_i^{(j)}$ by averaging these hard labels. In this case, $\xi_i = f(\eta_i) + \varepsilon_i$, where ε_i has mean zero and standard deviation at most $\sigma = 1/(2\sqrt{m})$. Therefore, the preceding bounds give estimation errors of order $m^{-1/2} + n^{-1/3}$. The additional $m^{-1/2}$ term diminishes as the number m of hard labels increases.

4 Real-world evaluation of optimal performance estimators

So far, we have discussed estimators of the optimal BER and AUC values and their theoretical guarantees. However, how can we empirically evaluate the performance of these estimators *on real-world datasets*? This is a non-trivial question because the true values of BER^* and AUC^* are unobservable for real-world data distributions. In this section, we first briefly review ***FeeBee*** (Renggli et al., 2021), an existing evaluation framework for estimators of the Bayes error Err^*. Then, we draw inspiration from FeeBee to propose a method to evaluate estimators of the optimal BER and AUC.

[6] $\mathbf{1}$ denotes the all-one vector.

Review of FeeBee The simplest way to evaluate Bayes error estimators on a real-world dataset would be to compare the estimates with the error rate u of the best available classifier on the dataset. Since we have a trivial bound $0 \leq \mathrm{Err}^* \leq u$ on the Bayes error, we can at least rule out estimators returning values greater than u. However, such a naive method is not very helpful because we cannot tell which of two estimators is better if both of them return estimates within the range $[0, u]$. Moreover, this approach cannot rule out an obviously useless estimator that always returns some constant between 0 and u.

The key idea of FeeBee is to inject label noise of a controlled level $\nu \in [0, 1)$ into the dataset. The injected noise increases the Bayes error Err^*_ν for the noise-injected distribution, which gives us another set of bounds $L(\nu) \leq \mathrm{Err}^*_\nu \leq U(\nu)$. We can then estimate Err^*_ν for the noise-injected dataset and test whether the estimate $\widehat{\mathrm{Err}^*_\nu}$ satisfies the bounds. Repeating this process for many different noise levels ν and aggregating the results, we can obtain a single score for the estimator.

Bounds on the optimal BER and AUC for noise-injected distributions In our proposed evaluation framework for the optimal BER and AUC, we inject label noise into the original distribution as follows: for a fixed noise level $\nu \in [0, 1)$, with probability ν, each original label y is replaced with an independent random label $w \in \{0, 1\}$ with a fixed mean $\beta \in (0, 1)$. The resulting noisy label is given by $y_\nu \coloneqq (1 - z) \cdot y + z \cdot w$, where $z \in \{0, 1\}$ is an independent Bernoulli random variable with mean ν. The class posterior for the noisy distribution is then related to the original posterior $\eta(x)$ by $\mathbb{P}(y_\nu = 1 \mid x) = \lambda_\nu(\eta(x))$, where $\lambda_\nu(t) \coloneqq (1 - \nu)t + \nu\beta$. Our noise model coincides with that of the original FeeBee when $\beta = \frac{1}{2}$, but we allow β to be chosen arbitrarily in $(0, 1)$.

The following theorem characterizes the optimal BER and AUC for the noise-injected distribution.

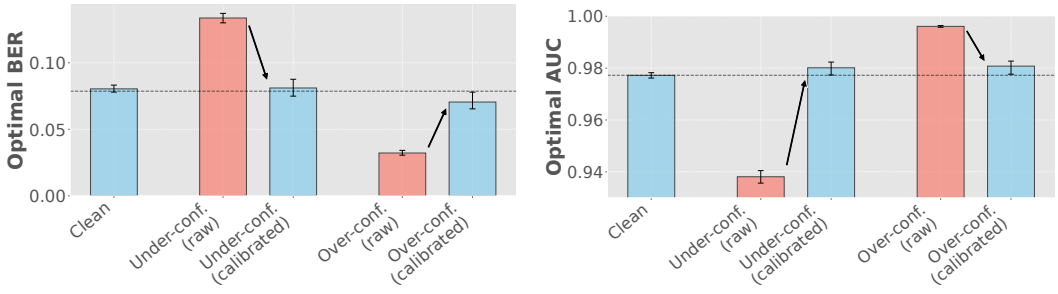
Theorem 4.1. *For $t \in [0, \frac{1}{2}]$, define $F_\nu(t) \coloneqq \frac{\lambda_\nu\left(2\theta(1-\theta)t+\theta^2\right)-\lambda_\nu(\theta)^2}{2\lambda_\nu(\theta)(1-\lambda_\nu(\theta))}$. For any noise level $\nu \in [0, 1)$, the optimal BER and AUC for the noise-injected distribution are given by $\mathrm{BER}^*_\nu = F_\nu(\mathrm{BER}^*)$ and $\mathrm{AUC}^*_\nu = 1 - F_\nu(1 - \mathrm{AUC}^*)$, respectively.*

Theorem 4.1 expresses the optimal BER value BER^*_ν for the noisy distribution as an increasing function of BER^*, and thus induces bounds on BER^*_ν as follows. Let u_{BER} be any upper bound on the optimal BER value BER^* for the original distribution (e.g., the lowest BER achieved by existing classifiers). Then, for each noise level $\nu \in [0, 1)$, BER^*_ν must fall between $L_{\mathrm{BER}}(\nu) \coloneqq F_\nu(0)$ and $U_{\mathrm{BER}}(\nu) \coloneqq F_\nu(u_{\mathrm{BER}})$. Similarly, the optimal AUC value AUC^*_ν for the noisy distribution must satisfy $L_{\mathrm{AUC}}(\nu) \coloneqq 1 - F_\nu(1 - l_{\mathrm{AUC}}) \leq \mathrm{AUC}^*_\nu \leq U_{\mathrm{AUC}}(\nu) \coloneqq 1 - F_\nu(0)$, where l_{AUC} is any lower bound on the optimal AUC value AUC^* for the original distribution.

Computing scores for BER and AUC Based on these bounds, we propose to evaluate estimators of the optimal BER on real-world datasets as follows. We first generate N noise-injected datasets with different noise levels $\nu_i \coloneqq \frac{i-1}{N}$ $(i = 1, \ldots, N)$ and evaluate a given estimator on each of them to obtain estimates $\widehat{\mathrm{BER}^*_{\nu_i}}$. Then, the estimator's score s_β for noise mean β is computed as $s_\beta \coloneqq \frac{1}{N}\sum_{i=1}^{N}\left[(\widehat{\mathrm{BER}^*_{\nu_i}} - U_{\mathrm{BER}}(\nu_i))_+ + (L_{\mathrm{BER}}(\nu_i) - \widehat{\mathrm{BER}^*_{\nu_i}})_+\right]$, where $(x)_+ \coloneqq \max\{x, 0\}$. The idea is that if the estimates fall outside their corresponding bounds $[L_{\mathrm{BER}}(\nu_i), U_{\mathrm{BER}}(\nu_i)]$, we penalize the estimator by the amount of violation; then the score is obtained by aggregating the penalties over all noise levels. The lower the score is, the better the estimator is. Note that the score s_β depends on the choice of β. Thus, how to choose β is an interesting question; possible choices include $\beta = \frac{1}{2}$ (the same as the original FeeBee) and $\beta = \theta$, in which case F_ν reduces simply to $F_\nu(t) = (1 - \nu)t + \frac{\nu}{2}$. Another option is averaging s_β over many different β values. In the experiments in Section 5.2, we will see that the last option aligns best with variance estimates, which suggests that it is a reasonable choice.[7]

The score for AUC can be computed in the same way by replacing L_{BER} and U_{BER} with L_{AUC} and U_{AUC}, respectively. This provides a practical way to evaluate estimators of the optimal BER and AUC on real-world datasets without requiring knowledge of the true optima. Note that the class prior θ has to be estimated from the data in order to compute the bounds. In many scenarios where the dataset contains hard labels $y_i \in \{0, 1\}$, this can be done by simply using their sample mean.

[7] This might be analogous to AUC, which aggregates performance over different classification thresholds.



(a) BER (b) AUC

Figure 2: The estimated optimal BER and AUC values with 95% bootstrap CIs on synthetic data; the dashed lines mark the ground-truth optimal values computed in closed form.

Bias sensitivity Without knowledge of the true optimal BER and AUC values, another way to assess the quality of an estimator is to compute the bootstrap variance of the estimator. However, the bootstrap variance only accounts for the variance of the estimator and does not reflect the estimator's bias. By contrast, our score can reflect certain types of estimator bias; see Appendix C.2 for details. Roughly speaking, any bias larger than the interval width $U_{\mathrm{BER}}(\nu) - L_{\mathrm{BER}}(\nu)$ produces a positive penalty in expectation even if the estimator has zero variance. The interval width scales as $1 - \nu$ and thus approaches zero as $\nu \to 1$. Therefore, for example, any constant additive bias that persists across noise levels is expected to be detected at a sufficiently large noise level. In this way, aggregating bound violations over different noise levels provides a meaningful signal of estimator quality.

5 Experiments

Here, we present experimental results to support our theoretical results. Further experimental details and additional results can be found in Appendix E.

5.1 Estimation on synthetic data

We first demonstrate the effectiveness of the proposed estimators by a simple experiment with synthetic data. Similarly to Ushio et al. (2026), we draw $n = 10{,}000$ points from a two-dimensional Gaussian mixture with class prior $\theta = 0.2$. We consider three settings: a *clean* one where the true soft labels η_i are given, and two corrupted ones where the observed corrupted soft labels ξ_i are either *under-confident* (pulled toward 0.5) or *over-confident* (pushed toward 0 or 1). We use the inverses of beta-calibration maps (Kull et al., 2017) as the underlying skew functions f following Ushio et al. (2026). For each setting and metric, we report the point estimate and 95% bootstrap confidence interval (CI) of the proposed estimator, namely, the estimator of Section 2 applied to η_i in the clean setting (*clean*), and the estimator of Section 3 based on isotonic regression in the corrupted settings (*calibrated*). For the corrupted settings, we also report a naive baseline (*raw*) that applies $\widehat{\mathrm{BER}}^*$ or $\widehat{\mathrm{AUC}}^*$ to the raw, uncalibrated ξ_i. In all cases, the class prior θ is treated as unknown and estimated from the data, as described earlier in Sections 2 and 3. We use the discriminant-based rules of Section 2.1 to pick between the min and max formulae on each sample. We also report the ground-truth optimal values, which can be computed in closed form using Proposition D.1 under this Gaussian-mixture setting.

Figure 2 shows the results. The outputs by the proposed estimators (blue bars) were clustered near the ground-truth values (dashed lines) for both BER and AUC while the naive baselines on corrupted soft labels (orange bars) had deviations an order of magnitude larger than the CI width, showing that the isotonic calibration of Section 3 plays an essential role in correcting the soft label corruption.

5.2 Real-data experiment: Discriminants and evaluation scores

We then perform estimation on different real-world datasets of soft labels from various domains to support (i) our discriminant theory in Section 2.1 and (ii) our proposed evaluation method in Section 4. Here we only discuss the results for BER (Table 1a), but the results are qualitatively similar for AUC

Table 1: Min vs. max formula estimators of BER^* and AUC^*. $\widehat{\Delta}_{\mathrm{BER}}, \widehat{\Delta}_{\mathrm{AUC}}$: the empirical discriminants $\pm$ its bootstrap standard errors; $\Delta\mathrm{SE}$, $\Delta\mathrm{Score}$: the $\max - \min$ gaps in the bootstrap standard error and evaluation score.

(a) BER

			ΔScore		
Dataset	$\widehat{\Delta}_{\mathrm{BER}}$	ΔSE	$\beta = 0.5$	$\beta = \theta$	β averaged
CIFAR-10	$+0.0093 \pm 0.0009$	$+2.62 \times 10^{-5}$	-0.0068	-2.53×10^{-4}	$+0.4236$
Fashion-MNIST	$(+0.00 \pm 9.12) \times 10^{-5}$	-4.03×10^{-5}	$+6.76 \times 10^{-5}$	$+8.71 \times 10^{-5}$	$+0.4332$
SNLI	$+0.0136 \pm 0.0023$	$+0.0044$	-0.0545	-0.0319	$+0.3868$
MNLI	$+0.0063 \pm 0.0020$	$+0.0047$	-0.0433	-0.0210	$+0.3567$
AbductiveNLI	$(+1.75 \pm 4.14) \times 10^{-4}$	$+1.71 \times 10^{-4}$	$+4.32 \times 10^{-4}$	$+7.52 \times 10^{-4}$	$+0.6099$
ICLR 2017–2026	$+0.0114 \pm 0.0003$	$+0.0027$	-0.0011	$+0.0331$	$+0.4380$

(b) AUC

			ΔScore		
Dataset	$\widehat{\Delta}_{\mathrm{AUC}}$	ΔSE	$\beta = 0.5$	$\beta = \theta$	β averaged
CIFAR-10	$+0.0094 \pm 0.0009$	$+2.08 \times 10^{-4}$	-0.0272	-5.34×10^{-4}	$+0.4168$
Fashion-MNIST	$(+0.00 \pm 5.09) \times 10^{-5}$	$+5.83 \times 10^{-5}$	-1.75×10^{-4}	-1.85×10^{-4}	$+0.4364$
SNLI	$+0.0182 \pm 0.0018$	$+0.0151$	-0.2275	-0.0820	$+0.1608$
MNLI	$+0.0091 \pm 0.0014$	$+0.0144$	-0.1951	-0.0503	$+0.1773$
AbductiveNLI	$(+1.10 \pm 2.91) \times 10^{-4}$	$+3.19 \times 10^{-4}$	$+7.99 \times 10^{-4}$	$+0.0012$	$+0.4303$
ICLR 2017–2026	$+0.0076 \pm 0.0002$	$+0.0011$	-0.0048	$+0.0530$	$+0.2072$

(Table 1b). First, on each dataset we tested which of the min/max formula estimators had smaller bootstrap variance estimates and then we compared this with the sign of the empirical discriminants. The result is shown in the $\widehat{\Delta}_{\mathrm{BER}}$ and $\Delta\mathrm{SE}$ columns of Table 1a, which aligns well with our theory that the sign of the discriminant can predict which estimator has smaller variance. The only sign mismatch is on Fashion-MNIST, where $\widehat{\Delta}$ is close to zero.

Next, we computed our evaluation score for each of the min/max estimators using the following choices of the parameter β: (i) $\beta = 0.5$, (ii) $\beta = \theta$ (estimated class prior), and (iii) average of the scores s_β over $\beta \in \{0.1, 0.2, \ldots, 0.9\}$, which are shown on the right side of Table 1a. We then compared them with the difference of bootstrapped standard errors. The results indicate that the scores averaged over different β values are far more consistent with the variance estimates than the scores at individual values ($\beta = 0.5$ or $\beta = \theta$). Note that this agreement does not imply that our score can be replaced by bootstrap variance estimates. As mentioned in Section 4, estimator bias can affect our score even when estimator variance remains unchanged. Bootstrap variance estimates, in contrast, quantify only estimator variance.

6 Conclusion

We studied the estimation of the Bayes-optimal BER and AUC, motivated by their practical importance. We proposed soft-label-based estimators which can be computed efficiently, and analyzed their theoretical properties. Our estimators handle not only clean soft labels but also corrupted soft labels, which makes them more practical. We also proposed a practical evaluation procedure for assessing the quality of estimators of the optimal BER and AUC without knowledge of the true optima.

Here, we mention a limitation of our work: our corruption model (Section 3) is quite simple and may not capture the full complexity of all possible data generation processes in practice. For example, it does not cover arbitrary instance-dependent corruption, which would require additional structure to recover the individual clean posterior values. Future work includes extending our results to other metrics popular under class imbalance or label noise, such as the F-measure (Gu et al., 2009; Parambath et al., 2014; Natarajan et al., 2015; Narasimhan et al., 2015). Most of such metrics are non-decomposable, i.e., their dataset-level values cannot be written as averages over individual data points, making their Bayes-optimal values harder to analyze and estimate.

Acknowledgments and Disclosure of Funding

RU was supported by the RIKEN Junior Research Associate Program, TI was supported by KAKENHI Grant Number 22K17946, and MS was supported by JST ASPIRE Grant Number JPMJAP25B1.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, pp. 641–647, 1955.

Berisha, V., Wisler, A., Hero, A. O., and Spanias, A. Empirically estimable classification bounds based on a nonparametric divergence measure. *IEEE Transactions on Signal Processing*, 64(3): 580–591, 2016.

Best, M. J. and Chakravarti, N. Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47(1):425–439, 1990.

Blom, G. Some properties of incomplete U-statistics. *Biometrika*, pp. 573–580, 1976.

Busing, F. M. Monotone regression: A simple and fast $O(n)$ PAVA implementation. *Journal of Statistical Software*, 102:1–25, 2022.

Callaert, H. and Veraverbeke, N. The order of the normal approximation for a studentized U-statistic. *The Annals of Statistics*, pp. 194–200, 1981.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pp. 961–970. PMLR, 2019.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

Clémençon, S., Colin, I., and Bellet, A. Scaling-up empirical risk minimization: Optimization of incomplete U-statistics. *Journal of Machine Learning Research*, 17(76):1–36, 2016.

Devijver, P. A. A multiclass, k-NN approach to Bayes risk estimation. *Pattern Recognition Letters*, 3 (1):1–6, 1985.

Fawcett, T. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Fukunaga, K. and Hostetler, L. k-nearest-neighbor Bayes-risk estimation. *IEEE Transactions on Information Theory*, 21(3):285–293, 1975.

Gu, Q., Zhu, L., and Cai, Z. Evaluation measures of the classification performance of imbalanced data sets. In *International Symposium on Intelligence Computation and Applications*, pp. 461–471. Springer, 2009.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Ishida, T., Yamane, I., Charoenphakdee, N., Niu, G., and Sugiyama, M. Is the performance of my deep network too good to be true? A direct approach to estimating the Bayes error in binary classification. In *International Conference on Learning Representations*, 2023.

Ishida, T., Lodkaew, T., and Yamane, I. CapBencher: Give your LLM benchmark a built-in alarm for test-set overfitting. In *Forty-third International Conference on Machine Learning*, 2026.

Jeong, M., Cardone, M., and Dytso, A. Demystifying the optimal performance of multi-class classification. *Advances in Neural Information Processing Systems*, 36:31638–31664, 2023.

Jeong, M., Cardone, M., and Dytso, A. Data-driven estimation of the false positive rate of the Bayes binary classifier via soft labels. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 368–373. IEEE, 2024.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kull, M., Silva Filho, T., and Flach, P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*, pp. 623–631. PMLR, 2017.

Lee, A. J. *U-Statistics: Theory and Practice*. Routledge, 2019.

Lehmann, E. L. and Casella, G. *Theory of Point Estimation*. Springer, 1998.

Ling, C. X. and Li, C. Data mining for direct marketing: Problems and solutions. In *The Fourth International Conference on Knowledge Discovery and Data Mining*, volume 98, pp. 73–79, 1998.

Luccioni, A. S., Viguier, S., and Ligozat, A.-L. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24(253):1–15, 2023.

Menon, A., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pp. 603–611. PMLR, 2013.

Menon, A., van Rooyen, B., Ong, C. S., and Williamson, B. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pp. 125–134. PMLR, 2015.

Menon, A. K., van Rooyen, B., and Natarajan, N. Learning from binary labels with instance-dependent noise. *Machine Learning*, 107(8):1561–1595, 2018.

Moon, K. R., Sricharan, K., Greenewald, K., and Hero III, A. O. Ensemble estimation of information divergence. *Entropy*, 20(8):560, 2018.

Narasimhan, H., Kar, P., and Jain, P. Optimizing non-decomposable performance measures: A tale of two classes. In *International Conference on Machine Learning*, pp. 199–208. PMLR, 2015.

Natarajan, N., Koyejo, O., Ravikumar, P., and Dhillon, I. S. Optimal decision-theoretic classification using non-decomposable performance metrics. *arXiv preprint arXiv:1505.01802*, 2015.

Neykov, M. Tossing coins under monotonicity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 21–30. PMLR, 2019.

Nie, Y., Zhou, X., and Bansal, M. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

Noshad, M., Xu, L., and Hero, A. Learning to benchmark: Determining best achievable misclassification error from training data. *arXiv preprint arXiv:1909.07192*, 2019.

Parambath, S. A., Usunier, N., and Grandvalet, Y. Optimizing F-measures by cost-sensitive classification. *Advances in Neural Information Processing Systems*, 27, 2014.

Peterson, J. C., Battleday, R. M., Griffiths, T. L., and Russakovsky, O. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9616–9625, 2019.

Renggli, C., Rimanic, L., Hollenstein, N., and Zhang, C. Evaluating Bayes error estimators on real-world datasets with FeeBee. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696, 2020.

Theisen, R., Wang, H., Varshney, L. R., Xiong, C., and Socher, R. Evaluating state-of-the-art classification models against Bayes optimality. *Advances in Neural Information Processing Systems*, 34:9367–9377, 2021.

Ushio, R., Ishida, T., and Sugiyama, M. Practical estimation of the optimal classification error with soft labels and calibration. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, 2017.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 694–699, 2002.

Zhou, X., Nie, Y., and Bansal, M. Distributed NLI: Learning to predict human opinion distributions for language reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022.

A Supplementary for Section 2

A.1 Deriving the optimal BER and AUC

Here, we prove Lemmas 2.1 and 2.2 that gives the expressions for the optimal BER and AUC. In the following, $\mathbb{1}[\cdot]$ denotes the indicator function.

Lemma 2.1. *The optimal BER is given by* $\mathrm{BER}^* = \mathbb{E}\left[\phi^{(1)}_{\mathrm{BER}}(\eta(x))\right] = \mathbb{E}\left[\phi^{(2)}_{\mathrm{BER}}(\eta(x))\right]$*, where* $\phi^{(1)}_{\mathrm{BER}}(z) \coloneqq \frac{1}{2}\min\left\{\frac{z}{\theta}, \frac{1-z}{1-\theta}\right\}$ *and* $\phi^{(2)}_{\mathrm{BER}}(z) \coloneqq 1 - \frac{1}{2}\max\left\{\frac{z}{\theta}, \frac{1-z}{1-\theta}\right\}$.

Proof. For any classifier $h : \mathcal{X} \to \{0, 1\}$, its false negative rate and false positive rate can be expressed as $\mathbb{E}\left[\frac{\eta(x)}{\theta}\mathbb{1}[h(x) = 0]\right]$ and $\mathbb{E}\left[\frac{1-\eta(x)}{1-\theta}\mathbb{1}[h(x) = 1]\right]$, respectively, by Bayes' theorem. Therefore, we have

$$\mathrm{BER}(h) = \frac{1}{2}\mathbb{E}\left[\frac{\eta(x)}{\theta}\mathbb{1}[h(x) = 0] + \frac{1-\eta(x)}{1-\theta}\mathbb{1}[h(x) = 1]\right],$$

which is minimized by taking $h(x) = \mathbb{1}[\eta(x) \geq \theta]$. This gives the optimal BER as

$$\mathrm{BER}^* = \frac{1}{2}\mathbb{E}\left[\min\left\{\frac{\eta(x)}{\theta}, \frac{1-\eta(x)}{1-\theta}\right\}\right] = \mathbb{E}\left[\phi^{(1)}_{\mathrm{BER}}(\eta(x))\right].$$

Next, since $\mathbb{E}[\eta(x)] = \theta$, we have

$$\mathbb{E}\left[\min\left\{\frac{\eta(x)}{\theta}, \frac{1-\eta(x)}{1-\theta}\right\} + \max\left\{\frac{\eta(x)}{\theta}, \frac{1-\eta(x)}{1-\theta}\right\}\right] = \mathbb{E}\left[\frac{\eta(x)}{\theta} + \frac{1-\eta(x)}{1-\theta}\right] = 2.$$

This implies $\mathrm{BER}^* = \mathbb{E}\left[\phi^{(2)}_{\mathrm{BER}}(\eta(x))\right]$. □

Lemma 2.2. *Let* x' *be an i.i.d. copy of* x*. Then, the optimal AUC is given by* $\mathrm{AUC}^* = \mathbb{E}\left[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right] = \mathbb{E}\left[\phi^{(2)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right]$*, where* $\phi^{(1)}_{\mathrm{AUC}}(z_1, z_2) \coloneqq 1 - \frac{1}{2\theta(1-\theta)}\min\{z_1(1-z_2), z_2(1-z_1)\}$ *and* $\phi^{(2)}_{\mathrm{AUC}}(z_1, z_2) \coloneqq \frac{1}{2\theta(1-\theta)}\max\{z_1(1-z_2), z_2(1-z_1)\}$.

Proof. Fix any scoring function $g : \mathcal{X} \to \mathbb{R}$. By Bayes' theorem, the expectation over $x_+ \sim \mathbb{P}(x \mid y = 1)$ and $x_- \sim \mathbb{P}(x \mid y = 0)$ can be rewritten as an expectation over the i.i.d. pair (x, x'):

$$\begin{aligned}\mathrm{AUC}(g) &= \mathbb{E}\left[\mathbb{1}[g(x_+) > g(x_-)] + \frac{1}{2}\mathbb{1}[g(x_+) = g(x_-)]\right] \\ &= \frac{1}{\theta(1-\theta)}\mathbb{E}\left[\eta(x)(1-\eta(x'))\left(\mathbb{1}[g(x) > g(x')] + \frac{1}{2}\mathbb{1}[g(x) = g(x')]\right)\right]. \quad (1)\end{aligned}$$

Noting x and x' are independent and identically distributed, we may exchange them to get

$$\mathrm{AUC}(g) = \frac{1}{\theta(1-\theta)}\mathbb{E}\left[\eta(x')(1-\eta(x))\left(\mathbb{1}[g(x) < g(x')] + \frac{1}{2}\mathbb{1}[g(x) = g(x')]\right)\right]. \quad (2)$$

Adding (1) and (2), we obtain

$$\begin{aligned}\mathrm{AUC}(g) = \frac{1}{2\theta(1-\theta)}\mathbb{E}\Big[&\eta(x)(1-\eta(x'))\,\mathbb{1}[g(x) > g(x')] \\ &+ \frac{\eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x))}{2}\mathbb{1}[g(x) = g(x')] \\ &+ \eta(x')(1-\eta(x))\,\mathbb{1}[g(x) < g(x')]\Big].\end{aligned}$$

Since $\eta(x)(1-\eta(x')) < \eta(x')(1-\eta(x)) \iff \eta(x) < \eta(x')$ (and similarly for the other direction), the expression inside the expectation is maximized by taking $g = \eta$ (or any strictly increasing transformation of η) and thus the maximum AUC is given by

$$\mathrm{AUC}^* = \frac{1}{2\theta(1-\theta)}\mathbb{E}\left[\max\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\}\right] = \mathbb{E}\left[\phi^{(2)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right].$$

To prove $\mathrm{AUC}^* = \mathbb{E}\left[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right]$, observe that

$$\begin{aligned}
&\mathbb{E}\left[\min\left\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\right\} + \max\left\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\right\}\right] \\
&= \mathbb{E}\left[\eta(x)(1-\eta(x')) + \eta(x')(1-\eta(x))\right] \\
&= 2\theta(1-\theta).
\end{aligned}$$

Here, we used $\mathbb{E}\left[\eta(x)\right] = \theta$ and the independence of x and x'. □

A.2 Discriminants for BER and AUC

First, we recall the definitions of the discriminants for BER and AUC. Let $\zeta \coloneqq \eta(x) - \theta$ be the centered version of the class posterior. The discriminant for BER is defined as $\Delta_{\mathrm{BER}} \coloneqq (1-2\theta)\,\mathbb{E}\left[\zeta\,|\zeta|\right]$. For AUC, we define $\Delta_{\mathrm{AUC}} \coloneqq (1-2\theta)\,\mathbb{E}\left[\zeta\,|\zeta - \zeta'|\right]$, where $\zeta' \coloneqq \eta(x') - \theta$ is an i.i.d. copy of ζ.

Now we prove Theorem 2.4, which we restate below:

Theorem 2.4.

(i) For all $n \in \mathbb{N}$, we have $\mathrm{Var}\left[\widehat{\mathrm{BER}}^*_1\right] \leq \mathrm{Var}\left[\widehat{\mathrm{BER}}^*_2\right]$ *if and only if* $\Delta_{\mathrm{BER}} \geq 0$.

(ii) Assume that the random variable $\eta(x)$ is not a.s. constant. Then, we have $\mathrm{Var}\left[\widehat{\mathrm{AUC}}^*_1\right] < \mathrm{Var}\left[\widehat{\mathrm{AUC}}^*_2\right]$ *for all $n \geq 2$ if and only if* $\Delta_{\mathrm{AUC}} \geq 0$.

We first prove the BER part:

Proposition A.1. *For all $n \in \mathbb{N}$, we have* $\mathrm{Var}\left[\widehat{\mathrm{BER}}^*_1\right] \leq \mathrm{Var}\left[\widehat{\mathrm{BER}}^*_2\right]$ *if and only if* $\Delta_{\mathrm{BER}} \geq 0$.

Proposition A.1 follows immediately from the next lemma.

Lemma A.2. *The variance gap between the two minimum BER estimators* $\widehat{\mathrm{BER}}^*_1$ *and* $\widehat{\mathrm{BER}}^*_2$ *is given by*

$$\mathrm{Var}\left[\widehat{\mathrm{BER}}^*_2\right] - \mathrm{Var}\left[\widehat{\mathrm{BER}}^*_1\right] = \frac{1-2\theta}{4\theta^2(1-\theta)^2 n}\,\mathbb{E}\left[\zeta|\zeta|\right].$$

Proof. Throughout the proof, we abuse the notation and write $\eta = \eta(x)$ for brevity. Let

$$M \coloneqq \max\left\{\eta(1-\theta), \theta(1-\eta)\right\}, \quad m \coloneqq \min\left\{\eta(1-\theta), \theta(1-\eta)\right\}.$$

Noting that

$$\begin{aligned}
M &= \frac{1}{2}\Big[\left\{\eta(1-\theta) + \theta(1-\eta)\right\} + |\eta - \theta|\Big], \\
m &= \frac{1}{2}\Big[\left\{\eta(1-\theta) + \theta(1-\eta)\right\} - |\eta - \theta|\Big],
\end{aligned}$$

it holds that

$$\begin{aligned}
\mathrm{Var}\left[M\right] - \mathrm{Var}\left[m\right] &= \mathrm{Cov}\left[\eta(1-\theta) + \theta(1-\eta), |\eta - \theta|\right] \\
&= \mathrm{Cov}\left[(1-2\theta)\eta + \theta, |\eta - \theta|\right] \\
&= (1-2\theta)\,\mathrm{Cov}\left[\eta - \theta, |\eta - \theta|\right] \\
&= (1-2\theta)\,\mathrm{Cov}\left[\zeta, |\zeta|\right].
\end{aligned}$$

Since ζ is a zero-mean random variable, we have $\mathrm{Cov}\left[\zeta, |\zeta|\right] = \mathbb{E}\left[\zeta|\zeta|\right]$. Also observe that $\phi^{(1)}_{\mathrm{BER}}(\eta) = \frac{1}{2\theta(1-\theta)}m$ and $\phi^{(2)}_{\mathrm{BER}}(\eta) = 1 - \frac{1}{2\theta(1-\theta)}M$. Hence, we obtain

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\mathrm{BER}}^*_2\right] - \mathrm{Var}\left[\widehat{\mathrm{BER}}^*_1\right] &= \frac{1}{n}\left(\mathrm{Var}\left[\phi^{(2)}_{\mathrm{BER}}(\eta)\right] - \mathrm{Var}\left[\phi^{(1)}_{\mathrm{BER}}(\eta)\right]\right) \\
&= \frac{1}{4\theta^2(1-\theta)^2 n}\left(\mathrm{Var}\left[M\right] - \mathrm{Var}\left[m\right]\right) \\
&= \frac{1-2\theta}{4\theta^2(1-\theta)^2 n}\,\mathbb{E}\left[\zeta|\zeta|\right].
\end{aligned}$$

□

Next, we prove the AUC part of Theorem 2.4 (restated as Proposition A.6). Recall that $\zeta \coloneqq \eta(x) - \theta$, $\zeta' \coloneqq \eta(x') - \theta$. ζ and ζ' are i.i.d. random variables with mean 0.

Lemma A.3. *The difference between the variances of the two maximum AUC estimators* $\widehat{\mathrm{AUC}_1^*}$ *and* $\widehat{\mathrm{AUC}_2^*}$ *is given by*

$$\mathrm{Var}\left[\widehat{\mathrm{AUC}_2^*}\right] - \mathrm{Var}\left[\widehat{\mathrm{AUC}_1^*}\right] = \frac{1-2\theta}{\theta^2(1-\theta)^2 n}\,\mathbb{E}\left[\zeta|\zeta-\zeta'|\right] - \frac{1}{\theta^2(1-\theta)^2 n(n-1)}\,\mathbb{E}\left[\zeta\zeta'|\zeta-\zeta'|\right].$$

Proof. Throughout the proof, we abuse the notation and write $\eta = \eta(x)$ and $\eta' = \eta(x')$ for brevity. Let

$$M(z,w) \coloneqq \max\left\{z(1-w), w(1-z)\right\}, \quad m(z,w) \coloneqq \min\left\{z(1-w), w(1-z)\right\},$$

and define

$$M_1 \coloneqq \mathbb{E}\left[M(\eta,\eta') \mid \eta\right], \quad m_1 \coloneqq \mathbb{E}\left[m(\eta,\eta') \mid \eta\right].$$

By the general theory of U-statistics (Hoeffding, 1948), we have

$$\mathrm{Var}\left[\frac{2}{n(n-1)}\sum_{i<j} M(\eta_i,\eta_j)\right] = \frac{2}{n(n-1)}\left(2(n-2)\,\mathrm{Var}\left[M_1\right] + \mathrm{Var}\left[M(\eta,\eta')\right]\right),$$
$$\mathrm{Var}\left[\frac{2}{n(n-1)}\sum_{i<j} m(\eta_i,\eta_j)\right] = \frac{2}{n(n-1)}\left(2(n-2)\,\mathrm{Var}\left[m_1\right] + \mathrm{Var}\left[m(\eta,\eta')\right]\right).$$

Noting that

$$M(\eta,\eta') = \frac{1}{2}\Big[\left\{\eta(1-\eta') + \eta'(1-\eta)\right\} + |\eta-\eta'|\Big],$$
$$m(\eta,\eta') = \frac{1}{2}\Big[\left\{\eta(1-\eta') + \eta'(1-\eta)\right\} - |\eta-\eta'|\Big],$$

we have

$$M_1 = \frac{1}{2}\Big[\left\{\eta(1-\theta) + \theta(1-\eta)\right\} + \mathbb{E}\left[|\eta-\eta'| \mid \eta\right]\Big],$$
$$m_1 = \frac{1}{2}\Big[\left\{\eta(1-\theta) + \theta(1-\eta)\right\} - \mathbb{E}\left[|\eta-\eta'| \mid \eta\right]\Big].$$

Therefore, it holds that

$$\begin{aligned}
\mathrm{Var}\left[M_1\right] - \mathrm{Var}\left[m_1\right] &= \mathrm{Cov}\left[\eta(1-\theta) + \theta(1-\eta), \mathbb{E}\left[|\eta-\eta'| \mid \eta\right]\right] \\
&= \mathrm{Cov}\left[(1-2\theta)\eta + \theta, \mathbb{E}\left[|\eta-\eta'| \mid \eta\right]\right] \\
&= (1-2\theta)\,\mathrm{Cov}\left[\eta, \mathbb{E}\left[|\eta-\eta'| \mid \eta\right]\right] \\
&= (1-2\theta)\,\mathrm{Cov}\left[\eta, |\eta-\eta'|\right] \\
&= (1-2\theta)\,\mathrm{Cov}\left[\zeta, |\zeta-\zeta'|\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}\left[M(\eta,\eta')\right] - \mathrm{Var}\left[m(\eta,\eta')\right] &= \mathrm{Cov}\left[\eta(1-\eta') + \eta'(1-\eta), |\eta-\eta'|\right] \\
&= \mathrm{Cov}\left[(1-2\theta)(\zeta+\zeta') - 2\zeta\zeta' + 2\theta(1-\theta), |\zeta-\zeta'|\right] \\
&= (1-2\theta)\,\mathrm{Cov}\left[(\zeta+\zeta'), |\zeta-\zeta'|\right] - 2\,\mathrm{Cov}\left[\zeta\zeta', |\zeta-\zeta'|\right] \\
&= 2(1-2\theta)\,\mathrm{Cov}\left[\zeta, |\zeta-\zeta'|\right] - 2\,\mathrm{Cov}\left[\zeta\zeta', |\zeta-\zeta'|\right].
\end{aligned}$$

Combining these results, we obtain

$$\begin{aligned}
&\operatorname{Var}\left[\widehat{\mathrm{AUC}_2^*}\right] - \operatorname{Var}\left[\widehat{\mathrm{AUC}_1^*}\right] \\
&= \frac{1}{4\theta^2(1-\theta)^2}\left(\operatorname{Var}\left[\frac{2}{n(n-1)}\sum_{i<j} M(\eta_i,\eta_j)\right] - \operatorname{Var}\left[\frac{2}{n(n-1)}\sum_{i<j} m(\eta_i,\eta_j)\right]\right) \\
&= \frac{1}{4\theta^2(1-\theta)^2}\frac{2}{n(n-1)}\Big[2(n-2)\cdot(1-2\theta)\operatorname{Cov}\left[\zeta,|\zeta-\zeta'|\right] \\
&\qquad\qquad + 2(1-2\theta)\operatorname{Cov}\left[\zeta,|\zeta-\zeta'|\right] - 2\operatorname{Cov}\left[\zeta\zeta',|\zeta-\zeta'|\right]\Big] \\
&= \frac{1}{\theta^2(1-\theta)^2}\frac{1}{n(n-1)}\Big[(n-1)\cdot(1-2\theta)\operatorname{Cov}\left[\zeta,|\zeta-\zeta'|\right] - \operatorname{Cov}\left[\zeta\zeta',|\zeta-\zeta'|\right]\Big].
\end{aligned}$$

Finally, since both ζ and $\zeta\zeta'$ have zero means, we have $\operatorname{Cov}\left[\zeta,|\zeta-\zeta'|\right] = \mathbb{E}\left[\zeta|\zeta-\zeta'|\right]$ and $\operatorname{Cov}\left[\zeta\zeta',|\zeta-\zeta'|\right] = \mathbb{E}\left[\zeta\zeta'|\zeta-\zeta'|\right]$. Now the lemma has been proved. □

Remark A.4 (Asymptotic variance difference). Let $\sigma_{i,n}^2 \coloneqq \operatorname{Var}\left[\sqrt{n}\left(\widehat{\mathrm{AUC}_i^*} - \mathrm{AUC}^*\right)\right]$. Lemma A.3 implies that[8]

$$\sigma_{2,n}^2 - \sigma_{1,n}^2 \xrightarrow{n\to\infty} \frac{1-2\theta}{\theta^2(1-\theta)^2}\mathbb{E}\left[\zeta|\zeta-\zeta'|\right].$$

Lemma A.5. *We have*

$$\mathbb{E}\left[\zeta\zeta'|\zeta-\zeta'|\right] < 0$$

unless ζ is a.s. constant.

Proof. If $\zeta\zeta' > 0$, we have $|\zeta-\zeta'| < |\zeta| + |\zeta'|$ and hence $\zeta\zeta'|\zeta-\zeta'| < \zeta\zeta'(|\zeta|+|\zeta'|)$. On the other hand, if $\zeta\zeta' \le 0$, it holds that $|\zeta-\zeta'| = |\zeta| + |\zeta'|$ and thus $\zeta\zeta'|\zeta-\zeta'| = \zeta\zeta'(|\zeta|+|\zeta'|)$. Therefore, we always have $\zeta\zeta'|\zeta-\zeta'| \le \zeta\zeta'(|\zeta|+|\zeta'|)$ and the equality holds only when $\zeta\zeta' \le 0$.

Taking the expectation on both sides, we obtain

$$\begin{aligned}
\mathbb{E}\left[\zeta\zeta'|\zeta-\zeta'|\right] &\le \mathbb{E}\left[\zeta\zeta'(|\zeta|+|\zeta'|)\right] \\
&= \mathbb{E}\left[\zeta'\right]\mathbb{E}\left[\zeta|\zeta|\right] + \mathbb{E}\left[\zeta\right]\mathbb{E}\left[\zeta'|\zeta'|\right] \\
&= 0.
\end{aligned}$$

Here, we used the independence of ζ and ζ' as well as their zero means.

To see that the strict inequality holds, note that

$$\begin{aligned}
\mathbb{P}\left(\zeta\zeta'|\zeta-\zeta'| \ne \zeta\zeta'(|\zeta|+|\zeta'|)\right) &= \mathbb{P}\left(\zeta\zeta' > 0\right) \\
&= \mathbb{P}\left((\zeta > 0 \text{ and } \zeta' > 0) \text{ or } (\zeta < 0 \text{ and } \zeta' < 0)\right) \\
&= \mathbb{P}\left(\zeta > 0\right)^2 + \mathbb{P}\left(\zeta < 0\right)^2.
\end{aligned}$$

We must have $\mathbb{P}\left(\zeta > 0\right)^2 + \mathbb{P}\left(\zeta < 0\right)^2 > 0$, since otherwise ζ would be a.s. constant. Therefore, we have a non-zero probability that $\zeta\zeta'|\zeta-\zeta'|$ disagrees with $\zeta\zeta'\left(|\zeta|+|\zeta'|\right)$, which implies the strict inequality. □

We can now prove the desired result, i.e., the AUC part of Theorem 2.4, by combining Lemma A.3 and Lemma A.5.

Proposition A.6. *Assume that the random variable $\eta(x)$ is not a.s. constant, and define $\Delta_{\mathrm{AUC}} \coloneqq (1-2\theta)\,\mathbb{E}\left[\zeta|\zeta-\zeta'|\right]$. Then, $\operatorname{Var}\left[\widehat{\mathrm{AUC}_1^*}\right] < \operatorname{Var}\left[\widehat{\mathrm{AUC}_2^*}\right]$ holds for all $n \ge 2$ if and only if $\Delta_{\mathrm{AUC}} \ge 0$.*

[8] Also, it can be shown that $\sqrt{n}\left(\widehat{\mathrm{AUC}_i^*} - \mathrm{AUC}^*\right) \xrightarrow{d} \mathcal{N}(0, \lim_{n\to\infty}\sigma_{i,n}^2)$ for each $i = 1, 2$ (Hoeffding, 1948).

Proof. If $\Delta_{\mathrm{AUC}} \geq 0$, then Lemmas A.3 and A.5 gives $\mathrm{Var}\left[\widehat{\mathrm{AUC}_2^*}\right] - \mathrm{Var}\left[\widehat{\mathrm{AUC}_1^*}\right] > 0$. Conversely, if $\Delta_{\mathrm{AUC}} < 0$, we have

$$\lim_{n\to\infty} n\left(\mathrm{Var}\left[\widehat{\mathrm{AUC}_2^*}\right] - \mathrm{Var}\left[\widehat{\mathrm{AUC}_1^*}\right]\right) = \frac{\Delta_{\mathrm{AUC}}}{\theta^2(1-\theta)^2} < 0,$$

and therefore we will end up with $\mathrm{Var}\left[\widehat{\mathrm{AUC}_1^*}\right] > \mathrm{Var}\left[\widehat{\mathrm{AUC}_2^*}\right]$ for sufficiently large n. □

A.2.1 How to choose the better estimator based on data

Define $h(x,y) \coloneqq \frac{1-2\theta}{2}(x+y)|x-y|$, so that we have $\Delta_{\mathrm{AUC}} = \mathbb{E}\left[h(\eta-\theta, \eta'-\theta)\right]$. Given a sample $\{\eta_i\}_{i=1}^n$, let

$$\bar{h}_i \coloneqq \frac{1}{n-1}\sum_{j:\, j\neq i} h(\eta_i-\theta, \eta_j-\theta)$$

for each $i \in \{1,\dots,n\}$. Define $\widehat{\Delta}_{\mathrm{AUC}}$ and $\widehat{v}_{\mathrm{AUC}}$ as

$$\widehat{\Delta}_{\mathrm{AUC}} \coloneqq \frac{1}{n}\sum_{i=1}^n \bar{h}_i. \tag{3}$$

Then, $\widehat{\Delta}_{\mathrm{AUC}}$ is clearly an unbiased estimator of Δ_{AUC}. Therefore, we can use the sign of $\widehat{\Delta}_{\mathrm{AUC}}$ as a data-dependent criterion to choose between $\widehat{\mathrm{AUC}_1^*}$ and $\widehat{\mathrm{AUC}_2^*}$: we choose $\widehat{\mathrm{AUC}_1^*}$ if $\widehat{\Delta}_{\mathrm{AUC}} \geq 0$, and choose $\widehat{\mathrm{AUC}_2^*}$ if $\widehat{\Delta}_{\mathrm{AUC}} < 0$.

Moreover, we can test the sign of Δ_{AUC} based on the asymptotic normality of $\widehat{\Delta}_{\mathrm{AUC}}$. by the general theory of U-statistics of order two (Callaert & Veraverbeke, 1981), we have

$$\frac{\widehat{\Delta}_{\mathrm{AUC}} - \Delta_{\mathrm{AUC}}}{\sqrt{\widehat{v}_{\mathrm{AUC}}/n}} \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\widehat{v}_{\mathrm{AUC}} \coloneqq \frac{4(n-1)}{(n-2)^2}\sum_{i=1}^n \left(\bar{h}_i - \widehat{\Delta}_{\mathrm{AUC}}\right)^2. \tag{4}$$

Based on this asymptotic normality, we can construct a Wald-type hypothesis test for the sign of Δ_{AUC}:

$$H_0\colon \Delta_{\mathrm{AUC}} < 0 \quad \text{vs.} \quad H_1\colon \Delta_{\mathrm{AUC}} \geq 0.$$

We define the test statistic as

$$T_{\mathrm{AUC}} \coloneqq \frac{\widehat{\Delta}_{\mathrm{AUC}}}{\sqrt{\widehat{v}_{\mathrm{AUC}}/n}}$$

and reject H_0 if $T_{\mathrm{AUC}} \geq z_\alpha$, where z_α is the upper α-quantile of the standard normal distribution. The p-value of this test is given by $1-\Phi(T_{\mathrm{AUC}})$, where Φ is the cumulative distribution function of the standard normal distribution. We can also consider the converse test

$$H_0\colon \Delta_{\mathrm{AUC}} \geq 0 \quad \text{vs.} \quad H_1\colon \Delta_{\mathrm{AUC}} < 0,$$

by rejecting H_0 if $T_{\mathrm{AUC}} \leq z_{1-\alpha}$. The p-value of this test is given by $\Phi(T_{\mathrm{AUC}})$.

A.2.2 Efficient computation of $\widehat{\Delta}_{\mathrm{AUC}}$ and $\widehat{v}_{\mathrm{AUC}}$

Naively, computing $\widehat{\Delta}_{\mathrm{AUC}}$ and $\widehat{v}_{\mathrm{AUC}}$ would take $O(n^2)$ time due to the double summation. However, we can construct an $O(n\log n)$-time algorithm, which is based on the following theorem.

Theorem A.7. *For each $j \in \{1,\dots,n\}$, let (j) denote the index of the j-th smallest element in $\{\eta_1,\dots,\eta_n\}$, i.e., $\eta_{(1)} \leq \cdots \leq \eta_{(n)}$, and let S_i be the i-th prefix sum of $\left\{(\eta_{(j)}-\theta)^2\right\}_{j=1}^n$, i.e., $S_i \coloneqq \sum_{j=1}^i (\eta_{(j)}-\theta)^2$. Then, $\bar{h}_{(i)}$ can be expressed as*

$$\bar{h}_{(i)} = \frac{1-2\theta}{2(n-1)}\left[(2i-n)(\eta_{(i)}-\theta)^2 + S_n - 2S_i\right]. \tag{5}$$

Algorithm A.1 An $O(n \log n)$-time algorithm for computing $\widehat{\Delta}_{\mathrm{AUC}}$ and $\widehat{v}_{\mathrm{AUC}}$

input Soft labels $\eta_1, \ldots, \eta_n$, class prior θ
$\eta_{(1)}, \ldots, \eta_{(n)} \leftarrow$ sort $\eta_1, \ldots, \eta_n$ in ascending order
$S_0 \leftarrow 0$
for $i = 1$ **to** n **do**
$S_i \leftarrow S_{i-1} + (\eta_{(i)} - \theta)^2$
end for
for $i = 1$ **to** n **do**
$\bar{h}_{(i)} \leftarrow \frac{1-2\theta}{2(n-1)} \left[(2i-n)(\eta_{(i)} - \theta)^2 + S_n - 2S_i\right]$
end for
$\widehat{\Delta}_{\mathrm{AUC}} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \bar{h}_i$
$\widehat{v}_{\mathrm{AUC}} \leftarrow \frac{4(n-1)}{(n-2)^2} \sum_{i=1}^{n} \left(\bar{h}_i - \widehat{\Delta}_{\mathrm{AUC}}\right)^2$
output $\widehat{\Delta}_{\mathrm{AUC}}, \widehat{v}_{\mathrm{AUC}}$

Proof. For ease of notation, let $a_i \coloneqq \frac{2(n-1)}{1-2\theta} \bar{h}_{(i)}$ and $\zeta_i \coloneqq \eta_{(i)} - \theta$. Then

$$\begin{aligned} a_i &= \sum_{j:j\neq i} (\zeta_i + \zeta_j)|\zeta_i - \zeta_j| \\ &= \sum_{j<i} (\zeta_i + \zeta_j)(\zeta_i - \zeta_j) + \sum_{j>i} (\zeta_i + \zeta_j)(\zeta_j - \zeta_i) \\ &= \sum_{j<i} \left(\zeta_i^2 - \zeta_j^2\right) + \sum_{j>i} \left(\zeta_j^2 - \zeta_i^2\right) \\ &= (i-1)\zeta_i^2 - \sum_{j<i} \zeta_j^2 + \sum_{j>i} \zeta_j^2 - (n-i)\zeta_i^2. \end{aligned}$$

Observe that

$$\sum_{j<i} \zeta_j^2 = S_{i-1} = S_i - \zeta_i^2, \quad \sum_{j>i} \zeta_j^2 = S_n - S_i.$$

Hence, we have

$$a_i = (2i-n)\zeta_i^2 + S_n - 2S_i,$$

which completes the proof. □

Theorem A.7 gives an algorithm that computes $\widehat{\Delta}_{\mathrm{AUC}}$ and $\widehat{v}_{\mathrm{AUC}}$ in $O(n \log n)$ time, as shown in Algorithm A.1. The steps are as follows:

1. Sort $\eta_1, \ldots, \eta_n$ in ascending order to obtain $\eta_{(1)}, \ldots, \eta_{(n)}$: $O(n \log n)$ time, assuming standard sorting algorithms such as Mergesort.
2. Compute the prefix sums $S_1, \ldots, S_n$: $O(n)$ time in total.
3. For each $i = 1, \ldots, n$, compute $\bar{h}_{(i)}$ using (5): constant time for each i, hence $O(n)$ time in total.
4. Compute $\widehat{\Delta}_{\mathrm{AUC}}$ and $\widehat{v}_{\mathrm{AUC}}$ by (3) and (4): $O(n)$ time.

A.2.3 Sufficient conditions for non-negative discriminants

Here, we discuss some sufficient conditions for $\Delta_{\mathrm{BER}}, \Delta_{\mathrm{AUC}} \geq 0$.

Proposition A.8. *Assume that the random variable $\eta(x)$ is not a.s. constant. Then, we have $\Delta_{\mathrm{BER}} \geq 0$ and $\Delta_{\mathrm{AUC}} \geq 0$ if one of the following conditions holds:*

(i) The two classes are balanced, i.e., $\theta = \frac{1}{2}$.

(ii) The law of $\eta(x)$ is symmetric around its mean θ (which may or may not be equal to $\frac{1}{2}$).

(iii) The Bayes error satisfies $\mathrm{Err}^* \leq \delta_\theta$*, where* $\delta_\theta > 0$ *is a certain constant depending only on* θ*.*

(i) is trivial by the definitions of Δ_{BER} and Δ_{AUC}. In what follows, we prove the rest of Proposition A.8. Specifically, (ii) follows from Lemma A.9; we prove (iii) as Lemma A.11.

Lemma A.9. *Assume that the distribution of* ζ *is symmetric around* 0*, i.e.,* ζ *and* $-\zeta$ *are identically distributed. Then, we have the following: (i)* $\mathbb{E}\left[\zeta|\zeta|\right] = 0$ *and (ii)* $\mathbb{E}\left[\zeta|\zeta - \zeta'|\right] = 0$*.*

Proof. (i) By symmetry, we have

$$\mathbb{E}\left[\zeta|\zeta|\right] = \mathbb{E}\left[-\zeta|-\zeta|\right] = -\mathbb{E}\left[\zeta|\zeta|\right],$$

which implies $\mathbb{E}\left[\zeta|\zeta|\right] = 0$.

(ii) By symmetry, (ζ, ζ') and $(-\zeta, -\zeta')$ are identically distributed, and hence

$$\mathbb{E}\left[\zeta|\zeta - \zeta'|\right] = \mathbb{E}\left[-\zeta|-\zeta + \zeta'|\right] = -\mathbb{E}\left[\zeta|\zeta - \zeta'|\right].$$

Therefore, we have $\mathbb{E}\left[\zeta|\zeta - \zeta'|\right] = 0$.

□

Example A.10. Consider the following two-point distribution:

$$\begin{cases} \eta(x) = a & \text{with probability } p, \\ \eta(x) = b & \text{with probability } 1-p, \end{cases}$$

where $0 \leq a < b \leq 1$ and $0 < p < 1$. Then, we have $\theta = \mathbb{E}\left[\eta(x)\right] = ap + b(1-p)$, and it can be shown that

$$\Delta_{\mathrm{BER}} = \Delta_{\mathrm{AUC}} = (b-a)^2 p(1-p) \cdot (1-2\theta)(2p-1).$$

Since the two discriminants coincide in this case, let us write $\Delta \coloneqq \Delta_{\mathrm{BER}} = \Delta_{\mathrm{AUC}}$ for ease of notation. Two immediate observations are:

(i) If $\theta = \frac{1}{2}$, we have $\Delta = 0$. This is consistent with Proposition A.8 (i).

(ii) If $p = \frac{1}{2}$, we have $\Delta = 0$ no matter how we choose a and b. This is the symmetric case in Proposition A.8 (ii).

In general, the sign of Δ coincides with that of

$$\frac{\Delta}{4(b-a)p(1-p)} = \left(\theta - \frac{1}{2}\right)\left(\theta - \frac{a+b}{2}\right)$$

since $4(b-a)p(1-p) > 0$.[9] Therefore, we have $\Delta < 0$ if

$$\min\left\{\frac{1}{2}, \frac{a+b}{2}\right\} < \theta < \max\left\{\frac{1}{2}, \frac{a+b}{2}\right\} \tag{6}$$

and $\Delta \geq 0$ otherwise. Some concrete examples are as follows:

(iii) For distributions with $a + b = 1$, including the Bernoulli case ($a = 0$, $b = 1$), we have $\Delta = 4(b-a)p(1-p) \cdot \left(\theta - \frac{1}{2}\right)^2 \geq 0$ regardless of p. In particular, $\Delta > 0$ if $\theta \neq \frac{1}{2}$.

(iv) To see that Δ can indeed be negative, let $a = 0.1$ and $b = 0.7$. Then, any $p \in (\frac{1}{3}, \frac{1}{2})$ satisfies (6) and hence gives $\Delta < 0$.

While Example A.10 (iii) is very simple, it has an interesting implication for a wider class of distributions; that is, it suggests that we have $\Delta_{\mathrm{BER}}, \Delta_{\mathrm{AUC}} \geq 0$ for any distribution with a sufficiently low label uncertainty (e.g., small Bayes error). The rationale is as follows. We have seen that $\Delta_{\mathrm{BER}} = \Delta_{\mathrm{AUC}} > 0$ holds for any Bernoulli distribution with mean $\theta \neq \frac{1}{2}$. Since Δ_{BER} and Δ_{AUC} are continuous as a functional of $\eta(x)$'s law (with respect to, e.g., the topology of weak convergence or Wasserstein distance), we can expect that they remain positive for any distribution sufficiently close to a Bernoulli distribution with mean θ. As Bernoulli distributions are the most "certain" distributions, being close to a Bernoulli distribution means that the distribution contains only a small amount of label uncertainty. Lemma A.11 formalizes this intuition.

[9] Note that $\frac{a+b}{2}$ is the class prior $\theta = ap + b(1-p)$ for $p = \frac{1}{2}$.

Lemma A.11. *We have* $\Delta_{\mathrm{BER}} \geq 0$ *and* $\Delta_{\mathrm{AUC}} \geq 0$ *for any distribution whose Bayes error* Err^* *and class prior* θ *satisfies*

$$\mathrm{Err}^* \leq \frac{1}{4} \min\{\theta, 1-\theta\} |1-2\theta|.$$

Proof. First, we prove the claim for Δ_{BER}. Let $h(x) = x|x|$ for $x \in [-\theta, 1-\theta]$. Then, we have $|h'(x)| = 2|x| \leq 2\max\{\theta, 1-\theta\}$, and hence h is $2\max\{\theta, 1-\theta\}$-Lipschitz continuous with respect to 1-norm. Therefore, for any $[0,1]$-valued random variable z, we have

$$\begin{aligned} |\mathbb{E}[h(\zeta)] - \mathbb{E}[h(z-\theta)]| &\leq 2\max\{\theta, 1-\theta\} \mathbb{E}[|\zeta - (z-\theta)|] \\ &= 2\max\{\theta, 1-\theta\} \mathbb{E}[|\eta(x) - z|]. \end{aligned} \tag{7}$$

Now, take z to be a random variable whose conditional distribution given $\eta(x)$ is a Bernoulli distribution with mean $\eta(x)$.[10] Then, z is a Bernoulli random variable with mean θ. Noting that $\mathbb{E}[|\eta(x) - z| \mid \eta(x)] = 2\eta(x)(1-\eta(x))$, we have

$$\begin{aligned} |\mathbb{E}[h(\zeta)] - \mathbb{E}[h(z-\theta)]| &\leq 4\max\{\theta, 1-\theta\} \mathbb{E}[\eta(x)(1-\eta(x))] \\ &\leq 4\max\{\theta, 1-\theta\} \mathbb{E}[\min\{\eta(x), 1-\eta(x)\}] \\ &= 4\max\{\theta, 1-\theta\} \mathrm{Err}^*. \end{aligned} \tag{8}$$

Observe that $(1-2\theta)\mathbb{E}[h(\zeta)] = \Delta_{\mathrm{BER}}$ and that $(1-2\theta)\mathbb{E}[h(z-\theta)] = \theta(1-\theta)|1-2\theta|^2$ (by Example A.10 (iii)). Hence, (8) implies

$$\begin{aligned} \Delta_{\mathrm{BER}} &\geq \theta(1-\theta)|1-2\theta|^2 - 4\max\{\theta, 1-\theta\}|1-2\theta|\mathrm{Err}^* \\ &= 4\max\{\theta, 1-\theta\}|1-2\theta| \left(\frac{1}{4}\min\{\theta, 1-\theta\}|1-2\theta| - \mathrm{Err}^* \right), \end{aligned}$$

which is non-negative if

$$\mathrm{Err}^* \leq \frac{1}{4} \min\{\theta, 1-\theta\} |1-2\theta|.$$

The proof for Δ_{AUC} is similar. Let $h(x,y) = \frac{1}{2}(x+y)|x-y|$ for $x, y \in [-\theta, 1-\theta]$. Then, we have $\|\nabla h(x,y)\|_\infty = \max\{|x|, |y|\} \leq \max\{\theta, 1-\theta\}$, and hence h is $\max\{\theta, 1-\theta\}$-Lipschitz continuous with respect to 1-norm. Therefore, for any $[0,1]$-valued random variable z and its i.i.d. copy z', we have

$$\begin{aligned} |\mathbb{E}[h(\zeta, \zeta')] - \mathbb{E}[h(z-\theta, z'-\theta)]| &\leq \max\{\theta, 1-\theta\} \mathbb{E}[|\zeta - (z-\theta)| + |\zeta' - (z'-\theta)|] \\ &= 2\max\{\theta, 1-\theta\} \mathbb{E}[|\eta(x) - z|]. \end{aligned} \tag{9}$$

Notice that the right-hand side of (9) exactly matches that of (7). Hence, the rest of the proof is almost identical. □

Remark A.12.

- It is possible to derive similar sufficient conditions in terms of the optimal BER or AUC instead of the Bayes error.
- This lemma implies that we can choose the constant δ_θ in Proposition A.8 (iii) to be

$$\delta_\theta = \begin{cases} \frac{1}{4}\theta(1-2\theta) & \text{if } \theta < \frac{1}{2}, \\ \infty & \text{if } \theta = \frac{1}{2}, \\ \frac{1}{4}(1-\theta)(2\theta-1) & \text{if } \theta > \frac{1}{2}, \end{cases}$$

  which might not be optimal.

[10] Let $u \sim \mathrm{Unif}([0,1])$ be independent of $\eta(x)$, and define $z := \mathbb{1}[u \leq \eta(x)]$. Then, $\mathbb{E}[z \mid \eta(x)] = \mathbb{E}[\mathbb{1}[u \leq \eta(x)] \mid \eta(x)] = \eta(x)$, hence $z \mid \eta(x) \sim \mathrm{Bernoulli}(\eta(x))$.

Algorithm A.2 An $O(n \log n)$-time algorithm for estimating AUC^* with the max formula

input Soft labels $\eta_1, \ldots, \eta_n$, class prior θ
  $\eta_{(1)}, \ldots, \eta_{(n)} \leftarrow$ sort $\eta_1, \ldots, \eta_n$ in ascending order
  $\mathrm{SUM} \leftarrow 0$
  $\mathrm{PREFIX} \leftarrow 0$
  for $j = 2$ **to** n **do**
    $\mathrm{PREFIX} \leftarrow \mathrm{PREFIX} + 1 - \eta_{(j-1)}$
    $\mathrm{SUM} \leftarrow \mathrm{SUM} + \eta_{(j)} \cdot \mathrm{PREFIX}$
  end for
output $\frac{\mathrm{SUM}}{n(n-1)\theta(1-\theta)}$

A.3 Near-linear time algorithm for computing the max formula $\widehat{\mathrm{AUC}}_2^*$

In Section 2.2, we presented an $O(n \log n)$-time algorithm for computing the min formula $\widehat{\mathrm{AUC}}_1^*$ or the optimal AUC. Here we provide the proof of the correctness and time complexity of Algorithm 2.1.

Theorem A.13. *Algorithm 2.1 computes $\widehat{\mathrm{AUC}}_1^*$ in $O(n \log n)$ time in the worst case, assuming any standard sorting algorithm with $O(n \log n)$ worst-case time complexity (e.g., Mergesort).*

Proof. First, we prove the correctness of Algorithm 2.1. Observe that $\min\{a(1-b), b(1-a)\} = \min\{a, b\}\,(1 - \max\{a, b\})$. Since $\eta_{(1)}, \ldots, \eta_{(n)}$ are sorted in ascending order, we have

$$\begin{aligned}\sum_{i<j} \min\left\{\eta_{(i)}(1-\eta_{(j)}), \eta_{(j)}(1-\eta_{(i)})\right\} &= \sum_{i<j} \eta_{(i)}(1-\eta_{(j)}) \\ &= \sum_{j=2}^{n} \left[(1-\eta_{(j)}) \sum_{i=1}^{j-1} \eta_{(i)}\right].\end{aligned}$$

At the j-th iteration of the for loop in Algorithm 2.1, the variable PREFIX stores the prefix sum $\sum_{i=1}^{j-1} \eta_{(i)}$. Therefore, Algorithm 2.1 correctly computes $\widehat{\mathrm{AUC}}_1^*$.

For the time complexity analysis, Algorithm 2.1 first sorts $\eta_1, \ldots, \eta_n$, which takes $O(n \log n)$ time by assumption. Then, the for loop iterates only $\Theta(n)$ times, and each iteration takes a constant time. Therefore, the total worse-case time complexity of Algorithm 2.1 is $O(n \log n)$. □

The max-based estimator $\widehat{\mathrm{AUC}}_2^*$ can also be computed in $O(n \log n)$ time by a similar algorithm, as shown in Algorithm A.2.

Theorem A.14. *Algorithm A.2 computes $\widehat{\mathrm{AUC}}_2^*$ in $O(n \log n)$ time in the worst case, assuming any standard comparison-based sorting algorithm with $O(n \log n)$ worst-case time complexity (e.g., Mergesort).*

Proof. We prove the correctness of Algorithm A.2. Observe that $\max\{a(1-b), b(1-a)\} = \max\{a, b\}\,(1 - \min\{a, b\})$. Since $\eta_{(1)}, \ldots, \eta_{(n)}$ are sorted in ascending order, we have

$$\begin{aligned}n(n-1)\theta(1-\theta)\widehat{\mathrm{AUC}}_2^* &= \sum_{i<j} \max\{\eta_i(1-\eta_j), \eta_j(1-\eta_i)\} \\ &= \sum_{i<j} \max\left\{\eta_{(i)}(1-\eta_{(j)}), \eta_{(j)}(1-\eta_{(i)})\right\} \\ &= \sum_{i<j} \eta_{(j)}(1-\eta_{(i)}) \\ &= \sum_{j=2}^{n} \sum_{i=1}^{j-1} \eta_{(j)}(1-\eta_{(i)}) \\ &= \sum_{j=2}^{n} \left[\eta_{(j)} \sum_{i=1}^{j-1} (1-\eta_{(i)})\right]\end{aligned}$$

At the j-th iteration of the for loop in Algorithm A.2, the variable PREFIX stores the prefix sum $\sum_{i=1}^{j-1} \eta_{(i)}$. Therefore, Algorithm A.2 correctly computes $\widehat{\mathrm{AUC}^*}$.

For the time complexity analysis, Algorithm A.2 first sorts $\eta_1, \ldots, \eta_n$, which takes $O(n \log n)$ time by assumption. Then, the for loop iterates $\Theta(n)$ times, and each iteration takes a constant time. Therefore, the total worse-case time complexity of Algorithm A.2 is $O(n \log n)$. □

Remark A.15. The algorithm for computing the discriminant estimator $\widehat{\Delta}_{\mathrm{AUC}}$ in $O(n \log n)$ time is shown in Algorithm A.1.

A.4 Estimation error bounds for estimators based on clean soft labels

A.4.1 Preliminaries: Hoeffding's and Bernstein's inequalities for U-statistics

U-statistics are a broad class of statistics introduced by Hoeffding (1948). Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a function that is symmetric, i.e., invariant to permutations of its d arguments. Given $n \geq d$ i.i.d. observations $Z_1, \ldots, Z_n$, consider the estimator

$$U_n \coloneqq \binom{n}{d}^{-1} \sum_{1 \leq i_1 < \ldots < i_d \leq n} \phi(Z_{i_1}, \ldots, Z_{i_d}). \tag{10}$$

U_n is called a U-statistic with a kernel ϕ of degree d.

For $d = 1$, U_n is simply an average of i.i.d. random variables, so the standard concentration inequalities such as the famous Hoeffding's inequality can be used to analyze its deviation from the mean. However, for $d \geq 2$, the summands in (10) are no longer independent. Nonetheless, in the same paper as the one that introduced his famous concentration inequality for sums of independent random variables, Hoeffding (1963) established a similar concentration bound for U-statistics: if the kernel ϕ is bounded in an interval $[a, b]$, then for any $\varepsilon > 0$,

$$\mathbb{P}\left(|U_n - \mathbb{E}\left[U_n\right]| > \varepsilon\right) \leq 2 \exp\left(-\frac{2(n/d)\varepsilon^2}{(b-a)^2}\right). \tag{11}$$

Or equivalently, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that

$$|U_n - \mathbb{E}\left[U_n\right]| \leq (b-a)\sqrt{\frac{\log(2/\delta)}{2n/d}}. \tag{12}$$

One thing to note is that, although we can also apply McDiarmid's inequality (also known as the bounded difference inequality) to obtain a similar concentration bound, Hoeffding's bound (12) is tighter by a multiplicative factor of $\sqrt{d}$.

Hoeffding (1963) also provided a Bernstein-type concentration bound for U-statistics, which takes into account the variance of the kernel: if the kernel ϕ is bounded as $|\phi(Z_1, \ldots, Z_d) - \mathbb{E}\left[\phi(Z_1, \ldots, Z_d)\right]| \leq c$, for any $\varepsilon > 0$, we have

$$\mathbb{P}\left(|U_n - \mathbb{E}\left[U_n\right]| > \varepsilon\right) \leq 2 \exp\left(-\frac{(n/d)\varepsilon^2}{2\left(\sigma^2 + c\varepsilon/3\right)}\right), \tag{13}$$

where $\sigma^2 \coloneqq \mathrm{Var}\left[\phi(Z_1, \ldots, Z_d)\right]$ is the variance of the kernel. It also can be rewritten as a tail bound: for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds that

$$|U_n - \mathbb{E}\left[U_n\right]| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n/d}} + \frac{2c \log(2/\delta)}{3(n/d)}. \tag{14}$$

Bernstein-type bounds (13), (14) are often tighter than Hoeffding-type bounds (11), (12) when the standard deviation σ is much smaller than the width $b - a$ of the kernel's range.

A.4.2 Proofs of the estimation error bounds

In the simplest setting of this paper where clean soft labels $\{\eta_i\}_{i=1}^n$ are available and the class prior θ is known, our estimators $\widehat{\mathrm{BER}^*}$ and $\widehat{\mathrm{AUC}^*}$ can be seen as U-statistics of degrees $d = 1, 2,$

respectively, with bounded kernels. Thus, the above concentration inequalities can be applied to obtain the bounds (Theorems 2.5 and 2.6) on the estimation errors of these estimators.

The bound for $\widehat{\mathrm{BER}^*}$ is a straightforward application of the Hoeffding bound (12), so we focus on the $\widehat{\mathrm{AUC}^*}$ case in what follows. The kernel

$$\phi^{(1)}_{\mathrm{AUC}}(z_1, z_2) = 1 - \frac{1}{2\theta(1-\theta)} \min\{z_1(1-z_2), z_2(1-z_1)\}$$

takes values between $a = 1 - \frac{1}{8\theta(1-\theta)}$ and $b = 1$, so we can use the Hoeffding bound (12) to obtain the following result:

$$\left|\widehat{\mathrm{AUC}^*} - \mathrm{AUC}^*\right| \leq \frac{1}{8\theta(1-\theta)} \sqrt{\frac{\log(2/\delta)}{n}} \quad \text{with probability at least } 1-\delta.$$

The θ-dependence of the right-hand side comes from the fact that, unlike in the case of $\widehat{\mathrm{BER}^*}$, the range $b - a = \frac{1}{8\theta(1-\theta)}$ of the kernel $\phi^{(1)}$ depends on the class prior θ. Under severe class imbalance where $\theta \to 0$ (or 1), the factor $\frac{1}{8\theta(1-\theta)}$ blows up at the rate of $O(1/\theta)$ (or $O(1/(1-\theta))$).

We can mitigate this issue and slow down the blow-up rate to $O(1/\sqrt{\theta})$ by using the Bernstein bound (14) instead of the Hoeffding bound. The resulting bound is stated in Theorem 2.6.

Lemma A.16. *The variance of the kernel $\phi^{(1)}_{\mathrm{AUC}}$ is bounded as*

$$\mathrm{Var}\left[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right] \leq \frac{1}{16\theta(1-\theta)}.$$

Proof. Let $m \coloneqq \min\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\}$. Then, we have $0 \leq m \leq \frac{1}{4}$ and $\mathbb{E}[m] = 2\theta(1-\theta)(1-\mathrm{AUC}^*) \leq \theta(1-\theta)$. So it follows that

$$\mathrm{Var}[m] \leq \mathbb{E}\left[m^2\right] \leq \frac{1}{4}\mathbb{E}[m] \leq \frac{1}{4}\theta(1-\theta).$$

Therefore, we have

$$\mathrm{Var}\left[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right] = \frac{1}{4\theta^2(1-\theta)^2}\mathrm{Var}[m] \leq \frac{1}{16\theta(1-\theta)}.$$

□

Lemma A.17. *The kernel $\phi^{(1)}_{\mathrm{AUC}}$ is bounded as*

$$\left|\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x')) - \mathbb{E}\left[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right]\right| \leq \frac{1}{8\theta(1-\theta)}.$$

Proof. Continuing from the proof of Lemma A.16, $m = \min\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\}$ is bounded as $0 \leq m \leq \frac{1}{4}$, and its mean satisfies $\mathbb{E}[m] = 2\theta(1-\theta)(1-\mathrm{AUC}^*) \in [0, \theta(1-\theta)]$. Thus, we have $m - \mathbb{E}[m] \leq \frac{1}{4}$ and $\mathbb{E}[m] - m \leq \theta(1-\theta) \leq \frac{1}{4}$, which implies

$$\left|\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x')) - \mathbb{E}\left[\phi^{(1)}_{\mathrm{AUC}}(\eta(x), \eta(x'))\right]\right| = \frac{1}{2\theta(1-\theta)}|m - \mathbb{E}[m]| \leq \frac{1}{8\theta(1-\theta)}.$$

□

Lemma A.18. *For any $\delta \in (0,1)$ and $n \geq \frac{\log(2/\delta)}{9\theta(1-\theta)}$, with probability at least $1-\delta$, it holds that*

$$\left|\widetilde{\mathrm{AUC}^*} - \mathrm{AUC}^*\right| \leq \left|\widehat{\mathrm{AUC}^*} - \mathrm{AUC}^*\right| \leq \sqrt{\frac{\log(2/\delta)}{\theta(1-\theta)n}}.$$

Proof. With Lemmas A.16 and A.17 proven above, we are now ready to apply the Bernstein bound (14) to the estimator $\widehat{\mathrm{AUC}^*}$. As a result, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\left|\widehat{\mathrm{AUC}^*} - \mathrm{AUC}^*\right| \leq \sqrt{\frac{\log(2/\delta)}{4\theta(1-\theta)n}} + \frac{\log(2/\delta)}{6\theta(1-\theta)n}.$$

Furthermore, the first term dominates the second term since $n \geq \frac{\log(2/\delta)}{9\theta(1-\theta)}$, so the above bound can be simplified as

$$\left|\widehat{\mathrm{AUC}^*} - \mathrm{AUC}^*\right| \leq \sqrt{\frac{\log(2/\delta)}{\theta(1-\theta)n}}.$$

Finally, the claim follows as we have $\left|\widetilde{\mathrm{AUC}^*} - \mathrm{AUC}^*\right| \leq \left|\widehat{\mathrm{AUC}^*} - \mathrm{AUC}^*\right|$ by the definition of the clipped estimator $\widetilde{\mathrm{AUC}^*}$. □

Theorem 2.6. *For any $\delta > 0$, with probability at least $1-\delta$, we have*

$$\left|\widetilde{\mathrm{AUC}^*} - \mathrm{AUC}^*\right| \leq \sqrt{\frac{\log(2/\delta)}{\theta(1-\theta)n}}.$$

Proof. If $n < \frac{4\log(2/\delta)}{\theta(1-\theta)}$, then the claim trivially holds because (RHS) $= \sqrt{\frac{\log(2/\delta)}{\theta(1-\theta)n}} > \frac{1}{2}$ and the left-hand side is at most $\frac{1}{2}$. If $n \geq \frac{4\log(2/\delta)}{\theta(1-\theta)}$, then the claim follows from Lemma A.18. □

A.5 Unknown class prior

A.5.1 Failure of the estimator by Jeong et al. (2024)

Jeong et al. (2024) discussed the estimation of the FPR and FNR of the Bayes optimal classifier from soft labels, and they proposed to estimate the class prior by $\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left[\eta_i \geq 0.5\right]$. However, this estimator relies on the naive assumption that $\mathbb{P}\left(y=1\right) = \mathbb{P}\left(\eta(x) \geq 0.5\right)$. It is easy to construct examples where this assumption does not hold.

Example A.19. Consider a simple distribution over $\mathcal{X} \times \mathcal{Y} = \{x_0\} \times \{0,1\}$ where $\mathbb{P}\left(y=0\right) = \mathbb{P}\left(y=1\right) = 0.5$. Then, $\mathbb{P}\left(y=1\right) = 0.5$ and $\mathbb{P}\left(\eta(x) \geq 0.5\right) = \mathbb{P}\left(0.5 \geq 0.5\right) = 1$ does not match. In this case, the estimator by Jeong et al. (2024) always returns 1 however large the sample size is, while the true class prior is 0.5. Therefore, their estimator is biased and statistically inconsistent.

Also, it is trivial to extend this example to allow $\mathcal{X}$ to have multiple (or even infinitely many) elements with various $\eta(x)$ values.

A.5.2 Clipped estimator of the class prior and plug-in estimators

Here, we first show that using Hoeffding's inequality to bound $|\widehat{\theta}_{\tau_n} - \theta|$ (Lemma A.20) results in a estimation error bound for the BER estimator of order $O_p(\frac{1}{\theta(1-\theta)\sqrt{n}})$ (Theorem A.21). Then, by employing Bernstein's inequality instead of Hoeffding's inequality (Lemma 2.9), we improve the θ-dependence to $O_p(\frac{1}{\sqrt{\theta(1-\theta)n}})$ (Theorem A.22). After that, a similar result for the AUC estimator (Theorem A.23) is presented. Finally, we prove the lemmas used in the proofs of these theorems.

Lemma A.20. *Under Assumption 2.8, for any $\delta \in (0,1)$ and $n \geq 1$, with probability at least $1-\delta$, we have $|\widehat{\theta}_{\tau_n} - \theta| \leq \sqrt{\frac{2\log(2/\delta)}{n}}$.*

Proof. We decompose the estimation error as

$$|\widehat{\theta}_{\tau_n} - \theta| \leq |\mathrm{clip}_{\tau_n}^{1-\tau_n}(\bar{\eta}) - \bar{\eta}| + |\bar{\eta} - \theta|.$$

By Hoeffding's inequality, the second term is bounded as

$$|\bar{\eta} - \theta| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability at least $1-\delta$. The first term is the error due to clipping, which is at most $\tau_n = \frac{c}{n}$. For all $n \geq n_0 \coloneqq \lceil \frac{2c^2}{\log(\frac{2}{\delta})} \rceil$, we have $\frac{c}{n} \leq \sqrt{\frac{\log(2/\delta)}{2n}}$ and hence

$$|\widehat{\theta}_{\tau_n} - \theta| \leq 2\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Finally, the proof is completed by noting that the above inequality holds for all $n \geq 1$ because $n_0 = 1$ for any $\delta \in (0,1)$ and $c \in (0, \frac{1}{2})$. □

Theorem A.21. *For any $\varepsilon > 0$, it holds that*

$$\mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| > \varepsilon\right) \leq 4\exp\left(-\frac{n\theta^2(1-\theta)^2\varepsilon^2}{8}\right).$$

Or equivalently, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \leq \frac{1}{\theta(1-\theta)}\sqrt{\frac{8\log(4/\delta)}{n}}.$$

Proof. We have

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| > \varepsilon\right) \\
\leq\ &\mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\theta) - \mathrm{BER}^*\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \widehat{\mathrm{BER}^*}(\theta)\right| > \frac{\varepsilon}{2}\right).
\end{aligned}$$

The first term is bounded by Hoeffding's inequality (11) as

$$\mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\theta) - \mathrm{BER}^*\right| > \frac{\varepsilon}{2}\right) \leq 2\exp\left(-2n\varepsilon^2\right).$$

As for the second term, we use Lemma A.26 to obtain

$$\mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \widehat{\mathrm{BER}^*}(\theta)\right| > \frac{\varepsilon}{2}\right) \leq \mathbb{P}\left(\frac{1}{\theta(1-\theta)}\left|\widehat{\theta}_{\tau_n} - \theta\right| > \frac{\varepsilon}{2}\right),$$

which can be further bounded by Lemma A.20 as

$$\begin{aligned}
\mathbb{P}\left(\left|\widehat{\theta}_{\tau_n} - \theta\right| > \frac{\theta(1-\theta)}{2}\varepsilon\right) &\leq 2\exp\left(-\frac{n\left(\frac{\theta(1-\theta)}{2}\varepsilon\right)^2}{2}\right) \\
&= 2\exp\left(-\frac{n\theta^2(1-\theta)^2\varepsilon^2}{8}\right).
\end{aligned}$$

Since $0 < \theta(1-\theta) \leq \frac{1}{4}$, we have $\exp(-2n\varepsilon^2) \leq \exp\left(-\frac{n\theta^2(1-\theta)^2\varepsilon^2}{8}\right)$, and thus

$$\mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| > \varepsilon\right) \leq 4\exp\left(-\frac{n\theta^2(1-\theta)^2\varepsilon^2}{8}\right).$$

□

The following lemma is similar to Lemma A.20, but it gives a tighter bound especially under severe class imbalance, i.e., when θ is close to 0 or 1.

Lemma 2.9. *Under Assumption 2.8, for any $\delta > 0$ and $n \geq \frac{(2\log(2/\delta)+3c)^2}{18\theta(1-\theta)\log(2/\delta)}$, with probability at least $1-\delta$, we have $|\widehat{\theta}_{\tau_n} - \theta| \leq \sqrt{\frac{8\theta(1-\theta)\log(2/\delta)}{n}}$.*

Proof. The proof is similar to that of Lemma A.20, but we use Bernstein's inequality instead of Hoeffding's inequality. Since η is a $[0,1]$-valued random variable with mean θ, its variance is at

most that of a Bernoulli random variable with the same mean, that is, $\theta(1-\theta)$. Also, $|\eta - \theta| \leq \max\{\theta, 1-\theta\} \leq 1$. Hence by Bernstein's inequality, with probability at least $1-\delta$, we have

$$|\bar{\eta} - \theta| \leq \sqrt{\frac{2\theta(1-\theta)\log(2/\delta)}{n}} + \frac{2\log(2/\delta)}{3n}.$$

and thus

$$\begin{aligned}
|\widehat{\theta}_{\tau_n} - \theta| &\leq |\mathrm{clip}_{\tau_n}^{1-\tau_n}(\bar{\eta}) - \bar{\eta}| + |\bar{\eta} - \theta| \\
&\leq \frac{c}{n} + |\bar{\eta} - \theta| \\
&\leq \sqrt{\frac{2\theta(1-\theta)\log(2/\delta)}{n}} + \frac{2\log(2/\delta) + 3c}{3n}.
\end{aligned}$$

If $n \geq \frac{(2\log(2/\delta)+3c)^2}{18\theta(1-\theta)\log(2/\delta)}$, the first term dominates the second term, and hence the above inequality is simplified as

$$|\widehat{\theta}_{\tau_n} - \theta| \leq 2\sqrt{\frac{2\theta(1-\theta)\log(2/\delta)}{n}}.$$

□

Theorem A.22. *Under Assumption 2.8, for any $\delta > 0$, the following holds with probability at least $1-\delta$:*

$$\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \leq \frac{17}{8}\sqrt{\frac{2\log(4/\delta)}{\theta(1-\theta)n}}.$$

Proof. Fix an arbitrary $\delta \in (0,1)$. If $n < \frac{4\log(4/\delta)}{\theta(1-\theta)}$, then the claim trivially holds because (RHS) $> \sqrt{\frac{\log(4/\delta)}{\theta(1-\theta)n}} > \frac{1}{2}$ and the left-hand side is at most $\frac{1}{2}$. Therefore, we assume $n \geq \frac{4\log(4/\delta)}{\theta(1-\theta)}$ from now on.

By Theorem 2.5, we have the following with probability at least $1 - \frac{\delta}{2}$:

$$\left|\widehat{\mathrm{BER}^*}(\theta) - \mathrm{BER}^*\right| \leq \sqrt{\frac{\log(4/\delta)}{8n}}.$$

Next, observe

$$\begin{aligned}
\frac{(2\log(4/\delta)+3c)^2}{18\theta(1-\theta)\log(4/\delta)} &\leq \frac{\left(\left(2 + \frac{3c}{2\log 2}\right)\log(4/\delta)\right)^2}{18\theta(1-\theta)\log(4/\delta)} \\
&= \frac{\left(2 + \frac{3c}{2\log 2}\right)^2}{18} \cdot \frac{\log(4/\delta)}{\theta(1-\theta)} \\
&< 0.53\frac{\log(4/\delta)}{\theta(1-\theta)} \\
&\leq n.
\end{aligned}$$

It follows that with probability at least $1 - \frac{\delta}{2}$, the gap between the plug-in estimator and the estimator with the true class prior can be bounded by Lemmas A.26 and 2.9 as

$$\begin{aligned}
\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \widehat{\mathrm{BER}^*}(\theta)\right| &\leq \frac{1}{\theta(1-\theta)}\left|\widehat{\theta}_{\tau_n} - \theta\right| \\
&\leq \frac{1}{\theta(1-\theta)}\sqrt{\frac{8\theta(1-\theta)\log(4/\delta)}{n}} \\
&= \sqrt{\frac{8\log(4/\delta)}{\theta(1-\theta)n}}.
\end{aligned}$$

Combining these two inequalities by union bound, with probability at least $1-\delta$, we have

$$\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \leq \sqrt{\frac{\log(4/\delta)}{8n}} + \sqrt{\frac{8\log(4/\delta)}{\theta(1-\theta)n}}.$$

We can further use $0 < \theta(1-\theta) \leq \frac{1}{4}$ to simplify the right-hand side as

$$\sqrt{\frac{\log(4/\delta)}{8n}} + \sqrt{\frac{8\log(4/\delta)}{\theta(1-\theta)n}} \leq \frac{1}{8}\sqrt{\frac{2\log(4/\delta)}{\theta(1-\theta)n}} + 2\sqrt{\frac{2\log(4/\delta)}{\theta(1-\theta)n}}$$
$$= \frac{17}{8}\sqrt{\frac{2\log(4/\delta)}{\theta(1-\theta)n}},$$

which completes the proof. □

Theorem A.23. *Under Assumption 2.8, for any $\delta > 0$, the following holds with probability at least $1-\delta$:*

$$\left|\widetilde{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \leq 10\sqrt{\frac{\log(4/\delta)}{\theta(1-\theta)n}}.$$

Proof. First, if $n < \frac{18\log(4/\delta)}{\theta(1-\theta)}$, then the claim follows trivially because (RHS) $= 10\sqrt{\frac{\log(4/\delta)}{\theta(1-\theta)n}} > \frac{5\sqrt{2}}{3} > \frac{1}{2}$ and the left-hand side is at most $\frac{1}{2}$. Therefore, we assume $n \geq \frac{18\log(4/\delta)}{\theta(1-\theta)}$ in the following.

By Theorem 2.6, with probability at least $1 - \frac{\delta}{2}$, we have

$$\left|\widehat{\mathrm{AUC}^*}(\theta) - \mathrm{AUC}^*\right| \leq \sqrt{\frac{\log(4/\delta)}{\theta(1-\theta)n}} \leq \frac{1}{2}.$$

In this event, we have

$$\left|\widehat{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n}) - \widehat{\mathrm{AUC}^*}(\theta)\right| \leq \frac{1/2+1/2}{\widehat{\theta}_{\tau_n}(1-\widehat{\theta}_{\tau_n})}|\widehat{\theta}_{\tau_n} - \theta| = \frac{1}{\widehat{\theta}_{\tau_n}(1-\widehat{\theta}_{\tau_n})}|\widehat{\theta}_{\tau_n} - \theta|.$$

by Lemma A.28. Also, by Lemma A.30, with probability at least $1 - \frac{\delta}{2}$, it holds that

$$\frac{1}{\widehat{\theta}_{\tau_n}(1-\widehat{\theta}_{\tau_n})}|\widehat{\theta}_{\tau_n} - \theta| \leq 6\sqrt{\frac{2\log(4/\delta)}{\theta(1-\theta)n}}.$$

Therefore, by union bound, with probability at least $1-\delta$, we have

$$\begin{aligned}
\left|\widetilde{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| &\leq \left|\widehat{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \\
&\leq \left|\widehat{\mathrm{AUC}^*}(\theta) - \mathrm{AUC}^*\right| + \left|\widehat{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n}) - \widehat{\mathrm{AUC}^*}(\theta)\right| \\
&\leq \sqrt{\frac{\log(4/\delta)}{\theta(1-\theta)n}} + 6\sqrt{\frac{2\log(4/\delta)}{\theta(1-\theta)n}} \\
&< 10\sqrt{\frac{\log(4/\delta)}{\theta(1-\theta)n}}.
\end{aligned}$$

□

Proposition A.24. *The mean squared error of $\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n})$ is*

$$\mathbb{E}\left[\left(\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right)^2\right] \leq \frac{C}{\theta(1-\theta)n}, \tag{15}$$

where $C > 0$ is a constant that does not depend on θ or n. Therefore $\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n})$ is asymptotically unbiased.

Proof. Take any $\varepsilon > 0$ and let $\delta \coloneqq 4\exp\left(-K\theta(1-\theta)n\varepsilon^2\right)$, where $K = \frac{8^2}{2\cdot 17^2}$ is a constant. By applying Theorem A.22 with the above δ, we have

$$\mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| > \varepsilon\right) \leq 4\exp\left(-K\theta(1-\theta)n\varepsilon^2\right).$$

Integrating the above tail bound over all $\varepsilon > 0$ gives

$$\begin{aligned}
\mathbb{E}\left[\left(\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right)^2\right] &= \int_0^\infty \mathbb{P}\left(\left(\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right)^2 > \varepsilon\right) \mathrm{d}\varepsilon \\
&= \int_0^\infty \mathbb{P}\left(\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right| > \sqrt{\varepsilon}\right) \mathrm{d}\varepsilon \\
&\leq \int_0^\infty 4\exp\left(-K\theta(1-\theta)n\varepsilon\right) \mathrm{d}\varepsilon \\
&= \frac{4/K}{\theta(1-\theta)n},
\end{aligned}$$

which proves (15) with $C = \frac{4}{K}$.

Asymptotic unbiasedness follows from

$$\begin{aligned}
\left|\mathbb{E}\left[\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n})\right] - \mathrm{BER}^*\right| &\leq \mathbb{E}\left[\left|\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right|\right] \\
&\leq \sqrt{\mathbb{E}\left[\left(\widehat{\mathrm{BER}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{BER}^*\right)^2\right]} \\
&\leq \sqrt{\frac{C}{\theta(1-\theta)n}} \\
&\xrightarrow{n\to\infty} 0.
\end{aligned}$$

□

Proposition A.25. *The mean squared error of* $\widehat{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n})$ *is*

$$\mathbb{E}\left[\left(\widehat{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n}) - \mathrm{AUC}^*\right)^2\right] \leq \frac{C}{\theta(1-\theta)n},$$

where $C > 0$ *is a constant that does not depend on* θ *or* n*. Therefore* $\widehat{\mathrm{AUC}^*}(\widehat{\theta}_{\tau_n})$ *is asymptotically unbiased.*

Proof. We omit the proof as it is almost the same as the one of Proposition A.24. □

In the following, we prove the lemmas used in the proof of Theorems A.22 and A.23.

Lemma A.26. *For any* $\theta' \in (0,1)$*, we have*

$$\left|\widehat{\mathrm{BER}^*}(\theta') - \widehat{\mathrm{BER}^*}(\theta)\right| \leq \frac{1}{\theta(1-\theta)}\left|\theta' - \theta\right|.$$

Proof. Fix an arbitrary $\eta \in [0,1]$ and let

$$N(\theta) \coloneqq \min\left\{\eta(1-\theta), \theta(1-\eta)\right\}, \quad D(\theta) \coloneqq \theta(1-\theta).$$

Then, we have

$$\min\left\{\frac{\eta}{\theta}, \frac{1-\eta}{1-\theta}\right\} = \frac{N(\theta)}{D(\theta)}.$$

Observe that any $\theta' \in (0,1)$ satisfies

$$\left|\frac{N(\theta')}{D(\theta')} - \frac{N(\theta)}{D(\theta)}\right| \leq \frac{1}{D(\theta)D(\theta')}\left(D(\theta')\left|N(\theta') - N(\theta)\right| + N(\theta')\left|D(\theta') - D(\theta)\right|\right)$$

and that

$$|N(\theta') - N(\theta)| \le |\theta' - \theta|, \quad |D(\theta') - D(\theta)| \le |\theta' - \theta|.$$

Therefore, we have

$$\left|\frac{N(\theta')}{D(\theta')} - \frac{N(\theta)}{D(\theta)}\right| \le \frac{D(\theta') + N(\theta')}{D(\theta)D(\theta')} |\theta' - \theta| \le \frac{2}{D(\theta)} |\theta' - \theta|,$$

where we used the fact that $N(\theta') \le D(\theta')$ for the last inequality.

Using the above inequality, we obtain

$$\begin{aligned}
\left|\widehat{\mathrm{BER}^*}(\theta') - \widehat{\mathrm{BER}^*}(\theta)\right| &\le \frac{1}{2n}\sum_{i=1}^{n}\left|\min\left\{\frac{\eta_i}{\theta'}, \frac{1-\eta_i}{1-\theta'}\right\} - \min\left\{\frac{\eta_i}{\theta}, \frac{1-\eta_i}{1-\theta}\right\}\right| \\
&\le \frac{1}{2n}\sum_{i=1}^{n}\frac{2}{\theta(1-\theta)}|\theta' - \theta| \\
&= \frac{1}{\theta(1-\theta)}|\theta' - \theta|.
\end{aligned}$$

□

Remark A.27. In Lemma A.26, we could have simply used the fact that $\theta \mapsto \min\left\{\frac{\eta}{\theta}, \frac{1-\eta}{1-\theta}\right\}$ is $\frac{1}{\min\{\theta,1-\theta\}^2}$-Lipschitz to obtain $\left|\widehat{\mathrm{BER}^*}(\theta') - \widehat{\mathrm{BER}^*}(\theta)\right| \le \frac{1}{\min\{\theta,1-\theta\}^2}|\theta' - \theta|$, but it would have given a worse θ-dependence.

Lemma A.28. *For any $\theta' \in (0,1)$, it holds that*

$$\left|\widehat{\mathrm{AUC}^*}(\theta') - \widehat{\mathrm{AUC}^*}(\theta)\right| \le \frac{1/2 + \left|\widehat{\mathrm{AUC}^*}(\theta) - \mathrm{AUC}^*\right|}{\theta'(1-\theta')}|\theta' - \theta|.$$

Proof. We first show that

$$\left|\widehat{\mathrm{AUC}^*}(\theta') - \widehat{\mathrm{AUC}^*}(\theta)\right| \le \frac{\left|1 - \widehat{\mathrm{AUC}^*}(\theta)\right|}{\theta'(1-\theta')}|\theta' - \theta|. \tag{16}$$

If $\widehat{\mathrm{AUC}^*}(\theta) = 1$, (16) trivially holds because we also have $\widehat{\mathrm{AUC}^*}(\theta') = 1$. If $\widehat{\mathrm{AUC}^*}(\theta) < 1$, we have

$$\begin{aligned}
\left|\frac{1 - \widehat{\mathrm{AUC}^*}(\theta')}{1 - \widehat{\mathrm{AUC}^*}(\theta)} - 1\right| &= \left|\frac{\theta(1-\theta)}{\theta'(1-\theta')} - 1\right| \\
&= \left|\frac{\theta'(1-\theta') - \theta(1-\theta)}{\theta'(1-\theta')}\right| \\
&\le \frac{|\theta' - \theta|}{\theta'(1-\theta')},
\end{aligned}$$

where the last inequality follows from the fact that $x \in (0,1) \mapsto x(1-x)$ is 1-Lipschitz. Now, (16) follows by multiplying both sides by $|1 - \widehat{\mathrm{AUC}^*}(\theta)|$.

Next, observe that the AUC of the trivial scoring function $g(x) \equiv 0$ is

$$\mathrm{AUC}(g) = \mathbb{E}\left[\mathbb{1}[0 > 0] + \frac{1}{2}\mathbb{1}[0 = 0]\right] = \frac{1}{2},$$

and hence the maximum AUC over all scoring functions satisfies $\mathrm{AUC}^* \in [\frac{1}{2}, 1]$. Using this fact, we have

$$\left|1 - \widehat{\mathrm{AUC}^*}(\theta)\right| \le |1 - \mathrm{AUC}^*| + \left|\widehat{\mathrm{AUC}^*}(\theta) - \mathrm{AUC}^*\right| \le \frac{1}{2} + \left|\widehat{\mathrm{AUC}^*}(\theta) - \mathrm{AUC}^*\right|.$$

Combining this with (16), we obtain the desired result. □

Remark A.29. In Lemma A.28, we could have argued more simply by using $\left|\frac{1}{\theta'(1-\theta')} - \frac{1}{\theta(1-\theta)}\right| \leq \frac{1}{\theta'(1-\theta')\theta(1-\theta)}\,|\theta' - \theta|$, but it would have given a worse θ-dependence in the final result.

Lemma A.30. *Suppose Assumption 2.8 holds. Then, for any $\delta \in (0,1)$ and $n \geq \frac{18\log(2/\delta)}{\theta(1-\theta)}$, with probability at least $1-\delta$, it holds that*

$$\frac{1}{\widehat{\theta}_{\tau_n}(1-\widehat{\theta}_{\tau_n})}|\widehat{\theta}_{\tau_n} - \theta| \leq 6\sqrt{\frac{2\log(2/\delta)}{\theta(1-\theta)n}}.$$

Proof. Some calculation shows that any $c \in (0, \frac{1}{2})$, $\delta \in (0,1)$ and $\theta \in (0,1)$ satisfy $\frac{18\log(2/\delta)}{\theta(1-\theta)} \geq \frac{(2\log(2/\delta)+3c)^2}{18\theta(1-\theta)\log(2/\delta)}$. Therefore, by Lemma 2.9, with probability at least $1-\delta$, we have

$$|\widehat{\theta}_{\tau_n} - \theta| \leq \sqrt{\frac{8\theta(1-\theta)\log(2/\delta)}{n}}. \tag{17}$$

Since $n \geq \frac{18\log(2/\delta)}{\theta(1-\theta)}$, it follows that

$$|\widehat{\theta}_{\tau_n} - \theta| \leq \sqrt{8\theta(1-\theta)\log(2/\delta)} \cdot \sqrt{\frac{\theta(1-\theta)}{18\log(2/\delta)}} \leq \frac{2\theta(1-\theta)}{3},$$

which implies

$$\widehat{\theta}_{\tau_n}(1-\widehat{\theta}_{\tau_n}) \geq \theta(1-\theta) - |\widehat{\theta}_{\tau_n} - \theta| \geq \frac{1}{3}\theta(1-\theta).$$

Therefore, again by (17), we have

$$\frac{1}{\widehat{\theta}_{\tau_n}(1-\widehat{\theta}_{\tau_n})}|\widehat{\theta}_{\tau_n} - \theta| \leq \frac{3}{\theta(1-\theta)}\sqrt{\frac{8\theta(1-\theta)\log(2/\delta)}{n}} = 6\sqrt{\frac{2\log(2/\delta)}{\theta(1-\theta)n}}.$$

□

B Supplementary for Section 3

Lemma B.1. *Under Assumption 2.8, for any $\delta \in (0,1)$ and $n \geq \frac{(2\log(2/\delta)+3c)^2}{18\theta(1-\theta)\log(2/\delta)}$, with probability at least $1-\delta$, we have $|\tilde{\theta}_{\tau_n} - \theta| \leq \sqrt{\frac{8\theta(1-\theta)\log(2/\delta)}{n}}$.*

Proof. The proof is almost identical to that of Lemma 2.9. □

Lemma B.2. *For any $\boldsymbol{z} = (z_1, \ldots, z_n) \in [0,1]^n$ and $\boldsymbol{w} = (w_1, \ldots, w_n) \in [0,1]^n$, we have*

$$\left|\widehat{\mathrm{BER}^*}(\boldsymbol{w}) - \widehat{\mathrm{BER}^*}(\boldsymbol{z})\right| \leq \frac{1}{2\min\{\theta, 1-\theta\}}\sqrt{\frac{1}{n}\sum_{i=1}^n (w_i - z_i)^2}, \tag{18}$$

$$\left|\widehat{\mathrm{AUC}^*}(\boldsymbol{w}) - \widehat{\mathrm{AUC}^*}(\boldsymbol{z})\right| \leq \frac{1}{\theta(1-\theta)}\sqrt{\frac{1}{n}\sum_{i=1}^n (w_i - z_i)^2}. \tag{19}$$

Proof. To prove (18), observe that the function $z \in [0,1] \mapsto \min\left\{\frac{z}{\theta}, \frac{1-z}{1-\theta}\right\}$ is $(\min\{\theta, 1-\theta\})^{-1}$-Lipschitz. Hence, we have

$$\begin{aligned}
\left|\widehat{\mathrm{BER}^*}(\boldsymbol{w}) - \widehat{\mathrm{BER}^*}(\boldsymbol{z})\right| &\leq \frac{1}{2n}\sum_{i=1}^n \left|\min\left\{\frac{w_i}{\theta}, \frac{1-w_i}{1-\theta}\right\} - \min\left\{\frac{z_i}{\theta}, \frac{1-z_i}{1-\theta}\right\}\right| \\
&\leq \frac{1}{2\min\{\theta, 1-\theta\}}\frac{1}{n}\sum_{i=1}^n |w_i - z_i| \\
&\leq \frac{1}{2\min\{\theta, 1-\theta\}}\sqrt{\frac{1}{n}\sum_{i=1}^n (w_i - z_i)^2},
\end{aligned}$$

where the last inequality follows from Jensen's inequality.

Next, we prove (19). Observe that $(z_1, z_2) \in [0,1]^2 \mapsto z_1(1-z_2)$ is 1-Lipschitz with respect to 1-norm. From this, we see that $(z_1, z_2) \in [0,1]^2 \mapsto \min\{z_1(1-z_2), z_2(1-z_1)\}$ is also 1-Lipschitz. Therefore,

$$\begin{aligned}
&\left|\widehat{\mathrm{AUC}^*}(\boldsymbol{w}) - \widehat{\mathrm{AUC}^*}(\boldsymbol{z})\right| \\
&\leq \frac{1}{\theta(1-\theta)n(n-1)} \sum_{i<j} \left|\min\{w_i(1-w_j), w_j(1-w_i)\} - \min\{z_i(1-z_j), z_j(1-z_i)\}\right| \\
&\leq \frac{1}{\theta(1-\theta)n(n-1)} \sum_{i<j} \left(|w_i - z_i| + |w_j - z_j|\right).
\end{aligned}$$

For ease of notation, let $\Delta_i = |w_i - z_i|$ for each $i = 1, \ldots, n$. Then, we have

$$\begin{aligned}
\sum_{i<j} \left(|w_i - z_i| + |w_j - z_j|\right) &= \frac{1}{2}\left(\sum_{i=1}^n \sum_{j=1}^n (\Delta_i + \Delta_j) - \sum_{i=1}^n (2\Delta_i)\right) \\
&= \frac{1}{2} \cdot 2n \sum_{i=1}^n \Delta_i - \sum_{i=1}^n \Delta_i \\
&= (n-1) \sum_{i=1}^n \Delta_i.
\end{aligned}$$

Therefore, we get

$$\left|\widehat{\mathrm{AUC}^*}(\boldsymbol{w}) - \widehat{\mathrm{AUC}^*}(\boldsymbol{z})\right| \leq \frac{1}{\theta(1-\theta)n} \sum_{i=1}^n \Delta_i \leq \frac{1}{\theta(1-\theta)} \sqrt{\frac{1}{n} \sum_{i=1}^n \Delta_i^2}.$$

The second inequality is again due to Jensen's inequality. Now the proof is complete. □

The following lemma bounds the estimation error of isotonic regression with binary observations. We state it in the form given by Ushio et al. (2026). An essentially equivalent bound, which differs only in constants, was established earlier by Neykov (2019, Theorem 2.3).

Lemma B.3 (Neykov (2019); Ushio et al. (2026))**.** *Let $\mathcal{M}_n = \{\boldsymbol{u} \in \mathbb{R}^n \mid u_1 \leq \cdots \leq u_n\}$ be the set of non-decreasing sequences of length n, and for constants $a < b$, let $\mathcal{M}_n(a,b) = \{\boldsymbol{u} \in \mathbb{R}^n \mid a \leq u_1 \leq \cdots \leq u_n \leq b\}$.*

Let $\boldsymbol{\mu} \in [0,1]^n$ be an unknown mean vector, which might not belong to $\mathcal{M}_n(0,1)$. Given a random vector $\boldsymbol{y} = (y_1, \ldots, y_n) \in \{0,1\}^n$ of n independent binary observations with mean $\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\mu}$, let $\widehat{\boldsymbol{\mu}}$ be the isotonic regression estimator of $\boldsymbol{\mu}$, i.e.,

$$\widehat{\boldsymbol{\mu}} \coloneqq \operatorname*{arg\,min}_{\boldsymbol{u} \in \mathcal{M}_n} \|\boldsymbol{y} - \boldsymbol{u}\|_2^2.$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\frac{1}{n}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2^2 \leq \min_{\boldsymbol{u} \in \mathcal{M}_n(0,1)} \frac{1}{n}\|\boldsymbol{u} - \boldsymbol{\mu}\|_2^2 + \left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right)^2,$$

where C is an absolute constant.

Lemma B.4. *Assume the skew function f is increasing. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\sqrt{\frac{1}{n}\sum_{i=1}^n (\widehat{\eta}_i - \eta_i)^2} \leq \frac{C}{n^{1/3}} + \sqrt{\frac{2\log(1/\delta)}{n}}.$$

Proof. By applying Lemma B.3 for $\boldsymbol{\mu} = (\eta_{(1)}, \dots, \eta_{(n)})$, with conditional probability at least $1-\delta$ given $(\eta_1, \dots, \eta_n)$, we have

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\eta}_i - \eta_i)^2 \le \min_{\boldsymbol{u}\in\mathcal{M}_n(0,1)} \frac{1}{n}\sum_{i=1}^{n}\left(u_i - \eta_{(i)}\right)^2 + \left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(1/\delta)}{n}}\right)^2.$$

Since f is increasing and $\xi_{(1)} = f(\eta_{(1)}) \le \cdots \le \xi_{(n)} = f(\eta_{(n)})$, we have $\eta_{(1)} \le \cdots \le \eta_{(n)}$, i.e., $(\eta_{(1)}, \dots, \eta_{(n)}) \in \mathcal{M}_n$. Therefore, the first term on the right-hand side is equal to zero. This completes the proof. □

Theorem 3.1. *Under Assumption 2.8, for any $\delta \in (0,1)$, each of the following holds with probability at least $1-\delta$:*

$$\left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \lesssim \frac{1}{\min\{\theta, 1-\theta\}}\left(\frac{1}{n^{1/3}} + \sqrt{\frac{\log(5/\delta)}{n}}\right)$$

and

$$\left|\widetilde{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \lesssim \frac{1}{\theta(1-\theta)}\left(\frac{1}{n^{1/3}} + \sqrt{\frac{\log(5/\delta)}{n}}\right).$$

Proof. **BER case.** By the same argument as in the proof of Lemma A.26 and Theorem A.22, with probability at least $1 - \frac{4}{5}\delta$, we have

$$\begin{aligned}\left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \theta)\right| + \left|\widehat{\mathrm{BER}^*}(\boldsymbol{\eta}, \theta) - \mathrm{BER}^*\right| &\le \frac{17}{8}\sqrt{\frac{2\log(5/\delta)}{\theta(1-\theta)n}} \\ &\le \frac{17}{8\min\{\theta, 1-\theta\}}\sqrt{\frac{2\log(5/\delta)}{n}}.\end{aligned}$$

Also, by Lemmas B.2 and B.4, with probability at least $1 - \frac{1}{5}\delta$,

$$\left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \theta) - \widehat{\mathrm{BER}^*}(\boldsymbol{\eta}, \theta)\right| \le \frac{1}{2\min\{\theta, 1-\theta\}}\left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(5/\delta)}{n}}\right)$$

Therefore, we have the following with probability at least $1-\delta$:

$$\begin{aligned}&\left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{BER}^*\right| \\ &\le \left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \theta)\right| + \left|\widehat{\mathrm{BER}^*}(\widehat{\boldsymbol{\eta}}, \theta) - \widehat{\mathrm{BER}^*}(\boldsymbol{\eta}, \theta)\right| + \left|\widehat{\mathrm{BER}^*}(\boldsymbol{\eta}, \theta) - \mathrm{BER}^*\right| \\ &\le \frac{17}{8\min\{\theta, 1-\theta\}}\sqrt{\frac{2\log(5/\delta)}{n}} + \frac{1}{2\min\{\theta, 1-\theta\}}\left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(5/\delta)}{n}}\right) \\ &= \frac{1}{2\min\{\theta, 1-\theta\}}\left(\frac{C}{n^{1/3}} + \frac{21}{4}\sqrt{\frac{2\log(5/\delta)}{n}}\right).\end{aligned}$$

AUC case. First, suppose that $2 \le n < \frac{18\log(5/\delta)}{\theta(1-\theta)}$. Since both $\widetilde{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n})$ and AUC^* belong to $[\frac{1}{2}, 1]$, we have

$$\left|\widetilde{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \le \frac{1}{2}.$$

On the other hand, since $n < \frac{18\log(5/\delta)}{\theta(1-\theta)}$ and $\theta(1-\theta) \le \frac{1}{4}$, we have

$$\begin{aligned}\frac{1}{\theta(1-\theta)}\left(\frac{1}{n^{1/3}} + \sqrt{\frac{\log(5/\delta)}{n}}\right) &> \frac{1}{\theta(1-\theta)}\sqrt{\frac{\theta(1-\theta)}{18}} \\ &= \frac{1}{\sqrt{18\theta(1-\theta)}} \\ &\ge \frac{\sqrt{2}}{3}.\end{aligned}$$

Therefore, the claimed bound holds in this case with the hidden constant at most $\frac{3}{2\sqrt{2}}$.

It remains to consider the large-n case: $n \geq \frac{18\log(5/\delta)}{\theta(1-\theta)}$. Let

$$A(\theta) \coloneqq \frac{1}{\theta(1-\theta)}, \quad B(\boldsymbol{\eta}) \coloneqq \frac{1}{n(n-1)} \sum_{i<j} \min\left\{\eta_i(1-\eta_j), \eta_j(1-\eta_i)\right\}.$$

Then, since $n \geq \frac{18\log(5/\delta)}{\theta(1-\theta)}$, by the same argument as the proof of Lemma A.30, with probability at least $1 - \frac{2}{5}\delta$, both (20) and (21) hold:

$$\begin{aligned}
\frac{\left|A(\tilde{\theta}_{\tau_n}) - A(\theta)\right|}{A(\theta)} &= \frac{\left|\tilde{\theta}_{\tau_n}(1-\tilde{\theta}_{\tau_n}) - \theta(1-\theta)\right|}{\tilde{\theta}_{\tau_n}(1-\tilde{\theta}_{\tau_n})} \\
&\leq \frac{\left|\tilde{\theta}_{\tau_n} - \theta\right|}{\tilde{\theta}_{\tau_n}(1-\tilde{\theta}_{\tau_n})} \\
&\leq 6\sqrt{\frac{2\log(5/\delta)}{\theta(1-\theta)n}}, \qquad (20) \\
A(\tilde{\theta}_{\tau_n}) &= \frac{1}{\tilde{\theta}_{\tau_n}(1-\tilde{\theta}_{\tau_n})} \leq \frac{3}{\theta(1-\theta)}. \qquad (21)
\end{aligned}$$

Also, by Lemmas B.2 and B.4, with probability at least $1 - \frac{1}{5}\delta$,

$$|B(\widehat{\boldsymbol{\eta}}) - B(\boldsymbol{\eta})| \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\eta}_i - \eta_i)^2} \leq \frac{C}{n^{1/3}} + \sqrt{\frac{2\log(5/\delta)}{n}}.$$

Therefore, we have

$$\begin{aligned}
&\left|\widehat{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \widehat{\mathrm{AUC}^*}(\boldsymbol{\eta}, \theta)\right| \\
&= \left|A(\tilde{\theta}_{\tau_n})B(\widehat{\boldsymbol{\eta}}) - A(\theta)B(\boldsymbol{\eta})\right| \\
&\leq A(\tilde{\theta}_{\tau_n})\left|B(\widehat{\boldsymbol{\eta}}) - B(\boldsymbol{\eta})\right| + A(\theta)B(\boldsymbol{\eta}) \cdot \frac{\left|A(\tilde{\theta}_{\tau_n}) - A(\theta)\right|}{A(\theta)} \\
&\leq \frac{3}{\theta(1-\theta)}\left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(5/\delta)}{n}}\right) + (1 - \widehat{\mathrm{AUC}^*}(\boldsymbol{\eta}, \theta)) \cdot 6\sqrt{\frac{2\log(5/\delta)}{\theta(1-\theta)n}}.
\end{aligned}$$

Also, by the assumption that $n \geq \frac{18\log(5/\delta)}{\theta(1-\theta)} > \frac{\log(5/\delta)}{9\theta(1-\theta)}$, we can apply Lemma A.18 to obtain that with probability at least $1 - \frac{2}{5}\delta$,

$$\left|\widehat{\mathrm{AUC}^*}(\boldsymbol{\eta}, \theta) - \mathrm{AUC}^*\right| \leq \sqrt{\frac{\log(5/\delta)}{\theta(1-\theta)n}}.$$

Combining these results, with probability at least $1-\delta$, we have

$$
\begin{aligned}
&\left|\widetilde{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \\
\leq\ &\left|\widehat{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \mathrm{AUC}^*\right| \\
\leq\ &\left|\widehat{\mathrm{AUC}^*}(\widehat{\boldsymbol{\eta}}, \tilde{\theta}_{\tau_n}) - \widehat{\mathrm{AUC}^*}(\boldsymbol{\eta}, \theta)\right| + \left|\widehat{\mathrm{AUC}^*}(\boldsymbol{\eta}, \theta) - \mathrm{AUC}^*\right| \\
\leq\ &\frac{3}{\theta(1-\theta)}\left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(5/\delta)}{n}}\right) \\
&\qquad + \left(1 - \mathrm{AUC}^* + \sqrt{\frac{\log(5/\delta)}{\theta(1-\theta)n}}\right) \cdot 6\sqrt{\frac{2\log(5/\delta)}{\theta(1-\theta)n}} + \sqrt{\frac{\log(5/\delta)}{\theta(1-\theta)n}} \\
\leq\ &\frac{3}{\theta(1-\theta)}\left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(5/\delta)}{n}}\right) + \frac{3\sqrt{2}+3}{2\theta(1-\theta)}\sqrt{\frac{\log(5/\delta)}{n}},
\end{aligned}
$$

where the last inequality follows from $\mathrm{AUC}^* \geq \frac{1}{2}$, $\frac{\log(5/\delta)}{\theta(1-\theta)n} \leq \frac{1}{3\sqrt{2}}\sqrt{\frac{\log(5/\delta)}{\theta(1-\theta)n}}$ (by $n \geq \frac{18\log(5/\delta)}{\theta(1-\theta)}$), and $\sqrt{\theta(1-\theta)} \geq 2\theta(1-\theta)$. Now the claim has been proved. □

Noisy corruption model Theorem 3.3 is proved by using the following lemma instead of Lemma B.4 in the proof of Theorem 3.1. This lemma is the counterpart of Lemma B.4 in the noisy setting.

Lemma B.5. *Under Assumption 3.2, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$
\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\eta}_i - \eta_i\right)^2} \leq \frac{\sigma}{\gamma} + \frac{C}{n^{1/3}} + (1+\sqrt{2})\sqrt{\frac{\log(2/\delta)}{n}}.
$$

Proof. By applying Lemma B.3 for $\boldsymbol{\mu} = (\eta_{(1)}, \ldots, \eta_{(n)})$, with conditional probability at least $1-\delta/2$ given $\eta_1, \ldots, \eta_n, \varepsilon_1, \ldots, \varepsilon_n$, we have

$$
\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{\eta}_i - \eta_i\right)^2 \leq \min_{\boldsymbol{u}\in\mathcal{M}_n(0,1)} \frac{1}{n}\sum_{i=1}^{n}\left(u_i - \eta_{(i)}\right)^2 + \left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(2/\delta)}{n}}\right)^2. \tag{22}
$$

We then proceed to bound the first term on the right-hand side of (22). Under the assumption, f is invertible, and f^{-1} is also differentiable and increasing. Since f^{-1} is increasing and $\xi_{(1)} \leq \cdots \leq \xi_{(n)}$, we have $\left(f^{-1}(\xi_{(1)}), \ldots, f^{-1}(\xi_{(n)})\right) \in \mathcal{M}_n(0,1)$. Therefore, we have

$$
\min_{\boldsymbol{u}\in\mathcal{M}_n(0,1)} \frac{1}{n}\sum_{i=1}^{n}(u_i - \eta_{(i)})^2 \leq \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\xi_{(i)}) - \eta_{(i)})^2 = \frac{1}{n}\sum_{i=1}^{n}(f^{-1}(\xi_i) - \eta_i)^2.
$$

Now, let $D_i \coloneqq (f^{-1}(\xi_i) - \eta_i)^2$. The mean value theorem and the assumption that $f' \geq \gamma$ imply that

$$
D_i = \left(f^{-1}(f(\eta_i) + \varepsilon_i) - f^{-1}(f(\eta_i))\right)^2 \leq \frac{\varepsilon_i^2}{\gamma^2}
$$

and hence $\mathbb{E}\left[D_i \mid \eta_1, \ldots, \eta_n\right] \leq \frac{\sigma^2}{\gamma^2}$. Also, since $D_i \in [0,1]$, $\mathrm{Var}\left[D_i \mid \eta_1, \ldots, \eta_n\right] \leq \mathbb{E}\left[D_i^2 \mid \eta_1, \ldots, \eta_n\right] \leq \mathbb{E}\left[D_i \mid \eta_1, \ldots, \eta_n\right] \leq \frac{\sigma^2}{\gamma^2}$. Moreover, $D_1, \ldots, D_n$ are conditionally independent given $\eta_1, \ldots, \eta_n$. Therefore, by Bernstein's inequality, we have

$$
\frac{1}{n}\sum_{i=1}^{n} D_i \leq \frac{\sigma^2}{\gamma^2} + \sqrt{\frac{2(\sigma^2/\gamma^2)\log(2/\delta)}{n}} + \frac{2\log(2/\delta)}{3n} \leq \left(\frac{\sigma}{\gamma} + \sqrt{\frac{\log(2/\delta)}{n}}\right)^2
$$

with conditional probability at least $1-\delta/2$ given $\eta_1, \ldots, \eta_n$. Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\eta}_i - \eta_i)^2} &\le \sqrt{\left(\frac{\sigma}{\gamma} + \sqrt{\frac{\log(2/\delta)}{n}}\right)^2 + \left(\frac{C}{n^{1/3}} + \sqrt{\frac{2\log(2/\delta)}{n}}\right)^2} \\ &\le \frac{\sigma}{\gamma} + \sqrt{\frac{\log(2/\delta)}{n}} + \frac{C}{n^{1/3}} + \sqrt{\frac{2\log(2/\delta)}{n}} \\ &= \frac{\sigma}{\gamma} + \frac{C}{n^{1/3}} + (1+\sqrt{2})\sqrt{\frac{\log(2/\delta)}{n}}.\end{aligned}$$

□

C Supplementary for Section 4

C.1 Proof of Theorem 4.1

Here we prove Theorem 4.1, which gives the exact expressions of the optimal BER and AUC values for the noise-injected distribution.

Theorem 4.1. *For $t \in [0, \frac{1}{2}]$, define $F_\nu(t) \coloneqq \frac{\lambda_\nu\left(2\theta(1-\theta)t+\theta^2\right)-\lambda_\nu(\theta)^2}{2\lambda_\nu(\theta)(1-\lambda_\nu(\theta))}$. For any noise level $\nu \in [0,1)$, the optimal BER and AUC for the noise-injected distribution are given by $\mathrm{BER}^*_\nu = F_\nu(\mathrm{BER}^*)$ and $\mathrm{AUC}^*_\nu = 1 - F_\nu(1-\mathrm{AUC}^*)$, respectively.*

The BER part is proved in Theorem C.2 and the AUC part is proved in Theorem C.4.

Lemma C.1. *The optimal BER value can be written as*

$$\mathrm{BER}^* = \frac{1}{2\theta(1-\theta)}\left(\mathbb{E}\left[\min\left\{\eta(x), \theta\right\}\right] - \theta^2\right).$$

Proof.

$$\begin{aligned}\mathrm{BER}^* &= \frac{1}{2}\,\mathbb{E}\left[\min\left\{\frac{\eta(x)}{\theta}, \frac{1-\eta(x)}{1-\theta}\right\}\right] \quad \text{(Lemma 2.1)} \\ &= \frac{1}{2\theta(1-\theta)}\,\mathbb{E}\left[\min\left\{\eta(x)(1-\theta), \theta(1-\eta(x))\right\}\right] \\ &= \frac{1}{2\theta(1-\theta)}\,\mathbb{E}\left[\min\left\{\eta(x), \theta\right\} - \theta\eta(x)\right] \\ &= \frac{1}{2\theta(1-\theta)}\left(\mathbb{E}\left[\min\left\{\eta(x), \theta\right\}\right] - \theta^2\right),\end{aligned}$$

where we used $\mathbb{E}\left[\eta(x)\right] = \theta$ in the last step. □

Theorem C.2. *For any noise level $\nu \in [0,1)$, the optimal BER for the noise-injected distribution is given by*

$$\mathrm{BER}^*_\nu = F_\nu(\mathrm{BER}^*).$$

Proof. Applying Lemma C.1 to the noisy distribution (and recalling that we assumed $\lambda_\nu(\theta) \in (0,1)$), we have

$$\mathrm{BER}^*_\nu = \frac{\mathbb{E}\left[\min\left\{\lambda_\nu(\eta(x)), \lambda_\nu(\theta)\right\}\right] - \lambda_\nu(\theta)^2}{2\lambda_\nu(\theta)(1-\lambda_\nu(\theta))}.$$

Since λ_ν is non-decreasing and affine, we can write

$$\mathbb{E}\left[\min\left\{\lambda_\nu(\eta(x)), \lambda_\nu(\theta)\right\}\right] = \mathbb{E}\left[\lambda_\nu(\min\left\{\eta(x), \theta\right\})\right] = \lambda_\nu(\mathbb{E}\left[\min\left\{\eta(x), \theta\right\}\right]).$$

Now, using Lemma C.1 again for the clean distribution, we have

$$\mathbb{E}\left[\min\left\{\eta(x), \theta\right\}\right] = 2\theta(1-\theta)\mathrm{BER}^* + \theta^2$$

which gives

$$\mathrm{BER}^*_\nu = \frac{\lambda_\nu\left(2\theta(1-\theta)\mathrm{BER}^* + \theta^2\right) - \lambda_\nu(\theta)^2}{2\lambda_\nu(\theta)(1-\lambda_\nu(\theta))},$$

which proves the theorem. □

Lemma C.3. *The optimal AUC value can be written as*

$$\mathrm{AUC}^* = 1 - \frac{1}{2\theta(1-\theta)}\left(\mathbb{E}\left[\min\left\{\eta(x), \eta(x')\right\}\right] - \theta^2\right),$$

where x' is an i.i.d. copy of x.

Proof.

$$\begin{aligned}
1 - \mathrm{AUC}^* &= \frac{1}{2\theta(1-\theta)}\,\mathbb{E}\left[\min\left\{\eta(x)(1-\eta(x')), \eta(x')(1-\eta(x))\right\}\right] \quad \text{(Lemma 2.2)} \\
&= \frac{1}{2\theta(1-\theta)}\,\mathbb{E}\left[\min\left\{\eta(x), \eta(x')\right\} - \eta(x)\eta(x')\right] \\
&= \frac{1}{2\theta(1-\theta)}\left(\mathbb{E}\left[\min\left\{\eta(x), \eta(x')\right\}\right] - \theta^2\right),
\end{aligned}$$

where we used $\mathbb{E}\left[\eta(x)\right] = \mathbb{E}\left[\eta(x')\right] = \theta$ and the independence of x and x' in the last step. □

Theorem C.4. *For any noise level $\nu \in [0, 1)$, the optimal AUC for the noise-injected distribution is given by*

$$\mathrm{AUC}^*_\nu = 1 - F_\nu(1 - \mathrm{AUC}^*).$$

Proof. The proof is analogous to that of Theorem C.2. Applying Lemma C.3 to the noisy distribution, we have

$$1 - \mathrm{AUC}^*_\nu = \frac{\mathbb{E}\left[\min\left\{\lambda_\nu(\eta(x)), \lambda_\nu(\eta(x'))\right\}\right] - \lambda_\nu(\theta)^2}{2\lambda_\nu(\theta)(1-\lambda_\nu(\theta))}.$$

Since λ_ν is non-decreasing and affine, we can write

$$\mathbb{E}\left[\min\left\{\lambda_\nu(\eta(x)), \lambda_\nu(\eta(x'))\right\}\right] = \mathbb{E}\left[\lambda_\nu(\min\left\{\eta(x), \eta(x')\right\})\right] = \lambda_\nu(\mathbb{E}\left[\min\left\{\eta(x), \eta(x')\right\}\right]).$$

Now, using Lemma C.3 again for the clean distribution, we have

$$\mathbb{E}\left[\min\left\{\eta(x), \eta(x')\right\}\right] = 2\theta(1-\theta)(1-\mathrm{AUC}^*) + \theta^2$$

which gives

$$1 - \mathrm{AUC}^*_\nu = \frac{\lambda_\nu\left(2\theta(1-\theta)(1-\mathrm{AUC}^*) + \theta^2\right) - \lambda_\nu(\theta)^2}{2\lambda_\nu(\theta)(1-\lambda_\nu(\theta))},$$

which proves the theorem. □

C.2 Bias sensitivity

Here we show that our score can reflect certain types of estimator bias, unlike bootstrap variance.

Theoretical analysis The following proposition roughly states that any bias larger than the interval width $U_{\mathrm{BER}}(\nu) - L_{\mathrm{BER}}(\nu)$ produces a positive expected penalty. This is in contrast to the bootstrap variance, which does not change if a constant is added to the estimator. Here we discuss the BER case, but the same argument also applies to AUC by replacing u_{BER} with $1 - l_{\mathrm{AUC}}$.

Proposition C.5. *Let* $\mathrm{bias}_\nu = \mathbb{E}\left[\widehat{\mathrm{BER}^*_\nu}\right] - \mathrm{BER}^*_\nu$ *denote the bias of the estimator* $\widehat{\mathrm{BER}^*_\nu}$ *at noise level ν. Then*

$$\mathbb{E}\left[\left(\widehat{\mathrm{BER}^*_\nu} - U_{\mathrm{BER}}(\nu)\right)_+ + \left(L_{\mathrm{BER}}(\nu) - \widehat{\mathrm{BER}^*_\nu}\right)_+\right] \geq \left[|\mathrm{bias}_\nu| - \frac{u_{\mathrm{BER}}}{K}(1-\nu)\right]_+,$$

where $K \coloneqq \min\left\{1, \frac{\beta(1-\beta)}{\theta(1-\theta)}\right\} \in (0, 1]$.

Proof. Let $d(z_1, z_2) = |z_1 - z_2|$ and $I_\nu = [L_{\mathrm{BER}}(\nu), U_{\mathrm{BER}}(\nu)]$. Then, the pointwise penalty $(\widehat{\mathrm{BER}}^*_\nu - U_{\mathrm{BER}}(\nu))_+ + (L_{\mathrm{BER}}(\nu) - \widehat{\mathrm{BER}}^*_\nu)_+$ at noise level ν is precisely $d(\widehat{\mathrm{BER}}^*_\nu, I_\nu) := \inf_{a \in I_\nu} d(\widehat{\mathrm{BER}}^*_\nu, a)$. By the convexity of $d(\cdot, I_\nu)$ and Jensen's inequality, the expected pointwise penalty is bounded as

$$\mathbb{E}\left[d\left(\widehat{\mathrm{BER}}^*_\nu, I_\nu\right)\right] \geq d\left(\mathbb{E}\left[\widehat{\mathrm{BER}}^*_\nu\right], I_\nu\right). \tag{23}$$

On the other hand, since $\mathrm{BER}^*_\nu \in I_\nu$, the bias of the estimator $\widehat{\mathrm{BER}}^*_\nu$ is bounded as follows:[11]

$$|\mathrm{bias}_\nu| = d\left(\mathbb{E}\left[\widehat{\mathrm{BER}}^*_\nu\right], \mathrm{BER}^*_\nu\right) \leq d\left(\mathbb{E}\left[\widehat{\mathrm{BER}}^*_\nu\right], I_\nu\right) + U_{\mathrm{BER}}(\nu) - L_{\mathrm{BER}}(\nu). \tag{24}$$

The interval width $U_{\mathrm{BER}}(\nu) - L_{\mathrm{BER}}(\nu)$ is bounded as

$$\begin{aligned} U_{\mathrm{BER}}(\nu) - L_{\mathrm{BER}}(\nu) &= F_\nu(u_{\mathrm{BER}}) - F_\nu(0) \\ &= \frac{(1-\nu)\theta(1-\theta)u_{\mathrm{BER}}}{\lambda_\nu(\theta)(1-\lambda_\nu(\theta))} \\ &\leq \frac{(1-\nu)\theta(1-\theta)u_{\mathrm{BER}}}{\min\{\theta(1-\theta), \beta(1-\beta)\}} \\ &= \frac{u_{\mathrm{BER}}}{K}(1-\nu). \end{aligned} \tag{25}$$

Combining (23), (24), and (25), we conclude that

$$\mathbb{E}\left[d\left(\widehat{\mathrm{BER}}^*_\nu, I_\nu\right)\right] \geq \left[|\mathrm{bias}_\nu| - \frac{u_{\mathrm{BER}}}{K}(1-\nu)\right]_+.$$

□

Thus, for each noise level ν, a bias larger than $\frac{u_{\mathrm{BER}}}{K}(1-\nu)$ produces a positive expected penalty. Moreover, since $\frac{u_{\mathrm{BER}}}{K}(1-\nu) \to 0$ as $\nu \to 1$, any fixed nonzero additive bias that persists across noise levels is expected to be detected at a sufficiently large noise level. It is worth noting that this is not a universal guarantee: a bias that shrinks sufficiently fast as $\nu \to 1$ may remain undetected.

Proposition C.5 quantifies the *pointwise* penalty for each noise level, and we can use it to discuss the overall score aggregated over different noise levels. As a simple example, consider an idealized version of s_β, obtained by replacing the finite average over N noise levels with an integral (informally, $N \to \infty$):

$$s_\beta = \int_0^1 d\left(\widehat{\mathrm{BER}}^*_\nu, I_\nu\right) \mathrm{d}\nu = \int_0^1 \left[\left(\widehat{\mathrm{BER}}^*_\nu - U_{\mathrm{BER}}(\nu)\right)_+ + \left(L_{\mathrm{BER}}(\nu) - \widehat{\mathrm{BER}}^*_\nu\right)_+\right] \mathrm{d}\nu.$$

The result above implies that the expected score is bounded as

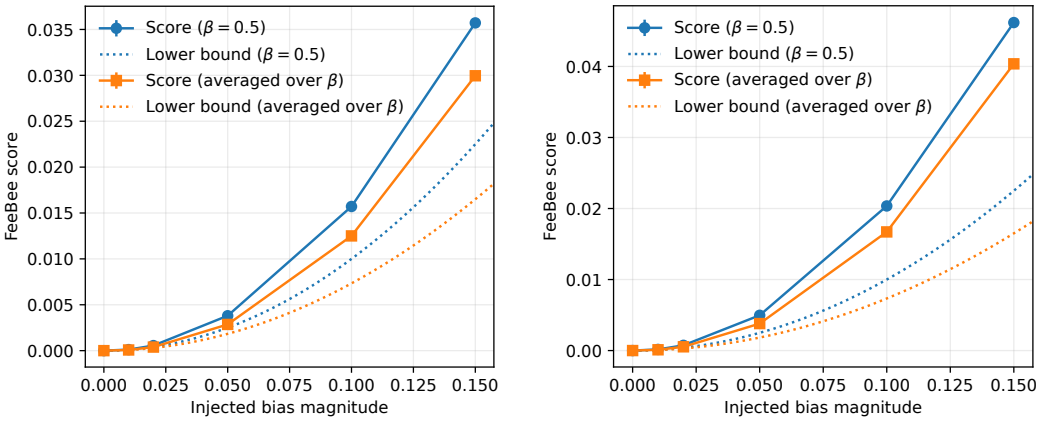
$$\mathbb{E}\left[s_\beta\right] \geq \int_0^1 \left[|\mathrm{bias}_\nu| - \frac{u_{\mathrm{BER}}}{K}(1-\nu)\right]_+ \mathrm{d}\nu.$$

For example, if $|\mathrm{bias}_\nu| \geq B \geq 0$ for all $\nu \in [0, 1)$, the expected score is lower-bounded by an increasing function of B as

$$\mathbb{E}\left[s_\beta\right] \geq \begin{cases} \frac{K}{2u_{\mathrm{BER}}} B^2 & \text{if } B \leq \frac{u_{\mathrm{BER}}}{K}, \\ B - \frac{u_{\mathrm{BER}}}{2K} & \text{otherwise.} \end{cases} \tag{26}$$

Empirical analysis We also present an experiment using synthetic data with controlled bias. We drew $n = 10{,}000$ points from a one-dimensional two-component Gaussian mixture with means -0.5 and 0.5, unit variances, and class prior $\theta = 0.5$. For each point, we computed the clean soft label η_i using the closed form obtained from the Gaussian PDF. By Proposition D.1, the true optimal values are $\mathrm{BER}^* = \Phi(-1/2) \approx 0.308538$ and $\mathrm{AUC}^* = \Phi(1/\sqrt{2}) \approx 0.760250$.

[11] This follows from the following fact: let $(\mathcal{Z}, d)$ be a metric space and let $A \subset \mathcal{Z}$. For any $z \in \mathcal{Z}$ and $a \in A$, we have $d(z, a) \leq d(z, A) + \mathrm{diam}(A)$, where $d(z, A) = \inf_{a' \in A} d(z, a')$ and $\mathrm{diam}(A) = \sup_{a_1, a_2 \in A} d(a_1, a_2)$.



(a) BER (b) AUC

Figure 3: Our evaluation scores versus injected bias magnitude for BER and AUC. Points show averages over 100 trials. Their 95% Monte Carlo confidence intervals are smaller than the markers. The blue curves use $\beta = 0.5$, whereas the orange curves average the scores over $\beta \in \{0.1, \ldots, 0.9\}$. The blue and orange dotted curves show the theoretical lower bounds in (26) for the corresponding β settings, evaluated at $B = b$. Both scores increase monotonically with the injected bias (although the estimator standard deviations remain unchanged).

We repeated the experiment 100 times. In each trial, we computed the min formula estimators $\widehat{\mathrm{BER}}^*_\nu$ and $\widehat{\mathrm{AUC}}^*_\nu$ at $N = 100$ noise levels. To change the bias without changing the variance, we defined shifted estimators at each noise level ν as $\widehat{\mathrm{BER}}^{*(b)}_\nu := \widehat{\mathrm{BER}}^*_\nu - b$ and $\widehat{\mathrm{AUC}}^{*(b)}_\nu := \widehat{\mathrm{AUC}}^*_\nu + b$. Because the same constant is added to or subtracted from the estimate in every trial, these changes do not change the variance of either estimator. We used $b \in \{0, 0.01, 0.02, 0.05, 0.10, 0.15\}$ and the same samples for every value of b. The parameters u_{BER} and l_{AUC} were both set to 0.5. We computed the scores s_β for $\beta = 0.5$ and the average scores over $\beta \in \{0.1, \ldots, 0.9\}$.

Figure 3 reports averages over the 100 trials (blue and orange solid curves) along with the theoretical lower bounds in (26) (blue and orange dotted curves). All standard errors of the scores were below 2.8×10^{-5}. The standard deviations (SDs) of the BER estimates and AUC estimates were 0.001241 and 0.001370, respectively, for every value of b (as expected, since the variance was kept constant). In contrast, the score averaged over β increased monotonically with b, from 0 to 0.029953 for BER and from 0 to 0.040368 for AUC. These results show that, unlike the standard deviations, our evaluation score successfully reflects the increase in bias even when the estimator variance is unchanged.

D Closed-form expressions for the optimal BER and AUC under Gaussian mixture models with a shared covariance matrix

Here, we derive the closed-form expressions for the optimal BER and AUC used as reference values in the Gaussian-mixture experiments of Appendix E.

Suppose $\mathcal{X} = \mathbb{R}^d$ and that the class-conditional distributions are given by Gaussian distributions with a shared covariance matrix:

$$x \mid \{y = 0\} \sim \mathcal{N}(\mu_0, \Sigma), \quad x \mid \{y = 1\} \sim \mathcal{N}(\mu_1, \Sigma),$$

where $\mu_0, \mu_1 \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is a positive definite matrix. For $k \in \{0, 1\}$, let $p_k(x)$ denote the class-conditional density function of x given $y = k$:

$$p_k(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^\top \Sigma^{-1}(x - \mu_k)\right).$$

The goal of this section is to prove the following proposition, which gives closed-form expressions for the optimal BER and AUC under this setting.

Proposition D.1. *Under the shared-covariance Gaussian mixture model described above, the optimal BER and AUC are given by*

$$\mathrm{BER}^* = \Phi\left(-\frac{J}{2}\right), \quad \mathrm{AUC}^* = \Phi\left(\frac{J}{\sqrt{2}}\right),$$

where Φ is the cumulative distribution function of the standard normal distribution and $J \coloneqq \sqrt{(\mu_1-\mu_0)^\top \Sigma^{-1}(\mu_1-\mu_0)}$ is the Mahalanobis distance between μ_0 and μ_1.

To prove Proposition D.1, define the score function $g^* : \mathcal{X} \to \mathbb{R}$ as

$$g^*(x) \coloneqq \log \frac{p_1(x)}{p_0(x)} = (\mu_1-\mu_0)^\top \Sigma^{-1}\left(x - \frac{\mu_0+\mu_1}{2}\right). \tag{27}$$

Proof. We first consider the nondegenerate case $J > 0$. We begin by proving the formula for BER^*. We saw in the proof of Lemma 2.1 that the BER-optimal classifier is given by $h^*_{\mathrm{BER}}(x) = \mathbb{1}\left[\eta(x) \geq \theta\right]$. Since the posterior probability $\eta(x)$ can be expressed as

$$\eta(x) = \frac{\theta p_1(x)}{\theta p_1(x) + (1-\theta)p_0(x)},$$

the condition $\eta(x) \geq \theta$ is equivalent to $p_1(x) \geq p_0(x)$, which is in turn equivalent to $g^*(x) \geq 0$. The class-conditional distributions of the score $g^*(x)$ are given by

$$g^*(x) \mid \{y=0\} \sim \mathcal{N}\left(-\frac{J^2}{2}, J^2\right), \quad g^*(x) \mid \{y=1\} \sim \mathcal{N}\left(\frac{J^2}{2}, J^2\right). \tag{28}$$

Therefore, the within-class error rates of h^*_{BER} are given by

$$\mathrm{FPR}(h^*_{\mathrm{BER}}) = \mathbb{P}\left(g^*(x) \geq 0 \mid y=0\right) = \Phi\left(-\frac{J}{2}\right),$$

$$\mathrm{FNR}(h^*_{\mathrm{BER}}) = \mathbb{P}\left(g^*(x) < 0 \mid y=1\right) = \Phi\left(-\frac{J}{2}\right),$$

and hence the optimal BER is given by

$$\mathrm{BER}^* = \frac{1}{2}\left(\mathrm{FPR}(h^*_{\mathrm{BER}}) + \mathrm{FNR}(h^*_{\mathrm{BER}})\right) = \Phi\left(-\frac{J}{2}\right).$$

Next, we prove the AUC case. In the proof of Lemma 2.2, we showed that the maximum AUC is attained by choosing a scoring function to be any strictly increasing transformation of $\eta(x)$. The scoring function g^* defined in (27) is an instance of such scores; to see this, note that

$$g^*(x) = \log \frac{\theta^{-1}-1}{\eta(x)^{-1}-1},$$

which is strictly increasing in $\eta(x)$. Thus, g^* achieves the maximum AUC.

Now, let $x_+ \sim \mathbb{P}\left(x \mid y=1\right)$ and $x_- \sim \mathbb{P}\left(x \mid y=0\right)$ be independent. Using the class-conditional distributions of $g^*(x)$ given in (28), we obtain

$$g^*(x_+) - g^*(x_-) \sim \mathcal{N}\left(J^2, 2J^2\right)$$

and hence the maximum AUC is given by

$$\begin{aligned}
\mathrm{AUC}^* &= \mathbb{P}\left(g^*(x_+) > g^*(x_-)\right) + \frac{1}{2}\mathbb{P}\left(g^*(x_+) = g^*(x_-)\right) \\
&= \mathbb{P}\left(g^*(x_+) - g^*(x_-) > 0\right) + 0 \\
&= \Phi\left(\frac{J}{\sqrt{2}}\right).
\end{aligned}$$

It remains to consider the degenerate case $J = 0$, in which g^* remains optimal. Since Σ is positive definite, $J = 0$ implies $\mu_0 = \mu_1$ and that the optimal classifier always predicts $h^*(x) = \mathbb{1}\left[g^*(x) \geq 0\right] = \mathbb{1}\left[0 \geq 0\right] = 1$. Hence, $\mathrm{BER}^* = (1+0)/2 = \Phi(-J/2)$. Similarly, we have $\mathrm{AUC}^* = 1/2 = \Phi(J/\sqrt{2})$ when $J = 0$ because $g^*(x_+) - g^*(x_-) = 0 - 0 = 0$ almost surely. This completes the proof. □

E Experiments

Here, we provide details on the experiments we conducted in Section 5. Our experiment code is based on that of Ushio et al. (2026), available at `https://github.com/RyotaUshio/bayes-error-estimation` and licensed under the MIT License.

E.1 Estimation on synthetic data

This appendix provides the full setup for the synthetic-data experiment of Section 5.1.

Data generation We followed the synthetic-data protocol of Ushio et al. (2026). Inputs $x_i \in \mathbb{R}^2$ were drawn from a two-component Gaussian mixture with means $(0, 0)$ and $(2, 2)$, identity covariances, and mixture weights $(0.8, 0.2)$, so that the class prior is $\theta = 0.2$. For each x_i, the clean soft label $\eta_i = \mathbb{P}(y = 1 \mid x = x_i)$ was computed in closed form from the mixture density, and a binary label y_i was sampled from $\mathrm{Bernoulli}(\eta_i)$. We used $n = 10{,}000$ samples. In this model, the optimal values can be computed analytically using Proposition D.1. The Mahalanobis distance J is $\sqrt{(2,2)^\top I_2^{-1}(2,2)} = 2\sqrt{2}$, where I_2 is the 2×2 identity matrix. Thus, $\mathrm{BER}^* = \Phi(-\sqrt{2}) \approx 0.07865$ and $\mathrm{AUC}^* = \Phi(2) \approx 0.97725$. The dashed lines in Figure 2 mark these values as references.

The corruption map $f\colon [0,1] \to [0,1]$ used here is the inverse beta-calibration (Kull et al., 2017) map also used in the experiments by Ushio et al. (2026),

$$f(\eta; a, b) = \frac{1}{1 + \left(\frac{1-\eta}{\eta}\right)^{1/a} \left(\frac{1-b}{b}\right)}, \tag{29}$$

parameterized by (a, b). The three regimes used in Figure 2 correspond to $(a, b) = (1, 0.5), (1.5, 0.5)$, and $(0.5, 0.5)$, respectively. Setting $a > 1$ produces under-confident soft labels (pulled toward 0.5), and $a < 1$ produces over-confident soft labels (pushed toward 0 or 1). The choice $b = 0.5$ keeps the corruption symmetric about 0.5.

Compared estimators For each metric, we compared the following estimates. For the clean setting, we simply applied the estimator $\widehat{\mathrm{BER}}^*$ or $\widehat{\mathrm{AUC}}^*$ directly to the soft labels η_i as discussed in Section 2 (**Clean**). For the corrupted settings (under- or over-confidence), we evaluated two estimators:

- **Corrupted, calibrated** (Section 3): fit isotonic regression (Ayer et al., 1955) of y_i on ξ_i to obtain $\widehat{\eta}_i$, then apply $\widehat{\mathrm{BER}}^*$ or $\widehat{\mathrm{AUC}}^*$ to $\widehat{\eta}_i$.
- **Corrupted, raw** (naive baseline): apply $\widehat{\mathrm{BER}}^*$ or $\widehat{\mathrm{AUC}}^*$ directly to the corrupted soft labels ξ_i without recalibration.

The class prior θ was treated as unknown and estimated from the given data, namely, by the sample mean of the soft labels η_i in the clean setting and that of the hard labels y_i in the corrupted settings. We used the rule of Section 2.1 to pick between the min and max formulae on each resample.

Bootstrap confidence intervals We computed 95% confidence intervals using the bias-corrected accelerated (BCa) bootstrap with $1{,}000$ resamples, as implemented by `scipy.stats.bootstrap` in SciPy 1.15.3 (Virtanen et al., 2020).

E.2 Real-data experiment: Discriminants and evaluation scores

Here we describe the detailed setup for the experiment on real-world datasets presented in Section 5.2.

E.2.1 Datasets

In our experiments, we used the following datasets from various domains, including computer vision, natural language processing and academic peer-review.

CIFAR-10 We used the test set of CIFAR-10 (Krizhevsky, 2009) with soft labels from CIFAR-10H (Peterson et al., 2019).[12] Following Ishida et al. (2023); Ushio et al. (2026), the original 10 classes were reorganized into two classes (animal vs. non-animal) by treating *bird*, *cat*, *deer*, *dog*, *frog* and *horse* as positive and the rest as negative.

Fashion-MNIST We also used the Fashion-MNIST dataset (Xiao et al., 2017) and its soft-labeled counterpart, Fashion-MNIST-H (Ishida et al., 2023).[13] As in CIFAR-10, we followed Ishida et al. (2023); Ushio et al. (2026) and relabeled the original 10 classes into two classes: positive (*T-shirt/top*, *pullover*, *dress*, *coat* and *shirt*) and negative (the rest).

ChaosNLI We used a natural language processing dataset called ChaosNLI (Nie et al., 2020; Zhou et al., 2022).[14] It contains 100 hard labels per data point, which can be used to construct soft labels by taking their average. It consists of three sub-datasets: SNLI ($n = 1,514$), MNLI ($n = 1,599$) and AbductiveNLI ($n = 1,532$).

ICLR 2017–2026 peer-review datasets Ushio et al. (2026) collected $n = 32,829$ instances of reviewer scores and confidences from the 2017–2025 ICLR conferences using the OpenReview API. Then they used them to construct a dataset consisting of pairs of a soft and hard label. Specifically, for each submitted paper x_i, they first calculated the averaged score weighted by the confidences, and then obtained a corrupted soft label ξ_i by normalizing the averaged score s_i into $[0, 1]$. The hard label is simply the final decision for the paper, i.e., accept ($y_i = 1$) or reject ($y_i = 0$). In this paper, we followed the same procedure to collect the 2017–2026 ICLR peer-review data (note that the 2026 data was not available at the time of Ushio et al. (2026)).

E.2.2 Experiment setup

For each metric (BER, AUC) and each dataset of Appendix E.2.1, we evaluated three quantities: the min formula ($\widehat{\mathrm{BER}}_1^*$ or $\widehat{\mathrm{AUC}}_1^*$), the max formula ($\widehat{\mathrm{BER}}_2^*$ or $\widehat{\mathrm{AUC}}_2^*$), and the empirical discriminant ($\widehat{\Delta}_{\mathrm{BER}}$ or $\widehat{\Delta}_{\mathrm{AUC}}$). The soft labels were used as is without calibration, and the class prior θ was estimated as the average of the hard labels y_i.

Bootstrap point estimate, CI, and SE. For each of the three estimators we report the point estimate, the bootstrap standard error, and a 95% bias-corrected accelerated (BCa) bootstrap confidence interval, all computed with $1{,}000$ resamples using `scipy.stats.bootstrap` in SciPy 1.15.3 (Virtanen et al., 2020).

Evaluation score s_β. For the min and max formula estimators, we additionally computed the score s_β of Section 4. s_β was computed with $N = 100$ different noise levels $\nu_i = (i-1)/N \in [0, 1)$ for $i = 1, \ldots, N$. The upper bound u_{BER} on BER^* (resp. the lower bound l_{AUC} on AUC^*) was set to its theoretical limit 0.5 on every dataset. We compared three choices for the noise mean β: (i) $\beta = 0.5$, recovering the original FeeBee setting of Renggli et al. (2021); (ii) $\beta = \theta$, the estimated class prior; and (iii) the average of s_β over the grid $\beta \in \{0.1, \ldots, 0.9\}$.

E.3 Approximately monotone corruption

Here, we present an additional experiment with synthetic data to investigate how the performance of our isotonic-regression-based estimator degrades when the corruption is only approximately monotone.

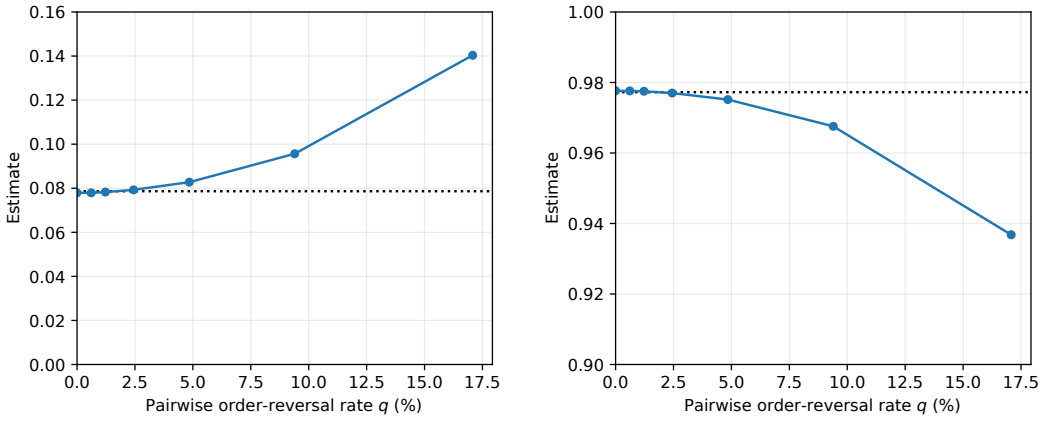
E.3.1 Experimental setup

Following the experimental protocol in Appendix D.5 of Ushio et al. (2026), we isolated order breakage by starting from the under-confident inverse-beta-calibration map f of (29) with parameters

[12] We were unable to find an explicit license stated for CIFAR-10, but it is widely used in the research community and is generally considered to be in the public domain. CIFAR-10H is licensed under BY-NC-SA 4.0.

[13] Fashion-MNIST is MIT-licensed. Fashion-MNIST-H is distributed as a part of their GPL-3.0 repository `https://github.com/ishida-lab/irreducible`, but no separate data-specific license is stated.

[14] Available at `https://github.com/easonnie/ChaosNLI` and distributed under the CC BY-NC 4.0 license.



(a) BER

(b) AUC

Figure 4: Estimates of BER^* and AUC^* under increasing violations of order preservation. The horizontal axis represents the mean pairwise order-reversal rate $q = (1-\tau)/2$. The points and error bars indicate means and 95% Monte Carlo CIs over 30 independent datasets. Note that the error bars are smaller than the marker size for most points. The dotted lines indicate the ground-truth optimal values.

$(a, b) = (1.5, 0.5)$ and adding Gaussian noise on the logit scale:

$$\xi_i = \mathrm{sigmoid}\left(\mathrm{logit}(f(\eta_i)) + z_i\right), \quad z_i \sim \mathcal{N}(0, \sigma^2)$$

Consequently, the corruption used here is nearly monotone but not exactly monotone. The parameter σ controls how far the corruption departs from monotonicity.

We used the same two-dimensional Gaussian mixture as in Section 5.1. We used a grid $\sigma \in \{0, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6\}$. For each σ, we used the proposed isotonic-regression-based estimator to compute the BER^* and AUC^* estimates. We also computed the Kendall rank correlation τ and $q = (1-\tau)/2$. The quantity q corresponds to the fraction of reversed pairs. We repeated each setting on 30 independently sampled datasets. By Proposition D.1, the ground-truth optimal values are $\mathrm{BER}^* = \Phi(-\sqrt{2}) = 0.078650$ and $\mathrm{AUC}^* = \Phi(2) = 0.977250$.

E.3.2 Results

Figure 4 shows the results. The estimates remained close to the ground-truth values as the pairwise order-reversal rate increased to 2.44%. Over this range, the 95% Monte Carlo confidence intervals (CIs) contained the ground-truth values. At $q = 4.85\%$ or higher, no CI contained the corresponding population value, and the deviations became substantially larger as q increased. Overall, introducing a small amount of rank violation translated to a mild increase in estimation error although the error increased progressively as the order-reversal rate grew.

E.4 Compute resources

All of the experiments presented in this paper were conducted on the CPU of a single Apple MacBook Pro (M1 chip, 16GB RAM).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the second half of Section 1, we summarized our claims and contributions referencing the relevant sections of the paper, and we made sure that they are accurate and reflect the scope of our work.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The appendices contain the proofs for all the theoretical results in this paper. We also made sure it is clear what assumptions are made in each theorem, proposition, lemma, and corollary.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Details about the experimental settings and results are provided in Section 5 and Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code and data are not released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The body text of the paper contains enough information to understand the experimental setting and results. The full details are provided with the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our experimental results are reported with error bars computed by bootstrapping.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details about the computer resources used in our experiments can be found in Appendix E.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics and made sure that this work conforms with every aspect of it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mention some positive impacts of this study in Section 1 (e.g., mitigation of environmental burden). As for the negative side, because our work is methodological and theoretical, it presents no direct path to malicious or unintended uses, nor to privacy or security risks, though potential indirect path may exist.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not involve the release of data or models that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The description of the licenses of the datasets we used can be found in Appendix E.2.1. Other work is properly credited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not include newly crowdsourced datasets or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core methods in this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.