

Multiclass Local Calibration With the Jensen-Shannon Distance

Cesare Barbera^{1,2}, Lorenzo Perini³, Giovanni De Toni⁴, Andrea Passerini¹, and Andrea Pugnana¹

¹DISI, University of Trento, Trento, Italy

{andrea.pugnana, cesare.babera, andrea.passerini}@unitn.it

²Univeristy of Pisa, Pisa, Italy

³Meta, London, United Kingdom, lorenzoperini@meta.com

⁴Fondazione Bruno Kessler, Trento, Italy gdetoni@fbk.eu

Abstract

Developing trustworthy Machine Learning (ML) models requires their predicted probabilities to be well-calibrated, meaning they should reflect true-class frequencies. Among calibration notions in multiclass classification, strong calibration is the most stringent, as it requires all predicted probabilities to be simultaneously calibrated across all classes. However, existing approaches to multiclass calibration lack a notion of distance among inputs, which makes them vulnerable to proximity bias: predictions in sparse regions of the feature space are systematically miscalibrated. This is especially relevant in high-stakes settings, such as healthcare, where the sparse instances are exactly those most at risk of biased treatment. In this work, we address this main shortcoming by introducing a local perspective on multiclass calibration. First, we formally define multiclass *local calibration* and establish its relationship with *strong calibration*. Second, we theoretically analyze the pitfalls of existing evaluation metrics when applied to multiclass local calibration. Third, we propose a practical method for enhancing *local calibration* in Neural Networks, which enforces alignment between predicted probabilities and local estimates of class frequencies using the Jensen-Shannon distance. Finally, we empirically validate our approach against existing multiclass calibration techniques.

1 Introduction

In many high-stakes applications, Machine Learning (ML) models are expected not only to be accurate but also well-calibrated [21, 25] - i.e., their predicted probabilities should reflect the true empirical frequencies of the corresponding classes. For instance, in the context of clinical decision-making, Van Calster et al. [24] compares two cardiovascular risk prediction models over two million patients from the UK. They show that the better-calibrated model, despite having a slightly lower AUC, provided more useful predictions, avoiding overestimation of risk.

However, existing works have mostly focused on *confidence calibration* [1, 14, 15, 28], which only looks at the top-predicted class to check a model’s calibration. While in binary classification tasks, a well-calibrated model on the top-predicted class ensures good calibration on the complementary class, this is not true for multiclass classification tasks.

Example. Consider an ML model trained to predict the stage of cancer in a cell. Suppose the model is calibrated only with respect to its most confident prediction, i.e., if it assigns a probability of $p\%$ to early-stage cancer, then about $p\%$ of the instances receiving such a score are indeed early-stage cancer. Although this might appear sufficient for decision-making [5], inaccurate probability estimates for less frequent classes can be critical: failing to approximate the likelihood of rare transitional states—such as cells halfway between benign and malignant—might hide patterns about tumor progression.

Hence, stronger notions for multiclass calibration have been proposed. One example is *strong calibration* [23], which enforces alignment across the full probability vector. However, existing multiclass calibration notions do not consider any form of distance among instances, making them prone to *proximity bias* [28], whereby predictions for instances in sparsely populated regions of the decision space are more likely to be poorly calibrated. In high-stakes settings, these are precisely the cases where one requires trustworthy predictions to avoid biased treatment.

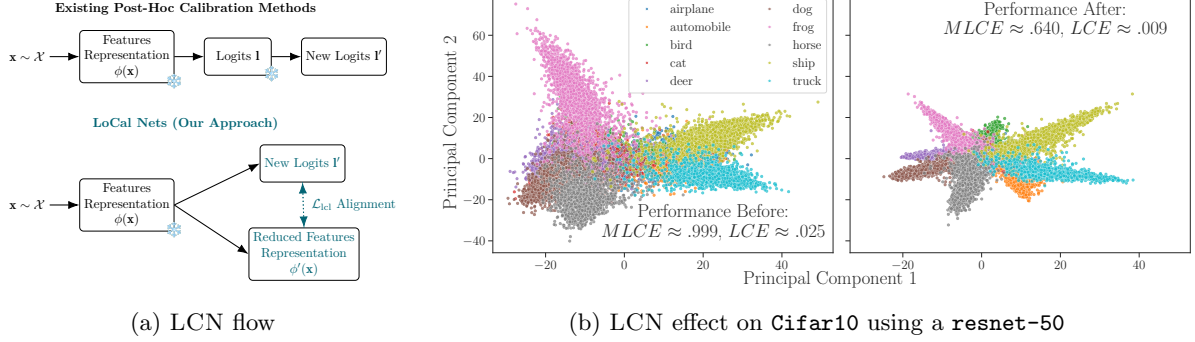


Figure 1: **Our LoCal Nets (LCN) provide local calibration through feature reshaping** (1a) Unlike post-hoc calibrators that rescale fixed logits, LCNs jointly (i) learn reduced feature representations $\phi'(x)$ and (ii) output new calibrated logits, aligning predictions with local class frequencies via Jensen–Shannon distance. (1b) On Cifar10 with resnet-50, LCNs yield tighter, better-separated class clusters and improved calibration ($\approx 64\%$ reduction in MLCE, 36% reduction in LCE).

Our Contributions. In this work, we tackle this shortcoming by framing multiclass calibration from a “local” perspective. More precisely, (i) in section 4 we introduce the notion of *Multi-Class Local-Calibration*, which leverages true empirical frequencies in the neighborhoods of inputs to assess model reliability across the decision-space, and we connect it to *strong calibration*; (ii) we provide theoretical insights on existing pitfalls of current evaluation metrics when applied to multiclass local calibration (Sections 5 - 6); (iii) in section 7 we propose a new neural network calibration method called LoCal Nets (LCNs), which exploits the Jensen-Shannon distance to align predicted probabilities and local estimates of class frequencies, while enforcing denser feature representations (Figure 1); (iv) we evaluate our approach against existing competitors (section 8), showing how our method improves over local calibration metrics and keeps competitive performance at the global level.

2 Related Work

Global Calibration Notions. Due to the stringent requirements of strong calibration and difficulty in both evaluation and enforcement, the literature has proposed several relaxed alternatives. Examples of such relaxations are *confidence calibration* [5] and *top-label calibration* [6]. While the former considers only the model’s maximum predicted confidence, the latter refines the definition of confidence calibration by conditioning the empirical frequencies on both the model confidence and the model prediction.

Class-wise calibration [13] evaluates calibration for each class marginally, by comparing predicted probabilities with empirical frequencies for that target alone. *Top-r calibration* [7] considers if the true class falls within the top- r predicted labels and if the cumulative confidence over the top- r classes aligns with observed frequencies. *Decision calibration* [32] defines calibration with respect to a decision-making policy, requiring predicted and empirical distributions to match according to a decision maker. Since these notions are global, they emphasize predicted probabilities while overlooking factors such as instance density or spatial position, failing to capture poor calibration for underrepresented groups.

Post-hoc Calibration Methods. When a trained model suffers from poorly calibrated outputs, many available model-agnostic post-hoc techniques can be employed to adjust its predicted probabilities. Popular methods include binning [30], isotonic regression [31], logistic scaling [20], temperature scaling [5] and parametric forms of scaling [12, 13]. We refer to de Menezes e Silva Filho et al. [1] for a complete overview of different calibration techniques. However, all these methods treat logits as fixed inputs to be rescaled. In contrast, our approach learns new feature representations that reshape the geometry of the representation space itself. Thus, these new neighborhoods better reflect true class frequencies and improve local calibration.

Local Calibration Notions. Most methods group instances with similar predicted confidences, but ignore that similar scores can be assigned to points in very different regions of the decision space. A first attempt to investigate the notion of *local calibration* is by Luo et al. [15], who propose to recalibrate a classifier’s confidence scores via kernel regressions. Although promising, their approach only tackles *confidence calibration* and considers fixed feature representations. As a consequence, the method significantly slows down inference times, requiring a kernel regression for each recalibrated probability score. Also [16] investigates the utility of kernel regressions for calibration, but their proposal directly applies to a model training process. As such, this approach can suffer from computational and data limitations. Finally, Perez-Lebel et al. [19] shows that whenever the classifier’s decision boundary is complex, it can lead to poor calibration of the scores predicted for less likely instances. They introduce (i) the *cancellation effect*, where miscalibration errors within a confidence group offset one another, and (ii) *proximity bias*, reflecting disparities in calibration quality for instances in sparsely populated regions of the decision space. Our approach differs as they focus on (i) data density rather than on the position of points in the decision space; (ii) their proposal addresses only the confidence score of the predicted class, not multiclass calibration. We refer readers to the Appendix C for a detailed analysis of the role of *proximity bias* in our proposal and an illustrative example of the dangers of density-based recalibration.

3 Background

Let us consider a multi-class classification setting, where $\mathcal{X} \subseteq \mathbb{R}^m$ is the feature space and $\mathcal{Y} = \{0, \dots, C-1\}$ is a finite target space with C distinct labels. Let us assume we have access to a given dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of input-output pairs drawn from an unknown joint distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. Each input $\mathbf{x}_i \in \mathcal{X}$ is a feature vector of m dimensions, and each label $y_i \in \mathcal{Y}$ has a corresponding one-hot encoded vector $\mathbf{y}_i$ indicating the correct class among the C possible classes. We consider a probabilistic classifier $f: \mathcal{X} \rightarrow \Delta^C$, where Δ^C is the $(C-1)$ -dimensional probability simplex. In words, a probabilistic classifier maps an input $\mathbf{x}$ to a probability distribution over classes, i.e., $f(\mathbf{x}) = \hat{\mathbf{p}} \in \Delta^C$, where each entry $\hat{\mathbf{p}}_k = f_k(\mathbf{x})$ of the predicted probability vector $\hat{\mathbf{p}}$ denotes the predicted probability of class k .

In the multiclass context, the weakest notion of calibration is *confidence calibration* [5], which requires that only the top-predicted probability $\max_{y \in \mathcal{Y}} \hat{\mathbf{p}}_y$ matches the frequency of correct predictions of the classifier. A more stringent notion is *strong calibration* [23], which requires the target class conditional distribution on any prediction of the classifier to match that prediction, i.e.,:

$$\mathbb{P}(\mathbf{y}_k = 1 \mid \hat{\mathbf{p}}) = \hat{\mathbf{p}}_k \quad \text{for all } k \in \{1, \dots, C\}. \quad (1)$$

Metrics. Because of the large number of calibration notions, various metrics have been proposed in the literature to evaluate the calibration of classifiers.

One of the most popular metrics is *Expected Calibration Error (ECE)* [18], which measures the calibration of a binary classifier. It bins predicted confidence scores into B intervals and compares the average predicted confidence with the empirical accuracy in each bin. A multiclass extension of *ECE* is *Class-wise ECE*, which calculates ECE separately for each class c and then averages the results. *Multidimensional Expected Calibration Error (MECE)* extends *ECE* and its class-wise version to multiclass, by using a multidimensional grid binning. Unfortunately, the combinatorial nature of such a binning hinders its practical application. *Expected Cumulative Calibration Error (ECCE)* [9] evaluates the cumulative discrepancy between confidence and accuracy across bins rather than averaging them.

Local Calibration Error (LCE) quantifies directly the state of *local calibration* of a probabilistic classifier. We provide here the multiclass extension from the original definition by Luo et al. [15]:

$$LCE = \frac{1}{C} \sum_{b=1}^{m_B} \frac{1}{n} \sum_{i \in B_b} \left\| \frac{\sum_{j \in B_b} (\hat{\mathbf{p}}_j - \mathbf{y}_j) k_\gamma(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j \in B_b} k_\gamma(\mathbf{x}_i, \mathbf{x}_j)} \right\|_1,$$

where $\|\cdot\|_1$ is an appropriate ℓ^1 norm, m_b the number of used bins, B_b is the set of instances in the b -th bin and $k_\gamma(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function that weights the influence of neighbouring points of the anchor $\mathbf{x}_i$ to its individual *LCE* score. Its appearance in the denominator as a normalization term ensures weights sum to 1 for every anchor. In practice, this metric captures the differences in the predicted probabilities and the corresponding ground truths for neighbours of an anchor point $\mathbf{x}_i$. Finally, the *Maximum Local*

Calibration Error is $MLCE = \max_{i \in D} \left\| \frac{\sum_{j \in b} (\hat{\mathbf{p}}_j - \mathbf{y}_j) k_\gamma(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j \in b} k_\gamma(\mathbf{x}_i, \mathbf{x}_j)} \right\|_1$. Notice that, in its original formulation by Luo et al. [15], *LCE* is limited to confidence calibration, making it insufficient to assess multiclass *local calibration*.

4 Defining Multiclass Local Calibration

Local Calibration requires introducing the concept of distance between samples to capture how one sample’s probabilistic outputs may influence or relate to another’s. Roughly speaking, our definition is based on the intuition that nearby instances should have more similar label distributions and affect each other’s calibration more strongly. More precisely, we require the model’s predicted probabilities for each sample to be consistent with the locally averaged estimates of the classes’ distribution. The degree of *locality* depends on a kernel function k and its bandwidth parameter γ , which controls the influence of neighboring points on the estimates for each instance.

Definition 1 (Multiclass Local Calibration). *Let f be a probabilistic classifier, and let D be an evaluation set drawn from $\mathcal{P}$. For each instance $i \in D$, let $k_\gamma(\mathbf{x}_i, \mathbf{x}_j)$, with $j \neq i \in D$ be a kernel function with bandwidth γ and consider an associated kernel estimator $\hat{\theta}(\mathbf{y}_i | \mathbf{x}_i) = \frac{\sum_{j \in D} W_j(\mathbf{x}_i) \mathbf{y}_j}{\sum_{j \in D} W_j(\mathbf{x}_i)}$ with $W_j(\mathbf{x}_i) \propto k_\gamma(\mathbf{x}_i, \mathbf{x}_j)$, a kernel induced weight, normalized to sum to 1. Finally, let $\hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)$ be consistent in the mean squared error (MSE) sense.*

*Then, f is **locally calibrated** on D if, for all $\mathbf{x}_i \in D$, the predicted probability vector $\hat{\mathbf{p}}_i = f(\mathbf{x}_i)$ is close to the kernel estimate up to a tolerance $\varepsilon \geq 0$, i.e.,*

$$\left\| \hat{\mathbf{p}}_i - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i) \right\|_1 \leq \varepsilon, \quad \forall i \in \{1, \dots, n\}. \quad (2)$$

*Moreover, when $\varepsilon = 0$, we say the classifier is **perfectly locally calibrated**.*

Notably, the multiclass notion of *local calibration* is closely related to *strong calibration*, and satisfying the former is sufficient for the latter. In the theorem below, we establish a bound in terms of the *continuous MECE*. We focus on the continuous form since, unlike binned variants, it integrates over the probability space and thus provides the most faithful and comprehensive measure of calibration.

Theorem 1 (Continuous MECE under Local Calibration). *Let D be an evaluation dataset drawn i.i.d. from a distribution $\mathcal{P}$. Define the continuous Multidimensional Expected Calibration Error (MECE) as:*

$$M(f) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \mathbb{E}_{\mathbf{y} \sim p(\mathbf{y})} [|\mathbb{E}[\mathbf{y} | \mathbf{x}] - f(\mathbf{x})|],$$

If a model f satisfies local calibration, then there exists $k \in [1/C, 1]$ such that continuous MECE is asymptotically upper bounded as:

$$M(f) \leq \varepsilon \cdot k \quad (3)$$

Theorem 1 shows that the continuous *MECE* is upper bounded for multiclass *locally calibrated* models and is minimal if the condition holds for small ε . A detailed proof is provided in the Appendix B.

5 Evaluating Multiclass Calibration

Although the continuous *MECE* allows for the derivation of theoretical results, such a metric is not computationally feasible in practice. We address this limitation by theoretically analyzing practical methods to evaluate the *local calibration* of probabilistic classifiers.

Definition 2 (General binning calibration metric). *Let f be a probabilistic classifier, and let D denotes an evaluation dataset. Let $\beta : \Delta^C \rightarrow \{1, \dots, m_B\}$ be a deterministic binning function that partitions the probability simplex Δ^C into m_B disjoint bins $\{B_b\}_{b=1}^{m_B}$ and let $\varphi : [0, 1] \times [0, 1] \rightarrow \mathbb{R}_{\geq 0}$ be a scalar comparator that measures discrepancy between an empirical frequency $\text{freq}_{b,c}$ and a predicted confidence*

$\text{conf}_{b,c}$ that is Lipschitz in both arguments with constant L_φ . The general multiclass bin-based calibration error metric is obtained as:

$$\mathcal{E}(D; \varphi; \beta) = \sum_{b=1}^{m_B} w_b \sum_{c=1}^C \pi_c \cdot \varphi(\text{freq}_{b,c}, \text{conf}_{b,c}). \quad (4)$$

where w_b are deterministic bin weights and π_c are deterministic class weights.

This general definition of calibration metrics provides a flexible framework that subsumes all calibration measures that leverage binning, including *ECE* and *MECE*. Interestingly, we can define a probabilistic bound for the value of any calibration metric satisfying definition 2 under *local calibration*.

Theorem 2 (Error decomposition of calibration metrics under Local Calibration). *Let $\mathcal{E}(D; \varphi; \beta)$ denote a calibration metric according to definition 2. If a classifier f satisfies local calibration for error ε , then, for any $\delta \in [0, 1]$, with at least probability $1 - \delta$ it holds:*

$$\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \cdot \left[\varepsilon + \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|\Psi(b, \mathcal{Y})|}} \right], \quad (5)$$

where $\Psi(\cdot; \cdot)$ selects a bin based on the index b and the set of labels $\mathcal{Y}$.

We provide the proof in the Appendix B. Notice that Theorem 2 establishes an upper bound on $\mathcal{E}(D; \varphi; \beta)$ in terms of ε , implying that a high value of the metric necessarily indicates poor *local calibration*. However, the converse does not hold in general, as binning-based metrics are known to suffer from *cancellation effects*, whereby opposing errors may offset each other and obscure miscalibration.

6 Evaluating Calibration in Neural Networks

In the following section, we focus on Neural Network classifiers. Exploiting their structure and the L -Lipschitz continuity property with respect to the norm $\|\cdot\|_1$ [4], we refine the general results presented in the previous sections and investigate further the properties of *Local Calibration Error*. Practically speaking, achieving *perfect local calibration* is unfeasible. In the ideal case where $\varepsilon = 0$, the distribution of points would be both perfectly centered around $\mathbf{x}_i$ and fully representative of the local data distribution. However, such conditions are never met in finite samples. Additionally, definition 1 does not take into account potential disparities in the miscalibration error of classes but only bounds the total sum of errors. For these two reasons, we relax definition 1 by bounding the admissible error in the predicted probability for class c by the maximum change in the output probabilities that can occur when moving within a ball of radius ρ in the feature space. Thus, decreasing ρ bounds the class-wise deviation a model can tolerate while approaching *perfect local calibration*. Note that our definition operates under the assumption that the kernel estimator refers to the Neural Network learned feature-representations learned as in [15].

Definition 3 (ρ -Perfect Local Calibration). *Let f be a probabilistic classifier composed of a feature extractor $\phi : \mathcal{X} \rightarrow \mathcal{F}$ and a final classification layer $g : \mathcal{F} \rightarrow \Delta^C$. Let D be an evaluation set. In addition, $\phi(\mathbf{x}_i)$ is assumed Lipschitz-continuous with respect to the softmax with constant $L > 0$. We say that the model f is ρ -perfectly locally calibrated if for every instance $\mathbf{x}_i \in D$ and for every class $c \in \{1, \dots, C\}$, the absolute calibration error is bounded as follows:*

$$|\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y} \mid \phi(\mathbf{x}_i))| \leq L \cdot \rho. \quad (6)$$

We exploit the ρ -perfect local calibration definition to further improve the theorem 2 bound. We formalize it in the following corollary:

Corollary 1 (Calibration measure under ρ -Perfect Local Calibration). *If a classifier f satisfies ρ -perfect local calibration, the error $\mathcal{E}(D; \varphi; \beta)$ becomes purely stochastic fluctuation. For any $\delta \in [0, 1]$, with probability at least $1 - \delta$ it holds that:*

$$\limsup_{\rho \rightarrow 0} \mathcal{E}(D; \varphi; \beta) \leq \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|\Psi(b, \mathcal{Y})|}} \quad (7)$$

Proof can be found in Appendix. This result underscores that the smaller the value of ρ for which a model exhibits perfect *local calibration*, the more the overall error is dominated by the stochastic component.

Our theoretical analysis on the role of *local calibration* in shaping the behavior of calibration metrics will next address the multiclass version of *LCE*. Its inclusion is essential as it is the only metric that directly quantifies the degree of *local calibration* of a probabilistic classifier. More precisely, we show that, unlike global calibration measures, LCE relies on kernel regression estimates, which introduces a bias-variance trade-off on top of the inherent calibration error. Roughly speaking, the bound we derive decomposes into three interpretable components: (i) a calibration term ε controlled by the *local calibration* property; (ii) a variance term growing as the kernel weights become concentrated on fewer neighbors; and (iii) a bias term which penalizes assigning large weights to distant samples.

Theorem 3 (Probabilistic bound for multiclass LCE under Local Calibration). *Let f be a probabilistic classifier, where f is composed of a feature extractor ϕ and a final classification layer g . Assume f to be locally calibrated up to error ε and let the LCE be computed on $\phi(\mathbf{x}_i)$. Then, there exists $k \in [1/C, 1]$ such that, for any $\delta \in [0, 1]$, with at least probability $1 - \delta$ the following holds:*

$$\text{LCE} \leq +k \left[\underbrace{\varepsilon + \frac{1}{n} \sum_{b=1}^{m_B} \sum_{i \in I_b} \sqrt{\frac{2 \log(\frac{n}{\delta})}{n_i^{\text{eff}}}}}_{\text{variance term}} + \underbrace{\frac{L}{n} \sum_{b=1}^{m_B} \mathbb{E} \left[\sum_{i \in I_b} \sum_{j \in I_b} w_{i,j} \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|_1 \right]}_{\text{bias term}} \right]$$

where $n_i^{\text{eff}} = \frac{1}{\sum_{j \in I_b} w_{i,j}^2}$, with $w_{i,j} = \frac{k_\gamma(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j))}{\sum_j k_\gamma(\phi(\mathbf{x}_i), \phi(\mathbf{x}_j))}$.

We provide the proof in Appendix. This decomposition highlights the fundamental bias-variance trade-off induced by kernel smoothing: tighter kernels reduce bias at the expense of increased variance, while broader kernels reduce variance but incur larger bias. Therefore, estimating the order of magnitude of these terms clarifies the state of *local calibration* for a model.

In conclusion, the results presented demonstrate that both binning-based metrics and kernel-based metrics like LCE admit a probabilistic decomposition under the assumption of *local calibration*. This theoretical perspective clarifies the role of *local calibration* in shaping the behavior of multiclass calibration metrics and highlights practical limitations of the metrics themselves in capturing the phenomenon.

7 Improving Local Calibration in Practice

We now focus on improving the *local calibration* of Neural Networks in practice. We first introduce *Local Calibration Networks* (LCN), i.e., a two-component neural network architecture, designed to produce representations that exhibit improved *local calibration*.

Figure 1a shows LCNs structure. While standard post hoc calibrators consider the features representation as fixed, one component of LCNs produces a new reduced-dimensionality feature representation $\phi'(\mathbf{x}_i)$ and the other component parametrizes new logits $\mathbf{l}'$. These two components can be trained to minimize the misalignment between the probabilistic outputs and the local estimates of the class distribution computed from $\phi'(\mathbf{x}_i)$. To obtain these estimates, one can resort to kernel-based methods, e.g., Nadaraya-Watson estimators [17, 27], which require specifying a bandwidth hyper-parameter γ . More precisely, LCNs minimize the following loss:

$$\mathcal{L}_{\text{icl}}(\mathbf{x}_i, \mathbf{y}_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{\text{d}_{\text{JSD}}(\hat{\mathbf{p}}_i, \hat{\theta}(\mathbf{y}_i | \phi'(\mathbf{x}_i)))}_{\text{Alignment term}} + \lambda \cdot \underbrace{\mathcal{L}_{\text{ce}}(\mathbf{y}_i, \hat{\theta}(\mathbf{y}_i | \phi'(\mathbf{x}_i)))}_{\text{Similarity term}}, \quad (8)$$

where $\text{d}_{\text{JSD}}(P, Q)$ is the Jensen-Shannon distance¹[2, 3] and $\mathcal{L}_{\text{ce}}$ is the categorical cross entropy. Notice that two distinct terms compose $\mathcal{L}_{\text{icl}}$, i.e., the *alignment term* and the *similarity term*.

On the one hand, the *alignment term* leverages the Jensen-Shannon distance between the model's predicted probability vector and the kernel estimates for each instance in the training batch. We prove

¹ $\text{d}_{\text{JSD}}(P, Q) := \sqrt{\frac{1}{2} \text{KL}(P \| \frac{P+Q}{2}) + \frac{1}{2} \text{KL}(Q \| \frac{P+Q}{2})}$

that, for a consistent estimator, the *alignment term* asymptotically converges to the divergence between the model prediction distribution and the true label distribution.

Theorem 4 (Asymptotic consistency of JSD). *For a probabilistic classifier and a consistent (in the MSE sense) estimator of point-wise conditional probabilities, the average Jensen-Shannon distance between the model confidences and the kernel estimates converges to the one computed using the true distribution $\mathbf{p}$, i.e.,:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n d_{\text{JSD}}(\hat{\mathbf{p}}_i, \hat{\theta}(\mathbf{y} \mid \mathbf{x}_i)) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n d_{\text{JSD}}(\hat{\mathbf{p}}_i, \mathbf{p}_i). \quad (9)$$

We provide the proof in the Appendix. Notably, the *alignment term* allows to learn probabilistic outputs that match the empirical neighbourhood frequencies.

On the other hand, the *similarity term* leverages the categorical cross entropy between the ground truth and the kernel estimates. Intuitively, this term encourages points with the same label to be attracted to nearby neighbours, but because the kernel is local, distant points of the same class exert little influence on each other. As a result, points that share fine-grained similarities are placed closer together, while more distinct variants remain further apart, but within the same class cluster. We provide a visual intuition of this behavior in Figure 1b. However, the *similarity term* could still collapse class representations if not regularized or if the kernel bandwidth is too wide. Hence, we introduce another regularization hyper-parameter λ (to be fine-tuned) to prevent this behaviour.

As a concluding remark, our method leverages kernel estimates during training but does not require them for inference, therefore fully maintaining the efficiency of feed-forward neural networks.

8 Experimental Evaluation

In this section, we address the following questions:

- **Q1:** Does our method match the performance of baselines in the global calibration metric?
- **Q2:** Does our proposal outperform baselines on *local calibration* metrics?
- **Q3:** Does our approach affect predictions?

Datasets and Methods. We evaluate our research questions over three multiclass datasets, i.e., `cifar10`, `cifar100` [11], and `tissuemnist` from the MedMNIST collection [29]. We evaluate our approach, LCN, against publicly available calibration baselines, including Temperature Scaling (TS) [5], Isotonic Regression (IR) [31], and Platt Scaling (PS) [20]. In addition, we compare with Dirichlet Calibration (DC) [13], the current state-of-the-art for multiclass calibration. We train a separate model for each dataset: a ResNet-50 backbone for CIFAR-10 and TissueMNIST, and a deeper ResNet-152 for CIFAR-100. Hyperparameter settings are provided in the Appendix D. The code to reproduce our results can be found at <https://github.com/Cesbar99/local-calibration/>.

Metrics. For **Q1**, we consider two global calibration metrics, i.e., *ECE* and *ECCE*. For **Q2**, we compute two *local calibration* metrics, i.e., *LCE* and *MLCE*. For **Q3**, we report two performance measures, i.e., accuracy (*Acc*) and Negative Log-Likelihood (*NLL*).

Experimental Setup. To evaluate calibration for all methods, we separate the data into three disjoint parts: (i) a training set (further split into training and validation); (ii) a calibration set (with an internal calibration/validation split); and (iii) a held-out test set used exclusively for evaluation. For CIFAR-10 and CIFAR-100, we use 45% of the original training data for training, 10% for validation, and 45% as test data, with the original test split serving as calibration data. For TissueMNIST, we use the pre-computed splits, but since calibration metrics computation requires larger sample sizes to be meaningful, we split the training data in half, yielding balanced training and test splits ($\approx 82.5k$ each). The remaining pre-computed test set ($\approx 40k$ instances) is used for calibration. For all the datasets, we split the calibration

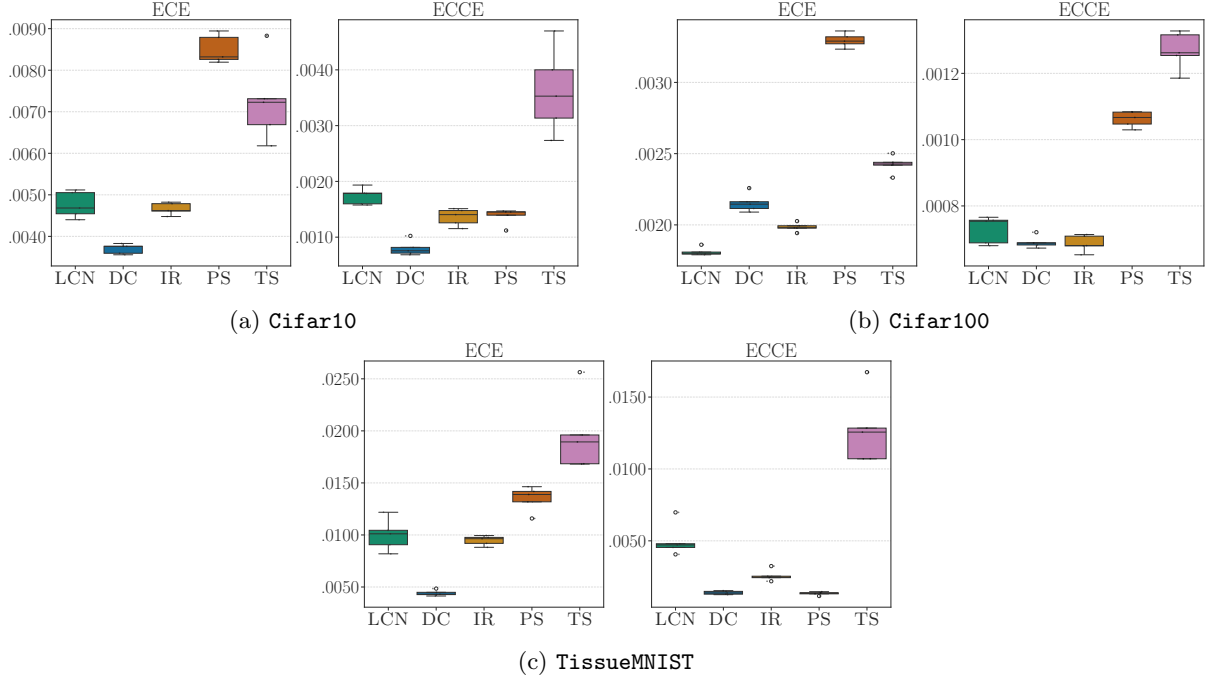


Figure 2: **Empirical global calibration metrics (Q1) over five runs.** The lower the better.

data into two sets, one to learn the calibration technique (90% of the calibration data) and one for validation (the remaining 10% of the calibration data). Then, we *(i)* train classifiers on the training set; *(ii)* calibrate using the different methods on the calibration set; *(iii)* compute the six metrics on the test set. We repeat this procedure using 5 different seeds and average results. Implementation details are provided in the Appendix D.

8.1 Experimental Results

Q1: LoCal Nets achieve competitive global calibration results across datasets. Figure 2 summarizes the performance of all methods on global calibration metrics across the three datasets. Our aim here is to show that LCN performs at a comparable level with the existing baselines. For *ECE*, DC is the strongest performer over *cifar10* with $ECE \approx .0037 \pm .0001$, followed by LCN with $ECE \approx .0047 \pm .0003$ (Figure 2a). For *ECCE*, DC again leads with $\approx .0008 \pm .0001$, while LCN achieves a level comparable to the other baselines. On *cifar100*, our method (LCN) achieves the best performance in terms of *ECE* and is tied with DC and IR on *ECCE* (Figure 2b). On *tissuemnist*, DC again performs best with $ECE \approx .0044 \pm .0002$, while LCN ranks second (Figure 2c). For *ECCE*, LCN performs slightly worse, obtaining $\approx .005 \pm .001$.

In summary, DC is the overall best-performing method for global calibration, with its most pronounced advantage on balanced, low-class datasets such as *cifar10*. Still, our proposed LCN consistently achieves competitive results across datasets.

Q2: LoCal Nets consistently achieve superior results on local calibration metrics. Figure 3 evaluates the methods in terms of *local calibration*. Across all datasets, LCN consistently emerges as the best-performing approach, providing substantial improvements over competing methods. On *cifar10*, LCN achieves the lowest *LCE* ($.0078 \pm .0002$), followed by DC ($.010 \pm .0003$), with TS and IR performing at a similar level to DC (Figure 3a). In terms of *MLCE*, LCN significantly outperforms all competitors with differences as large as .17. On *cifar100*, LCN again attains the best *LCE* setting an evident gap with all the competitors. (Figure 3b). Here, PS achieves a *MLCE* of $.7030 \pm .0366$, nearly matching LCN ($.7022 \pm .0063$). However, the lower variance of LCN highlights its reliability. The relative competitiveness of PS on this dataset is due to the limited per-class sample size (at most 600 instances), which favors parametric methods such as PS, whereas LCN benefits more from larger

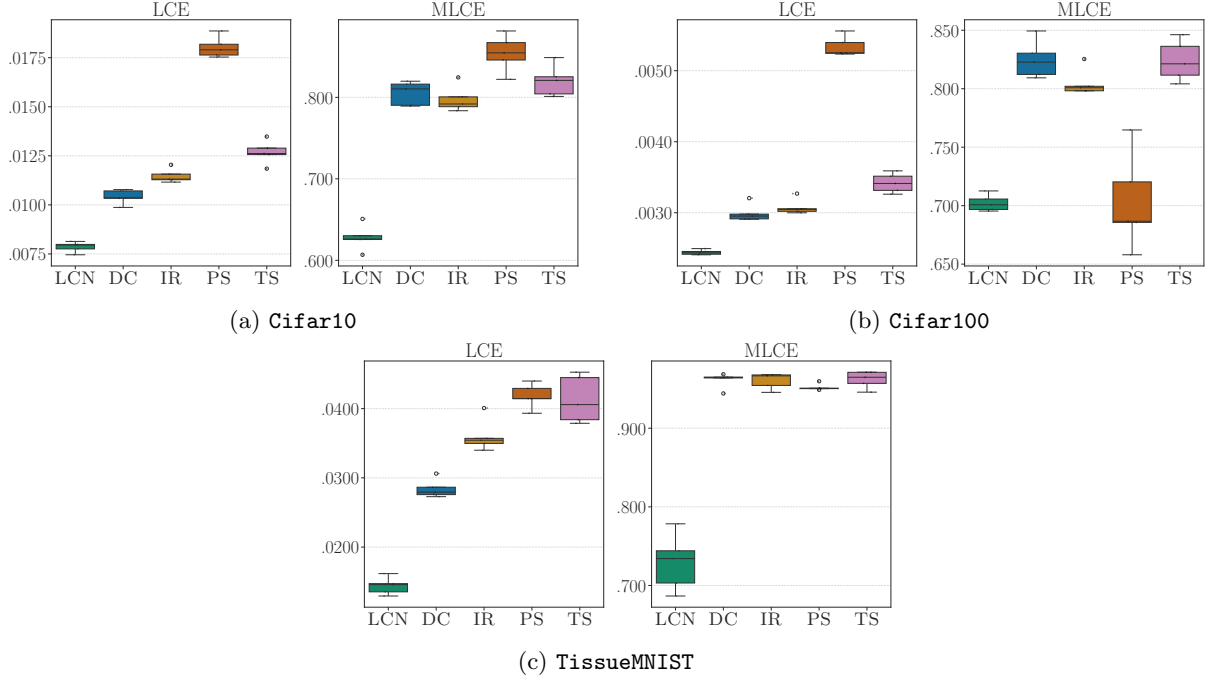


Figure 3: **Empirical local calibration metrics (Q2) over five runs.** The lower the better.

sample sizes. The largest relative gains are observed on `tissuemnist` (Figure 3c), where LCN achieves $.00143 \pm .0011$, markedly outperforming DC ($.0028 \pm .0011$) and all other baselines.

Overall, these results demonstrate that LCN provides significant improvements in *local calibration*, particularly in settings with sufficient sample size.

Q3: Beyond calibration, LoCal Nets enhance predictive performance. Table 1 reports results on predictive performance, measured by *NLL* and *Acc*. Our method achieves the largest reductions in *NLL* across datasets, with the sole exception of `cifar10`, where LCNs rank second ($.347 \pm .002$) compared to DC ($.333 \pm .007$). On `cifar100`, LCN clearly outperforms all competitors with the lowest *NLL* ($1.125 \pm .002$ vs. $1.265 \pm .016$ for DC). A similar trend is observed on `tissuemnist`, where LCN achieves $1.012 \pm .003$, again surpassing DC ($1.052 \pm .009$). Notably, LCN is the only method that improves model accuracy across datasets: $+0.4\%$ on `cifar10`, $+1.9\%$ on `cifar100`, and $+2.7\%$ on `tissuemnist`. This occurs as other methods are forced to work only on logits, while LCN can learn new feature representations, possibly improving the predictions’ quality.

These results highlight that beyond improving calibration, our approach translates into tangible gains in predictive performance, particularly on challenging datasets with larger class cardinality or imbalance.

9 Conclusions

In this work, we introduced a formal definition of multiclass *local calibration* and theoretically analyzed widely used calibration metrics under this assumption. Building on these insights, we proposed a novel post-hoc calibration method for neural networks that explicitly targets *local calibration* properties. Our empirical evaluation on both benchmarking and real-world datasets demonstrates that the proposed approach yields significant improvements in *local calibration*, while maintaining competitive performance with respect to global calibration metrics. These results highlight the importance of incorporating locality into the design of calibration methods.

Limitations and Future Work. Although Theorem 4 establishes the consistency of our loss function, this guarantee holds only asymptotically. In finite-sample regimes, kernel estimates may suffer from non-negligible bias, which can limit performance. An important research direction is to explore

Table 1: Results for **Q3**. We report *Acc* (the higher the better) and *NLL* (the lower the better) across datasets. We highlight in bold the best performer and we underscore the second-best method for each dataset and metric.

Method	Cifar10		Cifar100		TissueMNIST	
	<i>Acc</i>	<i>NLL</i>	<i>Acc</i>	<i>NLL</i>	<i>Acc</i>	<i>NLL</i>
LCN (Ours)	.888 ± .002	<u>.347 ± .002</u>	.688 ± .001	1.125 ± .002	.630 ± .001	1.012 ± .003
DC	.884 ± .001	.333 ± .007	.670 ± .002	<u>1.265 ± .016</u>	.603 ± .008	<u>1.052 ± .009</u>
IR	.884 ± .001	.364 ± .008	.670 ± .002	1.437 ± .029	.603 ± .008	1.096 ± .016
PS	.884 ± .001	.466 ± .004	.670 ± .002	1.618 ± .007	.603 ± .008	1.180 ± .008
TS	.884 ± .001	.362 ± .008	.670 ± .002	1.277 ± .009	.603 ± .008	1.112 ± .023

adaptive kernel choices and scalable training procedures that better manage the bias–variance tradeoff inherent to local estimation. Moreover, the hyperparameter γ requires careful tuning: small values yield unreliable estimates due to data sparsity, while large values obscure locality. A promising extension is to replace fixed kernels with adaptive or learned similarity functions, which may improve *local calibration* in high-dimensional or heterogeneous feature spaces. Finally, while our method is tailored to neural networks, extending it to enforce *local calibration* across other classes of models remains an important open direction.

Acknowledgements

The research of Cesare Barbera, Giovanni De Toni, Andrea Passerini, and Andrea Pugnana was partially supported by the following projects: Horizon Europe Programme, grants #101120237-ELIAS and #101120763-TANGO. Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The research of Cesare Barbera, Giovanni De Toni, Andrea Passerini, and Andrea Pugnana was also supported by Ministero delle Imprese e del Made in Italy (IPCEI Cloud DM 27 giugno 2022 – IPCEI-CL-0000007), PNRR-M4C2-Investimento 1.3, Partenariato Esteso PE00000013-“FAIR-Future Artificial Intelligence Research”, funded by the European Commission under the NextGeneration EU programme, and SoBigData.it, receiving funding from European Union – NextGenerationEU – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) – Project: “SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics” – Prot. IR0000013 – Avviso n. 3264 del 28/12/2021.

References

- [1] Telmo de Menezes e Silva Filho, Hao Song, Miquel Perelló-Nieto, Raúl Santos-Rodríguez, Meelis Kull, and Peter A. Flach. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach. Learn.*, 112(9):3211–3260, 2023.
- [2] Dominik Maria Endres and Johannes E. Schindelin. A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860, 2003.
- [3] Bent Fuglede and Flemming Topsøe. Jensen-shannon divergence and hilbert space embedding. In *ISIT*, page 31. IEEE, 2004.
- [4] Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *CoRR*, abs/1704.00805, 2017.
- [5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 2017.

- [6] Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In *ICLR*. OpenReview.net, 2022.
- [7] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *ICLR*. OpenReview.net, 2021.
- [8] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer, 2002.
- [9] Imanol Arrieta Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. Metrics of calibration for probabilistic predictions. *J. Mach. Learn. Res.*, 23:351:1–351:54, 2022.
- [10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.
- [11] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [12] Meelis Kull, Telmo de Menezes e Silva Filho, and Peter A. Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 623–631. PMLR, 2017.
- [13] Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, pages 12295–12305, 2019.
- [14] Adrien Le-Coz, Stéphane Herbin, and Faouzi Adjed. Confidence calibration of classifiers with many classes. In *NeurIPS*, 2024.
- [15] Rachel Luo, Aadyot Bhatnagar, Yu Bai, Shengjia Zhao, Huan Wang, Caiming Xiong, Silvio Savarese, Stefano Ermon, Edward Schmerling, and Marco Pavone. Local calibration: metrics and recalibration. In *UAI*, volume 180 of *Proceedings of Machine Learning Research*, pages 1286–1295. PMLR, 2022.
- [16] Charlie Marx, Sofian Zalouk, and Stefano Ermon. Calibration by distribution matching: Trainable kernel calibration metrics. In *NeurIPS*, 2023.
- [17] Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.
- [18] Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pages 2901–2907. AAAI Press, 2015.
- [19] Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. Beyond calibration: estimating the grouping loss of modern neural networks. In *ICLR*. OpenReview.net, 2023.
- [20] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.
- [21] Abhishek Singh Sambyal, Usma Niyaz, Narayanan C. Krishnan, and Deepti R. Bathula. Understanding calibration of deep neural networks for medical image classification. *Comput. Methods Programs Biomed.*, 242:107816, 2023.
- [22] Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.
- [23] Juozas Vaicenavicius, David Widmann, Carl R. Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. Evaluating model calibration in classification. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR, 2019.
- [24] Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, Ewout W Steyerberg, Topic Group ‘Evaluating diagnostic tests, and prediction models’ of the STRATOS initiative Bossuyt Patrick Collins Gary S. Macaskill Petra McLernon David J. Moons Karel GM Steyerberg Ewout W. Van Calster Ben van Smeden Maarten Vickers Andrew J. Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230, 2019.

- [25] Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *CoRR*, abs/2308.01222, 2023.
- [26] Larry Wasserman. *All of nonparametric statistics*. Springer, 2006.
- [27] Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372, 1964.
- [28] Miao Xiong, Ailin Deng, Pang Wei W. Koh, Jiaying Wu, Shen Li, Jianqing Xu, and Bryan Hooi. Proximity-informed calibration for deep neural networks. In *NeurIPS*, 2023.
- [29] Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *ISBI*, pages 191–195. IEEE, 2021.
- [30] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, pages 609–616. Morgan Kaufmann, 2001.
- [31] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, pages 694–699. ACM, 2002.
- [32] Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *NeurIPS*, pages 22313–22324, 2021.

A General Assumptions

Before presenting individual proofs and continuing further the discussion, we first outline the general assumptions that hold throughout this Appendix. These assumptions simplify notation while preserving full generality of the results and hold unless otherwise specified.

A.1 Notation and Binning Assumptions.

We consider a multi-class classification setting, where $\mathcal{X} \subseteq \mathbb{R}^m$ is the feature space and $\mathcal{Y} = \{0, \dots, C-1\}$ is a finite target space with C distinct labels. Let us assume we have access to a given dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ of input-output pairs drawn from an unknown joint distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. Each input $\mathbf{x}_i \in \mathcal{X}$ is a feature vector of m dimensions, and each label $y_i \in \mathcal{Y}$ has a corresponding one-hot encoded vector $\mathbf{y}_i$ indicating the correct class among the C possible classes. We consider a probabilistic classifier $f: \mathcal{X} \rightarrow \Delta^C$, where Δ^C is the $(C-1)$ -dimensional probability simplex. In words, a probabilistic classifier maps an input $\mathbf{x}$ to a probability distribution over classes, i.e., $f(\mathbf{x}) = \hat{\mathbf{p}} \in \Delta^C$, where each entry $\hat{p}_k = f_k(\mathbf{x})$ of the predicted probability vector $\hat{\mathbf{p}}$ denotes the predicted probability of class k .

Throughout this appendix, we present the theoretical results and corresponding proofs under a unified framework that simultaneously covers *multi-dimensional binning*-based calibration metrics and *class-wise binning*-based metrics constructed using *equal-frequency binning*. This formulation is adopted for clarity and compactness, as it allows for a seamless integration of both frameworks under a shared notation. For completeness, at the end of each proof that requires it, we explicitly discuss how the same analytical results extend to the case of class-wise calibration metrics based on *fixed-confidence-threshold binning*, which involves only minor technical adjustments. Hence, the presented analysis extends naturally to both binning schemes.

A.2 Conditioning Assumptions.

All probabilistic bounds are derived under the following conditioning setup:

- **Conditioning on features.** We condition on the observed feature values $\{\mathbf{x}_i\}_{i=1}^n$. Since the binning rule $\beta(\cdot)$ is deterministic, this also fixes the bin memberships $\{I_b\}_{b=1}^B$. Hence, after conditioning, the sets of indices per bin are deterministic.
- **Conditioning on kernel estimates.** We condition on the kernel estimates $\hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)$, which are computed on an independent, unlimited disjoint dataset. For the evaluation set, these estimates depend only on the observed $\mathbf{x}_i$; thus, once features are fixed, the estimates are deterministic as well.

A.3 Kernel Consistency Assumptions

In the following, we provide the underlying assumptions for the kernel estimator consistency:

1. **Data Assumptions.** The evaluation set $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is drawn independently and identically distributed (i.i.d.) from a joint distribution over a compact subset of $\mathcal{X} \times \mathcal{Y}$. The marginal probability density function $p(\mathbf{x})$ and the true conditional probability function $p(\mathbf{y}|\mathbf{x})$ are continuous and bounded on $\mathcal{X}$.
2. **Kernel Assumptions.** The kernel function k is a non-negative, symmetric, and bounded function that integrates to one, i.e., $\int_{\mathbb{R}^d} k(u) du = 1$.
3. **Bandwidth Assumptions.** The bandwidth parameter γ_n is a positive sequence that depends on the sample size n and satisfies the following conditions as $n \rightarrow \infty$:
 - (a) $\gamma_n \rightarrow 0$ (the bandwidth shrinks).
 - (b) $n\gamma_n^d \rightarrow \infty$, where d is the dimensionality of $\mathcal{X}$

It is important to acknowledge that estimates do not constitute a point-wise unbiased approximation of the true conditional distribution. Kernel estimators are known to suffer from both *design bias*—a

form of bias introduced by the distribution of the covariates x —and *boundary bias*, which is particularly pronounced near the edges of the support or in low-density regions of the input space [26]. Nevertheless, NW estimators are known to be consistent in the mean squared error (MSE) sense under mild conditions on the kernel and bandwidth sequence [8, 26]. Note that the MSE consistency is a *stronger* notion that implies Mean Integrated Squared Error (MISE) and Mean Absolute Error (MAE) consistencies. As a consequence, the estimator asymptotically yields a statistically meaningful approximation of the target function:

$$\mathbb{E}\left[\left(\hat{\theta}(\mathbf{y} \mid \mathbf{x}) - \Pr(\mathbf{y} \mid \mathbf{x})\right)^2\right] \xrightarrow{n \rightarrow \infty} 0, \quad (10)$$

B Proofs

B.1 Proof of Theorem 1

Given a *locally calibrated* classifier f , meaning that for instance $i \in D$, $\left\|f(\mathbf{x}_i) - \hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)\right\|_1 \leq \varepsilon$, where $\hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)$ is a kernel estimator of the true conditional expectation $\mathbb{E}[\mathbf{y}_i \mid \mathbf{x}_i]$, we aim to bound the *Multidimensional Expected Calibration Error (MECE)*, defined as:

$$\mathbb{E}_{x \sim p(x)} \mathbb{E}_{y \sim p(y)} [|\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})|],$$

by showing there exists $k \in [1/C, 1]$ such that the following holds:

$$MECE \leq \varepsilon \cdot k.$$

Proof. We can rewrite this quantity by marginalizing over the input and output spaces:

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \mathbb{E}_{y \sim p(y)} [|\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})|] &= \int_{\mathcal{X}} \left[\sum_{y \in \mathcal{Y}} p(y) \cdot |\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})| \right] p(\mathbf{x}) d\mathbf{x} \\ &\leq \int_{\mathcal{X}} \left[\sum_{y \in \mathcal{Y}} \max_{y \in \mathcal{Y}} p(y) \cdot |\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})| \right] p(\mathbf{x}) d\mathbf{x} \\ &\leq \max_{y \in \mathcal{Y}} \int_{\mathcal{X}} \left[\sum_{y \in \mathcal{Y}} \cdot |\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})| \right] p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

By the triangle inequality we obtain:

$$\mathbb{E}_{x \sim p(x)} \mathbb{E}_{y \sim p(y)} [|\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})|] \leq \max_{y \in \mathcal{Y}} p(y) \int_{\mathcal{X}} \left(\left\| \hat{\theta}(\mathbf{y} \mid \mathbf{x}) - f(\mathbf{x}) \right\|_1 + \left\| \mathbb{E}[\mathbf{y} \mid \mathbf{x}] - \hat{\theta}(\mathbf{y} \mid \mathbf{x}) \right\|_1 \right) p(x) dx.$$

Since f is locally calibrated, the first integrand term is bounded pointwise:

$$\left\| \hat{\theta}(\mathbf{y} \mid \mathbf{x}) - f(\mathbf{x}) \right\|_1 \leq \varepsilon.$$

The second term, $\int_{\mathcal{X}} \left\| \mathbb{E}[\mathbf{y} \mid \mathbf{x}] - \hat{\theta}(\mathbf{y} \mid \mathbf{x}) \right\|_1 p(\mathbf{x}) d\mathbf{x}$, vanishes asymptotically under mild regularity assumptions on the kernel and the data distribution (refer to Appendix A.3 for all the details):

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X}} \left\| \mathbb{E}[\mathbf{y} \mid \mathbf{x}] - \hat{\theta}(\mathbf{y} \mid \mathbf{x}) \right\|_1 p(x) dx = 0.$$

Hence, for sufficiently large n , this second term can be made arbitrarily small. Denoting it by $\delta_n \rightarrow 0$, we provide the final bound for a constant $k \geq \max_{y \in \mathcal{Y}} p(y)$:

$$\mathbb{E}_{x \sim p(x)} \mathbb{E}_{y \sim p(y)} [|\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})|] \leq (\varepsilon + \delta_n) \cdot k.$$

As $n \rightarrow \infty$, this yields:

$$\mathbb{E}_{x \sim p(x)} \mathbb{E}_{y \sim p(y)} [|\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(x)|] \leq \varepsilon \cdot k. \quad \square \quad (11)$$

B.1.1 Proof of Theorem 2

In the following we prove that the value of a calibration metric satisfying the requisites of Definition 2 for a model that satisfies *local calibration* is bounded. More precisely, the proof will show that for a generic bin used to compute the metric, the calibration error in the bin is bounded. Finally, we will extend this to all bins simultaneously to bound the calibration error captured by the metric for a model that exhibits *local calibration*.

Let a deterministic binning function $\beta : \Delta_c \rightarrow \{1, \dots, m_B\}$ partition the probability simplex Δ^C into B disjoint bins $\{B_b\}_{b=1}^{m_B}$. For each bin B_b , define the index set of points that fall into it as $I_b = \{i : \hat{\mathbf{p}}_i \in B_b\}$, and let $|B_b|$ denote its cardinality. Define the per-bin frequency and confidence:

$$\text{freq}_{b,c} = \frac{1}{|B_b|} \sum_{i \in I_b} \mathbf{1}\{y_i = c\}, \quad \text{conf}_{b,c} = \frac{1}{|B_b|} \sum_{i \in I_b} f_c(\mathbf{x}_i),$$

Let $\varphi : [0, 1] \times [0, 1] \rightarrow \mathbb{R}_{\geq 0}$ a scalar comparator that measure discrepancy between an empirical frequency and a predicted confidence that is Lipschitz in both arguments:

$$|\varphi(a, b) - \varphi(a', b')| \leq L_\varphi(|a - a'| + |b - b'|) \quad \forall a, a', b, b' \in [0, 1]$$

The *general multiclass bin-based calibration error metric* is obtained as:

$$\mathcal{E}(D; \varphi; \beta) = \sum_{b=1}^{m_B} w_b \sum_{c=1}^C \pi_c \cdot \varphi(\text{freq}_{b,c}, \text{conf}_{b,c}). \quad (12)$$

where w_b are deterministic bin class weights typically set to $|B_b|/n$ and π_c are deterministic class weights typically set to $1/C$ for balanced data. We now prove that if the model satisfies *local calibration*, i.e.,

$$\|f(\mathbf{x}_i) - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)\|_1 \leq \varepsilon, \quad \forall i \in \{1, \dots, n\},$$

then, with probability at least as high as $1 - \delta \in [0, 1]$, the metric is bounded as follows:

$$\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \left[\varepsilon + \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|\Psi(b, \mathcal{Y})|}} \right].$$

Where $\Psi(\cdot; \cdot)$ is a function that selects a bin based on index b and labels $\mathcal{Y}$.

Proof. We begin our proof by fixing a bin b and a class c and, given the per-instance kernel estimates $\hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)$, we define the local estimator average:

$$\hat{\theta}_{b,c} := \frac{1}{|B_b|} \sum_{i \in I_b} \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i).$$

Then, for any fixed bin b and class c ,

$$\varphi(\text{freq}_{b,c}, \text{conf}_{b,c}) = \varphi(\text{freq}_{b,c}, \hat{\theta}_{b,c}) + [\varphi(\text{freq}_{b,c}, \text{conf}_{b,c}) - \varphi(\text{freq}_{b,c}, \hat{\theta}_{b,c})].$$

By the Lipschitz property,

$$|\varphi(\text{freq}_{b,c}, \text{conf}_{b,c}) - \varphi(\text{freq}_{b,c}, \hat{\theta}_{b,c})| \leq L_\varphi |\text{conf}_{b,c} - \hat{\theta}_{b,c}|.$$

Moreover, since φ is also Lipschitz in its first argument,

$$\varphi(\text{freq}_{b,c}, \hat{\theta}_{b,c}) \leq |\varphi(\text{freq}_{b,c}, \hat{\theta}_{b,c}) - \varphi(\hat{\theta}_{b,c}, \hat{\theta}_{b,c})| \leq L_\varphi |\text{freq}_{b,c} - \hat{\theta}_{b,c}|,$$

as a direct consequence of $\varphi(t, t) = 0$.

Combining the two we obtain:

$$\varphi(\text{freq}_{b,c}, \text{conf}_{b,c}) \leq L_\varphi (|\text{freq}_{b,c} - \hat{\theta}_{b,c}| + |\text{conf}_{b,c} - \hat{\theta}_{b,c}|). \quad (13)$$

Using (12) and (13), by pulling constants outside,

$$\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \sum_{b=1}^B w_b \sum_{c=1}^C \pi_c (|\text{freq}_{b,c} - \hat{\theta}_{b,c}|) + L_\varphi \sum_{b=1}^B w_b \sum_{c=1}^C \pi_c (|\text{conf}_{b,c} - \hat{\theta}_{b,c}|). \quad (14)$$

We bound the two terms inside parentheses separately.

(ii) **Miscalibration:** $|\text{conf}_{b,c} - \hat{\theta}_{b,c}|$.

$$\text{conf}_{b,c} - \hat{\theta}_{b,c} = \frac{1}{|B_b|} \sum_{i \in I_b} (\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)).$$

By the local calibration assumption, for every instance i we have $\|\hat{\mathbf{p}}_i - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)\|_1 \leq \varepsilon$. Consequently each coordinate satisfies $|\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)| \leq \varepsilon$, hence

$$|\text{conf}_{b,c} - \hat{\theta}_{b,c}| \leq \frac{1}{|B_b|} \sum_{i \in I_b} |\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)| \leq \varepsilon. \quad (15)$$

(i) **Empirical fluctuation:** $|\text{freq}_{b,c} - \hat{\theta}_{b,c}|$. For fixed b, c ,

$$\text{freq}_{b,c} - \hat{\theta}_{b,c} = \frac{1}{|B_b|} \sum_{i \in I_b} (\mathbf{1}\{y_i = c\} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)).$$

In the following we apply Hoeffding inequality to bound the **Empirical fluctuation** term. More precisely, to apply Hoeffding, we need independent, bounded, zero-mean summands.

Under the **conditioning assumptions** (Appendix A.2), the only source of randomness in the term

$$\text{freq}_{b,c} - \hat{\theta}_{b,c} = \frac{1}{|B_b|} \sum_{i \in I_b} (\mathbf{1}\{y_i = c\} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i))$$

is the label variables $\{y_i\}_{i \in I_b}$. For each i , the summand satisfies:

1. **Zero mean:**

$$\mathbb{E}[\mathbf{1}\{y_i = c\} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i) \mid \mathbf{x}_i] = P(y_i = c \mid \mathbf{x}_i) - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i) \approx 0,$$

and exactly zero if the estimator is consistent in mean absolute error (MAE) (refer to Appendix A.3 for all the details).

2. **Independence:** the pairs $(\mathbf{x}_i, y_i)$ are i.i.d., and conditioning on the $\mathbf{x}_i$ leaves the labels $\{y_i\}$ independent.

3. **Boundedness:** both $\mathbf{1}\{y_i = c\}$ and $\hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)$ lie in $[0, 1]$, hence their difference lies in $[-1, 1]$.

Hoeffding's inequality applies to their average, yielding the desired concentration bound. More precisely, Hoeffding's inequality gives, for any $\tau > 0$,

$$\Pr\left(|\text{freq}_{b,c} - \hat{\theta}_{b,c}| > \tau\right) \leq 2 \exp(-2|B_b|\tau^2).$$

Choosing $\tau_b = \sqrt{\frac{\log(2Cm_B/\delta)}{2|B_b|}}$ and applying the union bound over the $m_B \cdot C$ bin-class pairs yields: with probability at least $1 - \delta$,

$$\forall b, c: \quad |\text{freq}_{b,c} - \hat{\theta}_{b,c}| \leq \sqrt{\frac{\log(2Cm_B/\delta)}{2|B_b|}}. \quad (16)$$

Now insert (15) and (16) into (14). With probability at least $1 - \delta$,

$$\begin{aligned} \mathcal{E}(D; \varphi; \beta) &\leq L_\varphi \sum_{b=1}^{m_B} w_b \sum_{c=1}^C \pi_c \left(\sqrt{\frac{\log(2Cm_B/\delta)}{2|B_b|}} + \varepsilon \right) \\ &= L_\varphi \left(\varepsilon \sum_{c=1}^C \pi_c \sum_{b=1}^{m_B} w_b + \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|B_b|}} \sum_c \pi_c \right). \end{aligned}$$

Using $\sum_b w_b = 1$ and $\sum_c \pi_c = 1$, this simplifies to

$$\Pr\left(\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \left[\varepsilon + \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|B_b|}} \right]\right) \geq 1 - \delta. \quad (17)$$

Therefore achieving the form of theorem 2 when $\Psi(b, \mathcal{Y}) = B_b$.

Confidence Based Binning In the context of class-wise metrics that use confidence based binning, the only difference is the introduction of a dependence between bins' cardinalities and the classes: $|B_b| \rightarrow |B_{b,c}|$. Then, with probability at least $1 - \delta$, the bound applies with the same exact form under the following minor changes:

$$\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \sum_{b=1}^B w_b \sum_{c=1}^C \pi_c \left(\sqrt{\frac{\log(2Cm_B/\delta)}{2|B_{b,c}|}} + \varepsilon \right) \leq L_\varphi \left[\varepsilon + \sum_{b=1}^B w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2 \min_c |B_{b,c}|}} \right] \quad (18)$$

and by setting $\Psi(b, \mathcal{Y}) = B_{b,c^*}$ and $c^* = \arg \min_{c \in \mathcal{Y}} |B_{b,c}|$ we obtain a bound of the same form of theorem 2. $\square$

Proof of Corollary 1.

Proof. Recall that by the local calibration assumption, for every instance $i \in D$ we have $\|\hat{\mathbf{p}}_i - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)\|_1 \leq \varepsilon$. Consequently, each coordinate satisfies $|\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)| \leq \varepsilon$, hence:

$$|\text{conf}_{b,c} - \hat{\theta}_{b,c}| \leq \frac{1}{|B_b|} \sum_{i \in I_b} |\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)| \leq \varepsilon.$$

If additionally the model f satisfies ρ -perfect uniform local calibration. Then, for every instance $i \in D$ and class $c \in \{1, \dots, C\}$, the absolute calibration error is bounded:

$$|\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i | \mathbf{x}_i)| \leq L \cdot \rho$$

where $L \cdot \rho$ is the maximum variation in the predicted probability for class c within the isotropic neighborhood of radius ρ . In this context the miscalibration error can be further reduced:

$$\|\hat{\theta}(\mathbf{y}_i | \mathbf{x}_i) - \hat{\mathbf{p}}_i\|_1 \leq C \cdot L \cdot \rho$$

Substituting back into (6) we obtain:

$$\Pr \left(\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \left[C \cdot L \cdot \rho + \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|\Psi(b, \mathcal{Y})|}} \right] \right) \geq 1 - \delta.$$

Conclusion: In the limit of ρ -perfect local calibration, the calibration error reduces to pure stochastic fluctuation:

$$\limsup_{\rho \rightarrow 0} \Pr \left(\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|\Psi(b, \mathcal{Y})|}} \right) \geq 1 - \delta. \quad (7)$$

$\square$

B.1.2 Proof of Theorem 3

Proof. Let $f : \mathcal{X} \rightarrow \Delta^C$ be a probabilistic classifier, where f is composed of a feature extractor $\phi : \mathcal{X} \rightarrow \mathcal{F}$ and a final classification layer $g : \mathcal{F} \rightarrow \Delta^C$. Assume f to be locally calibrated up to error ε and let $k_\gamma(\mathbf{x}_i, \mathbf{x}_j)$ kernel functions to obtain the kernel-weighted mean of both the empirical frequencies and the predicted probabilities for given anchor point $\mathbf{x}_i$:

$$\begin{aligned} \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i) &:= \sum_{j \in I_b} \frac{k_\gamma(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j \in I_b} k_\gamma(\mathbf{x}_i, \mathbf{x}_j)} \mathbf{y}_j, \\ \hat{\theta}(\hat{\mathbf{p}}_i | \mathbf{x}_i) &:= \sum_{j \in I_b} \frac{k_\gamma(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j \in I_b} k_\gamma(\mathbf{x}_i, \mathbf{x}_j)} \hat{\mathbf{p}}_j. \end{aligned}$$

For a given bin b we write the value of LCE:

$$\text{LCE} = \frac{1}{C} \sum_{b=1}^B \frac{|B_b|}{n} \frac{1}{|B_b|} \sum_{i \in b} \left\| \frac{\sum_{j \in b} (\hat{\mathbf{p}}_j - \mathbf{y}_j) k_\gamma(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{j \in b} k_\gamma(\mathbf{x}_i, \mathbf{x}_j)} \right\|_1 = \frac{1}{C} \sum_{b=1}^B \frac{|B_b|}{n} \frac{1}{|B_b|} \sum_{i \in b} \|\hat{\theta}(\hat{\mathbf{p}}_i | \mathbf{x}_i) - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)\|_1$$

We can rewrite as follows:

$$\begin{aligned} \frac{1}{|B_b|} \sum_{i \in b} \|\hat{\theta}(\hat{\mathbf{p}}_i | \mathbf{x}_i) - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)\|_1 &= \frac{1}{|B_b|} \sum_{i \in b} \|\hat{\theta}(\hat{\mathbf{p}}_i | \mathbf{x}_i) - \hat{\mathbf{p}}_i - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i) + \hat{\mathbf{p}}_i\|_1 \leq \\ \frac{1}{|B_b|} \sum_{i \in b} \|\hat{\theta}(\hat{\mathbf{p}}_i | \mathbf{x}_i) - \hat{\mathbf{p}}_i\|_1 &+ \frac{1}{|B_b|} \sum_{i \in b} \|\hat{\mathbf{p}}_i - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)\|_1 \end{aligned}$$

According to same *local calibration* assumption of Theorem 2, the last component is bounded as follows:

$$\frac{1}{|B_b|} \sum_{i \in I_b} \|\hat{\theta}(\mathbf{y}_i | \mathbf{x}_i) - \hat{\mathbf{p}}_i\|_1 \leq \varepsilon.$$

Before proceeding let us rewrite:

$$\begin{aligned} \|\hat{\theta}(\hat{\mathbf{p}}_i | \mathbf{x}_i) - \hat{\mathbf{p}}_i\|_1 &= \left\| \sum_{j \in I_b} w_{ij} \hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i \right\|_1 = \left\| \sum_{j \in I_b} w_{ij} (\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i) + \sum_{j \in I_b} w_{ij} \hat{\mathbf{p}}_i - \hat{\mathbf{p}}_i \right\|_1 = \\ \left\| \sum_{j \in I_b} w_{ij} (\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i) + \underbrace{\left(\sum_{j \in I_b} w_{ij} - 1 \right) \hat{\mathbf{p}}_i}_{=1} \right\|_1 &= \left\| \sum_{j \in I_b} w_{ij} (\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i) \right\|_1 \end{aligned}$$

In addition note that:

$$\frac{1}{|B_b|} \sum_{i \in I_b} \left\| \sum_{j \in I_b} w_{ij} (\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i) \right\|_1 \leq \frac{1}{|B_b|} \sum_{i \in I_b} \underbrace{\sum_{j \in I_b} w_{i,j} \|\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i\|_1}_{Z_i}$$

Again we follow the **conditioning assumptions** of Appendix A.2. Please note that each coordinate-wise distance satisfies $\|\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i\|_1 \in [0, 2]$, hence $Z_i \in [0, 2]$. We can now apply the weighted version of Hoeffding's inequality to each centered quantity $Z_i - \mathbb{E}[Z_i]$, conditioning on the kernel weights and features to obtain zero-mean summands. The union bound then yields a simultaneous statement over anchors. We proceed fixing a bin b with index set I_b of size $|B_b|$. For each anchor $i \in I_b$ define the *effective sample size* associated with the weights:

$$n_i^{\text{eff}} := \frac{1}{\sum_{j \in I_b} w_{i,j}^2}$$

Applying weighted Hoeffding's to $Z_i - \mathbb{E}[Z_i]$ for zero mean Hoeffding assumption:

$$\Pr \left(Z_i - \mathbb{E}[Z_i] \geq \tau \right) \leq \exp \left(- \frac{n_i^{\text{eff}} \tau^2}{2} \right)$$

And, by setting $\delta' = \delta/n$ we have:

$$\Pr \left(Z_i - \mathbb{E}[Z_i] \leq \sqrt{\frac{2 \log(1/\delta')}{n_i^{\text{eff}}}} \right) \geq 1 - \delta'$$

The union bound over all n anchors gives that with probability at least $1 - \delta$ for every anchor:

$$Z_i - \mathbb{E}[Z_i] \leq \sqrt{\frac{2 \log(n/\delta)}{n_i^{\text{eff}}}}$$

Averaging for all anchors in the bin:

$$\Pr \left(\frac{1}{|B_b|} \sum_{i \in I_b} Z_i \leq \frac{1}{|B_b|} \sum_{i \in I_b} \mathbb{E}[Z_i] + \frac{1}{|B_b|} \sum_{i \in I_b} \sqrt{\frac{2 \log(n/\delta)}{n_i^{\text{eff}}}} \right) \geq 1 - \delta$$

We conclude the proof by providing a bound for the expectation in the context of a neural network classifier $\phi(\cdot)$. More precisely, by the Lipschitz continuity of the softmax [4],

$$\frac{1}{|B_b|} \sum_{i \in I_b} \sum_{j \in I_b} w_{i,j} \|\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_i\|_1 \leq \frac{L}{|B_b|} \sum_{i \in I_b} \sum_{j \in I_b} w_{i,j} \|\phi(\mathbf{x}_j) - \phi(\mathbf{x}_i)\|_1.$$

More precisely, $L \leq 1$ if the kernel estimates are obtained using logits as inputs. If instead the kernel is applied to $\phi(\cdot)$ mapped to logits via $z = Wh + b$ then $L \leq \max_{1 \leq j \leq n} \sum_{i=1}^m |W_{ij}|$. We now define the kernel-weighted local radius or $\phi(\mathbf{x}_i)$:

$$R_i := \sum_{j \in I_b} w_{i,j} \|\phi(\mathbf{x}_j) - \phi(\mathbf{x}_i)\|_1.$$

Thus $Z_i \leq LR_i$. Taking expectation over the sampling of points in the bin:

$$\mathbb{E}[Z_i] \leq L \mathbb{E}[R_i].$$

Combining we obtain that with probability at least $1 - \delta$,

$$\frac{1}{|B_b|} \sum_{i \in I_b} \|\hat{\theta}(\hat{\mathbf{p}}_i | \mathbf{x}_i) - \hat{\mathbf{p}}_i\|_1 \leq \frac{L}{|B_b|} \mathbb{E} \left[\sum_{i \in I_b} \sum_{j \in I_b} w_{i,j} \|\phi(\mathbf{x}_j) - \phi(\mathbf{x}_i)\|_1 \right] + \frac{1}{|B_b|} \sum_{i \in I_b} \sqrt{\frac{2 \log(\frac{n}{\delta})}{n_i^{\text{eff}}}}$$

Conclusion: for $k \geq \max_{y \in \mathcal{Y}} y \in [1/C, 1]$, averaging over bins (weight $|B_b|/n$) yields the final bound:

$$\Pr \left(\text{LCE} \leq k \left[\varepsilon + \frac{L}{n} \sum_{b=1}^{m_B} \mathbb{E} \left[\sum_{i \in I_b} \sum_{j \in I_b} w_{i,j} \|\phi(\mathbf{x}_j) - \phi(\mathbf{x}_i)\|_1 \right] + \frac{1}{n} \sum_{b=1}^{m_B} \sum_{i \in I_b} \sqrt{\frac{2 \log(\frac{n}{\delta})}{n_i^{\text{eff}}}} \right] \right) \geq 1 - \delta, \quad (19)$$

□

The bound decomposes into a *bias term* $\mathbb{E}[Z]$, which depends on the kernel radius through the weights $w_{i,j}$, and an average *variance term* that scales as $1/\sqrt{n_i^{\text{eff}}}$. Smaller kernel radii yield more concentrated weights: this reduces bias but also decreases n_i^{eff} , thereby inflating the average variance. Conversely, larger kernels spread the weights more evenly, which decreases variance at the expense of bias. This captures the bias-variance tradeoff.

B.1.3 Proof of Theorem 4

Let $\hat{P}_i \in \Delta^C$ be the softmax prediction for input $\mathbf{x}_i$, and let $\hat{Q}_i \in \Delta^C$ be a consistent estimator in the mean integrated squared error sense (MISE) (refer to Appendix A.3 for a detailed description of the underlying assumptions) of the true conditional distribution $Q_i = \Pr(\mathbf{y}_i | \mathbf{x}_i)$, meaning:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_i - Q_i\|_1 = 0.$$

Then, the average Jensen-Shannon distance computed using $\hat{Q}_i$ converges to the one computed using the true distribution:

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n d_{\text{JSD}}(\hat{P}_i \| \hat{Q}_i) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n d_{\text{JSD}}(\hat{P}_i \| Q_i), \quad (20)$$

where $d_{\text{JSD}}(P \| Q) := \sqrt{\text{JSD}(P \| Q)}$ denotes the Jensen-Shannon distance.

Proof. Since the Jensen-Shannon distance d_{JSD} is a metric, it satisfies the triangle inequality:

$$d_{\text{JSD}}(\hat{P}_i \| \hat{Q}_i) \leq d_{\text{JSD}}(\hat{P}_i \| Q_i) + d_{\text{JSD}}(Q_i \| \hat{Q}_i).$$

Averaging over i , we obtain:

$$\frac{1}{n} \sum_{i=1}^n d_{\text{JSD}}(\hat{P}_i \| \hat{Q}_i) \leq \frac{1}{n} \sum_{i=1}^n d_{\text{JSD}}(\hat{P}_i \| Q_i) + \frac{1}{n} \sum_{i=1}^n d_{\text{JSD}}(Q_i \| \hat{Q}_i).$$

We now apply an inequality that relates the Jensen-Shannon divergence to the total variation distance. For any pair of categorical distributions $Q, \hat{Q}$, it holds that:

$$\text{JSD}(Q\|\hat{Q}) \leq \frac{\log_b(2)}{2} \|Q - \hat{Q}\|_1.$$

which depends on the log basis b used to compute JSD. Taking square roots and averaging, and using Jensen's inequality for the concave square root function:

$$\frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(Q_i\|\hat{Q}_i) = \frac{1}{n} \sum_{i=1}^n \sqrt{\text{JSD}(Q_i\|\hat{Q}_i)} \leq \sqrt{\frac{1}{n} \sum_{i=1}^n \text{JSD}(Q_i\|\hat{Q}_i)} \leq \sqrt{\frac{\log_b(2)}{2n} \sum_{i=1}^n \|Q_i - \hat{Q}_i\|_1}.$$

By the consistency assumption of kernel estimator,

$$\frac{1}{n} \sum_{i=1}^n \|Q_i - \hat{Q}_i\|_1 \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

and therefore,

$$\frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(Q_i\|\hat{Q}_i) \rightarrow 0.$$

Combining we obtain:

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|\hat{Q}_i) \leq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|Q_i).$$

We now prove the reverse inequality. Again, using the triangle inequality:

$$\text{d}_{\text{JSD}}(\hat{P}_i\|Q_i) \leq \text{d}_{\text{JSD}}(\hat{P}_i\|\hat{Q}_i) + \text{d}_{\text{JSD}}(\hat{Q}_i\|Q_i),$$

and therefore:

$$\frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|Q_i) \leq \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|\hat{Q}_i) + \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{Q}_i\|Q_i).$$

As before, by symmetry:

$$\frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{Q}_i\|Q_i) \rightarrow 0.$$

Combining we obtain:

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|\hat{Q}_i) \geq \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|Q_i).$$

Which concludes our proof:

$$\boxed{\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|\hat{Q}_i) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \text{d}_{\text{JSD}}(\hat{P}_i\|Q_i)}. \quad \square \quad (21)$$

C Further Discussion

C.1 Extension of Theorem 2 to ECCE

The subsequent analysis aims to extend the applicability of Theorem 2 to the specific class of cumulative binning-based metrics, with a focus on *ECCE* (Expected Cumulative Calibration Error). Unlike standard binning metrics, which directly compare per-bin statistics (as defined in definition 2), cumulative binning metrics operate on the cumulative sums of per-bin statistics. Despite this systematic difference,

we demonstrate that cumulative binning metrics, specifically *ECCE*, admit an upper bound of an analogous form to that presented in Theorem 2. This establishes cumulative binning metrics as a special case under the unifying bound structure provided.

Let us consider a multi-class classification setting with label space $\mathcal{Y} = \{0, \dots, C-1\}$, and assume that the dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is drawn from an unknown joint distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$, with each $\mathbf{x}_i \in \mathbb{R}^m$ and $\mathbf{y}_i \in \{0, 1\}^C$ being the one-hot encoding of label y_i . We consider a probabilistic classifier $f: \mathcal{X} \rightarrow \Delta^C$, where Δ^C is the $(C-1)$ -dimensional probability simplex. Let $\hat{\mathbf{p}}_i = f(\mathbf{x}_i)$ denote the predicted class probabilities for $\mathbf{x}_i$ and let γ be the bandwidth parameter used to compute the kernel estimates $\hat{\theta}(\mathbf{y} \mid \mathbf{x})$ on a disjoint set of instances.

Let a deterministic binning function $\beta: \Delta_c \rightarrow \{1, \dots, m_B\}$ partition the probability simplex Δ^C into m_B disjoint bins $\{B_b\}_{b=1}^{m_B}$. For each bin B_b , define the index set of points that fall into it as $I_b = \{i: \hat{\mathbf{p}}_i \in B_b\}$. Finally, let $|B_b|$ denote the bin cardinality and their cumulative sums $S_b = \sum_{i \leq b} |B_i|$. The *class-wise ECCE* is:

$$\text{class-wise ECCE} = \sum_{c=1}^C \pi_c \sum_{b=1}^{m_B} \frac{|B_b|}{n} \left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in \mathcal{I}_i} \left(\mathbf{1}\{y_j = c\} - f_c(\mathbf{x}_j) \right) \right|.$$

And, with at least probability $1 - \delta \in [0, 1]$, a bound of the same form of the one of Theorem 2 applies:

$$\text{class-wise ECCE} \leq \varepsilon + \sum_{b=1}^{m_B} w_b \sqrt{\frac{\log(2Cm_B/\delta)}{2|\Psi(b, \mathcal{Y})|}}, \quad (22)$$

Proof. We rewrite class-wise *ECCE* with the use of per-instance kernel estimates $\hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i)$:

$$\begin{aligned} & \sum_{c=1}^C \pi_c \sum_{b=1}^{m_B} \frac{|B_b|}{n} \left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in \mathcal{I}_i} \left(\mathbf{1}\{y_j = c\} - f_c(\mathbf{x}_j) \right) \right| \\ &= \sum_{c=1}^C \pi_c \sum_{b=1}^{m_B} \frac{|B_b|}{n} \left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in \mathcal{I}_i} \left(\mathbf{1}\{y_j = c\} - \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) + \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) - f_c(\mathbf{x}_j) \right) \right| \\ &\leq \sum_{c=1}^C \pi_c \sum_{b=1}^{m_B} \frac{|B_b|}{n} \left[\underbrace{\left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in \mathcal{I}_i} \left(\mathbf{1}\{y_j = c\} - \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) \right) \right|}_{\text{empirical fluctuation}} + \underbrace{\left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in \mathcal{I}_i} \left(\hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) - f_c(\mathbf{x}_j) \right) \right|}_{\text{miscalibration}} \right]. \end{aligned}$$

Recall that by the local calibration assumption, for every instance i we have $\|f(\mathbf{x}_i) - \hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)\|_1 \leq \varepsilon$. Consequently each coordinate satisfies $|f_c(\mathbf{x}_i) - \hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i)| \leq \varepsilon$, for the **miscalibration** component:

$$\left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in \mathcal{I}_i} \left(\hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) - f_c(\mathbf{x}_j) \right) \right| \leq \varepsilon.$$

In the following we apply Hoeffding inequality to bound the **Empirical fluctuation** term, we clarify the underlying assumptions our bound. More precisely, to apply Hoeffding, we need independent, bounded, zero-mean summands.

Under the **conditioning assumptions** (Appendix A.2), the only source of randomness in the term $\frac{1}{|B_i|} \sum_{j \in \mathcal{I}_i} (\mathbf{1}\{y_j = c\} - \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j))$ is the label variables $\{y_j\}_{j \in \mathcal{I}_i}$. For each j , the summand satisfies:

1. **Zero mean:**

$$\mathbb{E} \left[\mathbf{1}\{y_j = c\} - \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) \mid \mathbf{x}_j \right] = P(y_j = c \mid \mathbf{x}_j) - \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) \approx 0,$$

and exactly zero if the estimator is consistent in mean absolute error (MAE) (refer to Appendix A.3 for all the details).

2. **Independence:** the pairs $(\mathbf{x}_j, y_j)$ are i.i.d., and conditioning on the $\mathbf{x}_j$ leaves the labels $\{y_j\}$ independent.

3. **Boundedness:** both $\mathbf{1}\{y_j = c\}$ and $\hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j)$ lie in $[0, 1]$, hence their difference lies in $[-1, 1]$.

Therefore, the summands are independent, bounded in $[-1, 1]$, and zero-mean. Hoeffding's inequality applies to their average, yielding the desired concentration bound. Choosing $\tau_b = \sqrt{\frac{\log(2Cm_B/\delta)}{2|B_i|}}$ and applying the union bound over the $m_B \cdot C$ bin-class pairs yields: with probability at least $1 - \delta$,

$$\forall b, c: \quad \frac{1}{|B_i|} \sum_{j \in I_i} (\mathbf{1}\{y_j = c\} - \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j)) \leq \sqrt{\frac{\log(Cm_B/\delta)}{2|B_i|}}.$$

Then with high probability:

$$\begin{aligned} & \sum_{c=1}^C \pi_c \sum_{b=1}^{m_B} \frac{|B_b|}{n} \left[\underbrace{\left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in I_i} \left(\mathbf{1}\{y_j = c\} - \hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) \right) \right|}_{\text{empirical fluctuation}} + \underbrace{\left| \sum_{i=1}^b \frac{|B_i|}{S_b} \frac{1}{|B_i|} \sum_{j \in I_i} \left(\hat{\theta}_c(\mathbf{y}_j \mid \mathbf{x}_j) - f_c(\mathbf{x}_j) \right) \right|}_{\text{miscalibration}} \right] \\ & \leq \sum_{c=1}^C \pi_c \sum_{b=1}^{m_B} \frac{|B_b|}{n} \left[\sum_{i=1}^b \frac{|B_i|}{S_b} \sqrt{\frac{\log(Cm_B/\delta)}{2|B_i|}} + \sum_{i=1}^b \frac{|B_i|}{S_b} \varepsilon \right] \end{aligned}$$

Since $\sum_c \pi_c = \sum_{b=1}^{m_B} |B_b|/n = \sum_{i=1}^b |B_i|/S_b = 1$, the bound simplifies as follows:

$$\Pr \left(\text{class-wise ECCE} \leq \varepsilon + \sum_{b=1}^{m_B} \frac{|B_b|}{n} \sqrt{\frac{\log(Cm_B/\delta)}{2|\Psi(b, \mathcal{Y})|}} \right) \geq 1 - \delta$$

with $\Psi(b, \mathcal{Y}) = B_{i^*}$ and $i^* = \arg \min_{i \leq b} |B_i|$

Confidence Based Binning In the context of class-wise metrics that use confidence based binning, the only difference is the introduction of a dependence between bins' cardinalities and the classes: $|B_b| \rightarrow |B_{b,c}|$. Then, with probability at least $1 - \delta$, the bound applies with the same exact form under the following minor changes:

$$\text{class-wise ECCE} \leq \varepsilon + \sum_{b=1}^{m_B} \frac{|B_b|}{n} \sum_{c=1}^C \pi_c \sqrt{\frac{\log(Cm_B/\delta)}{2 \min_{i \leq b} |B_{i,c}|}} \leq \varepsilon + \sum_{b=1}^{m_B} \frac{|B_b|}{n} \sqrt{\frac{\log(Cm_B/\delta)}{2 \min_{c, i \leq b} |B_{i,c}|}} \quad (23)$$

and by setting $\Psi(b, \mathcal{Y}) = B_{i^*, c^*}$ with $i^*, c^* = \arg \min_{c \in \mathcal{Y}, i \leq b} |B_{i,c}|$ we obtain a bound of the same form of theorem 2. $\square$

C.2 Local Calibration and Proximity Bias

Proximity bias is a well-documented phenomenon in probabilistic classifiers [28], where models tend to exhibit systematic miscalibration on sparsely represented instances. This behavior is particularly concerning, as it can introduce unintended biases against underrepresented subpopulations. Addressing *proximity bias* is therefore critical to ensuring fairness in algorithmic decision-making, especially in high-stakes domains such as law and medicine, where equitable and reliable predictions are essential. The most effective way to characterize this phenomenon is by directly comparing the class frequency distributions of two subgroups that share similar model confidence scores but differ in input-space density.

We leverage this approach to examine how *local calibration* may mitigate *proximity bias*. Specifically, we provide a theoretical decomposition of the change in class frequencies when transitioning from high-density to low-density regions and use this framework to derive a probabilistic upper bound on *proximity bias* under the assumption of *local calibration*. More precisely, the total error can be decomposed into three components: a stochastic fluctuation, a calibration error and a distribution shift term respectively. The latter captures the extent to which the score distributions vary across different regions of the input space—particularly when transitioning from densely to sparsely represented instances.

Theorem 5 (Error Decomposition of Proximity Bias). *Let S_1 and S_2 be two proximity-based sub-bins drawn from the same score-based bin, with cardinalities $|S_1|$ and $|S_2|$. Define:*

$$\text{freq}(S_s) := \frac{1}{|S_s|} \sum_{i \in S_s} \mathbf{y}_i, \quad \text{conf}(S_s) := \frac{1}{|S_s|} \sum_{i \in S_s} f(\mathbf{x}_i)$$

If a classifier f satisfies local calibration, then with probability at least $1 - \delta \in [0, 1]$ the difference in class frequencies between the two sub-bins is bounded as follows:

$$\Pr \left(\|\text{freq}(S_1) - \text{freq}(S_2)\|_1 \leq 2\varepsilon + \sqrt{\frac{2 \log(4C/\delta)}{\min(|S_1|, |S_2|)}} + \|\text{conf}(S_1) - \text{conf}(S_2)\|_1 \right) \geq 1 - \delta. \quad (24)$$

A detailed proof is provided:

Proof of Theorem 5 Suppose the simplex Δ^C is partitioned into m_B score-based disjoint bins $\{B_b\}_{b=1}^{m_B}$. Each score-based bin is further subdivided by grouping points with similar feature-space proximity. For each point $\mathbf{x}_i$, define its proximity score:

$$\pi_k(\mathbf{x}_i) := \frac{1}{k} \sum_{j=1}^k \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_{(i,j)})\|_2,$$

where $\mathbf{x}_{(i,1)}, \dots, \mathbf{x}_{(i,k)}$ are the k nearest neighbors of $\mathbf{x}_i$ in feature space (excluding $\mathbf{x}_i$ itself). We aim to bound the quantity $\|\text{freq}(S_1) - \text{freq}(S_2)\|_1$, which measures the difference in empirical label distributions between the two sub-bins, under the assumption that the model satisfies *local calibration*:

$$\|f(\mathbf{x}_i) - \hat{\theta}(\mathbf{y}_i | \mathbf{x}_i)\|_1 \leq \varepsilon, \quad \forall i \in \{1, \dots, n\}.$$

Proof. We begin by applying the triangle inequality:

$$\begin{aligned} \|\text{freq}(S_1) - \text{freq}(S_2)\|_1 &= \left\| \left(\text{freq}(S_1) - \text{conf}(S_1) \right) + \left(\text{conf}(S_1) - \text{conf}(S_2) \right) + \left(\text{conf}(S_2) - \text{freq}(S_2) \right) \right\|_1 \\ &\leq \left\| \left(\text{freq}(S_1) - \text{conf}(S_1) \right) + \left(\text{conf}(S_1) - \text{conf}(S_2) \right) \right\|_1 + \|\text{conf}(S_2) - \text{freq}(S_2)\|_1 \\ &\leq \|\text{freq}(S_1) - \text{conf}(S_1)\|_1 + \|\text{conf}(S_1) - \text{conf}(S_2)\|_1 + \|\text{conf}(S_2) - \text{freq}(S_2)\|_1. \end{aligned} \quad (25)$$

From Theorem 2, which applies identically to any bin or subset under the same **conditioning assumptions** of Appendix A.2, we have that for any $\delta' \in [0, 1]$ the following probabilistic bound for each sub-bin S_s :

$$\Pr \left(\|\text{freq}(S_s) - \text{conf}(S_s)\|_1 \leq \varepsilon + \sqrt{\frac{\log(2C/\delta')}{2|S_s|}} \right) \geq 1 - \delta'.$$

Now define the following events:

$$\begin{aligned} A &:= \{ \|\text{freq}(S_1) - \text{conf}(S_1)\|_1 > \varepsilon + \eta_{\delta_1} \}, \quad \text{where } \eta_{\delta_1} := \sqrt{\frac{\log(2C/\delta')}{2|S_1|}}, \\ B &:= \{ \|\text{freq}(S_2) - \text{conf}(S_2)\|_1 > \varepsilon + \eta_{\delta_2} \}, \quad \text{where } \eta_{\delta_2} := \sqrt{\frac{\log(2C/\delta')}{2|S_2|}}. \end{aligned}$$

Applying the union bound:

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B) \leq 2\delta'.$$

Thus, with probability at least $1 - 2\delta'$, both events do not occur:

$$\Pr(\bar{A} \cap \bar{B}) \geq 1 - 2\delta'.$$

Under this event, we can bound eq. (25). More precisely, both the first and third terms is bounded by $\varepsilon + \sqrt{\frac{\log(2C/\delta')}{2n_{S_s}}}$, and by choosing $\delta' = \frac{\delta}{2}$ we can conclude:

$$\Pr\left(\|\text{freq}(S_1) - \text{freq}(S_2)\|_1 \leq 2\varepsilon + \sqrt{\frac{2\log(4C/\delta)}{\min(|S_1|, |S_2|)}} + \|\text{conf}(S_1) - \text{conf}(S_2)\|_1\right) \geq 1 - \delta. \quad (26)$$

□

Intuitively, under local calibration, predicted scores approximate true class frequencies. Therefore, any shift in the score distribution within a bin implies a corresponding shift in the underlying class frequencies. The error due to this distributional inconsistency can be reduced by refining the density-based bins, but finer binning leads to smaller sample sizes per bin, thereby increasing the stochastic fluctuation error. This trade-off highlights an inherent tension in binning procedures: reducing distribution shift comes at the cost of increased variance.

Although limited availability of data is problematic in capturing *proximity bias* of a *locally calibrated* model, we can investigate the phenomenon from a theoretical perspective in the limit of infinite data availability. This allows us to schedule the bin width reduction, bounding the admissible score change, while keeping a sufficient bin cardinality to workaround the inherent trade-off between the two. This analysis leads to the conclusion that, in the presence of *local calibration*, the value of *proximity bias*—if it could be computed with access to infinite data—would be tightly bounded by the model’s calibration error, which is explicitly controlled by the *local calibration* property.

Corollary 2 (Infinite Limit of Proximity bias under Local Calibration). *Let the assumptions of Theorem 5 hold, and assume moreover that the conditional density $h(\mathbf{x} \mid \hat{\mathbf{p}})$ and marginal density $q_f(\hat{\mathbf{p}})$ be continuous in a neighborhood of $(\mathbf{x}_i, \hat{\mathbf{p}}_i)$, with $h(\mathbf{x}_i \mid \hat{\mathbf{p}}_i) > 0$ and $q_f(\hat{\mathbf{p}}_i) > 0$ for some $i \in S_s$.*

Then, the bound on the proximity bias asymptotically simplifies to:

$$\lim_{n \rightarrow \infty} \Pr\left(\|\text{freq}(S_1) - \text{freq}(S_2)\|_1 \leq 2\varepsilon\right) = 1. \quad (27)$$

Thus, the empirical difference in class frequencies between proximity sub-bins becomes entirely determined by ε . This theoretical result allows to infer that *proximity bias* is directly controlled by the *local calibration* property of a model and in the following we provide proof:

Proof of Corollary 2. Let the setting and notation be as in Theorem 5. Additionally, fix a target confidence vector $\hat{\mathbf{p}}_0 \in (0, 1)^C$ and let the confidence (scores) bin centered at $\hat{\mathbf{p}}_0$ be:

$$B_n(\hat{\mathbf{p}}_0) = \{\hat{\mathbf{p}} : |\hat{\mathbf{p}} - \hat{\mathbf{p}}_0| \leq w_n\},$$

with radius $w_n \rightarrow 0$. Define the set of indices of samples in this bin as:

$$\mathcal{I}_n(\hat{\mathbf{p}}_0) = \{i : \hat{\mathbf{p}}_i \in B_n(\hat{\mathbf{p}}_0)\}.$$

Within this set, consider two disjoint density-based sub-bins $S_1, S_2 \subseteq \mathcal{I}_n(\hat{\mathbf{p}}_0)$ corresponding to local neighborhoods in feature space. Additionally, let the conditional density $h(\mathbf{x} \mid \hat{\mathbf{p}})$ and marginal density $q_f(\hat{\mathbf{p}})$ be continuous in a neighborhood of $(\mathbf{x}_i, \hat{\mathbf{p}}_i)$, with $h(\mathbf{x}_i \mid \hat{\mathbf{p}}_i) > 0$ and $q_f(\hat{\mathbf{p}}_i) > 0$ for some $i \in S_s$.

By the triangle inequality (as in (25)),

$$\|\text{freq}(S_1) - \text{freq}(S_2)\|_1 \leq \|\text{freq}(S_1) - \text{conf}(S_1)\|_1 + \|\text{conf}(S_1) - \text{conf}(S_2)\|_1 + \|\text{conf}(S_2) - \text{freq}(S_2)\|_1.$$

From Theorem 5 (coordinate-wise Hoeffding + union bound) we have, for any fixed $\delta = 2 \cdot \delta'$,

$$\Pr\left(\|\text{freq}(S_s) - \text{conf}(S_s)\|_1 \leq \varepsilon + \sqrt{\frac{\log(2C/\delta')}{2\min(|S_1|, |S_2|)}}\right) \geq 1 - \delta'$$

Applying the union bound gives that with probability at least $1 - \delta$ both deviations are bounded simultaneously:

$$\|\text{freq}(S_1) - \text{freq}(S_2)\|_1 \leq 2\varepsilon + \sqrt{\frac{2\log(4C/\delta)}{\min(|S_1|, |S_2|)}} + \|\text{conf}(S_1) - \text{conf}(S_2)\|_1$$

Then, under the local calibration assumption $\|f(\mathbf{x}_j) - \hat{\theta}(\mathbf{y}_j | \mathbf{x}_j)\|_1 \leq \varepsilon$ for all $j \in D$, we have:

$$\lim_{n \rightarrow \infty} \Pr\left(\|\text{freq}(S_1) - \text{freq}(S_2)\|_1 \leq 2\varepsilon\right) = 1. \quad (28)$$

Proof. Step 1. Shrinking confidence bins controls differences in predicted scores. By construction, for every sample $j \in \mathcal{I}_n(\hat{\mathbf{p}}_0)$,

$$\hat{\mathbf{p}}_j \in [\hat{\mathbf{p}}_0 - w_n, \hat{\mathbf{p}}_0 + w_n].$$

and thus, for any two sub-bins $S_1, S_2 \subseteq \mathcal{I}_n(\hat{\mathbf{p}}_0)$,

$$\|\text{conf}(S_1) - \text{conf}(S_2)\|_1 \leq 2Cw_n.$$

Therefore, as $w_n \rightarrow 0$, the difference in average predicted confidences between any two density-based sub-bins within the same score bin also vanishes:

$$\|\text{conf}(S_1) - \text{conf}(S_2)\|_1 \xrightarrow{n \rightarrow \infty} 0. \quad (29)$$

Step 2. Maintaining infinite data within shrinking bins. We now show that it is possible to shrink both the confidence-bin width $2w_n$ and the density sub-bin radius r_n simultaneously, while guaranteeing that each sub-bin still contains infinitely many samples with high probability. As a consequence the stochastic square root term asymptotically vanishes in probability.

Fix a ball $B(\mathbf{x}_i, r_n) \subseteq \mathcal{R}$ centered in $\mathbf{x}_i$ and with volume $\text{vol}(B(0, r_n))$ where $\mathcal{R}$ is the region of space associated to a sub-bin. Likewise, fix a ball $B_n(\hat{\mathbf{p}}_i, w_i) \subseteq B_n(\hat{\mathbf{p}}_0)$. The joint probability that a sample lies in both balls is:

$$\Pr(\mathbf{X} \in B(\mathbf{x}_i, r_n), f(\mathbf{x}) \in B_n(\hat{\mathbf{p}}_i, w_i)) = \int_{B_n(\hat{\mathbf{p}}_i, w_i)} \int_{B(\mathbf{x}_i, r_n)} h(\mathbf{x} | \hat{\mathbf{p}}) q_f(\hat{\mathbf{p}}) d\mathbf{x} d\hat{\mathbf{p}}.$$

Then, by the Lebesgue differentiation theorem (see e.g. Theorem 1.3 in [22]), for sufficiently small r_n and w_i ,

$$\Pr(\mathbf{X} \in B(\mathbf{x}_i, r_n), f(\mathbf{X}) \in B_n(\hat{\mathbf{p}}_i, w_i)) = h(\mathbf{x}_i | \hat{\mathbf{p}}_i) q_f(\hat{\mathbf{p}}_i) \text{vol}(B(0, r_n)) \text{vol}(B(0, w_i)) + o(r_n^d w_i^{C-1}),$$

where $o(r_n^d w_i^{C-1})$ denotes a term that becomes negligible compared to the product of the volumes $r_n^d w_i^{C-1}$. Note that this result leverages the norm equivalence for finite-dimensional spaces like $\Delta^C \subset \mathbb{R}^{C-1}$. As a consequence, using L_1 or L_2 balls only changes bounds by constant factors which do not affect asymptotic rates. Moreover, by the (ε, δ) -definition of continuity (Weierstrass-Jordan), there exist finite positive constants $c_h, C_h, q_{\min}, q_{\max}$ and a neighborhood U of $(\mathbf{x}_i, \hat{\mathbf{p}}_i)$ such that:

$$0 < c_h \leq h(\mathbf{x} | \hat{\mathbf{p}}) \leq C_h < \infty, \quad 0 < q_{\min} \leq q_f(\hat{\mathbf{p}}) \leq q_{\max} < \infty, \quad \forall (\mathbf{x}, \hat{\mathbf{p}}) \in U.$$

For all sufficiently small r_n, w_i , so that $B(\mathbf{x}_i, r_n) \times B_n(\hat{\mathbf{p}}_i, w_i) \subset U$, the joint probability admits a two-sided bound:

$$c_h q_{\min} \text{vol}(B(0, r_n)) \text{vol}(B(0, w_i)) \leq \Pr(\mathbf{X} \in B(\mathbf{x}_i, r_n), f(\mathbf{X}) \in B_n(\hat{\mathbf{p}}_i, w_i)) \leq C_h q_{\max} \text{vol}(B(0, r_n)) \text{vol}(B(0, w_i)).$$

Hence the expected number of points in a sub-bin satisfies:

$$\mathbb{E}[|S_s|] \asymp n r_n^d w_i^{C-1},$$

where $\asymp$ indicates asymptotic proportionality, meaning that $\mathbb{E}[|S_s|]$ grows at the same rate as $n r_n^d w_i^{C-1}$ up to constant factors. Choosing sequences $w_i = n^{-\alpha}$ and $r_n = n^{-\beta}$ with:

$$0 < \alpha < \frac{1}{C-1}, \quad 0 < \beta < \frac{1 - \alpha(C-1)}{d},$$

we obtain:

$$n r_n^d w_i^{C-1} = n^{1 - \alpha(C-1) - \beta d} \rightarrow \infty,$$

while both $w_i, r_n \rightarrow 0$. Since each sample $\{(\mathbf{x}_j, f(\mathbf{x}_j))\}_{j=1}^n$ is drawn i.i.d., the number of samples in a local sub-bin S_s follows a binomial distribution $|S_s| \sim \text{Binomial}(n, p_n)$ with success probability $p_n = \Pr((\mathbf{x}, f(\mathbf{x})) \in B(\mathbf{x}_i, r_n) \times B_n(\hat{\mathbf{p}}_i, w_i))$. As a consequence, for any $\eta \in (0, 1)$, the Chernoff bound gives:

$$\Pr(|S_s| \geq (1 - \eta) \mathbb{E}[|S_s|]) \geq 1 - \exp\left(-\frac{\eta^2}{2} \mathbb{E}[|S_s|]\right).$$

Since $\mathbb{E}[|S_s|] \rightarrow \infty$ as $n \rightarrow \infty$, with probability tending to one,

$$|S_s| \geq (1 - \eta) \mathbb{E}[|S_s|] \rightarrow \infty, \quad \text{i.e.,} \quad \Pr(|S_s| \rightarrow \infty) \rightarrow 1. \quad (30)$$

Which allows us to conclude:

$$\lim_{n \rightarrow \infty} \Pr\left(\|\text{freq}(S_1) - \text{freq}(S_2)\|_1 \leq 2\varepsilon\right) = 1. \quad (31)$$

□

Remark. The argument readily extends to the case where each proximity-based sub-bin S_s is a finite union of disjoint regions $\{R_{s,k}\}_{k=1}^{K_s}$. Under the same regularity and consistency assumptions applied component-wise (shrinking diameters and diverging per-component sample sizes), the concentration and continuity arguments hold uniformly over components, and the aggregate deviation remains bounded by 2ε with maximum probability. We restrict the proof to a single region per sub-bin for notational simplicity.

C.3 Illustrative Example

In this section, we present a toy example to highlight potential pitfalls of density-based calibration. Specifically, we show that the choice of bin width plays a critical role: overly wide bins may lead to ineffective recalibration, while overly fine bins require large sample sizes and can become computationally prohibitive. Our goal here is to raise awareness on risks that can arise in practice.

Consider a binary probabilistic classifier $f(\cdot)$ and a dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where each input $\mathbf{x}_i \in \mathbb{R}^m$ takes one of six distinct sets of values (here $m = 2$ for visualization purposes). Figure 4 provides a visual representation of these points in the decision space (left), as well as their corresponding locations in the density-confidence space used for calibration (right).

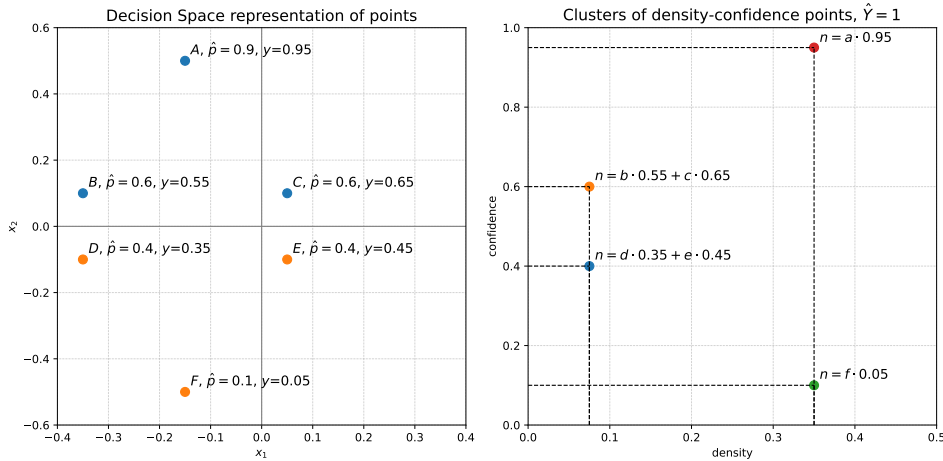


Figure 4: Points in the decision space (right) and their mapping to density-confidence space for calibration (left).

Grouping points by their coordinates yields six disjoint regions. Within each region, the classifier assigns a constant predicted probability (all inputs have same values within region). Table 2 reports the size, density, predicted probability, and empirical label frequency for each region.

Table 2: The six disjoint regions characterized by density, predicted probability, and empirical frequency.

Set	Size	Density	$\mathbf{p}$	$\mathbf{Y}$
A	a	0.35	0.9	0.95
B	b	0.075	0.6	0.55
C	c	0.075	0.6	0.65
D	d	0.075	0.4	0.35
E	e	0.075	0.4	0.45
F	f	0.35	0.1	0.05

In Figure 4 (right), each point is plotted according to its predicted probability and estimated local density. A most fine-grained approach in calibration is to aggregate predictions based on proximity in this joint space. That is, the calibrated probability for a point is conditioned on both its score $\hat{P}$ and density estimate $\hat{D}$, and is computed as:

$$p_{\text{cal}} = \Pr(\hat{Y} = Y \mid \hat{P}, \hat{D}) = \frac{\Pr(\hat{P}, \hat{D} \mid \hat{Y} = Y) \cdot \Pr(\hat{Y} = Y)}{\Pr(\hat{P}, \hat{D} \mid \hat{Y} = Y) \cdot \Pr(\hat{Y} = Y) + \Pr(\hat{P}, \hat{D} \mid \hat{Y} \neq Y) \cdot \Pr(\hat{Y} \neq Y)}$$

We focus on calibrating predictions for points with $\hat{P} = 0.6$ and $\hat{D} = 0.075$. Among the six regions, regions b and c match this pair of values. Then:

$$\begin{aligned} \Pr(\hat{P} = 0.6, \hat{D} = 0.075 \mid \hat{Y} = 1) &= \frac{b \cdot 0.55 + c \cdot 0.65}{a \cdot 0.95 + f \cdot 0.05 + d \cdot 0.35 + e \cdot 0.45 + b \cdot 0.55 + c \cdot 0.65} = \frac{\text{NUM}_1}{\text{DEN}_1} \\ \Pr(\hat{P} = 0.6, \hat{D} = 0.075 \mid \hat{Y} = 0) &= \frac{b \cdot 0.45 + c \cdot 0.35}{a \cdot 0.05 + f \cdot 0.95 + d \cdot 0.65 + e \cdot 0.55 + b \cdot 0.45 + c \cdot 0.35} = \frac{\text{NUM}_2}{\text{DEN}_2} \end{aligned}$$

Since

$$\frac{\Pr(\hat{Y} \neq Y)}{\Pr(\hat{Y} = Y)} = \frac{\text{DEN}_2}{\text{DEN}_1},$$

we simplify the calibrated probability as:

$$p_{\text{cal}} = \frac{\text{NUM}_1}{\text{NUM}_1 + \text{NUM}_2} = \frac{b \cdot 0.55 + c \cdot 0.65}{b + c}$$

Since regions b and c have similar sizes (i.e., $b \approx c$), then:

$$p_{\text{cal}} \approx \frac{0.55 + 0.65}{2} = 0.6 \neq [0.55, 0.65]$$

This recalibrated probability equals a weighted average of the empirical frequencies of regions b and c . As such, it cannot simultaneously correct both, and whichever side is under/overconfident remains miscalibrated after recalibration. The magnitude of miscalibration aggravates when one region is underconfident and the other is overconfident. Such heterogeneous calibration errors do not exclusively occur when regions with similar predicted confidences differ in density but can also arise in presence of differences in class overlap or representation smoothness, as commonly observed in deep neural networks [5, 13]. Consequently, grouping them within the same density–confidence bin averages incompatible local behaviors, masking or amplifying miscalibration. While narrower bins could mitigate this effect, they quickly become sample-inefficient and computationally demanding.

D Experimental Details

In this section, we provide all the details that should allow exact reproducibility of our results.

D.1 Training of Classifiers

We report here all training configurations for the classifiers used in the calibration experiments. All classifiers used categorical cross-entropy, and no batch normalization layers or weight decay were applied during fine-tuning.

CIFAR-10. We use a **ResNet-50** architecture initialized with **IMAGENET1K_V2** pre-trained weights. A dropout layer with rate 0.2 is appended to the final backbone layer, followed by a linear classification head. During fine-tuning, all layers are frozen except for the last backbone block and the classification head. Optimization is performed for 9 epochs using the **Adam** optimizer [10] with a learning rate of 3×10^{-4} .

CIFAR-100. We adopt a **ResNet-152** model pre-trained on **IMAGENET1K_V2**. A dropout layer with a rate 0.5 is inserted before the classification layer. All layers except the last backbone block and the classifier are frozen during training. We optimize for 9 epochs using **Adam** with a learning rate of 3×10^{-4} .

TissueMNIST. We employ a **ResNet-50** backbone initialized with **IMAGENET1K_V2** weights. A dropout layer with a rate 0.2 is applied before the linear classification layer. As in the previous setups, all layers except the last backbone block and the classification layer are trainable. We train for 10 epochs using the **Adam** optimizer with a learning rate of 3×10^{-4} .

D.2 Training of LoCal Nets

We report here all training configurations for LCN in the calibration experiments.

Residual Modelling. The LCN operates in a residual fashion. Given the intermediate representations $\phi(\mathbf{x})$ extracted from a pre-trained backbone, let $\phi_{\text{PCA}}(\mathbf{x})$ denote the reduced feature representation obtained via Principal Component Analysis (PCA). The LCN processes $\phi(\mathbf{x})$ through its hidden layer to produce refined features $\tilde{\phi}_{\text{PCA}}(\mathbf{x})$ and logits $\tilde{g}(\mathbf{x})$. The final representations $\phi'_{\text{PCA}}(\mathbf{x})$ and $g'(\mathbf{x})$ are obtained through a weighted residual combination:

$$\phi'_{\text{PCA}}(\mathbf{x}) = \tilde{\phi}_{\text{PCA}}(\mathbf{x}) + w_{\phi} \cdot \phi_{\text{PCA}}(\mathbf{x}) + b_{\phi}, \quad g'(\mathbf{x}) = \tilde{g}(\mathbf{x}) + w_g \cdot g(\mathbf{x}) + b_g, \quad (32)$$

where w_{ϕ}, b_{ϕ}, w_g and b_g are learnable scalar weights and biases that adaptively control the contribution of the original features and logits, respectively. Weights are initialized as 1 and biases are randomly sampled from normal distributions with 0. location and 0.01 scale parameters. The residual formulation provides strong initialization for LCN outputs, enables preserving the semantic content of the backbone features while introducing locally calibrated corrections, and improves both stability and convergence to meaningful solutions.

CIFAR-10. The LCN is implemented as a fully connected network with a single hidden layer of size 64 and dropout rate 0.3. It has two output heads: one of dimension 10 (corresponding to the number of classes) and one of dimension 50, used for the PCA-reduced feature representations. The loss weighting hyperparameter λ is set to 1, ensuring equal contribution of both components of the objective. The fixed kernel bandwidth γ is set to 10, chosen to be as small as possible to preserve locality while maintaining stable convergence of the cross-entropy component of the loss, as excessively small values lead to training collapse. This choice is validated empirically using a held-out validation set. Optimization is performed using the **Adam** optimizer with learning rate 1×10^{-3} , for 22 epochs and a batch size of 1024.

CIFAR-100. The LCN uses a fully connected architecture with one hidden layer of size 128 and dropout rate 0.5. As before, it has two output heads: one of dimension 100 (matching the number of classes) and one of dimension 50 for the PCA-reduced features. We set $\lambda = 1$ for equal loss weighting and fix $\gamma = 10$ for consistency with the other datasets. In this case, slightly smaller bandwidth values were found feasible, but $\gamma = 10$ was retained for coherence across experiments. Optimization uses **Adam** with learning rate 1×10^{-3} , trained for 30 epochs with a batch size of 1024.

TissueMNIST. The LCN is a single-hidden-layer fully connected network with hidden dimension 256 and dropout rate 0.3. It includes two output heads: one of dimension 8 (the number of classes) and one of dimension 50 for the PCA-reduced representations. The hyperparameter λ is set to 1, and the bandwidth γ is fixed at 10, following the same locality–stability trade-off principle described above, validated via a held-out set. Optimization uses **Adam** with a learning rate of 1×10^{-3} for 60 epochs with a batch size of 1024.

D.3 Metrics

In the following, we provide implementation details for the calibration metrics and the associated hyperparameter configurations used in our experiments to allow full reproducibility of our results.

Class-wise Binning Metrics. For both the *Expected Calibration Error* (ECE) and the *Expected Cumulative Calibration Error* (ECCE), we partitioned $f_c(\mathbf{x})$ into 15 bins based on predicted confidence scores. Empty bins, when present, were excluded from the computation. Class-wise calibration errors were first computed independently for each class and subsequently aggregated using class-frequency weights estimated from the training data. While **CIFAR-10** and **CIFAR-100** are both balanced datasets, **TissueMNIST** exhibits class imbalance, with priors ranging approximately from 0.32 to 0.04.

Class-wise Kernel Metrics. We employ two kernel-based calibration metrics: the *multiclass Local Calibration Error* (LCE) and its maximum variant (*MLCE*). To extend *LCE* to the multiclass setting, we adopt a class-wise formulation analogous to that used for binning-based metrics. Specifically, $f_c(\mathbf{x})$ is partitioned into 15 confidence-based bins, and for each fixed class, we use the corresponding bins to identify the neighborhood of each anchor point for kernel estimation. Bins with fewer than 20 elements were discarded to prevent using unstable kernel estimates. For each $i \in D$, the *LCE* is computed as the absolute difference between kernel-weighted estimates of predicted confidences and empirical labels of instances in the same confidence bin. Per-sample deviations are then averaged, and the values for each class are combined using priors to obtain the final *LCE*. The kernel bandwidth parameter was set to $\gamma = 10$, consistent with the bandwidth used during the training of the LCN.

D.4 Hardware and Training Time

For our experiments, we use a 16-core machine with an AMD Ryzen 9 7950X CPU and 2 NVIDIA GeForce RTX 4090 GDDR6X with 24GB of memory, OS Ubuntu 22.04.4 LTS.