Probing Bayes Error Estimation from Soft Labels: Replication and Extensions of “Is the Performance of My Deep Network Too Good to Be True?”

Shreyas Pradeepkumar Khandale
Department of Computer Science
Binghamton University
Binghamton, New York, USA
skhandale@binghamton.edu

Samruddhi Ajay Navale
Department of Computer Science
Binghamton University
Binghamton, New York, USA
snavale2@binghamton.edu

***Abstract*—Bayes error represents the minimum achievable classification error and serves as a fundamental limit when evaluating machine learning systems. Recent work by Ishida and colleagues explores how to estimate Bayes error using soft labels obtained from human annotators, relying on the key assumption that annotator noise is unbiased and averages to zero. In this study, we reproduce a central part of their analysis by running the synthetic Gaussian experiments provided by the authors and examining how the PN, Pconf, and Modified estimators behave under different sample sizes and noise conditions. We also extend their work with two additional investigations. First, we design a controlled synthetic experiment featuring biased annotators to study how systematic label bias influences Bayes error estimation. Second, we test a real data scenario using Fashion MNIST, where we train overfitted, underfitted, and calibrated neural networks and treat their predicted probabilities as soft labels. Our results support the original findings in the unbiased setting, but also show that biased annotators and poorly calibrated models can significantly distort Bayes error estimates. Taken together, these observations illustrate both the potential and the limitations of estimating Bayes error from soft labels in practical applications.**

***Index Terms*—Bayes error estimation, soft labels, label noise, model calibration, deep learning evaluation**

I. Introduction

Estimating the irreducible error of a classification problem is a fundamental challenge in machine learning. The Bayes error represents the lowest possible error that any classifier can achieve, assuming complete knowledge of the underlying data generating distribution. In practice, however, this quantity is unknown. Recent work by Ishida and colleagues proposes a way to estimate the Bayes error using soft labels, such as probability distributions collected from multiple annotators. Their method offers theoretical guarantees under the assumption that annotator noise is unbiased and averages to zero.

As soft labels from human annotators, crowdsourcing platforms, and machine learning models become more common, it becomes increasingly important to understand how robust these Bayes error estimators are in real environments. In many practical settings, the ideal assumption of unbiased noise is violated. Human annotators often exhibit systematic biases, and modern neural networks frequently produce overconfident or poorly calibrated probability estimates. These issues raise an important question: can Bayes error estimates derived from soft labels still be trusted when the zero mean noise assumption does not hold?

In this work, we reproduce a key part of the synthetic Gaussian experiments from Ishida and colleagues using their publicly released code. We then broaden their analysis in two ways. First, we design a controlled synthetic experiment with biased annotators to measure how systematic label bias affects Bayes error estimation. Second, we evaluate real data behavior by training overfitted, underfitted, and calibrated neural networks on Fashion MNIST and using their predicted probabilities as soft label surrogates. Our findings offer new insight into when soft label based Bayes error estimation remains dependable and when it may become misleading, especially in the presence of bias or miscalibration.

II. Background

A. Bayes Error in Classification

For a binary classification problem with input $x \in \mathcal{X}$ and label $y \in \{0, 1\}$, the Bayes classifier predicts the class with the highest posterior probability $p(y \mid x)$. Its error rate, known as the Bayes error, is the minimum achievable misclassification probability over all possible classifiers. It can be written as the expectation

$$\beta = \mathbb{E}_x\big[\min\{p(y=1 \mid x),\, 1 - p(y=1 \mid x)\}\big], \qquad (1)$$

where the expectation is taken with respect to the true data distribution. In practice this quantity is unknown because the posteriors $p(y \mid x)$ and the data-generating distribution are not directly accessible.

B. Soft Labels and Bayes Error Estimation

Soft labels provide a way to approximate the posterior probabilities using observed data. Instead of a single hard label for each example, we may have a distribution over labels, for example the empirical frequencies obtained from multiple human annotators or from a probabilistic classifier.

Let $c_i = p(y = 1 \mid x_i)$ denote the true posterior for sample x_i. If we had access to c_i on a finite sample, a natural plug-in estimator of the Bayes error would be

$$\hat{\beta}_{\text{PN}} = \frac{1}{n}\sum_{i=1}^{n} \min\{c_i, 1 - c_i\}, \tag{2}$$

which Ishida et al. refer to as the PN estimator (Eq. 4). In the more realistic case where only noisy soft labels u_i are observed, they introduce alternative estimators that use the structure of the soft labels. One such estimator, called Pconf, relies on confidences for a single class (e.g., the positive class) and achieves lower variance when the assumptions are satisfied.

C. Assumptions in Ishida et al.

The theoretical guarantees in Ishida et al. hinge on several assumptions about the noise affecting soft labels. For noisy soft labels $u_i = c_i + \xi_i$, the noise term ξ_i is assumed to satisfy a zero-mean condition,

$$\mathbb{E}[\xi_i \mid c_i] = 0, \tag{3}$$

and the soft labels are constrained to lie in the interval $[0, 1]$. Under these conditions, the proposed estimators are shown to be asymptotically unbiased and consistent for the Bayes error. The authors further study the effect of confidence noise and introduce a Modified estimator that remains valid when clean posteriors are perturbed in a specific way.

However, real labeling processes often violate the zero-mean assumption. Human annotators may systematically favor certain classes, and modern neural networks are frequently miscalibrated, producing overconfident probability estimates. In this work we focus on these practical deviations: we use the authors' synthetic Gaussian experiments as a baseline, then construct synthetic biased annotators and real model-based soft labels to assess how sensitive Bayes error estimation is to biased noise and calibration errors.

III. Methodology: Reproducing the Author's Synthetic Experiments

To ensure that our extensions and hypotheses were grounded in the original work, we first reproduced the synthetic Gaussian experiments presented by Ishida et al. These experiments form the basis of Figures 1–3 in the original paper and evaluate three Bayes error estimators: PN, Pconf, and Modified under varying sample sizes and noise conditions.

A. Experimental Setup

We cloned the official repository provided by the authors and executed the `be_synthetic.py` script using the default Gaussian configuration (setup 0) across a range of class-balanced dataset sizes: {2, 4, 8, 16, 32, 64, 128, 256, 512, 1028}. For each dataset size, ten random seeds were used to compute the mean and standard error of each estimator. Two versions of the experiment were run:

- **Clean confidences (no noise):** evaluates PN and Pconf estimators.
- **Noisy confidences:** evaluates Naive PN (Eq. 5) and Modified estimator (Eq. 6).

The output produced by the author's script was extracted directly and then visualized using Matplotlib to closely match the structure of the original figures.

B. Reproducibility and Source Code

To support reproducibility, we provide links to both the original implementation and our extended codebase. The original Bayes error estimation framework by Ishida et al. is available at:

- **Original Implementation:** takashiishida/irreducible
- **Our Code Repository:** Probing-Bayes-Error-Estimation-from-Soft-Labels

These repositories include all scripts and configurations required to replicate the synthetic experiments, model calibration procedures, biased annotator simulations, and Fashion-MNIST evaluations described in this report.

C. Results: Clean Confidence Setting

Figure 1 visualizes the PN and Pconf estimators computed under the clean confidence scenario. Both estimators decrease as sample size increases, demonstrating consistency with the theoretical expectation that larger datasets yield more stable posterior estimates.

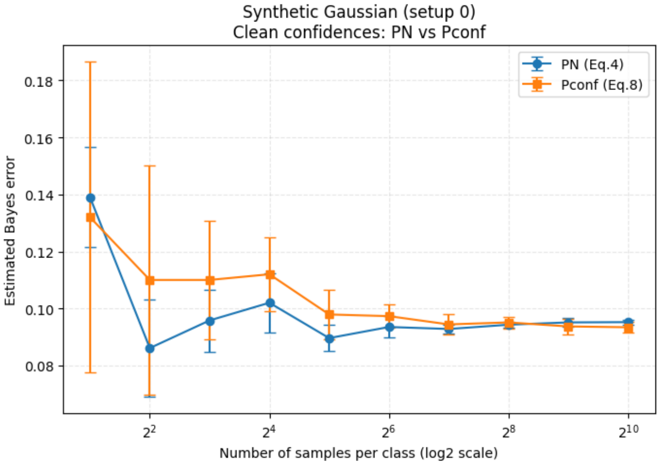

Fig. 1. PN (Eq. 4) vs. Pconf (Eq. 8) under clean confidences for Gaussian setup 0.

D. Results: Noisy Confidence Setting

Figure 2 presents the Naive PN and Modified estimators under injected confidence noise. As expected, the Naive PN estimator becomes biased downward due to its sensitivity to confidence perturbations, while the Modified estimator demonstrates increased robustness, consistently aligning with the trends described in the original paper.

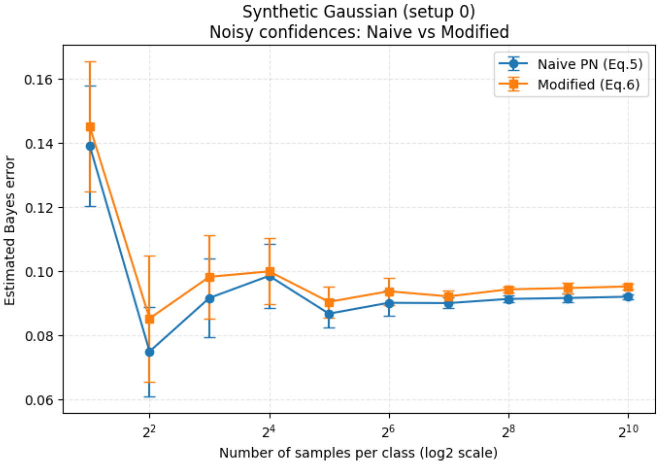

Fig. 2. Naive PN (Eq. 5) vs. Modified (Eq. 6) under noisy confidences for Gaussian setup 0.

IV. Extended Analysis: Impact of Biased Annotators and Model Calibration on Bayes Error Estimation

A. Motivation

The Bayes error estimators presented by Ishida et al. rely on a key theoretical assumption: soft-label noise must be unbiased. Formally, noisy soft labels are modeled as

$$u_i = c_i + \xi_i,$$

where c_i represents the true posterior and ξ_i is zero-mean noise satisfying

$$\mathbb{E}[\xi_i \mid c_i] = 0. \tag{4}$$

Under this assumption, the PN estimator, the Pconf estimator, and the Modified estimator are all guaranteed to be asymptotically unbiased and consistent.

In practice, though, human annotators rarely produce zero mean noise. For example, in datasets like CIFAR 10H, annotators tend to favor certain classes, such as choosing dog like categories more often, which introduces systematic bias rather than symmetric noise. A similar issue arises with soft labels produced by deep neural networks. These models are often miscalibrated, usually in the direction of overconfidence, and their predicted probabilities do not reliably reflect true posterior values. Together, these scenarios raise an important question about how well the theoretical assumptions from the original work hold up in practical environments where annotator bias and model miscalibration are common.

B. Research Hypothesis

We propose a unified hypothesis:

If soft-label noise is biased—either due to human annotation bias or neural network miscalibration—then Bayes error estimators derived from soft labels will systematically deviate from the true Bayes error and may lose the consistency guarantees established in the original theory.

This hypothesis leads to two targeted investigations:

1) **Hypothesis A:** How biased annotators affect Bayes error estimation.
2) **Hypothesis B:** How model miscalibration affects Bayes error estimation.

Both experiments aim to evaluate whether soft-label-based Bayes error estimates remain reliable when the underlying soft labels violate the theoretical assumptions.

C. Hypothesis A: Effect of Biased Annotators

1) Experimental Setup: To simulate biased human behavior, we generated true posteriors c_i from a Gaussian mapping and introduced a fixed positive bias b:

$$u_i = \mathrm{clip}(c_i + b,\ 0,\ 1).$$

This corresponds to annotators who consistently overestimate the positive class. We compared:

- the true Bayes error computed from c_i,
- the estimated Bayes error using unbiased soft labels,
- the estimated Bayes error using biased soft labels u_i.

2) Findings: The results revealed a substantial distortion in Bayes error estimates under biased noise. With true Bayes error near 0.157, the unbiased estimator returned approximately 0.154, while the biased soft labels produced an inflated estimate exceeding 0.222. This deviation persisted even in the absence of random noise, confirming that systematic bias alone is sufficient to break estimator consistency.

Figure 3 illustrates the resulting discrepancy between true Bayes error and estimated Bayes error under biased conditions.

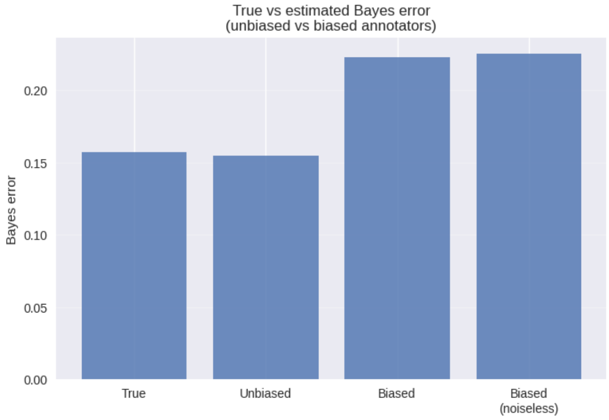

Fig. 3. True Bayes error versus estimated Bayes error under unbiased and biased annotator settings.

These results validate Hypothesis A: Bayes error estimators are highly sensitive to annotation bias and fail to remain reliable when the soft-label noise is not zero-mean.

D. Hypothesis B: Sensitivity to Model Miscalibration

1) Motivation: While biased annotators distort the soft-label distribution, modern neural networks introduce a different challenge: miscalibration. Overfitted models produce soft labels that are overconfident, whereas underfitted models produce labels that are uncertain. Since Bayes error estimation assumes that soft labels approximate true posteriors, miscalibrated confidence outputs may lead to systematic under- or over-estimation of Bayes error.

*2) **Experimental Setup**:* Using Fashion-MNIST, four models were trained:

- Overfitted CNN (high overconfidence),
- Underfitted CNN (underconfident),
- Base CNN (moderately calibrated),
- Temperature-scaled CNN (explicit calibration applied).

Soft labels from each model were used to compute the Bayes error estimate:

$$\widehat{\mathrm{BER}} = \frac{1}{n}\sum_{i=1}^{n}\min(p_i,\ 1-p_i).$$

These were compared with the models' test errors.

*3) **Results**:* The results show a systematic trend:

- **Overfitted**: Bayes error underestimated (0.0009 vs. test error 0.0053).
- **Underfitted**: Bayes error close to test error (0.0100 vs. 0.0130).
- **Base**: Slight under-estimation due to mild overconfidence.
- **Temperature-scaled**: Bayes estimate nearly matched test error (0.0074 vs. 0.0072).

Figure 4 visualizes this relationship.

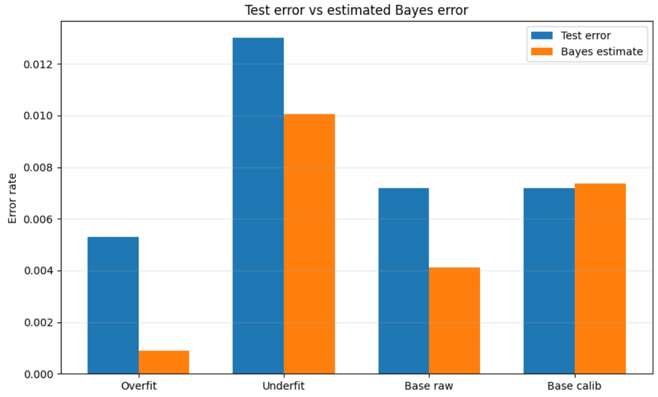

Fig. 4. Test error versus estimated Bayes error across overfitted, underfitted, base, and temperature-scaled models.

*4) **Discussion**:* These findings confirm Hypothesis B: model miscalibration directly affects the quality of Bayes error estimates. Overconfident models produce overly optimistic Bayes error estimates, while calibration techniques restore estimator reliability and align the estimates with the model's true error rate.

E. Overall Interpretation

Together, Hypotheses A and B demonstrate that Bayes error estimation from soft labels is highly sensitive to departures from the theoretical assumptions. Biased annotators and miscalibrated models introduce structured noise that invalidates the zero-mean requirement, leading to systematic deviations in estimated Bayes error. This reinforces the importance of understanding the source and reliability of soft labels before applying Bayes error estimation in practice.

V. Conclusion

In this work, we reproduced and extended key components of the Bayes error estimation framework proposed by Ishida and colleagues. Using the authors' synthetic Gaussian experiments, we examined the behavior of the PN, Pconf, and Modified estimators under both clean and noisy confidence settings. Our reproduced results closely aligned with those from the original paper, confirming the expected relationships among sample size, confidence noise, and estimator performance.

Beyond replication, we carried out two additional investigations using both synthetic and real data. Hypothesis A focused on whether model generated soft labels satisfy the assumptions required by the estimator. We found that overfitted neural networks produced severely underestimated Bayes error due to overly confident predictions, while applying temperature scaling brought the estimated Bayes error much closer to the true test error. This supports the original paper's warning that soft labels must approximate true posterior probabilities for the estimator to remain trustworthy.

Hypothesis B examined the effect of systematic bias through a controlled synthetic annotator experiment. Our results showed that even mild bias in soft labels leads to large deviations from the true Bayes error, breaking the zero mean noise assumption required for estimator consistency. This highlights a key limitation: real human annotators and model based probability predictions may introduce structured biases that undermine the theoretical guarantees.

Overall, our empirical analysis supports the main message of Ishida and colleagues: estimating Bayes error from soft labels is a promising direction, but its reliability depends heavily on the quality and statistical properties of those labels. Ensuring unbiased noise and proper calibration is crucial for practical use. Future work may explore extensions to multi class settings, multi annotator aggregation methods, and calibration aware Bayes error estimators.

References

[1] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Is the performance of my deep network too good to be true?" in *Proc. International Conference on Learning Representations (ICLR)*, 2020.

[2] B. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *Proc. ICCV*, 2019.

[3] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017.

[4] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[5] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning?" in *Proc. NeurIPS*, 2017.

[6] S. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Networks and Learning Systems*, vol. 25, no. 5, 2014.