Bandwidth Selection in Kernel Density Estimation for Model Calibration

Han Zhou[1] **Teodora Popordanoska**[1] **Matthew B. Blaschko**[1]

[1]Department ESAT, Center for Processing Speech and Images, KU Leuven, Leuven, Belgium

Abstract

As deep learning models are increasingly deployed in high-stakes applications, providing well-calibrated uncertainty estimates has become as critical as achieving high predictive accuracy. While Kernel Density Estimation (KDE) has emerged as a smooth and continuous alternative to traditional binning for quantifying miscalibration, its reliability is heavily dependent on the choice of the kernel bandwidth. Standard selection techniques, such as Maximum Likelihood Estimation (MLE), often fail to produce optimal bandwidths for calibration tasks. In this work, we introduce Risk Alignment (RA), a novel optimization framework that determines the optimal bandwidth by aligning KDE-reconstructed risk with empirical risk. We theoretically demonstrate that this alignment minimizes calibration estimation bias across the data distribution, establishing a principled bandwidth selection criterion applicable to various metrics, including the challenging case of canonical calibration error. Extensive experiments across multiple architectures and datasets show that RA consistently outperforms standard bandwidth selection methods, yielding more reliable calibration assessments. Our implementation is available at `https://github.com/han678/KDE-Bandwidth-Learning`.

1 INTRODUCTION

Over the past decade, deep learning has propelled neural networks to achieve, and occasionally exceed, human-level proficiency in domains ranging from image recognition [Haggenmüller et al., 2021] to natural language understanding [Achiam et al., 2023, Liu et al., 2024]. Despite these gains in predictive power, high accuracy does not inherently guarantee that a model's confidence scores are well-aligned with the true likelihood of its outcomes. This discrepancy, termed miscalibration [Guo et al., 2017], poses significant risks in high-stakes environments where reliable uncertainty quantification is as critical as the prediction.

To evaluate this alignment, various metrics have been proposed to quantify calibration error [Naeini et al., 2015, Guo et al., 2017, Ding et al., 2020, Popordanoska et al., 2022]. However, the inherent challenge in evaluating these metrics stems from our lack of access to the true conditional probability $R = \mathbb{E}[y \mid f(x)]$, which necessitates reliance on empirical approximations. The most prevalent approach, the binning method [Guo et al., 2017], partitions the unit interval into discrete segments to approximate the conditional probability. While straightforward, binning introduces artificial discontinuities and suffers from high variance, fundamentally limiting its estimation performance. Specifically, these estimators are often asymptotically inconsistent [Widmann et al., 2019] and highly sensitive to the partitioning scheme [Nixon et al., 2019]. Moreover, they suffer from the curse of dimensionality in multi-class settings, where exponential bin growth leads to severe data sparsity [Ding et al., 2020, Arrieta-Ibarra et al., 2022]. Recognizing these limitations, another line of research leverages Kernel Density Estimation (KDE) to provide a differentiable alternative [Popordanoska et al., 2022]. By employing a kernel function to aggregate information across the probability space, this approach yields a smooth representation of the conditional probability R, thereby circumventing the information loss inherent in discretization.

However, the reliability of these KDE-based estimators remains highly contingent on a critical tuning parameter: the kernel bandwidth [Buch-Larsen et al., 2005, Silverman, 2018, Blasiok and Nakkiran, 2024, Gruber and Bach, 2025]. While standard MLE methods [Duin, 1976] are optimal for general density estimation, they do not explicitly account for the estimation variance inherent in calibration metrics. In practice, the likelihood-based objective often favors narrow bandwidths that overfit to local sampling fluctuations

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:8270–8291.

rather than capturing the underlying calibration profile. This mismatch leads to noisy estimates that limit the reliability of the resulting calibration surface. In this work, we tackle the bandwidth selection challenge by proposing the Risk Alignment objective (RA) framework. Unlike density-centric methods, RA utilizes a principled optimization objective grounded in the decomposition of proper scoring rules to determine an optimal, variance-robust bandwidth. This objective is applicable across all notions of calibration metrics, including the notoriously difficult canonical calibration. Our contributions are as follows:

- We introduce a novel bandwidth selection framework for KDE-based calibration estimators in Section 5.1, formulated as a risk estimation alignment (RA) optimization objective.
- We provide a theoretical analysis of the bias-variance trade-off in calibration estimation, demonstrating that the RA objective recovers the underlying conditional probability by minimizing aggregate bias.
- We empirically demonstrate that our method consistently selects superior bandwidths compared to MLE, leading to better calibration estimation.

2 RELATED WORK

In this section, we provide an overview of existing literature on quantifying model calibration errors.

Binning-Based Estimation. A prominent line of work involves binning-based methods, such as the widely used Expected Calibration Error (ECE) [Guo et al., 2017], which estimates R by partitioning the prediction space into discrete intervals and averaging labels within each bin. Early techniques utilized equal-width or adaptive binning [Ding et al., 2020], but these suffer from sensitivity to the binning scheme [Kumar et al., 2019]. However, a fundamental limitation persists: these approaches treat R as a *piecewise constant* function, which introduces notable statistical bias [Roelofs et al., 2022] and renders the metric non-differentiable. This lack of smoothness precludes their integration into gradient-based optimization frameworks, preventing their utility as differentiable calibration regularizers during model training.

Kernel Density Estimation (KDE). Another line of work utilizes KDE to circumvent the limitations of discrete binning by providing smooth, differentiable approximations of conditional probability. While early approaches [Zhang et al., 2020] faced scalability issues in multi-class settings, recent methods have introduced specific kernel choices, such as RBF [Blasiok and Nakkiran, 2024] and Dirichlet kernel [Popordanoska et al., 2022, 2024], to estimate calibration error. However, these estimators remain highly sensitive to kernel bandwidth. Current practices often rely on maximum likelihood estimation (MLE) or manual visual inspection. This often leads to suboptimal bandwidth values that fail to generalize, causing the estimated conditional probability R to overfit to local sampling noise rather than capturing the true underlying probability across the data distribution. A recent advancement [Gruber and Bach, 2025] introduces an MSE-based risk framework to optimize calibration estimators by reformulating the task as a regression problem. However, this approach requires expensive cross-validation for its regularization parameter and entails $\mathcal{O}(n^3)$ complexity due to eigenvalue decompositions—limiting its scalability compared to the $\mathcal{O}(n^2)$ efficiency of KDE method [Popordanoska et al., 2024]. To address this, RA retains the KDE framework but derives a direct optimization objective from the bregman decomposition of the proper scoring rules. This approach remains naturally robust to sampling noise, achieving $\mathcal{O}(n^2)$ efficiency while bypassing the expensive cross-validation required by prior methods.

3 NOTATION AND PRELIMINARIES

Multiclass Classification. Consider a multiclass classification task with k classes. We aim to learn a classifier $f : \mathcal{X} \to \Delta^{K-1}$ that maps an input x from the feature space $\mathcal{X} \subseteq \mathbb{R}^d$ to the probability simplex $\Delta^{K-1} = \{f^{(k)}(x) \in [0,1] : \sum_{i=1}^{k} f^{(k)}(x_i) = 1\}$. The output vector $f(x)$ represents the predicted class probabilities. Let $y \in \{0,1\}^k$ be the one-hot encoded label, where (x, y) follows a joint distribution $\mathbb{P}(x, y)$.

The performance of a classifier is evaluated by the expected risk:

$$\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim\mathbb{P}(x,y)}[\ell(f(x), y)], \tag{1}$$

where ℓ is a loss function. A loss ℓ is called **proper** if the risk is minimized by the Bayes classifier $f^*(x) := \mathbb{E}[y|x]$ [Gneiting and Raftery, 2007].

Bregman Divergence. Every differentiable proper loss is uniquely associated with a Bregman divergence [Bregman, 1967]. Given a continuously differentiable and strictly convex function $F : \Delta^{K-1} \to \mathbb{R}$, the associated Bregman divergence is defined as:

$$D_F(p, q) := F(p) - F(q) - \langle \nabla F(q), p - q \rangle \tag{2}$$

which measures the discrepancy between the value of F at p and its linear approximation anchored at q. Notable special cases include the squared Euclidean distance $D_F(p, q) = \|p - q\|_2^2$ induced by the L_2 norm $F(p) = \|p\|_2^2$ and the Kullback–Leibler (KL) divergence $D_F(p, q) = \mathrm{KL}(p\|q)$ induced by the negative Shannon entropy $F(p) = \sum_{i=1}^{k} p_i \ln p_i$.

Bregman Decomposition. A well-known result [Bröcker, 2008, Kull and Flach, 2015, Ovcharov, 2015, Popordanoska et al., 2022, Gruber and Buettner, 2022a, Berta et al., 2025]

in the study of proper scoring rules is that the risk of a proper loss can be decomposed into calibration error and refinement. Let $R = \mathbb{E}[y|f(x)]$ denote the true conditional probability, the risk $\mathcal{R}(f)$ is decomposed as follows:

$$\mathcal{R}(f) = \underbrace{\mathbb{E}\left[D_F\big(R, f(x)\big)\right]}_{\mathrm{CE}_F(f)\text{ (Proper Calibration Error)}} + \underbrace{\mathbb{E}\left[F(y) - F\big(R\big)\right]}_{\mathrm{REF}_F(f)\text{ (Refinement)}} \tag{3}$$

where the calibration error measures how well the predicted probabilities match the true conditional probabilities; the refinement reflects the model's ability to *distinguish* samples. We note that the Generalized Entropy term $\mathbb{E}[F(Y)]$ [Gneiting and Raftery, 2007] is independent of the predictor f, so it often excluded for optimization and model comparison. In this framework, the proper calibration error (CE_F) can be related to the specific strictly convex function F that defines the corresponding Bregman divergence. One special case is the L_2 calibration error [Murphy, 1973], defined as:

$$\mathrm{CE}_2^2(f) = \mathbb{E}\left[\|R - f(x)\|_2^2\right]. \tag{4}$$

While canonical calibration assesses the model's overall reliability, it is often desirable to evaluate calibration at a per-class level, particularly in the presence of class imbalance. Under the L_2 loss, the class-wise calibration error is defined as the sum of squared binary calibration errors:

$$\mathrm{CWCE}_2^2(f) = \sum_{k=1}^{K} \mathbb{E}\left[\left(R^{(k)} - f^{(k)}(x)\right)^2\right]. \tag{5}$$

where $R^{(k)}$ and $f^{(k)}(x)$ denote the true and predicted probabilities for class k, respectively. This multi-class evaluation [Nixon et al., 2019, Panchenko et al., 2022, Popordanoska et al., 2024] can be done by decomposing the problem into K independent binary tasks via a one-versus-rest (OvR) approach. Under this decomposition, the expected class-wise risk is defined as the sum of these binary risks. Consequently, for any proper scoring rule induced by F, the class-wise risk admits an individual Bregman decomposition for each class, the details of which are provided in Section A.4.

4 KERNEL DENSITY BASED CALIBRATION ERROR ESTIMATOR

4.1 BREGMAN DECOMPOSITION

Based on the risk decomposition in Eq. (3), we can derive empirical estimators for calibration and refinement. Given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ and assuming that the true conditional probability R is known, then it leads to the the Bregman formulation of calibration error estimator [Gruber and Buettner, 2022b, Popordanoska et al., 2024]:

$$\widehat{\mathrm{CE}}_F(f) \approx \frac{1}{n}\sum_{i=1}^{n}\Big(F(R_i) - F(f(x_i)) - \langle \nabla F(f(x_i)), R_i - f(x_i)\rangle\Big) \tag{6}$$

where R_i is the true conditional probability at $f(x_i)$. Similarly, the refinement estimator is given by

$$\widehat{\mathrm{REF}}_F(f) = \frac{1}{n}\sum_{i=1}^{n}\big(F(y_i) - F(R_i)\big). \tag{7}$$

In particular, under the squared L_2 loss, the resulting calibration error estimator reduces to

$$\widehat{\mathrm{CE}}_2^2(f) = \frac{1}{n}\sum_{i=1}^{n}\|R_i - f(x_i)\|_2^2 \tag{8}$$

while for KL divergence, it takes the form

$$\widehat{\mathrm{CE}}_{\mathrm{KL}}(f) = \frac{1}{n}\sum_{i=1}^{n}\left\langle R_i, \log\left(\frac{R_i}{f(x_i)}\right)\right\rangle. \tag{9}$$

4.2 KERNEL DENSITY ESTIMATOR FOR R

In practice, the true conditional expectation R is unknown and must be inferred from observed data. The fidelity of any calibration error estimator—whether based on discrete binning or kernel-based methods—fundamentally relies on the accurate recovery of the conditional label distribution R, as the discrepancy between R and the predicted confidence $f(x)$ defines the calibration error. Popordanoska et al. [2022] propose a KDE approach that estimates R as a locally weighted average of observed labels, yielding a smooth, differentiable surface that reflects the underlying conditional label distribution.

Kernel Density Estimation. While the Beta kernel is typically employed for binary classification, the multi-class setting—where predictions reside on the probability simplex Δ^{K-1}—is naturally modeled with a density estimator based on the Dirichlet kernel, defined as [Popordanoska et al., 2022, 2024]:

$$k_{\mathrm{Dir}}(f(x_i), f(x_j); h) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_{jk})}{\prod_{k=1}^{K}\Gamma(\alpha_{jk})}\prod_{k=1}^{K} f(x_i)_k^{\alpha_{jk}-1}, \tag{10}$$

where $\alpha_j = \frac{f(x_j)}{h} + 1$ and $h > 0$ is the smoothing bandwidth. To ensure a consistent and asymptotically unbiased estimation of R_i at each observed point $f(x_i)$, we utilize a leave-one-out (LOO) estimator:

$$\hat{R}_i = \frac{\sum_{j\neq i} k_{\mathrm{Dir}}(f(x_i), f(x_j); h) y_j}{\sum_{j\neq i} k_{\mathrm{Dir}}(f(x_i), f(x_j); h)}. \tag{11}$$

By substituting $\hat{R}_i$ for the true conditional probability R_i in Eq. (6) and (7), we obtain a kernel density-based estimators for both calibration error and refinement. However, the fidelity of these estimators is intrinsically tied to the smoothing bandwidth h. This parameter dictates the balance between the local sensitivity and the global stability of the conditional probability manifold. Consequently, identifying an optimal h is not merely a numerical necessity but a prerequisite for ensuring that the empirical metrics reflect the true underlying calibration.

4.3 BANDWIDTH SELECTION

Choosing the bandwidth parameter h is a long-recognized and critical step [Loader, 1999, Popordanoska et al., 2022, Gruber and Bach, 2025], as it governs the capacity of the selected kernel to aggregate information from neighboring samples. Improperly tuned h can lead to either over-smoothing (high bias) or excessive volatility (high variance) in the conditional probability estimate.

Log-likelihood Maximization (MLE). A standard heuristic for bandwidth selection, adopted by Popordanoska et al. [2022], is the maximization of the leave-one-out (LOO) log-likelihood. This nonparametric method focuses on the generative probability of the features in the prediction space. Given bandwidth h, the LOO density estimate at $f(x_i)$ is defined as

$$\widehat{p}_{-i}(f(x_i); h) = \frac{1}{n-1} \sum_{j \neq i} k_{\text{Dir}}(f(x_i), f(x_j); h)\,. \quad (12)$$

The optimal bandwidth $h^\star$ is then sought by maximizing the aggregate log-likelihood:

$$\mathcal{L}_{\text{MLE}}(h) = \sum_{i=1}^{n} \log \widehat{p}_{-i}(f(x_i); h)\,. \quad (13)$$

While MLE is effective for general density estimation, it fails in calibration due to a critical loss of generality. As the sample size n increases, the MLE objective favors pathologically small bandwidths (See Fig. 1a). By pursuing local density maximization, MLE selects infinitesimal h that causes the kernel to contract around individual data points, which often leads to overfitting. Consequently, the estimator $\hat{R}_i$ degenerates from the true conditional probability into a nearest-neighbor label sampler. Instead of recovering a smooth, generalizable reliability curve, it produces sharp spikes that capture stochastic sampling noise. As derived in Section 5.2, this overfitting creates a "variance-bias trap": the excessive estimation variance is misattributed to the calibration error estimator, yielding a persistent positive bias.

5 CALIBRATION-ORIENTED BANDWIDTH SELECTION

We revisit bandwidth optimization through a calibration-oriented lens by utilizing the decomposition of proper scoring rules. We show that the estimated risk bias is aligned with the conditional probability bias through a scaling factor related to the predicted probability. This insight allows us to develop Risk Alignment (RA), a framework that identifies the optimal bandwidth h by ensuring consistency between the KDE-reconstructed risk and the empirical risk observed from samples.

5.1 RISK ALIGNMENT

For the binary case under the L_2 risk (Brier Score) with predicted confidence $f(x) \in [0, 1]$ and label $y \in \{0, 1\}$, let $R = \mathbb{E}[y|f(x)]$ denote the true conditional probability. The expected L_2 risk has the form:

$$\mathcal{R}_{L_2}(f) = \underbrace{\mathbb{E}[(R - f(x))^2]}_{\text{CE}_2^2(f)} + \underbrace{\mathbb{E}[R(1-R)]}_{\text{REF}_2(f)}\,. \quad (14)$$

In practice, as R is inaccessible, we substitute it with the LOO-KDE estimator $\hat{R}_i$ from Eq. (11). For a given h, we define the *pointwise KDE-reconstructed risk* as:

$$\hat{r}_{\text{kde},i}(h) = \widehat{\text{ce}}_i(h) + \widehat{\text{ref}}_i(h), \quad (15)$$

where

$$\widehat{\text{ce}}_i(h) = (\hat{R}_i - f(x_i))^2, \quad \widehat{\text{ref}}_i(h) = \hat{R}_i(1 - \hat{R}_i)$$

denote the sample-wise calibration and refinement estimators, respectively. Correspondingly, let $r_i = (y_i - f(x_i))^2$ be the observed risk at sample i. We use lowercase notation to denote pointwise quantities.

Algorithm 1 Risk-based Bandwidth Selection

1: **Input:** Dataset $\mathcal{D} = \{(f_i, y_i)\}_{i=1}^n$, grid of candidates $\mathcal{H} = \{h_1, h_2, \ldots, h_m\}$
2: **Output:** Optimal bandwidth h^*
3: Compute the point-wise risk r_i
4: **for** each candidate $h \in \mathcal{H}$ **do**
5: Compute the LOO estimates $\widehat{R}_i$ using Eq. (11)
6: Evaluate $\mathcal{L}_{\text{RA}}(h)$ in Eq. (16)
7: **end for**
8: **Return:** $h^* = \arg\min_{h \in \mathcal{H}} \mathcal{L}_{\text{RA}}(h)$

Risk alignment (RA) objective. In this work, we propose an objective to recover R by minimizing the aggregate discrepancy between the reconstructed and observed risks:

$$\mathcal{L}_{\text{RA}}(h) = \sum_{i=1}^{n} |\hat{r}_{\text{kde},i}(h) - r_i|^p\,. \quad (16)$$

By aligning the KDE-reconstructed risk with the empirical ground truth, this framework facilitates bandwidth selection conducive to out-of-sample generalization. Unlike the MLE criterion—which is prone to the "variance-bias trap" and tends to select pathologically small bandwidths by overfitting local density fluctuations—this risk-minimization strategy favors a more conservative (larger) bandwidth. This leads to a marked reduction in estimation variance and, crucially, effectively mitigates the calibration bias induced by high-variance fluctuations (see Fig. 1d). As we will demonstrate in the following analysis, by suppressing these variance-related artifacts, RA ensures a consistent estimation of R across the entire prediction space.

5.2 DECOMPOSITION OF ESTIMATION BIAS

To reveal the internal structural imbalance of risk-based alignment, we analyze the bias of the reconstructed risk at each individual sample x_i. For calibration error, we decompose its expectation of by centering $\hat{R}_i$ around the oracle value R_i:

$$\mathbb{E}[\widehat{\text{ce}}_i] = \mathbb{E}[((\hat{R}_i - R_i) + (R_i - f(x_i)))^2] \tag{17}$$

Using the identity $\mathbb{E}[(\hat{R}_i - R_i)^2] = \text{Var}(\hat{R}_i) + (\mathbb{E}[\hat{R}_i] - R_i)^2$, the local calibration estimation bias relative to the oracle value $(R_i - f(x_i))^2$ is:

$$\begin{aligned}\text{Bias}(\widehat{\text{ce}}_i) =& 2(R_i - f(x_i))(\mathbb{E}[\hat{R}_i] - R_i) + \text{Var}(\hat{R}_i) \\ &+ (\mathbb{E}[\hat{R}_i] - R_i)^2.\end{aligned} \tag{18}$$

Since the bias of the KDE estimator $\hat{R}_i$ is known to be $\mathcal{O}(h^2)$ [Dehnad, 1987, Wand and Jones, 1994], the squared bias term $(\mathbb{E}[\hat{R}_i] - R_i)^2$ is of order $\mathcal{O}(h^4)$. Neglecting this higher-order term, we obtain:

$$\text{Bias}(\widehat{\text{ce}}_i) \approx 2(R_i - f(x_i))(\mathbb{E}[\hat{R}_i] - R_i) + \text{Var}(\hat{R}_i). \tag{19}$$

Similarly, for the refinement component $\widehat{\text{ref}}_i = \hat{R}_i(1 - \hat{R}_i)$, we analyze its expectation by utilizing the identity $\mathbb{E}[\hat{R}_i^2] = \text{Var}(\hat{R}_i) + (\mathbb{E}[\hat{R}_i])^2$:

$$\mathbb{E}[\widehat{\text{ref}}_i] = \mathbb{E}[\hat{R}_i] - \mathbb{E}[\hat{R}_i^2] = \mathbb{E}[\hat{R}_i] - \left(\text{Var}(\hat{R}_i) + (\mathbb{E}[\hat{R}_i])^2\right).$$

Subtracting the oracle refinement value $R_i(1 - R_i)$, the bias is given by:

$$\text{Bias}(\widehat{\text{ref}}_i) = (\mathbb{E}[\hat{R}_i] - R_i) - \left((\mathbb{E}[\hat{R}_i])^2 - R_i^2\right) - \text{Var}(\hat{R}_i).$$

Applying the identity $(\mathbb{E}[\hat{R}_i])^2 - R_i^2 = (\mathbb{E}[\hat{R}_i] - R_i)(\mathbb{E}[\hat{R}_i] + R_i)$ and substituting $\mathbb{E}[\hat{R}_i] + R_i = 2R_i + (\mathbb{E}[\hat{R}_i] - R_i)$, we expand the term:

$$\begin{aligned}(\mathbb{E}[\hat{R}_i])^2 - R_i^2 &= (\mathbb{E}[\hat{R}_i] - R_i)(2R_i + \mathbb{E}[\hat{R}_i] - R_i) \\ &= 2R_i(\mathbb{E}[\hat{R}_i] - R_i) + (\mathbb{E}[\hat{R}_i] - R_i)^2.\end{aligned}$$

Substituting this back into the refinement bias yields:

$$\begin{aligned}\text{Bias}(\widehat{\text{ref}}_i) =& (1 - 2R_i)(\mathbb{E}[\hat{R}_i] - R_i) - \text{Var}(\hat{R}_i) \\ &- (\mathbb{E}[\hat{R}_i] - R_i)^2.\end{aligned} \tag{20}$$

Again, neglecting the $\mathcal{O}(h^4)$ term, we obtain the refinement bias:

$$\text{Bias}(\widehat{\text{ref}}_i) \approx (1 - 2R_i)(\mathbb{E}[\hat{R}_i] - R_i) - \text{Var}(\hat{R}_i). \tag{21}$$

Equations (19) and (21) highlight a critical trade-off: individual estimators are contaminated by the variance of the estimator of R. In the calibration term, this variance acts as a spurious positive bias, whereas in the refinement term, it manifests as a negative bias. Summing these components leads to the exact cancellation of the second-order terms, yielding the risk bias:

$$\text{Bias}(\hat{r}_{\text{kde},i}) \approx w_i(\mathbb{E}[\hat{R}_i] - R_i), \tag{22}$$

where $w_i = 1 - 2f(x_i)$ for the binary L_2 case. Thus the bias of the aggregate reconstructed risk in Eq. (15) is given by:

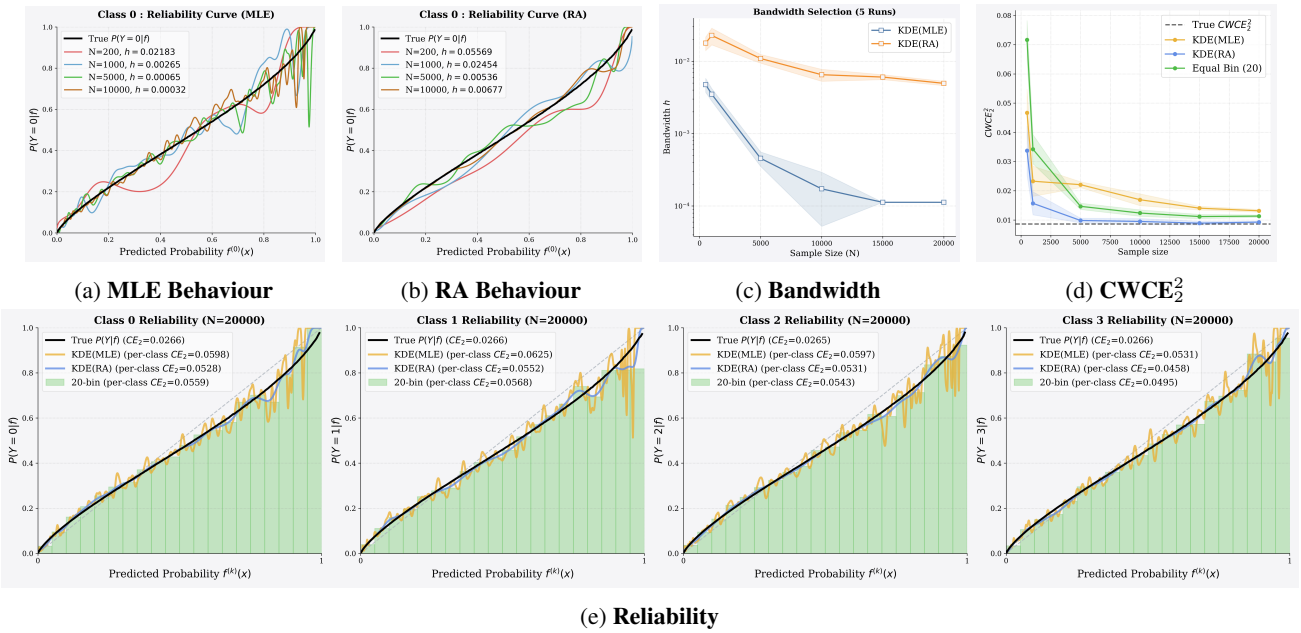
$$\text{Bias}\left(\sum_{i=1}^{n} \hat{r}_{\text{kde},i}\right) \approx \sum_{i=1}^{n} w_i(\mathbb{E}[\hat{R}_i] - R_i). \tag{23}$$

This internal cancellation of variance terms is central to the proposed risk alignment objective. We demonstrate that this property generalizes across various classification settings and proper scoring rules in Appendix A. The functional forms of the resulting weights w_i are summarized in Table 1, with complete derivations provided in the Appendix.

Table 1: Weights w_i for various proper scoring rules.

Scoring Rule	**Setting**	**Weight** (w_i)
Brier Score (L_2)	Binary	$1 - 2f(x_i)$
	Multi-class	$-2f^{(k)}(x_i)$
	Class-wise	$1 - 2f^{(k)}(x_i)$
Log Loss (KL)	Binary	$\log \frac{1-f(x_i)}{f(x_i)}$
	Multi-class	$-\log f^{(k)}(x_i)$
	Class-wise	$\log \frac{1-f^{(k)}(x_i)}{f^{(k)}(x_i)}$

Variance-Bias Trap. The RA objective exploits the structural stability of the total risk to identify an optimal bandwidth. While MLE is a standard criterion for density estimation, it frequently fails in calibration because its objective is maximized when the estimated density peaks at observed sample locations. As n increases, this causes the kernel to contract around individual data points, where the KDE estimator $\hat{R}_i$ degenerates into a nearest-neighbor label sampler. Although this fits local density, it maximizes



(a) **MLE Behaviour** (b) **RA Behaviour** (c) **Bandwidth** (d) **CWCE$_2^2$**

(e) **Reliability**

Figure 1: Bandwidth selection behavior on synthetic data. 1a and 1b illustrates the results of $\hat{R}_i$ using the bandwidth h selected by MLE or RA objective. While MLE is designed for density recovery, it tends to favor pathologically smaller bandwidths compared with RA objective. This preference leads to the behavior shown in the reliability plots (1a, 1b and 1e), where the MLE-selected bandwidth causes the estimated bias of $\hat{R}_i$ to manifest as sharp spikes—reflecting overfitting—rather than the robust, smooth distribution produced by our Risk-based approach. Consequently, as demonstrated in 1d, this poorly tuned bandwidth results in suboptimal calibration error estimation.

the estimation variance $\mathrm{Var}(\hat{R}_i)$. As revealed by Eq. (19), the individual calibration estimator $\widehat{\mathrm{ce}}_i$ is biased upward by this variance, which explains the failure of MLE-selected bandwidths. In contrast, Eq. (22) demonstrates that the total risk $\hat{r}_{\mathrm{kde},i}$ remains robust because $\mathrm{Var}(\hat{R}_i)$ is neutralized by the refinement component $\widehat{\mathrm{ref}}_i$ through internal cancellation. This mechanism explains the stability of classical binning: by averaging local labels, it acts as a primitive low-pass filter that mitigates estimation variance, albeit while introducing discretization bias. By utilizing the RA objective, we achieve a similar suppression of local stochastic noise while preserving the smoothness of the estimator. As shown in Eq. (23), the aggregate risk bias remains coupled with the estimation bias of R_i only through the weight w_i. By decoupling the bias from variance terms, the RA framework circumvents the bias-variance tradeoff and recovers the underlying calibration mapping, as illustrated in Fig. 1b.

6 EXPERIMENTS

6.1 SYNTHETIC DATA

Following Popordanoska et al. [2022], we evaluate different calibration error estimators on synthetic data where the oracle $R_i^\star = \mathbb{E}[y \mid f(x_i)]$ is known by design, allowing for exact calibration error quantification.

Data generation. We simulate $N = 20,000$ i.i.d. samples for K classes. For each instance, a latent vector $u_i \in \Delta^{K-1}$ is sampled uniformly via the Kraemer algorithm [Smith and Tromble, 2004]. We define the oracle distribution as $R_i^\star = \sigma(\log u_i / t_1)$, from which labels are drawn $y_i \sim \mathrm{Categorical}(R_i^\star)$ to ensure $\mathbb{E}[y \mid f(x)] = R^\star$. Miscalibration is then induced by setting the model output to $f(x_i) = \mathrm{softmax}(\log R_i^\star / t_2)$. We set $t_1 = 1.0$ and $t_2 = 0.8$ throughout.

Ground-truth calibration error. Since $R^\star$ and $f(x)$ are known, the ground-truth canonical calibration error (Eq. (3)) can be computed over the dataset as $\mathrm{CE}_F^\star(f) = \frac{1}{N}\sum_{i=1}^{N} D_F\big(R_i^\star, f(x_i)\big)$. Analogously, the ground-truth classwise calibration error $\mathrm{CWCE}_F^\star(f)$ is evaluated by applying the same principle across all classes.

Baselines. We evaluate KDE-based calibration error estimators through different bandwidth selection strategies, such as MLE and our RA method. We set $p = 2$ for RA objective in Eq. (16). Both employ a grid search across the same domain $\mathcal{H}$ to ensure a fair comparison. Specifically, the bandwidth h is optimized per class in classwise settings, whereas in canonical settings, a single global bandwidth is optimized. For binning-based baselines, we adopt the joint simplex binned estimator for canonical calibration [Popordanoska et al., 2022], which partitions the $(K-1)$-dimensional probability simplex into a fixed-width

grid to approximate the oracle $R_i^\star$. Consistent with Popordanoska et al. [2022], we employ leave-one-out (LOO) unbiased estimation within each partition. For the L_2 setting, we additionally compare against the Kernel Ridge Regression-based (KRR) estimator [Gruber and Bach, 2025]. For classwise calibration, we adopt the adaptive binning scheme [Ding et al., 2020], which adjusts bin edges based on class-specific confidence distributions.

Results. Fig. 2 (Appendix Fig. 4) and Fig. 3 (Appendix Fig. 5) illustrate the performance of the classwise estimator $\widehat{\mathrm{CWCE}}_F(f)$ and canonical estimator $\widehat{\mathrm{CE}}_F(f)$, respectively, as a function of n. These results are commputed over 10 subsamples for scenarios involving 4, 8, 16, and 32 classes. As shown in Fig. 2 and 4, the KDE estimator utilizing the RA objective achieves the best performance compared to all existing methods in the classwise settings. Furthermore, the reliability plots in Fig. 1e demonstrate that our proposed method aligns more closely with the ground truth $R^\star$ than either the MLE-based or binning approaches. As illustrated in Fig. 3 and 5, binning methods demonstrate poor performance in canonical cases, significantly underperforming KDE methods. Despite outperforming MLE-based KDE, KDE-based estimators are sensitive to the increasing dimensionality of the probability simplex in the canonical setting. As the number of classes K grows, the resulting data sparsity on the high-dimensional simplex makes it increasingly difficult for the Dirichlet kernel to accurately fit the underlying distribution. Thus KDE methods become less competitive than KKR in that setting.

6.2 REAL-WORLD DATA

Datasets. Evaluation is conducted on image classification benchmarks, including CIFAR-10/100 [Krizhevsky, 2009] and ImageNet [Deng et al., 2009], alongside the Amazon Reviews text dataset [Ni et al., 2019]. The Amazon dataset comprises review text paired with five-star ordinal ratings as labels. For CIFAR-10/100, we aggregate the training and test sets to serve as the population, whereas for ImageNet, we utilize the validation set.

Models. For experiments on CIFAR10/100, we report the results on the VGG16 [Simonyan and Zisserman, 2014] model with batch norm layers, WideResNet28x10 [Zagoruyko and Komodakis, 2016], and ResNet [He et al., 2016] models with different depths $(20, 56)$. We evaluate the calibration error estimators using models that are pre-trained on the CIFAR10/100 dataset. For experiments on Amazon dataset, we use pre-trained transformer-based models – BERT [Kenton and Toutanova, 2019], RoBERTa [Liu et al., 2021b], Distill-Bert (D-BERT), and Distill-Roberta (D-RoBERTa) [Sanh, 2019]. For experiments on the ImageNet dataset, we use the pre-trained models from *timm* [Wightman, 2019] package, including two Vision Transformer (ViT) variants [Dosovitskiy et al., 2020], ViT-Small and ViT-Base, along with a Swin Transformer-Small [Liu et al., 2021a] and a Data-efficient Image Transformer (DeiT-Base) [Touvron et al., 2021].

Proxy ground-truth calibration error. For real datasets, the ground truth cannot be directly derived and must be inferred. We therefore adopted the best-performing estimators identified from our synthetic experiments—KRR for canonical L_2 error and RA for all other metrics. Crucially, we computed these proxies over the largest available data pools (e.g., the combined training and test sets for CIFAR-10/100). The theoretical validity of our KDE method is supported by Ouimet and Tolosana-Delgado [2022], who establishes that Dirichlet KDE achieves optimal convergence under specific bandwidth constraints. Satisfying these constraints depends heavily on dimensionality. At our population scale, 1D classwise calibration strictly meets these conditions, yielding a mathematically robust proxy. Conversely, because canonical calibration in high-dimensional spaces (e.g., 1,000 ImageNet classes) is statistically prohibitive due to the curse of dimensionality, we exclude it from our ImageNet evaluation.

Experimental setup. To assess estimator performance across varying sample sizes, we report the mean and 95% confidence intervals calculated over 10 independent subsets. Additionally, the standard deviations are derived from those subsets by measuring the absolute error of each estimate relative to the ground truth. Additional results can be found in Appendix B.

Results. Consistent with our synthetic findings, the RA method demonstrate a clear superiority over the MLE-based approach, particularly in estimating class-wise calibration error as shown in Table 2. In contrast, among binning techniques, the adaptive binning method underperforms the equal-width binning method in class-wise settings. This underperformance is likely due to the overconfidence of pre-trained networks; the resulting concentration of probabilities near 1 forces adaptive binning to create extremely narrow bins at the extremes and overly coarse bins in the middle. Consequently, this non-uniform resolution fails to capture calibration gaps in sparse regions, whereas equal-width binning maintains a stable and representative resolution across the entire unit interval.

7 DISCUSSION AND CONCLUSION

In this paper, we revisited the fundamental challenge of bandwidth selection for KDE-based calibration error estimators. While KDE offers a differentiable and smooth alternative to traditional binning, its reliability is notoriously sensitive to the kernel bandwidth. The MLE objective often collapses toward pathologically small bandwidths, causing the estimator to overfit to local noise.

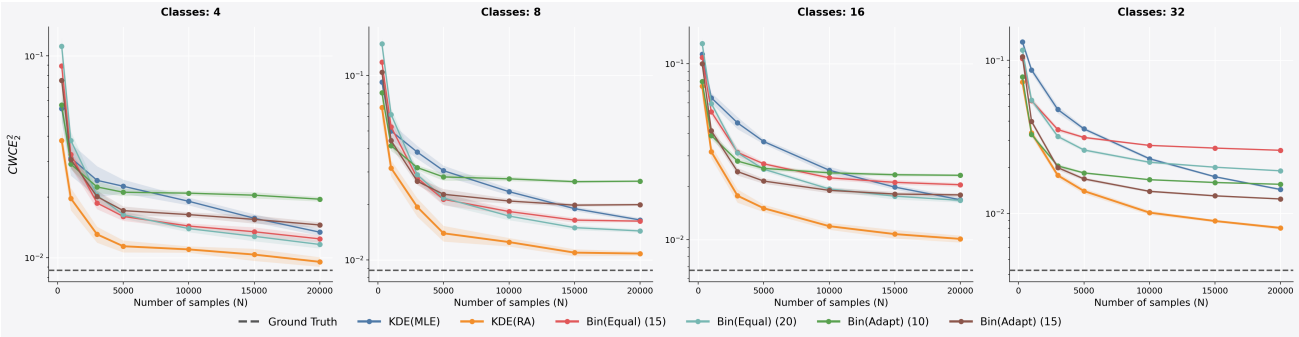


Figure 2: (**Synthetic data**) Finite sample estimator of CWCE_2^2: mean and 95% confidence interval computation across all subsamples of synthetic data.

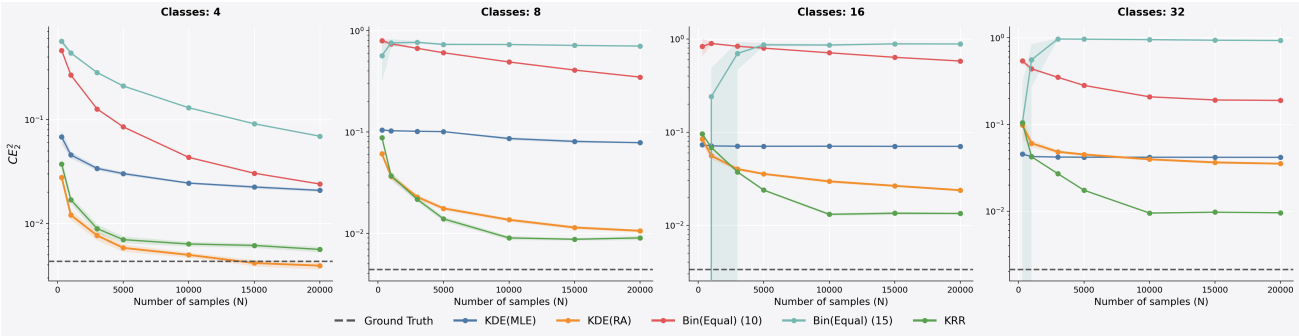


Figure 3: (**Synthetic data**) Finite sample estimator of CE_2^2: mean and 95% confidence interval computation across all subsamples of synthetic data.

Table 2: Summary of MAE of $\text{CWCE}_2^2 \times 10^2$ (↓) with standard deviations. Best values are in red, and second-best in teal. Results are reported at $N = 20,000$ for CIFAR-10/100, Amazon and ImageNet.

Dataset	**Model**	**Equal Binning**		**Adaptive Binning**		**KDE Methods**	
		15	**20**	**10**	**15**	**MLE**	**RA**
CIFAR10	PreResNet-20	$0.19_{\pm 0.03}$	$0.28_{\pm 0.05}$	$1.98_{\pm 0.58}$	$30.61_{\pm 0.09}$	$1.15_{\pm 0.08}$	$0.07_{\pm 0.05}$
	PreResNet-56	$0.19_{\pm 0.05}$	$0.30_{\pm 0.05}$	$1.98_{\pm 0.70}$	$31.44_{\pm 0.08}$	$1.10_{\pm 0.09}$	$0.08_{\pm 0.04}$
	VGG-16 (BN)	$0.06_{\pm 0.03}$	$0.16_{\pm 0.04}$	$1.99_{\pm 0.66}$	$31.92_{\pm 0.06}$	$1.15_{\pm 0.07}$	$0.10_{\pm 0.08}$
	WRN-28-10	$0.20_{\pm 0.02}$	$0.25_{\pm 0.02}$	$1.80_{\pm 0.37}$	$30.81_{\pm 0.08}$	$0.88_{\pm 0.03}$	$0.08_{\pm 0.02}$
CIFAR100	PreResNet-20	$1.55_{\pm 0.11}$	$2.15_{\pm 0.09}$	$129.58_{\pm 0.26}$	$119.85_{\pm 0.26}$	$9.03_{\pm 0.16}$	$0.85_{\pm 0.08}$
	PreResNet-56	$1.57_{\pm 0.08}$	$2.03_{\pm 0.09}$	$167.13_{\pm 0.17}$	$157.14_{\pm 0.17}$	$5.78_{\pm 0.09}$	$0.85_{\pm 0.08}$
	VGG-16 (BN)	$0.36_{\pm 0.06}$	$0.83_{\pm 0.06}$	$172.27_{\pm 0.11}$	$162.28_{\pm 0.12}$	$4.76_{\pm 0.15}$	$0.82_{\pm 0.14}$
	WRN-28-10	$1.41_{\pm 0.08}$	$1.87_{\pm 0.05}$	$171.58_{\pm 0.14}$	$161.60_{\pm 0.14}$	$4.79_{\pm 0.17}$	$0.61_{\pm 0.10}$
AMAZON	Bert	$2.07_{\pm 0.45}$	$1.83_{\pm 0.42}$	$4.83_{\pm 0.47}$	$3.38_{\pm 0.45}$	$0.85_{\pm 0.50}$	$0.38_{\pm 0.29}$
	Distill Bert	$1.18_{\pm 0.31}$	$1.04_{\pm 0.33}$	$3.93_{\pm 0.29}$	$2.57_{\pm 0.33}$	$0.99_{\pm 0.32}$	$0.23_{\pm 0.25}$
	Roberta	$0.31_{\pm 0.25}$	$0.23_{\pm 0.20}$	$3.80_{\pm 0.22}$	$2.27_{\pm 0.26}$	$0.95_{\pm 0.30}$	$0.27_{\pm 0.15}$
	Distill Roberta	$1.65_{\pm 0.26}$	$1.43_{\pm 0.26}$	$4.23_{\pm 0.34}$	$2.96_{\pm 0.30}$	$1.10_{\pm 0.30}$	$0.23_{\pm 0.15}$
IMAGENET	ViT-Small	$11.24_{\pm 0.26}$	$14.16_{\pm 0.29}$	$125.55_{\pm 0.20}$	$124.56_{\pm 0.20}$	$34.03_{\pm 0.21}$	$5.45_{\pm 0.35}$
	ViT-Base	$9.86_{\pm 0.19}$	$12.32_{\pm 0.22}$	$121.47_{\pm 0.22}$	$120.62_{\pm 0.22}$	$28.76_{\pm 0.34}$	$4.27_{\pm 0.24}$
	Swin-Small	$9.30_{\pm 0.11}$	$11.68_{\pm 0.16}$	$117.68_{\pm 0.19}$	$116.85_{\pm 0.19}$	$27.11_{\pm 0.36}$	$4.29_{\pm 0.24}$
	DeiT-Base	$9.08_{\pm 0.15}$	$11.48_{\pm 0.19}$	$114.44_{\pm 0.21}$	$113.68_{\pm 0.21}$	$27.22_{\pm 0.23}$	$3.77_{\pm 0.06}$

To address this, we introduce Risk Alignment (RA) objective, a novel optimization framework grounded in the Bregman decomposition of proper scoring rules. Our theoretical analysis demonstrates that while the biases of individual calibration and refinement estimators are coupled with the estimation variance of the conditional probability R, their sum—the total risk—enables a systematic variance cancellation. This property ensures that the RA objective is decoupled from high-variance artifacts and is primarily driven by the bias of the conditional probability, scaled by a confidence-dependent weight. By overcoming the inherent variance-bias tradeoff in calibration estimation, RA facilitates a robust recovery of the conditional probability manifold. Extensive experiments show that RA consistently outperforms MLE in calibration estimation, particularly in class-wise settings.

Despite its advantages, our study identifies several remaining challenges. Experiments show that KDE methods, including RA, can degrade in extremely high-dimensional settings, such as canonical calibration with many classes, primarily due to data sparsity on the simplex. In these regimes, the Dirichlet kernel may be less effective than the RBF kernel, suggesting that while RA optimizes bandwidth, the choice of kernel remains critical. Since RA is grounded in Bregman decomposition, its objective can be readily adapted to alternative kernels like RBF to mitigate high-dimensional issues. Furthermore, integrating RA-optimized KDE as a differentiable loss term is a promising direction. By providing stable and less biased gradients, RA could facilitate more effective calibration-aware training.

ACKNOWLEDGMENTS

This research received funding from the Flemish Government (AI Research Program) and the Research Foundation - Flanders (FWO) through project number G0G2921N. HZ is supported by the China Scholarship Council. We acknowledge EuroHPC JU for awarding the project ID EHPC-DEV-2025D06-055 and EHPC-DEV-2025D12-054 access to the EuroHPC supercomputer LEONARDO, hosted by CINECA (Italy) and the LEONARDO consortium.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Imanol Arrieta-Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 23(351):1–54, 2022.

Eugene Berta, David Holzmüller, Michael I Jordan, and Francis R Bach. Rethinking early stopping: Refine, then calibrate. *CoRR*, 2025.

Jaroslaw Blasiok and Preetum Nakkiran. Smooth ece: Principled reliability diagrams via kernel smoothing. In *The Twelfth International Conference on Learning Representations*, 2024.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(67)90040-7. URL https://www.sciencedirect.com/science/article/pii/0041555367900407.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 134(632):1529–1535, 2008. doi: 10.1002/qj.301.

Tine Buch-Larsen, Jens Perch Nielsen, Montserrat Guillén, and Catalina Bolancé. Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6):503–516, 2005.

Khosrow Dehnad. Density estimation for statistics and data analysis, 1987.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Yukun Ding, Jinglan Liu, Jinjun Xiong, and Yiyu Shi. Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4–5, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

Duin. On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Transactions on computers*, 100(11):1175–1179, 1976.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Sebastian Gruber and Francis Bach. Optimizing estimators of squared calibration errors in classification. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=BPDVZajOW5.

Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8618–8632. Curran Associates, Inc., 2022a.

Sebastian Gruber and Florian Buettner. Better uncertainty calibration via proper scores for classification and beyond. *Advances in Neural Information Processing Systems*, 35: 8618–8632, 2022b.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Sarah Haggenmüller, Roman C Maron, Achim Hekler, Jochen S Utikal, Catarina Barata, Raymond L Barnhill, Helmut Beltraminelli, Carola Berking, Brigid Betz-Stablein, Andreas Blum, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer*, 156:202–216, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

Meelis Kull and Peter Flach. Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15*, pages 68–85. Springer, 2015.

Ananya Kumar, Percy S Liang, and Tengyu Ma. Verified uncertainty calibration. *Advances in neural information processing systems*, 32, 2019.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv e-prints*, pages arXiv–2412, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021a.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer, 2021b.

Clive Loader. Local regression and likelihood. *Statistics and Computing*, 1999.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

Frédéric Ouimet and Raimon Tolosana-Delgado. Asymptotic properties of dirichlet kernel density estimators. *Journal of Multivariate Analysis*, 187:104832, 2022.

Evgeni Y Ovcharov. Existence and uniqueness of proper scoring rules. *J. Mach. Learn. Res.*, 16:2207–2230, 2015.

Michael Panchenko, Anes Benmerzoug, and Miguel de Benito Delgado. Class-wise and reduced calibration methods. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1093–1100. IEEE, 2022.

Teodora Popordanoska, Raphael Sayer, and Matthew Blaschko. A consistent and differentiable lp canonical calibration error estimator. *Advances in Neural Information Processing Systems*, 35:7933–7946, 2022.

Teodora Popordanoska, Sebastian Gregor Gruber, Aleksei Tiulpin, Florian Buettner, and Matthew B Blaschko. Consistent and asymptotically unbiased estimation of proper calibration errors. In *International Conference on Artificial Intelligence and Statistics*, pages 3466–3474. PMLR, 2024.

Rebecca Roelofs, Nicholas Cain, Jonathon Shlens, and Michael C Mozer. Mitigating bias in calibration error estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 4036–4054. PMLR, 2022.

V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*, 2019.

Bernard W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Noah A Smith and Roy W Tromble. Sampling uniformly from the unit simplex. *Johns Hopkins University, Tech. Rep*, 29, 2004.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021.

Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC press, 1994.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. *Advances in neural information processing systems*, 32, 2019.

Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Procedings of the British Machine Vision Conference 2016*, pages 87–1. British Machine Vision Association, 2016.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*, pages 11117–11128. PMLR, 2020.

Bandwidth Selection in Kernel Density Estimation for Model Calibration (Supplementary Material)

Han Zhou[1] **Teodora Popordanoska**[1] **Matthew B. Blaschko**[1]

[1]Department ESAT, Center for Processing Speech and Images, KU Leuven, Leuven, Belgium

A APPENDIX: EXTENSION TO MULTI-CLASS AND KL DIVERGENCE

While the main text focuses on the binary L_2 case for clarity, the risk-consistency principle is fundamentally rooted in the properties of Bregman divergences. In this section, we demonstrate that the internal cancellation of estimation variance generalizes to multi-class settings and alternative proper scoring rules, such as the KL divergence (Cross-Entropy).

A.1 MULTI-CLASS CANONICAL L_2 RISK (BRIER SCORE)

We consider a model output $f(x)$ on the K-dimensional unit simplex Δ^{K-1}. Let $f^{(k)}(x_i) \in [0,1]$ be the predicted probability for class $k \in \{1,\dots,K\}$ and $y^{(k)} \in \{0,1\}$ be the ground truth label.

Canonical L_2 Bregman Decomposition: Following the quadratic properties of the L_2 scoring rule, the multi-class Brier score admits an orthogonal decomposition:

$$\mathcal{R}_{L_2}(f) = \underbrace{\mathbb{E}\left[\sum_{k=1}^{K}(R^{(k)} - f^{(k)}(x))^2\right]}_{\mathrm{CE}_2^2(f)} + \underbrace{\mathbb{E}\left[\sum_{k=1}^{K} R^{(k)}(1-R^{(k)})\right]}_{\mathrm{REF}_2(f)}, \tag{24}$$

where $R^{(k)}$ denotes the true conditional probability $P(Y = k|f(x))$. We reconstruct the pointwise risk as $\hat{r}_{\mathrm{kde},i} = \sum_{k=1}^{K}(\widehat{\mathrm{ce}}_i^{(k)} + \widehat{\mathrm{ref}}_i^{(k)})$, where $\widehat{\mathrm{ce}}_i^{(k)} = (\hat{R}_i^{(k)} - f^{(k)}(x_i))^2$ and $\widehat{\mathrm{ref}}_i^{(k)} = \hat{R}_i^{(k)}(1-\hat{R}_i^{(k)})$.

Multi-class Bias Derivation: To analyze the stability of this estimator, we consider the bias for a single sample i and class k. Centering the calibration term around the oracle value $R_i^{(k)}$:

$$\mathbb{E}\left[\widehat{\mathrm{ce}}_i^{(k)}\right] = \mathbb{E}\left[(\hat{R}_i^{(k)} - R_i^{(k)})^2 + 2(\hat{R}_i^{(k)} - R_i^{(k)})(R_i^{(k)} - f^{(k)}(x_i)) + (R_i^{(k)} - f^{(k)}(x_i))^2\right]. \tag{25}$$

Using the identity $\mathbb{E}[(\hat{R}_i^{(k)} - R_i^{(k)})^2] = \mathrm{Var}(\hat{R}_i^{(k)}) + (\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)})^2$, the bias is:

$$\mathrm{Bias}\big(\widehat{\mathrm{ce}}_i^{(k)}\big) = 2(R_i^{(k)} - f^{(k)}(x_i))(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) + \mathrm{Var}(\hat{R}_i^{(k)}) + (\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)})^2. \tag{26}$$

Neglecting the $\mathcal{O}(h^4)$ squared bias term $(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)})^2$ [Dehnad, 1987], the bias of the calibration error estimator is:

$$\mathrm{Bias}\big(\widehat{\mathrm{ce}}_i^{(k)}\big) \approx 2(R_i^{(k)} - f^{(k)}(x_i))(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) + \mathrm{Var}(\hat{R}_i^{(k)}). \tag{27}$$

Similarly, for the refinement component, expanding $\mathbb{E}[(\hat{R}_i^{(k)})^2] = \text{Var}(\hat{R}_i^{(k)}) + (\mathbb{E}[\hat{R}_i^{(k)}])^2$ and applying the difference of squares $(\mathbb{E}[\hat{R}_i^{(k)}])^2 - (R_i^{(k)})^2 = (\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)})(\mathbb{E}[\hat{R}_i^{(k)}] + R_i^{(k)})$, we have:

$$\begin{aligned}\text{Bias}\big(\widehat{\text{ref}}_i^{(k)}\big) &= (\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) - \Big((\mathbb{E}[\hat{R}_i^{(k)}])^2 - (R_i^{(k)})^2\Big) - \text{Var}(\hat{R}_i^{(k)}) \\ &= (1 - (\mathbb{E}[\hat{R}_i^{(k)}] + R_i^{(k)}))(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) - \text{Var}(\hat{R}_i^{(k)}) \\ &= (1 - (2R_i^{(k)} + \mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}))(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) - \text{Var}(\hat{R}_i^{(k)}). \end{aligned} \tag{28}$$

Neglecting the $\mathcal{O}(h^4)$ term, the refinement bias is:

$$\text{Bias}\big(\widehat{\text{ref}}_i^{(k)}\big) \approx (1 - 2R_i^{(k)})(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) - \text{Var}(\hat{R}_i^{(k)}). \tag{29}$$

Summing these across all classes k for each sample i, both the variance terms $\text{Var}(\hat{R}_i^{(k)})$ and the $\mathcal{O}(h^4)$ cancel out exactly. The resulting pointwise risk bias is:

$$\text{Bias}(\hat{r}_{\text{kde},i}) = \sum_{k=1}^{K} \Big[(1 - 2f^{(k)}(x_i))\Big] (\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}). \tag{30}$$

By applying the simplex constraint $\sum_k R_i^{(k)} = 1$ and the model constraint $\sum_k f^{(k)}(x_i) = 1$, we expand the summation:

$$\begin{aligned}\text{Bias}(\hat{r}_{\text{kde},i}) &\approx \left(\sum_{k=1}^{K} \mathbb{E}[\hat{R}_i^{(k)}] - \sum_{k=1}^{K} R_i^{(k)}\right) - 2\sum_{k=1}^{K} f^{(k)}(x_i)(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) \\ &= \left(\sum_{k=1}^{K} \mathbb{E}[\hat{R}_i^{(k)}] - 1\right) - 2\sum_{k=1}^{K} f^{(k)}(x_i)(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}). \end{aligned} \tag{31}$$

Under the assumption that the KDE estimator preserves the sum-to-one constraint, i.e., $\sum_k \mathbb{E}[\hat{R}_i^{(k)}] = 1$, the first term vanishes, and the pointwise risk bias simplifies to:

$$\text{Bias}(\hat{r}_{\text{kde},i}) \approx -2\sum_{k=1}^{K} f^{(k)}(x_i)(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}). \tag{32}$$

A.2 BINARY KL RISK

In the context of Kullback-Leibler (KL) divergence, we initially focus our analysis on the foundational case of binary classification for a predicted confidence $f(x) \in [0, 1]$ and ground-truth label $y \in \{0, 1\}$.

Binary KL Bregman Decomposition: The expected KL risk admits the following decomposition:

$$\mathcal{R}_{\text{KL}}(f) = \underbrace{\mathbb{E}[H(R)]}_{\text{REF}_{\text{KL}}(f)} + \underbrace{\mathbb{E}[D_{\text{KL}}(R, f(x))]}_{\text{CE}_{\text{KL}}(f)}, \tag{33}$$

where $H(R) = -R\log R - (1 - R)\log(1 - R)$ is the refinement component. We reconstruct the pointwise risk as $\hat{r}_{\text{kde},i}(h) = \widehat{\text{ce}}_i(h) + \widehat{\text{ref}}_i(h)$, using the LOO-KDE estimates $\hat{R}_i$.

Binary Bias Derivation: We analyze the bias via a second-order Taylor expansion around the true conditional probability R_i. For the calibration component $\widehat{\text{ce}}_i$, let $g(R) = R\log\frac{R}{f(x_i)} + (1 - R)\log\frac{1-R}{1-f(x_i)}$. Its derivatives are:

$$g'(R) = \log\frac{R(1 - f(x_i))}{f(x_i)(1 - R)}, \quad g''(R) = \frac{1}{R(1 - R)}. \tag{34}$$

The Taylor expansion $\mathbb{E}[g(\hat{R}_i)] \approx g(R_i) + g'(R_i)(\mathbb{E}[\hat{R}_i] - R_i) + \frac{1}{2}g''(R_i)\mathbb{E}[(\hat{R}_i - R_i)^2]$ yields the bias:

$$\text{Bias}\big(\widehat{\text{ce}}_i\big) \approx g'(R_i)(\mathbb{E}[\hat{R}_i] - R_i) + \frac{\text{Var}(\hat{R}_i) + (\mathbb{E}[\hat{R}_i] - R_i)^2}{2R_i(1 - R_i)}. \tag{35}$$

Similarly, for the refinement component $\widehat{\text{ref}}_i = H(\hat{R}_i) = -\hat{R}_i \log \hat{R}_i - (1-\hat{R}_i)\log(1-\hat{R}_i)$, the derivatives are $H'(R) = \log \frac{1-R}{R}$ and $H''(R) = -\frac{1}{R(1-R)}$. The expansion yields:

$$\text{Bias}\big(\widehat{\text{ref}}_i\big) \approx H'(R_i)(\mathbb{E}[\hat{R}_i] - R_i) - \frac{\text{Var}(\hat{R}_i) + (\mathbb{E}[\hat{R}_i] - R_i)^2}{2R_i(1-R_i)}. \tag{36}$$

Since the bias of the KDE estimator $\hat{R}_i$ is known to be $\mathcal{O}(h^2)$ [Dehnad, 1987, Wand and Jones, 1994], the squared bias term $(\mathbb{E}[\hat{R}_i] - R_i)^2$ in both Equations (35) and (36) is of order $\mathcal{O}(h^4)$. Neglecting these higher-order terms, we obtain the calibration bias:

$$\text{Bias}\big(\widehat{\text{ce}}_i\big) \approx \log \frac{R_i(1-f(x_i))}{f(x_i)(1-R_i)}(\mathbb{E}[\hat{R}_i] - R_i) + \frac{\text{Var}(\hat{R}_i)}{2R_i(1-R_i)}, \tag{37}$$

$$\text{Bias}\big(\widehat{\text{ref}}_i\big) \approx \log \frac{1-R_i}{R_i}(\mathbb{E}[\hat{R}_i] - R_i) - \frac{\text{Var}(\hat{R}i)}{2R_i(1-R_i)}. \tag{38}$$

Standard KL-based calibration estimation is notoriously sensitive to the variance $\text{Var}(\hat{R}_i)$ near the boundaries. As shown in Eqs. (37) and (38), the second-order bias coefficients diverge as $R_i \to 0$ or 1, causing individual components to become numerically unstable. However, summing these equations leads to the exact cancellation of these boundary-sensitive terms. Combining the first-order terms, we have:

$$\text{Bias}(\hat{r}_{\text{kde},i}) \approx \left(\log \frac{R_i(1-f(x_i))}{f(x_i)(1-R_i)} + \log \frac{1-R_i}{R_i}\right)(\mathbb{E}[\hat{R}_i] - R_i) = \left(\log \frac{1-f(x_i)}{f(x_i)}\right)(\mathbb{E}[\hat{R}_i] - R_i). \tag{39}$$

The resulting pointwise risk bias is thus stabilized as a first-order functional.

A.3 MULTI-CLASS CANONICAL KL RISK

Extending the analysis to the multi-class setting, we consider a model output $f(x)$ on the K-dimensional unit simplex Δ^{K-1}.

Multi-class KL Bregman Decomposition: The multi-class KL risk admits a decomposition into multi-class refinement and calibration components:

$$\mathcal{R}_{\text{KL}}(f) = \underbrace{\mathbb{E}\left[\sum_{k=1}^{K} -R^{(k)} \log R^{(k)}\right]}_{\text{REF}_{\text{KL}}(f)} + \underbrace{\mathbb{E}\left[\sum_{k=1}^{K} R^{(k)} \log \frac{R^{(k)}}{f^{(k)}(x)}\right]}_{\text{CE}_{\text{KL}}(f)}. \tag{40}$$

We reconstruct the pointwise risk as $\hat{r}_{\text{kde},i} = \sum_k \widehat{\text{ce}}_i^{(k)} + \sum_k \widehat{\text{ref}}_i^{(k)}$.

Multi-class Bias Derivation: We perform a second-order multivariate Taylor expansion around $\mathbf{R}_i$. Let $\delta_i^{(k)} = \mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}$. For $g_{CE}(\mathbf{R}) = \sum_k R^{(k)}(\log R^{(k)} - \log f^{(k)})$, the partial derivative $\frac{\partial g_{CE}}{\partial R^{(k)}} = \log \frac{R^{(k)}}{f^{(k)}} + 1$, and Hessian entries $1/R^{(k)}$:

$$\text{Bias}\Big(\sum_k \widehat{\text{ce}}_i^{(k)}\Big) \approx \sum_{k=1}^{K}\left[\left(\log \frac{R_i^{(k)}}{f^{(k)}(x_i)} + 1\right)\delta_i^{(k)} + \frac{\text{Var}(\hat{R}_i^{(k)}) + (\delta_i^{(k)})^2}{2R_i^{(k)}}\right]. \tag{41}$$

For $g_{REF}(\mathbf{R}) = \sum_k -R^{(k)} \log R^{(k)}$, the derivatives are $-\log R^{(k)} - 1$ and Hessian entries $-1/R^{(k)}$:

$$\text{Bias}\Big(\sum_k \widehat{\text{ref}}_i^{(k)}\Big) \approx \sum_{k=1}^{K}\left[\left(-\log R_i^{(k)} - 1\right)\delta_i^{(k)} - \frac{\text{Var}(\hat{R}_i^{(k)}) + (\delta_i^{(k)})^2}{2R_i^{(k)}}\right]. \tag{42}$$

Summing Equations (41) and (42) and neglecting $\mathcal{O}(h^4)$ terms, the boundary-sensitive variance terms and $\log R_i^{(k)}$ terms cancel. Using the assumption that $\sum_k \delta_i^{(k)} = 0$ (simplex constraint preservation), the constant shifts also vanish:

$$\text{Bias}(\hat{r}_{\text{kde},i}) \approx \sum_{k=1}^{K}\left(\log R_i^{(k)} - \log f^{(k)}(x_i) + 1 - \log R_i^{(k)} - 1\right)\delta_i^{(k)} = \sum_{k=1}^{K}\big(-\log f^{(k)}(x_i)\big)\delta_i^{(k)}. \tag{43}$$

A.4 CLASS-WISE ONE-VERSUS-REST DECOMPOSITION

In addition to canonical calibration, it is often desirable to evaluate calibration on a per-class basis, especially in scenarios with class imbalance. Following Popordanoska et al. [2024], the multi-class problem can be decomposed into K independent binary tasks using a one-versus-rest (OvR) approach. We consider a model output $f(x)$ residing on the $(K-1)$-dimensional unit simplex Δ^{K-1}. For each class $k \in \{1, \dots, K\}$, we denote the individual predicted probability as $f^{(k)}(x)$, which represents the model's confidence in class k such that $\sum_{k=1}^{K} f^{(k)}(x) = 1$. The corresponding ground-truth label is represented by a one-hot vector $\mathbf{y}$, with binary indicators $y^{(k)} \in \{0, 1\}$ satisfying $\sum_{k=1}^{K} y^{(k)} = 1$.

The expected class-wise risk $\mathcal{R}_{\text{cw}}$ is the sum of these binary risks. For any proper scoring rule induced by F, the class-wise risk admits an individual Bregman decomposition for each class:

$$\mathcal{R}_{\text{cw}}(f) = \sum_{k=1}^{K} \mathbb{E}[\ell(f^{(k)}(x), y^{(k)})] = \sum_{k=1}^{K} \left(\text{CWCE}_F^{(k)}(f) + \text{REF}_F^{(k)}(f) \right). \tag{44}$$

Class-wise L_2 Calibration Error (CWCE_2^2): Under the L_2 loss, the total class-wise calibration error is defined as the sum of squared binary calibration errors:

$$\text{CWCE}_2^2(f) = \sum_{k=1}^{K} \mathbb{E}\left[\left(R^{(k)} - f^{(k)}(x) \right)^2 \right]. \tag{45}$$

Using the LOO-KDE estimates $\hat{R}_i^{(k)}$, the bias of the class-wise estimator follows the variance-cancellation mechanism derived in the binary L_2 case. For each class k, the positive variance term $\text{Var}(\hat{R}_i^{(k)})$ from the calibration component is neutralized by the corresponding negative term in the refinement component. By neglecting the $\mathcal{O}(h^4)$ squared bias terms for each class, the resulting total class-wise risk bias remains a stable first-order functional:

$$\begin{aligned} \text{Bias}\big(\widehat{\mathcal{R}}_{\text{cw},L_2}(f)\big) &\approx \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left[(2(R_i^{(k)} - f^{(k)}(x_i)) + (1 - 2R_i^{(k)}))(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (1 - 2f^{(k)}(x_i))(\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}). \end{aligned} \tag{46}$$

Class-wise KL Calibration Error (CWCE_{KL}): Similarly, the class-wise calibration error induced by the KL divergence (Log loss) is the sum of binary KL divergences between the class-specific marginals:

$$\text{CWCE}_{\text{KL}}(f) = \sum_{k=1}^{K} \mathbb{E}\left[D_{\text{KL}}\big(R^{(k)}, f^{(k)}(x)\big) \right]. \tag{47}$$

Applying the second-order Taylor expansion to each binary KL term, we observe that the boundary-sensitive variance terms $\text{Var}(\hat{R}_i^{(k)})/(2R_i^{(k)}(1 - R_i^{(k)}))$ cancel out exactly within each class k when aggregated into the total risk. Neglecting higher-order terms, the stabilized bias for $\widehat{\mathcal{R}}_{\text{cw,KL}}$ is given by:

$$\begin{aligned} \text{Bias}\big(\widehat{\mathcal{R}}_{\text{cw,KL}}(f)\big) &\approx \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\left(\log \frac{R_i^{(k)}(1 - f^{(k)}(x_i))}{f^{(k)}(x_i)(1 - R_i^{(k)})} + \log \frac{1 - R_i^{(k)}}{R_i^{(k)}} \right) (\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}) \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \left(\log \frac{1 - f^{(k)}(x_i)}{f^{(k)}(x_i)} \right) (\mathbb{E}[\hat{R}_i^{(k)}] - R_i^{(k)}). \end{aligned} \tag{48}$$

B ADDITIONAL RESULTS

Below we present additional results on the synthetic data, as shown in Table 3-6 and Figs 4-5. RA consistently outperforms both discrete binning and standard KDE methods. In the classwise setting, RA achieves the lowest MAE in every scenario, often reducing error by an order of magnitude. This advantage is most pronounced at $K = 32$, where RA maintains stable estimates despite the data sparsity that compromises traditional binning. Similarly, in the canonical case, RA dominates the KL divergence results. While KRR emerges as a strong baseline for canonical L_2 error in higher dimensions, RA offers a more versatile framework that generalizes across different proper scoring rules and evaluation metrics.

Table 3: Summary of MAE for Calibration Error estimators $\times 10^2$ ($\downarrow$) ($t_1 = 1.0, t_2 = 0.8$). Best values are in red, and second-best in teal. Results are reported at $N = 20,000$. Note that Equal Binning uses $n_{\text{bins}} \in \{10, 15\}$ for canonical and $n_{\text{bins}} \in \{15, 20\}$ for classwise setting.

Mode	Metric	# Classes	Equal Binning		Adaptive Binning		KDE Methods		
			10/15	15/20	**10**	**15**	**MLE**	**KRR**	**RA**
Canonical	$\mathbf{CE}_{\mathbf{KL}}$	4	7.677	17.613	-	-	3.903	-	0.663
		8	113.732	168.737	-	-	24.049	-	6.242
		16	229.377	260.033	-	-	63.142	-	23.321
		32	161.158	367.625	-	-	71.926	-	51.287
	$\mathbf{CE}_2^2$	4	1.970	6.488	-	-	1.654	0.132	0.049
		8	34.362	70.000	-	-	7.401	0.462	0.620
		16	57.777	88.249	-	-	6.707	1.005	2.047
		32	18.749	92.894	-	-	3.968	0.747	3.320
Classwise	$\mathbf{CWCE}_{\mathbf{KL}}$	4	2.732	1.882	2.235	1.323	0.778	-	0.176
		8	7.658	5.271	2.528	1.606	1.216	-	0.157
		16	18.982	13.443	2.762	1.951	1.868	-	0.349
		32	40.411	30.510	3.306	2.713	2.858	-	0.794
	$\mathbf{CWCE}_2^2$	4	0.374	0.298	1.085	0.587	0.473	-	0.102
		8	0.745	0.558	1.794	1.116	0.769	-	0.205
		16	1.383	1.003	1.653	1.127	1.017	-	0.341
		32	2.158	1.473	1.125	0.818	1.008	-	0.379

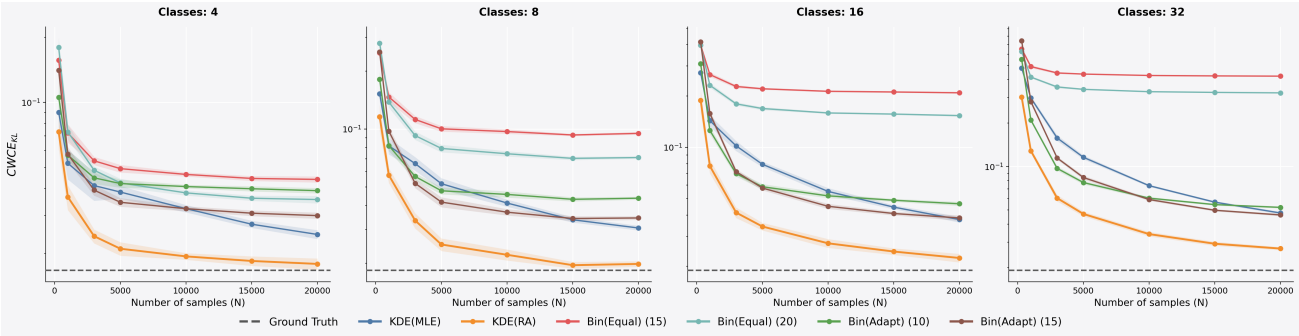


Figure 4: (**Synthetic data** ($t_1 = 1.0, t_2 = 0.8$)) Finite sample estimator of CWCE_{KL}: mean and 95% confidence interval computation across all subsamples of synthetic data.

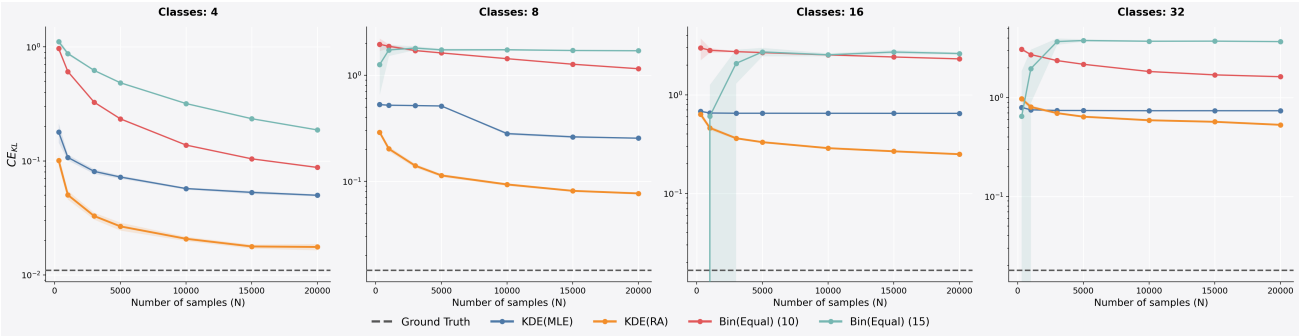


Figure 5: (**Synthetic data** ($t_1 = 1.0, t_2 = 0.8$)) Finite sample estimator of CE_{KL}^2: mean and 95% confidence interval computation across all subsamples of synthetic data.

Table 4: Summary of MAE for Calibration Error estimators $\times 10^2$ ($\downarrow$) ($t_1 = 0.9, t_2 = 0.6$). Best values are in red, and second-best in teal. Results are reported at $N = 20,000$. Note that Equal Binning uses $n_{\text{bins}} \in \{10, 15\}$ for canonical and $n_{\text{bins}} \in \{15, 20\}$ for classwise setting.

Mode	Metric	# Classes	Equal Binning		Adaptive Binning		KDE Methods		
			10/15	15/20	**10**	**15**	**MLE**	**KRR**	**RA**
Canonical	$\textbf{CE}_{\text{KL}}$	4	14.558	21.008	-	-	11.575	-	0.791
		8	109.049	143.685	-	-	51.455	-	9.287
		16	249.979	270.066	-	-	114.399	-	35.361
		32	329.954	391.260	-	-	117.565	-	79.370
	$\textbf{CE}_2^2$	4	2.094	6.127	-	-	5.381	0.140	0.406
		8	28.053	48.029	-	-	17.594	0.464	0.049
		16	60.181	76.850	-	-	26.899	0.967	1.379
		32	52.826	93.387	-	-	6.776	1.462	3.313
Classwise	$\textbf{CWCE}_{\text{KL}}$	4	8.670	6.533	3.720	2.117	0.764	-	0.492
		8	18.642	14.312	4.232	2.483	1.069	-	0.991
		16	37.257	29.019	4.694	2.893	1.896	-	1.085
		32	65.637	53.519	5.544	3.919	3.322	-	0.717
	$\textbf{CWCE}_2^2$	4	0.194	0.177	0.989	0.414	0.449	-	0.227
		8	0.467	0.316	2.783	1.538	0.877	-	0.166
		16	1.230	0.944	3.569	2.262	1.475	-	0.236
		32	2.037	1.505	2.865	1.994	1.641	-	0.374

Table 5: Summary of MAE for Calibration Error estimators $\times 10^2$ ($\downarrow$) ($t_1 = 0.9, t_2 = 0.7$). Best values are in red, and second-best in teal. Results are reported at $N = 20,000$. Note that Equal Binning uses $n_{\text{bins}} \in \{10, 15\}$ for canonical and $n_{\text{bins}} \in \{15, 20\}$ for classwise setting.

Mode	Metric	# Classes	Equal Binning		Adaptive Binning		KDE Methods		
			10/15	15/20	**10**	**15**	**MLE**	**KRR**	**RA**
Canonical	$\textbf{CE}_{\text{KL}}$	4	10.656	19.042	-	-	8.052	-	0.250
		8	110.227	151.600	-	-	38.832	-	7.510
		16	246.570	260.973	-	-	86.389	-	28.865
		32	266.295	384.407	-	-	98.759	-	66.036
	$\textbf{CE}_2^2$	4	1.997	6.406	-	-	3.587	0.133	0.234
		8	31.402	56.564	-	-	12.783	0.490	0.219
		16	63.588	82.642	-	-	19.480	1.018	1.674
		32	37.539	94.876	-	-	5.718	1.155	3.273
Classwise	$\textbf{CWCE}_{\text{KL}}$	4	5.297	3.866	3.034	1.757	0.816	-	0.237
		8	12.734	9.250	3.530	2.145	1.243	-	0.307
		16	27.723	20.684	3.935	2.546	1.996	-	0.329
		32	53.301	41.988	4.640	3.438	3.304	-	0.243
	$\textbf{CWCE}_2^2$	4	0.297	0.242	1.104	0.538	0.484	-	0.100
		8	0.656	0.500	2.503	1.461	0.865	-	0.099
		16	1.344	1.003	2.811	1.834	1.290	-	0.314
		32	2.153	1.546	2.121	1.494	1.424	-	0.439

Table 6: Summary of MAE for Calibration Error estimators $\times 10^2$ ($\downarrow$) ($t_1 = 0.9, t_2 = 0.8$). Best values are in red, and second-best in teal. Results are reported at $N = 20,000$. Note that Equal Binning uses $n_{\text{bins}} \in \{10, 15\}$ for canonical and $n_{\text{bins}} \in \{15, 20\}$ for classwise setting.

Mode	Metric	# Classes	Equal Binning		Adaptive Binning		KDE Methods		
			10/15	15/20	**10**	**15**	**MLE**	**KRR**	**RA**
Canonical	$\mathbf{CE}_{\mathbf{KL}}$	4	8.636	17.911	-	-	5.501	-	0.511
		8	110.441	157.466	-	-	29.843	-	6.237
		16	236.318	255.321	-	-	71.770	-	24.707
		32	205.838	371.927	-	-	83.543	-	56.561
	$\mathbf{CE}_2^2$	4	1.996	6.459	-	-	2.367	0.137	0.085
		8	33.175	63.510	-	-	9.604	0.483	0.444
		16	61.803	84.961	-	-	7.814	1.026	1.889
		32	26.573	94.227	-	-	4.789	0.922	3.324
Classwise	$\mathbf{CWCE}_{\mathbf{KL}}$	4	3.428	2.416	2.504	1.467	0.817	-	0.166
		8	8.973	6.299	2.959	1.852	1.283	-	0.148
		16	21.506	15.449	3.344	2.261	1.988	-	0.295
		32	44.723	34.119	3.918	3.056	3.108	-	0.669
	$\mathbf{CWCE}_2^2$	4	0.400	0.304	1.144	0.602	0.491	-	0.092
		8	0.745	0.567	2.167	1.319	0.831	-	0.202
		16	1.397	1.031	2.197	1.469	1.153	-	0.356
		32	2.213	1.554	1.577	1.130	1.243	-	0.427

B.1 REAL-WORLD DATA

Below we present the result of CWCE_2^2, $\text{CWCE}_{\text{KL}}^2$, CE_{KL} and CE_2^2 estimators across different datasets and architectures. Consistent with our synthetic findings, RA demonstrates clear superiority over the MLE-based approach and traditional binning schemes. Across benchmarks ranging from ResNets to Vision Transformers, RA aligns more closely with the ground truth computed over the full population than those methods, particularly in class-wise settings, as shown in Figs. 6-13. We further observe that KDE performance degrades in canonical settings with many classes, such as CIFAR-100, due to data sparsity on the high-dimensional simplex. In these regimes, KRR remains a strong baseline.

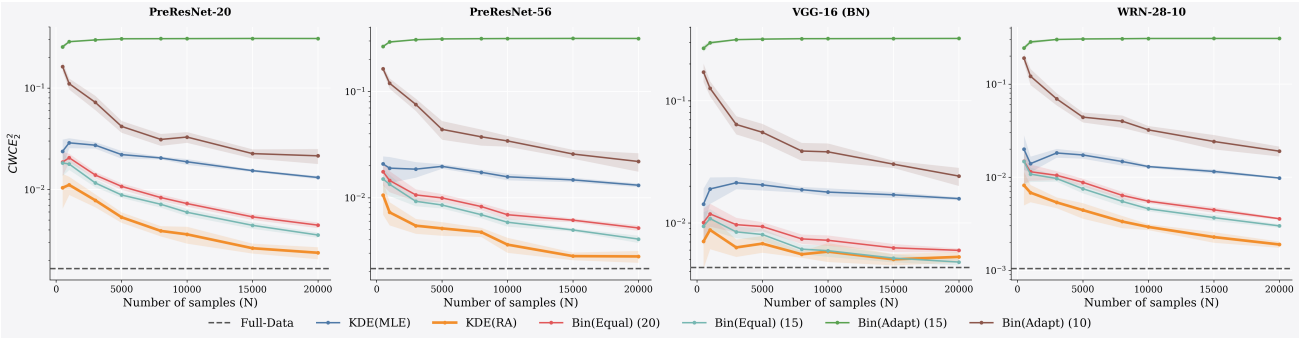


Figure 6: (**CIFAR10**) Finite sample estimator of CWCE_2^2: mean and 95% confidence interval computation across all subsamples.

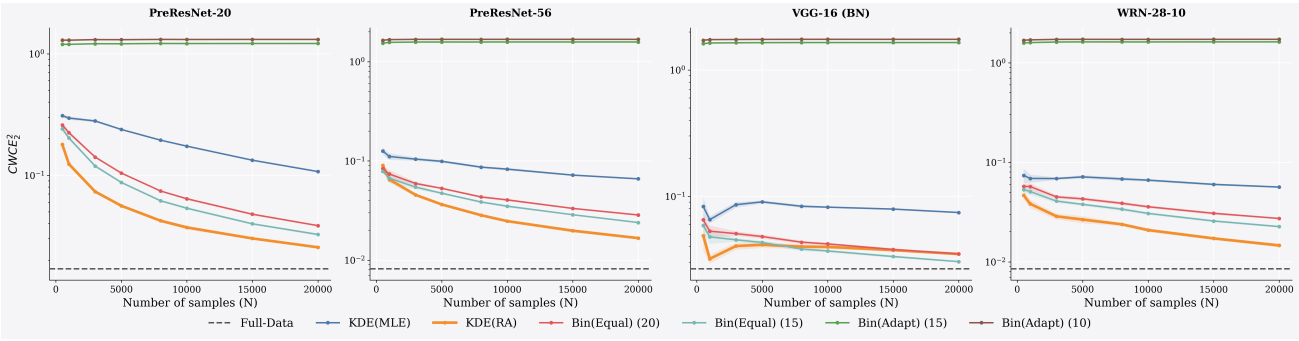


Figure 7: (**CIFAR100**) Finite sample estimator of CWCE_2^2: mean and 95% confidence interval computation across all subsamples.

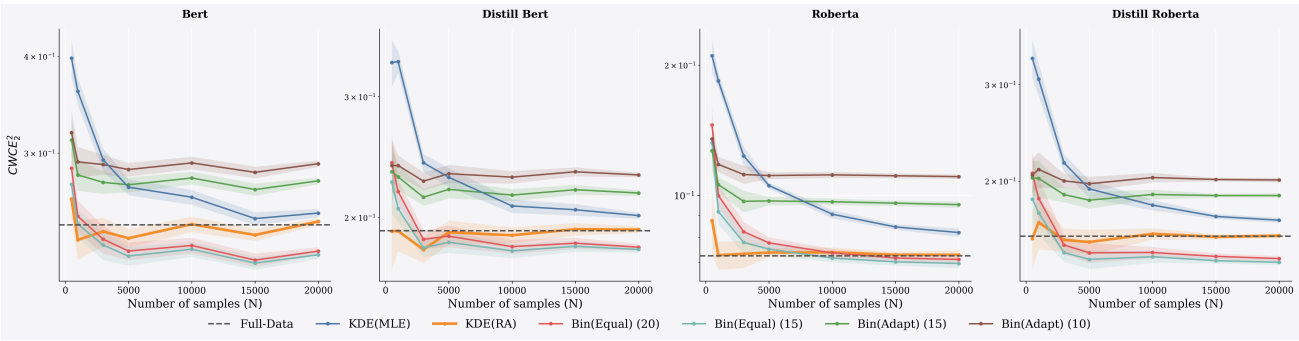


Figure 8: (**Amazon**) Finite sample estimator of CWCE_2^2: mean and 95% confidence interval computation across all subsamples.

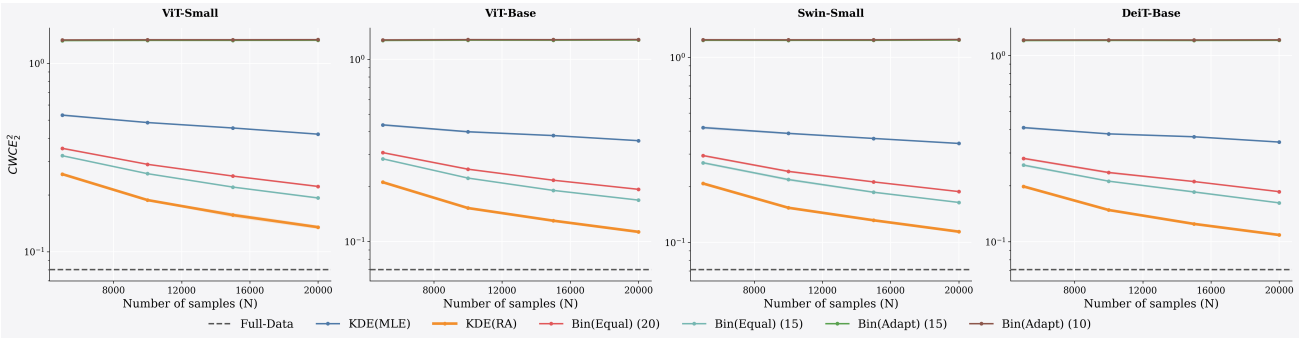


Figure 9: (**ImageNet**) Finite sample estimator of CWCE_2^2: mean and 95% confidence interval computation across all subsamples.

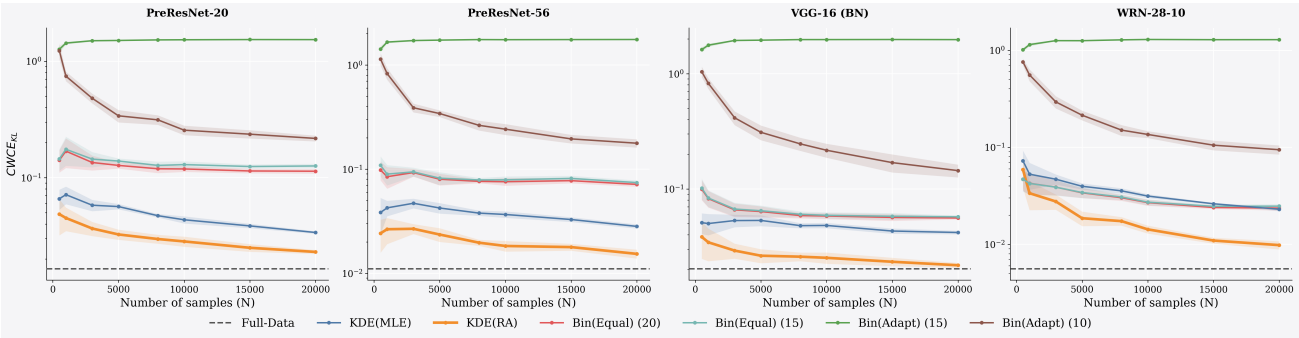


Figure 10: (**CIFAR10**) Finite sample estimator of $\text{CWCE}_{\text{KL}}^2$: mean and 95% confidence interval computation across all subsamples.

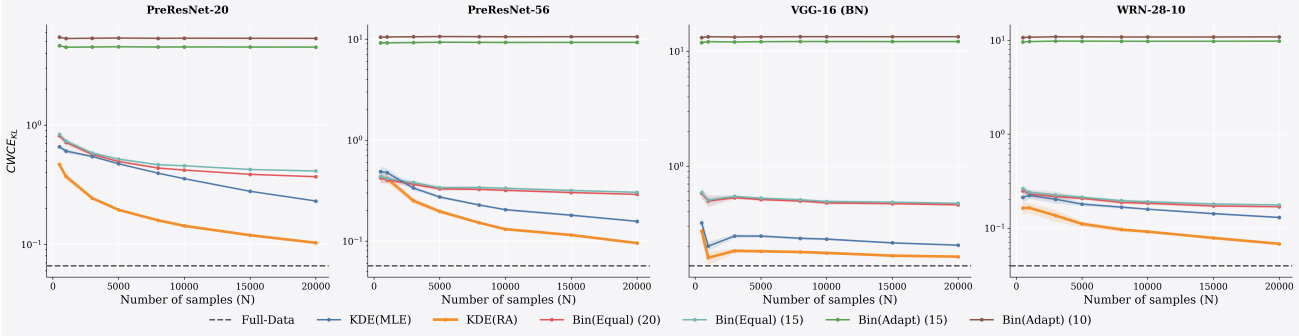


Figure 11: (**CIFAR100**) Finite sample estimator of $\text{CWCE}^2_{\text{KL}}$: mean and 95% confidence interval computation across all subsamples.

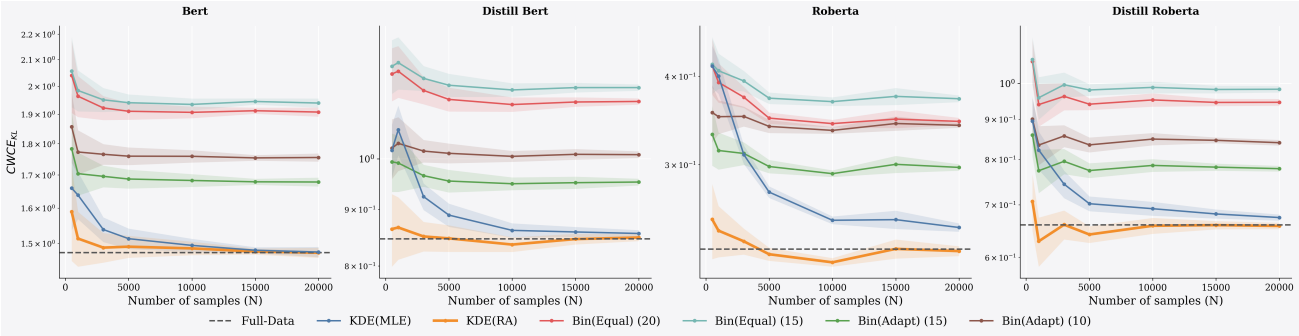


Figure 12: (**Amazon**) Finite sample estimator of $\text{CWCE}^2_{\text{KL}}$: mean and 95% confidence interval computation across all subsamples.

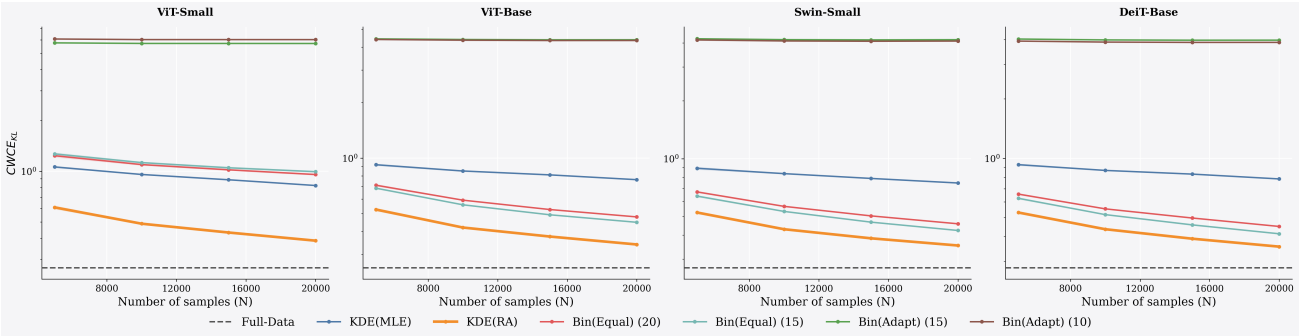


Figure 13: (**ImageNet**) Finite sample estimator of $\text{CWCE}^2_{\text{KL}}$: mean and 95% confidence interval computation across all subsamples.

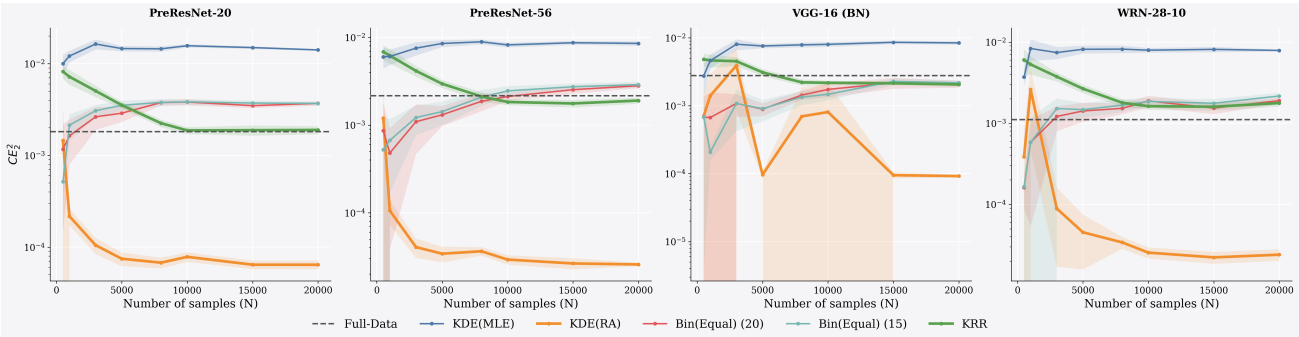


Figure 14: (**CIFAR10**) Finite sample estimator of CE^2_2: mean and 95% confidence interval computation across all subsamples.

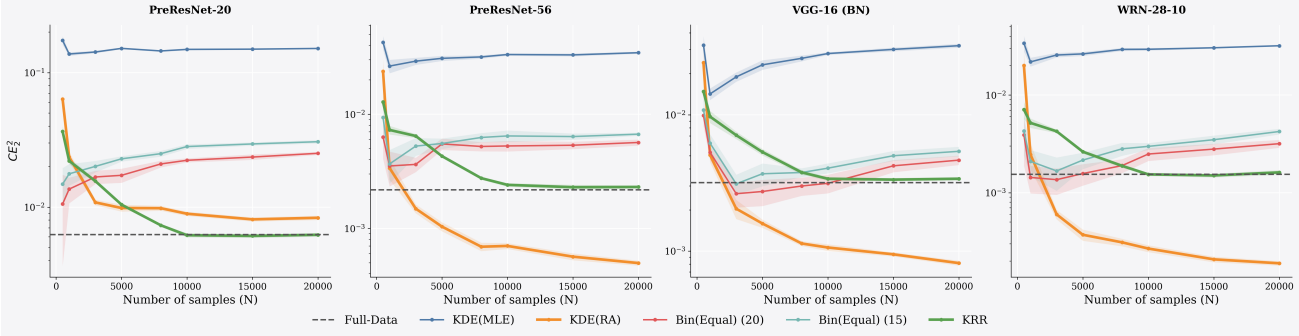


Figure 15: (**CIFAR100**) Finite sample estimator of CE_2^2: mean and 95% confidence interval computation across all subsamples.

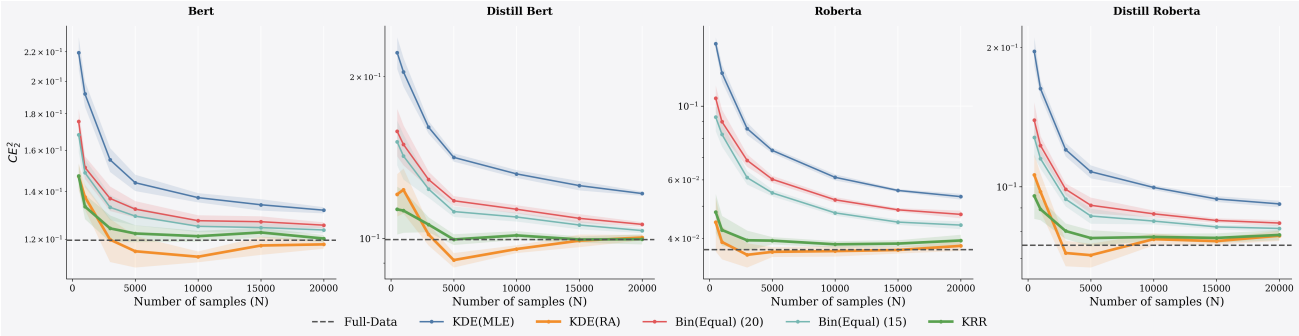


Figure 16: (**Amazon**) Finite sample estimator of CE_2^2: mean and 95% confidence interval computation across all subsamples.

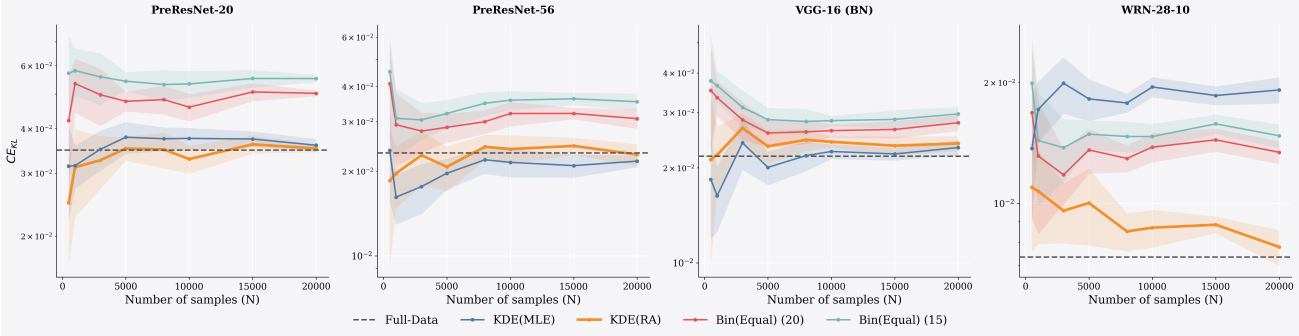


Figure 17: (**CIFAR10**) Finite sample estimator of $\mathrm{CE}_{\mathrm{KL}}^2$: mean and 95% confidence interval computation across all subsamples.

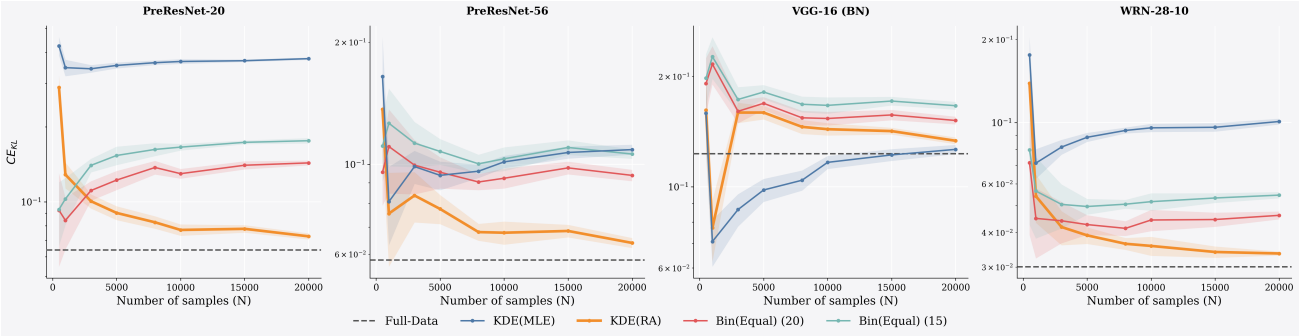


Figure 18: (**CIFAR100**) Finite sample estimator of $\mathrm{CE}_{\mathrm{KL}}^2$: mean and 95% confidence interval computation across all subsamples.

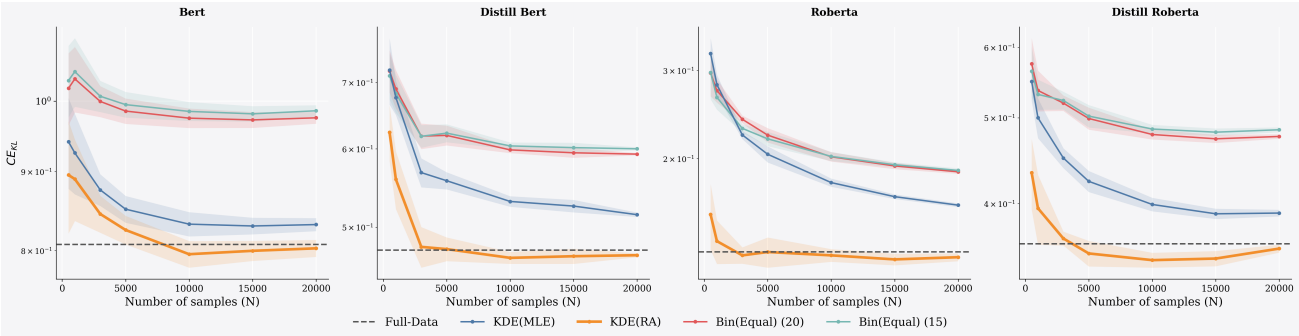


Figure 19: (**Amazon**) Finite sample estimator of $\mathrm{CE}^2_{\mathrm{KL}}$: mean and 95% confidence interval computation across all subsamples.