Multiclass Local Calibration with the Jensen-Shannon Distance

Cesare Barbera[◇,*,†] **Lorenzo Perini**[⋆] **Giovanni De Toni**[‡]

Andrea Passerini[*] **Andrea Pugnana**[◇,*]

[*]University of Trento [†]University of Pisa [⋆]Meta [‡]Fondazione Bruno Kessler

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. 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 to enhance *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 (Wang, 2023; Sambyal et al., 2023; Cianci et al., 2023; Ruggieri and Pugnana, 2025) - i.e., their predicted probabilities should reflect the true empirical frequencies of the corresponding classes. For instance, in the healthcare context, Van Calster et al. (2019) compares two cardiovascular risk prediction models over two million patients from the UK. They show that the better-calibrated model, despite having a lower AUC, avoided overestimation of risk.

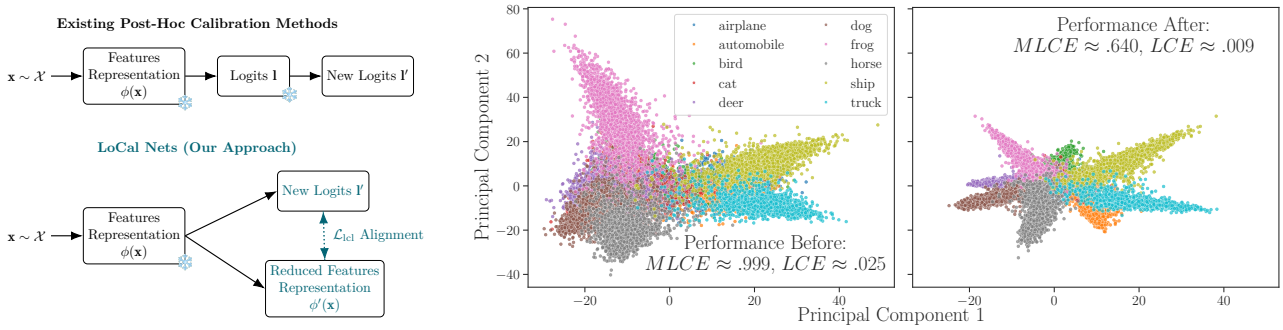
However, existing works have mostly focused on *confidence calibration* (Le-Coz et al., 2024; Luo et al., 2022; Xiong et al., 2023; de Menezes e Silva Filho et al., 2023), which only looks at the top-predicted class to check a model's calibration. While in binary classification tasks, a well-calibrated model on the top-predicted class ensures good calibration, 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 (Guo et al., 2017), 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* (Vaicenavicius et al., 2019), which requires alignment across the full probability vector. However, existing multiclass calibration notions do not consider distances among instances, making them prone to *proximity bias* (Xiong et al., 2023), whereby predictions for instances in sparsely populated regions of the decision space are more likely to be poorly calibrated. This reveals a structural limitation of existing multiclass calibration notions and motivates the need for definitions

[◇]Corresponding authors. `{name}.{surname}@unitn.it`.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).



(a) LN flow

(b) LN effect on `Cifar10` using a `resnet-50`

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

that explicitly incorporate geometric structure.

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 (LNs), 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* (Guo et al., 2017) and *top-label calibration* (Gupta and Ramdas, 2022). 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* (Kull et al., 2019) evaluates calibration for each class marginally, by comparing predicted probabilities with empirical frequencies for that target alone. *Top-r calibration* (Gupta et al., 2021) 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* (Zhao et al., 2021) 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 (Zadrozny and Elkan, 2001), isotonic regression (Zadrozny and Elkan, 2002), logistic scaling (Platt et al., 1999), temperature scaling (Guo et al., 2017) and parametric forms of scaling (Kull et al., 2017, 2019; Berta et al., 2026). We refer to de Menezes e Silva Filho et al. (2023) 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. Hence, researchers focused on approaches that incorporate a

notion of distance between inputs. Luo et al. (2022), Lin et al. (2023), and Valk and Kull (2023) propose recalibration approaches based on kernel regression. Luo et al. (2022) focus on confidence calibration, leveraging kernel smoothing to locally adjust predicted confidences. In contrast, Lin et al. (2023) and Valk and Kull (2023) extend kernel-based recalibration to the multiclass setting. A key distinction concerns how feature representations are treated. Both Luo et al. (2022) and Valk and Kull (2023) operate on fixed representations. In contrast, Lin et al. (2023) learn new feature representations tailored to kernel-based estimates. Our approach is closest in spirit to the latter, yet it differs fundamentally in its objective. Rather than using kernel regression as the final predictor, we use kernel-based local frequency estimates only as a supervision signal to induce locality in the model's probabilistic output. Moreover, these approaches do not explicitly formalize or evaluate multiclass calibration as a local property. In contrast, our work introduces a formal notion of multiclass local calibration and directly studies its theoretical and empirical implications. Also Marx et al. (2023) 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. (2023) 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 only addresses the confidence score of the predicted class. We provide in Appendix C.2 a detailed analysis of the role of *proximity bias* in our proposal and an illustrative example of the pitfalls of density-based recalibration.

Multicalibration. Another notion, known as *Multicalibration*, requires the model to be calibrated not just on the entire population, but on virtually any sub-population identifiable by a specific hypothesis class (Hébert-Johnson et al., 2018; Jung et al., 2021; Baldeschi et al., 2025; Jin et al., 2025; Tax et al., 2026). Thus, the goal of multicalibration approaches is to ensure that models do *not* make biased predictions on any of these sub-populations (Globus-Harris et al., 2023; Noarov and Roth, 2023).

While related, multicalibration differs from our setting in two key aspects. First, it is mostly studied for binary/regression tasks, whereas we address *multi-class* calibration and the complexities of the probability simplex. Second, it identifies semantically meaningful subgroups (*e.g.*, `User Age > 18`) using expressive features in tabular data. In contrast, our approach operates in the high-dimensional representation space of neural networks, where dimensions are abstract and "auditing" for subgroups via standard multicalibration techniques is computationally intractable.

3 BACKGROUND

In this section, we formally introduce the main definitions and metrics used to evaluate calibration.

Definitions. 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 $f\colon \mathcal{X} \to \Delta^C$, 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* (Guo et al., 2017), 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* (Vaicenavicius et al., 2019), which requires the target class conditional distribution on any prediction of the classifier to match that prediction, i.e.,:

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

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)* (Naeini et al., 2015), 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 multi-class, by using a multidimensional grid binning. Unfortunately,

the combinatorial nature of such a binning hinders its practical application. *Expected Cumulative Calibration Error (ECCE)* (Ibarra et al., 2022) 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. (2022):

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

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 neighbors of an anchor point $\mathbf{x}_i$. The shown formalization assumes balanced classes for simplicity and is presented in the most general form leveraging multidimensional binning. In practice, a trivial class-wise adaptation can be implemented. Finally, the *Maximum Local Calibration Error* is $MLCE = \max_{i\in D}\left\|\frac{\sum_{j\in b}\left(\hat{\mathbf{p}}_j-\mathbf{y}_j\right)k_\gamma(\mathbf{x}_i,\mathbf{x}_j)}{\sum_{j\in b}k_\gamma(\mathbf{x}_i,\mathbf{x}_j)}\right\|_1$. Notice that, in its original formulation by Luo et al. (2022), 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)**.** *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. We consider bandwidth sequences $\{\gamma_n\}$ satisfying standard consistency conditions. Define the associated kernel estimator:*

$$\hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i) = \frac{\sum_{j\in D}W_j(\mathbf{x}_i)\mathbf{y}_j}{\sum_{j\in D}W_j(\mathbf{x}_i)}, \qquad W_j(\mathbf{x}_i) \propto k_\gamma(\mathbf{x}_i,\mathbf{x}_j), \tag{2}$$

*where the weights are normalized to sum to 1. 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.,*

$$\left\|\hat{\mathbf{p}}_i - \hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)\right\|_1 \leq \varepsilon, \quad \forall i \in \{1,\dots,n\} \tag{3}$$

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

Please refer to Appendix A.2 for a comprehensive description of the *consistency conditions*. We stress that Definition 1 is inherently finite-sample and bandwidth-dependent as it requires alignment with the statistically optimal local frequency estimate at the chosen resolution γ, not with the true conditional distribution. The consistency assumptions on the kernel estimator relate this empirical notion to population-level calibration measures. Notably, the notion of multiclass local calibration is structurally related to strong calibration. On the one hand, strong calibration enforces the alignment of predicted probabilities with empirical frequencies conditional on the predicted probability vector. On the other hand, local calibration constrains this alignment to be pointwise in the feature space. As we show in Theorem 1, satisfying local calibration is asymptotically sufficient to guarantee strong calibration. More precisely, we establish an upper bound on the continuous Multidimensional Expected Calibration Error (MECE). We focus on this formulation since, unlike its binned counterparts, it integrates over the probability simplex and provides a distribution-level characterization of calibration.

Theorem 1 (Continuous MECE under Local Calibration)**.** *Define the continuous Multidimensional Expected Calibration Error (MECE) as:*

$$M(f) = \sum_{y\in\mathcal{Y}} p(y)\,\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[|\mathbb{E}[\mathbf{y}\mid\mathbf{x}]_y - f_y(\mathbf{x})|\right],$$

If a model f satisfies local calibration, under standard regularity conditions there exists $k \in (0,1]$ such that continuous $MECE$ is asymptotically upper bounded:

$$M(f) \leq \varepsilon \cdot k \tag{4}$$

Thus, in the limit, local calibration can be interpreted as a geometric refinement of strong calibration, ensuring that global reliability emerges from consistent local behavior. A detailed proof is in Appendix B.1.

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 $\beta : \Delta^C \to \{1, \ldots, m_B\}$ 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 $\varphi : [0,1] \times [0,1] \to \mathbb{R}_{\geq 0}$ be a scalar comparator that measures discrepancy between an empirical frequency $freq_{b,c}$ and a predicted confidence $conf_{b,c}$ that is Lipschitz in both arguments with constant L_φ. The general multiclass bin-based calibration error metric is obtained as:*

$$\mathcal{E}(D; \varphi; \beta) = \sum_{b=1}^{m_B} \sum_{c=1}^{C} w_{b,c} \cdot \pi_c \cdot \varphi(freq_{b,c}, conf_{b,c}) \quad (5)$$

where $w_{b,c}$ are possibly class dependent 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*. See Appendix A for a detailed description of the assumptions.

Theorem 2 (Error decomposition of calibration metrics under Local Calibration)**.** *Let $\mathcal{E}(D; \varphi; \beta)$ be a calibration metric satisfying Definition 2, evaluated on a sample $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$. Let f be locally calibrated with error ε relative to kernel estimates $\hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)$ computed on an independent arbitrarily large auxiliary dataset. Conditioning on the evaluation inputs $\{\mathbf{x}_i\}_{i=1}^n$ and the corresponding consistent kernel estimates, for any $\delta \in (0,1)$, with probability at least $1 - \delta$:*

$$\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \sum_{b=1}^{m_B} \sum_{c=1}^{C} w_{b,c} \pi_c \left(\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|\Psi(b,c)|}} \; + \; \varepsilon \right) \quad (6)$$

where $\Psi(\cdot, \cdot)$ selects a bin based on the index b and possibly the label c.

We provide the proof in the Appendix B.2 together with a discussion of the result under limited cardinality of the auxiliary dataset leveraged for kernel estimates. We also show the result extends to cumulative metrics like $ECCE$ in Appendix C.1. The bound reveals a natural decomposition of calibration error into two components: a deterministic term proportional to the local calibration error ε, and a stochastic term that arises inevitably from finite-sample estimation of empirical frequencies within bins. Even when a model is perfectly locally calibrated ($\varepsilon = 0$), the calibration metric remains subject to sampling fluctuations that vanish only as bin cardinalities grow. 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 $\|\cdot\|_1$ (Gao and Pavel, 2017), 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 as in Luo et al. (2022).

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

$$\left| \hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y} \mid \phi(\mathbf{x}_i)) \right| \leq L \cdot \rho. \quad (7)$$

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:*

$$\limsup_{\rho\to 0} \mathcal{E}(D;\varphi;\beta) \leq L_\varphi \sum_{b=1}^{m_B}\sum_{c=1}^{C} w_{b,c}\pi_c \sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|\Psi(b,c)|}} \tag{8}$$

Proof is in Appendix B.3. 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. We discuss in detail assumptions in Appendix A.

Theorem 3 (Probabilistic bound for multiclass LCE under Local Calibration)**.** *Let f be composed of a feature extractor ϕ and classification layer g. Let f be locally calibrated with error ε relative to kernel estimates $\hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)$ computed on an independent arbitrarily large auxiliary dataset. Let the LCE be computed on $\phi(\mathbf{x}_i)$. Then, conditioning on the evaluation inputs $\{\mathbf{x}_i\}_{i=1}^n$ and on the corresponding consistent kernel estimates, for any $\delta \in [0,1]$, with at least probability $1-\delta$:*

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

where $n_i^{\mathrm{eff}} = \frac{1}{\sum_{j\in I_b} w_{i,j}^2}$, *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 B.4. 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* (LN), i.e., a two-component neural network architecture, designed to produce representations that exhibit improved *local calibration*.

Figure 1a shows the structure of LN. While standard post hoc calibrators consider the features representation as fixed, one component of LNs produces a new reduced-dimensionality feature representation $\phi'(\mathbf{x}_i)$ and the other component parametrizes new logits $\mathbf{l}'$. From this perspective LNs can be viewed as a representation-level post-hoc calibration method. The two components are 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 (Nadaraya, 1964; Watson, 1964), which require specifying a bandwidth hyper-parameter γ. More precisely, LNs minimize the following loss:

$$\mathcal{L}_{\mathrm{LN}}(\mathbf{x}_i,\mathbf{y}_i) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{\mathrm{d}_{\mathrm{JSD}}\left(\hat{\mathbf{p}}_i, \hat{\theta}(\mathbf{y}_i \mid \phi'(\mathbf{x}_i)\right)}^{\text{Alignment term}} + \lambda \cdot \underbrace{\mathcal{L}_{\mathrm{ce}}\left(\mathbf{y}_i, \hat{\theta}(\mathbf{y}_i \mid \phi'(\mathbf{x}_i)\right)}_{\text{Similarity term}}, \tag{9}$$

where $\mathrm{d}_{JSD}(P,Q)$ is the Jensen-Shannon distance[1](Fuglede and Topsøe, 2004; Endres and Schindelin, 2003) and $\mathcal{L}_{\mathrm{ce}}$ is the categorical cross entropy. Notice that two distinct terms compose $\mathcal{L}_{\mathrm{LN}}$, 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 that, for

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

a consistent estimator, the *alignment term* asymptotically converges to the divergence between the model prediction distribution and the true label distribution. This result relies on *kernel consistency* assumptions described in detail in Appendix A.

Theorem 4 (Asymptotic consistency of JSD)**.** *Let f be a probabilistic classifier and let $\hat{\theta}(\cdot \mid \mathbf{x})$ be a kernel estimator of $p(\cdot \mid \mathbf{x})$ that is consistent in mean squared error. Then, under standard regularity conditions:*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} d_{\mathrm{JSD}} \left(\hat{\mathbf{p}}_i, \hat{\theta}(\mathbf{y} \mid \mathbf{x}_i \right) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} d_{\mathrm{JSD}} \left(\hat{\mathbf{p}}_i, \mathbf{p}_i \right). \tag{10}$$

We provide proof in the Appendix B.5. 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. This term serves a similar scope to the one of Lin et al. (2023). Intuitively, it encourages points with the same label to be attracted to nearby neighbors, 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 proposal outperform baselines on *local calibration* metrics?
- **Q2:** Does our approach match the performance of baselines in global calibration metrics?
- **Q3:** Does our method affect predictive performance?

The code[2] to reproduce our results can be found at `https://github.com/Cesbar99/multiclass-local-calibration`

[2]Code developed by University of Trento authors.

8.1 Experimental Settings

Datasets. We evaluate our research questions over three multiclass datasets, i.e., `cifar10`, `cifar100` (Krizhevsky et al., 2009) and `tissuemnist` from the `MedMNIST` collection (Yang et al., 2021).

Methods and Architectures. We evaluate our approach, LN, against different calibration baselines, including Temperature Scaling (TS) (Guo et al., 2017), Isotonic Regression (IR) (Zadrozny and Elkan, 2002), and Platt Scaling (PS) (Platt et al., 1999). In addition, we compare with Dirichlet Calibration (DC) (Kull et al., 2019), the current state-of-the-art for multiclass calibration. Moreover, we consider two K-Cal implementations (Lin et al., 2023): the first one considers the kernel classifier trained using the same loss as LN (KO); the second one does not use the extra Jensen-Shannon regularization as in the original version (KC). We also report the performance of the original non-calibrated classifier (NC).

All methods are based on a `ResNet-50` backbone for `CIFAR-10` and `TissueMNIST`, and a `ResNet-152` for `CIFAR-100`. We detail hyperparameters in Appendix.

Metrics. For **Q1**, we compute two *local calibration* metrics, i.e., class-wise LCE and $MLCE$. For **Q2**, we consider two global calibration metrics, i.e., ECE and $ECCE$. For **Q3**, we report two performance measures, i.e., accuracy (Acc) and Negative Log-Likelihood (NLL). Implementation details are provided in Appendix D.3.

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 data into two sets, one to learn the calibration technique (90% of the calibration data) and one for validation (the re-

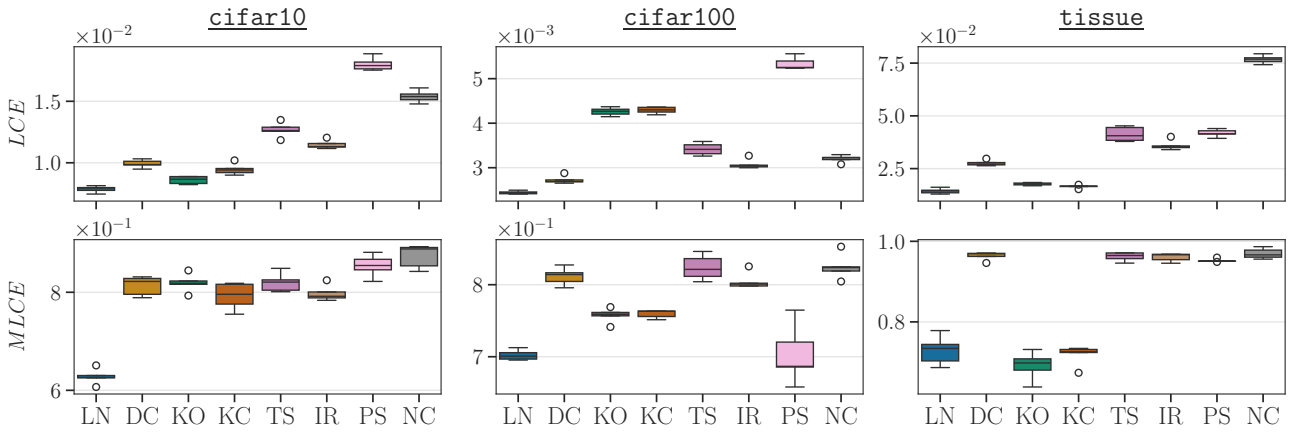


Figure 2: **Empirical local calibration metrics (Q1) over five runs.** The lower the better.

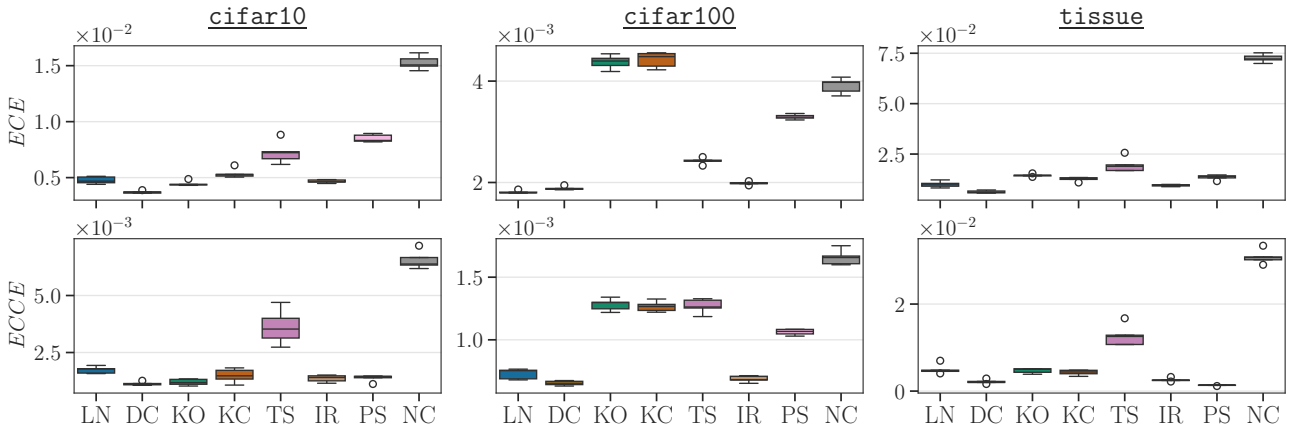


Figure 3: **Empirical global calibration metrics (Q2) over five runs.** The lower the better.

maining 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.

8.2 Experimental Results

Q1: Local Nets consistently achieve superior results on local calibration metrics. Figure 2 evaluates the methods in terms of *local calibration*. Across all datasets, LN consistently emerges as the best-performing approach, providing substantial improvements over competing methods.

On `cifar10`, LN achieves the lowest LCE ($.0079 \pm .0003$), followed by KO ($.0086 \pm .0003$), with KC performing at a similar level to KO. In terms of $MLCE$, LN significantly outperforms all competitors with differences as large as $\approx .16$.

On `cifar100`, LN again attains the best LCE, setting a clear gap with all competitors. Here, LN achieves $MLCE \approx .7022 \pm .0070$, with PS nearly matching its result ($MLCE \approx .7031 \pm .0409$). However, the lower variance of LN 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 LN benefits more from larger sample sizes.

Notably, LN achieves the best performance in terms of LCE also on `tissuemnist`, with $LCE \approx .0144 \pm .0012$, outperforming KC ($LCE \approx .0165 \pm .0008$) and all other baselines. In terms of $MLCE$, KO achieves the best performance with $.6913 \pm .0349$ while LN reaches similar results ($MLCE \approx .7293 \pm .0359$).

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

Q2: Local Nets achieve competitive global calibration results across datasets. Figure 3 summa-

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.

	$ACC \uparrow$			$NLL \downarrow$		
Method	`cifar10`	`cifar100`	`tissue`	`cifar10`	`cifar100`	`tissue`
LN	$.888 \pm .002$	$\underline{.688 \pm .001}$	$.630 \pm .001$	$.347 \pm .002$	$\mathbf{1.125 \pm .002}$	$\mathbf{1.012 \pm .003}$
DC	$.890 \pm .002$	$\mathbf{.690 \pm .002}$	$.617 \pm .003$	$\mathbf{.332 \pm .007}$	$\underline{1.154 \pm .009}$	$1.052 \pm .009$
KO	$\underline{.893 \pm .001}$	$.679 \pm .003$	$\underline{.632 \pm .001}$	$.340 \pm .002$	$1.352 \pm .009$	$1.027 \pm .002$
KC	$\mathbf{.893 \pm .001}$	$.679 \pm .004$	$\mathbf{.632 \pm .001}$	$\underline{.340 \pm .003}$	$1.351 \pm .007$	$\underline{1.021 \pm .002}$
PS	$.884 \pm .001$	$.670 \pm .002$	$.603 \pm .008$	$.466 \pm .004$	$1.618 \pm .007$	$1.180 \pm .008$
IR	$.884 \pm .001$	$.670 \pm .002$	$.603 \pm .008$	$.364 \pm .008$	$1.437 \pm .029$	$1.096 \pm .016$
TS	$.884 \pm .001$	$.670 \pm .002$	$.603 \pm .008$	$.362 \pm .008$	$1.277 \pm .009$	$1.112 \pm .023$
NC	$.884 \pm .001$	$.670 \pm .002$	$.603 \pm .008$	$.494 \pm .018$	$1.502 \pm .036$	$2.100 \pm .101$

rizes the performance of all methods on global calibration metrics. We aim to show that LN performs at a level comparable to the existing baselines.

For ECE, DC is the strongest performer over `cifar10` with $ECE \approx .0037 \pm .0001$, followed by KO with $ECE \approx .0045 \pm .0002$. Still, LN reaches comparable performance ($ECE \approx .0048 \pm .0003$). For $ECCE$, DC again leads with $\approx .0011 \pm .0001$, while LN achieves a level comparable to the other baselines. On `cifar100`, LN achieves the best performance in terms of ECE and is tied with DC and IR on $ECCE$ (Figure 3). On `tissuemnist`, DC again performs best with $ECE \approx .0062 \pm .0007$, while LN ranks third (Figure 3). For $ECCE$, LN performs slightly worse, obtaining $\approx .0050 \pm .0011$.

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 LN consistently achieves competitive results across datasets.

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 LNs rank second ($.347 \pm .002$) compared to DC ($.332 \pm .007$). On `cifar100`, LN 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 LN achieves $1.012 \pm .003$, surpassing KC ($1.021 \pm .002$).

Like all the non-accuracy-preserving methods, we observe gains for LN: +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 LN 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 AND FUTURE WORK

Conclusion. 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 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, potentially improving *local calibration*. 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

We thank the anonymous reviewers for their feedback.

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.

References

Baldeschi, R. C., Di Gregorio, S., Fioravanti, S., Fusco, F., Guy, I., Haimovich, D., Leonardi, S., Linder, F., Perini, L., Russo, M., et al. (2025). Multicalibration yields better matchings. *arXiv preprint arXiv:2511.11413*.

Berta, E., Holzmüller, D., Jordan, M. I., and Bach, F. (2026). Structured matrix scaling for multi-class calibration. In *AISTATS*.

Cianci, G., Goglia, R., Guidotti, R., Kapllaj, M., Mosca, R., Pugnana, A., Ricotti, F., and Ruggieri, S. (2023). Applied data science for leasing score prediction. In *IEEE Big Data*, pages 1687–1696. IEEE.

de Menezes e Silva Filho, T., Song, H., Perelló-Nieto, M., Santos-Rodríguez, R., Kull, M., and Flach, P. A. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach. Learn.*, 112(9):3211–3260.

Endres, D. M. and Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Trans. Inf. Theory*, 49(7):1858–1860.

Fuglede, B. and Topsøe, F. (2004). Jensen-shannon divergence and hilbert space embedding. In *ISIT*, page 31. IEEE.

Gao, B. and Pavel, L. (2017). On the properties of the softmax function with application in game theory and reinforcement learning. *CoRR*, abs/1704.00805.

Globus-Harris, I., Gupta, V., Jung, C., Kearns, M., Morgenstern, J., and Roth, A. (2023). Multicalibrated regression for downstream fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 259–286.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Gupta, C. and Ramdas, A. (2022). Top-label calibration and multiclass-to-binary reductions. In *ICLR*. OpenReview.net.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. (2021). Calibration of neural networks using splines. In *ICLR*. OpenReview.net.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer series in statistics. Springer.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR.

Ibarra, I. A., Gujral, P., Tannen, J., Tygert, M., and Xu, C. (2022). Metrics of calibration for probabilistic predictions. *J. Mach. Learn. Res.*, 23:351:1–351:54.

Jin, H. H., Ding, Z., Ngo, D. D., and Wu, Z. S. (2025). Discretization-free multicalibration through loss minimization over tree ensembles. *arXiv preprint arXiv:2505.17435*.

Jung, C., Lee, C., Pai, M., Roth, A., and Vohra, R. (2021). Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kull, M., de Menezes e Silva Filho, T., and Flach, P. A. (2017). 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.

Kull, M., Perelló-Nieto, M., Kängsepp, M., de Menezes e Silva Filho, T., Song, H., and Flach, P. A. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, pages 12295–12305.

Le-Coz, A., Herbin, S., and Adjed, F. (2024). Confidence calibration of classifiers with many classes. In *NeurIPS*.

Lin, Z., Trivedi, S., and Sun, J. (2023). Taking a step back with kcal: Multi-class kernel-based calibration for deep neural networks. In *ICLR*. OpenReview.net.

Luo, R., Bhatnagar, A., Bai, Y., Zhao, S., Wang, H., Xiong, C., Savarese, S., Ermon, S., Schmerling, E., and Pavone, M. (2022). Local calibration: metrics and recalibration. In *UAI*, volume 180 of *Proceedings of Machine Learning Research*, pages 1286–1295. PMLR.

Marx, C., Zalouk, S., and Ermon, S. (2023). Calibration by distribution matching: Trainable kernel calibration metrics. In *NeurIPS*.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Naeini, M. P., Cooper, G. F., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, pages 2901–2907. AAAI Press.

Noarov, G. and Roth, A. (2023). The statistical scope of multicalibration. In *International Conference on Machine Learning*, pages 26283–26310. PMLR.

Perez-Lebel, A., Morvan, M. L., and Varoquaux, G. (2023). Beyond calibration: estimating the grouping loss of modern neural networks. In *ICLR*. OpenReview.net.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Ruggieri, S. and Pugnana, A. (2025). Things machine learning models know that they don't know. In *AAAI*, pages 28684–28693. AAAI Press.

Sambyal, A. S., Niyaz, U., Krishnan, N. C., and Bathula, D. R. (2023). Understanding calibration of deep neural networks for medical image classification. *Comput. Methods Programs Biomed.*, 242:107816.

Stein, E. M. and Shakarchi, R. (2009). *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press.

Tax, N., Perini, L., Linder, F., Haimovich, D., Karamshuk, D., Okati, N., Vojnovic, M., and Apostolopoulos, P. A. (2026). MCGrad: Multicalibration at Web Scale. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD 2026)*.

Vaicenavicius, J., Widmann, D., Andersson, C. R., Lindsten, F., Roll, J., and Schön, T. B. (2019). Evaluating model calibration in classification. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR.

Valk, K. and Kull, M. (2023). Assuming locally equal calibration errors for non-parametric multiclass calibration. *Trans. Mach. Learn. Res.*, 2023.

Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L., Steyerberg, E. W., diagnostic tests, T. G. E., 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. (2019). Calibration: the achilles heel of predictive analytics. *BMC medicine*, 17(1):230.

Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. *CoRR*, abs/2308.01222.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 359–372.

Xiong, M., Deng, A., Koh, P. W. W., Wu, J., Li, S., Xu, J., and Hooi, B. (2023). Proximity-informed calibration for deep neural networks. In *NeurIPS*.

Yang, J., Shi, R., and Ni, B. (2021). Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In *ISBI*, pages 191–195. IEEE.

Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, pages 609–616. Morgan Kaufmann.

Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *KDD*, pages 694–699. ACM.

Zhao, S., Kim, M. P., Sahoo, R., Ma, T., and Ermon, S. (2021). Calibrating predictions to decisions: A novel approach to multi-class calibration. In *NeurIPS*, pages 22313–22324.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **[Yes]** We clearly state the settings and assumptions of our method in Section 7.

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Not Applicable**] We do not study these properties in our work.

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**] The code to reproduce our results can be found at `https://anonymous.4open.science/r/local-calibration-25A3`.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**] All our theorems and definitions in Sections 4, 5, 6, 7 clearly state assumptions.

   (b) Complete proofs of all theoretical results. [**Yes**] The proofs can be found in Appendix (Section Proofs)

   (c) Clear explanations of any assumptions. [Yes] See Sections 4, 5, 6, 7.

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**] The code to reproduce our results can be found at `https://anonymous.4open.science/r/local-calibration-25A3`.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**] See Section 8.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**] We provide results using boxplots and $avg \pm std$.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**] We provide this information in Appendix (Section Experimental Details).

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [**Yes**] We cite all the datasets owners.

   (b) The license information of the assets, if applicable. [**Not Applicable**]

   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]

   (d) Information about consent from data providers/curators. [**Not Applicable**]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

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 $\mathcal{Y} = \{0, \ldots, C-1\}$ 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 $f\colon \mathcal{X} \to \Delta^C$, 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.

Throughout this appendix, we present the theoretical results and corresponding proofs under a unified framework that covers *multi-dimensional binning*-based calibration metrics This formulation is adopted for clarity and compactness. 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, 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 $\hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i)$, which are computed on an independent, arbitrarily large 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 $n \to \infty$:
   (a) $\gamma_n \to 0$ (the bandwidth shrinks).
   (b) $n\gamma_n^d \to \infty$, 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 (Wasserman, 2006). Nevertheless, NW estimators are known to be consistent in the mean squared error (MSE) sense under mild conditions on the kernel and

bandwidth sequence (Györfi et al., 2002; Wasserman, 2006). 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:

$$\mathbb{E}\Big[\big(\hat{\theta}(\mathbf{y} \mid \mathbf{x}) - \Pr(\mathbf{y} \mid \mathbf{x})\big)^2\Big] \xrightarrow[n\to\infty]{} 0, \tag{11}$$

B Proofs

B.1 Proof of Theorem 1

Statement. Define the continuous Multidimensional Expected Calibration Error (MECE) as:

$$M(f) = \sum_{y\in\mathcal{Y}} p(y)\, \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[|\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(\mathbf{x})|\right],$$

If a model f satisfies local calibration, under standard regularity conditions there exists $k \in (0, 1]$ such that continuous $MECE$ is asymptotically upper bounded:

$$M(f) \leq \varepsilon \cdot k \tag{12}$$

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

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

By the triangle inequality we obtain:

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

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):

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

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

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

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

$$\sum_{y\in\mathcal{Y}} p(y)\, \mathbb{E}_{x\sim p(x)}\left[|\mathbb{E}[\mathbf{y} \mid \mathbf{x}]_y - f_y(x)|\right] \leq \varepsilon \cdot k. \tag{13}$$

□

B.2 Proof of Theorem 2

Statement. Let $\mathcal{E}(D;\varphi;\beta)$ be a calibration metric satisfying Definition 2, evaluated on a sample $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$. Let f be locally calibrated with error ε relative to kernel estimates $\hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)$ computed on an independent arbitrarily large auxiliary dataset. Conditioning on the evaluation inputs $\{\mathbf{x}_i\}_{i=1}^n$ and corresponding consistent kernel estimates, for any $\delta \in (0,1)$, with probability at least $1-\delta$:

$$\mathcal{E}(D;\varphi;\beta) \leq L_\varphi \sum_{b=1}^{m_B}\sum_{c=1}^{C} w_{b,c} \cdot \pi_c\left(\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|\Psi(b,c)|}} + \varepsilon\right) \tag{14}$$

where $\Psi(\cdot,\cdot)$ secelts a bin based on the index b and possibly the label c.

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

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

Then, for any fixed bin b and class c,

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

By the Lipschitz property,

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

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

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

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

Combining the two we obtain:

$$\varphi(\mathrm{freq}_{b,c},\mathrm{conf}_{b,c}) \leq L_\varphi\Big(\big|\mathrm{freq}_{b,c} - \hat{\theta}_{b,c}\big| + \big|\mathrm{conf}_{b,c} - \hat{\theta}_{b,c}\big|\Big). \tag{15}$$

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

$$\mathcal{E}(D;\varphi;\beta) \leq L_\varphi \sum_{b=1}^{B} w_b \sum_{c=1}^{C}\pi_c\Big(\big|\mathrm{freq}_{b,c} - \hat{\theta}_{b,c}\big|\Big) + L_\varphi \sum_{b=1}^{B} w_b \sum_{c=1}^{C}\pi_c\Big(\big|\mathrm{conf}_{b,c} - \hat{\theta}_{b,c}\big|\Big). \tag{16}$$

We bound the two terms inside parentheses separately.

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

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

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

$$\big|\mathrm{conf}_{b,c} - \hat{\theta}_{b,c}\big| \leq \frac{1}{|B_b|}\sum_{i\in I_b}|\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i)| \leq \varepsilon. \tag{17}$$

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

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

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

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

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

1. **Zero mean:**
$$\mathbb{E}\Big[\mathbf{1}\{y_i = c\} - \hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i) \;\Big|\; \mathbf{x}_i\Big] = P(y_i = c \mid \mathbf{x}_i) - \hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i) \approx 0,$$
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 $\hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i)$ 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 $\tau > 0$,

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

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

$$\forall\, b, c: \qquad \left|\mathrm{freq}_{b,c} - \hat{\theta}_{b,c}\right| \leq \sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|B_b|}}. \tag{18}$$

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

$$\boxed{\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \sum_{b=1}^{m_B} \sum_{c=1}^{C} w_b \cdot \pi_c \left(\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|B_b|}} + \varepsilon\right)} \tag{19}$$

Therefore achieving the form of Theorem 2 when $\Psi(b, c) = B_b$ and $w_{b,c} = w_b$.

Class-wise metrics. In the context of class-wise metrics, the only difference is the introduction of a dependence between bins' cardinalities and the classes: $|B_b| \to |B_{b,c}|$. Then, a bound of same style applies and with probability at least $1 - \delta$:

$$\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \sum_{b=1}^{m_B} \sum_{c=1}^{C} w_{b,c} \cdot \pi_c \left(\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|B_{b,c}|}} + \varepsilon\right)$$

and by setting $\Psi(b, c) = B_{b,c}$ we obtain a bound of the same form of Theorem 2. □

Discussion on finite-sample kernel estimates. The proof above assumes that the reference kernel estimates are computed on an arbitrarily large disjoint set, so that they may be treated as deterministic. When the Nadaraya–Watson estimator is instead computed on a finite sample, additional stochastic error arises.

Fix a bin b, class c, and anchor point $\mathbf{x}_i$. The kernel estimator can be written as

$$\hat{\theta}_{i,c} = \sum_{j \in I_b} w_{ij}\, \mathbf{1}\{y_j = c\}, \qquad w_{ij} = \frac{k_{\gamma_n}(\mathbf{x}_i, \mathbf{x}_j)}{\sum_{\ell \in I_b} k_{\gamma_n}(\mathbf{x}_i, \mathbf{x}_\ell)}. \tag{20}$$

Conditioning on the inputs $\{\mathbf{x}_j\}_{j \in I_b}$, the weights w_{ij} are deterministic and the only randomness stems from the labels $\{y_j\}_{j \in I_b}$. Since the label indicators $\mathbf{1}\{y_j = c\}$ are independent and bounded in $[0, 1]$, a weighted Hoeffding inequality applies. In particular,

$$\Pr\Big(\big|\hat{\theta}_{i,c} - \mathbb{E}[\hat{\theta}_{i,c} \mid X]\big| \geq \tau \;\Big|\; X\Big) \leq 2\exp\big(-2\tau^2\, n_i^{\text{eff}}\big)\,, \tag{21}$$

where the effective sample size is

$$n_i^{\text{eff}} := \frac{1}{\sum_{j \in I_b} w_{ij}^2}. \tag{22}$$

Applying a union bound over all $m_B C$ bin–class pairs yields that, with probability at least $1 - \delta$,

$$\forall b, c: \quad \big|\hat{\theta}_{b,c} - \mathbb{E}[\hat{\theta}_{b,c} \mid X]\big| \leq \sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2\, n_b^{\text{eff}}}}, \tag{23}$$

where n_b^{eff} denotes a lower bound on the effective sample size within bin b.

Now decompose

$$\big|\text{freq}_{b,c} - \hat{\theta}_{b,c}\big| = \big|\text{freq}_{b,c} - \mathbb{E}[\hat{\theta}_{b,c} \mid X] - \hat{\theta}_{b,c} + \mathbb{E}[\hat{\theta}_{b,c} \mid X]\big| \tag{24}$$

and apply the triangle inequality:

$$\big|\text{freq}_{b,c} - \hat{\theta}_{b,c}\big| \leq \big|\text{freq}_{b,c} - \mathbb{E}[\hat{\theta}_{b,c} \mid X]\big| + \big|\hat{\theta}_{b,c} - \mathbb{E}[\hat{\theta}_{b,c} \mid X]\big|. \tag{25}$$

The first term is the standard empirical fluctuation term and satisfies

$$\big|\text{freq}_{b,c} - \mathbb{E}[\hat{\theta}_{b,c} \mid X]\big| \leq \sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2\,|B_b|}}. \tag{26}$$

The second term satisfies

$$\big|\hat{\theta}_{b,c} - \mathbb{E}[\hat{\theta}_{b,c} \mid X]\big| \leq \sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2\, n_b^{\text{eff}}}}. \tag{27}$$

Since $n_b^{\text{eff}} \leq |B_b|$, the two terms are of the same order and we may upper bound their sum by

$$\big|\text{freq}_{b,c} - \hat{\theta}_{b,c}\big| \leq 2\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2\, n_b^{\text{eff}}}}. \tag{28}$$

Moreover, under standard smoothness assumptions on the conditional class probabilities (Appendix A.3), the kernel-smoothed conditional probability differs from the true conditional probability by a bias of order $\mathcal{O}(\gamma_n^2)$ (Wasserman, 2006). Combining stochastic concentration and smoothing bias yields

$$\big|\text{freq}_{b,c} - \hat{\theta}_{b,c}\big| \leq 2\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2\, n_b^{\text{eff}}}} + \mathcal{O}(\gamma_n^2). \tag{29}$$

Therefore the bound in Eq. (19) becomes:

$$\Pr\left(\mathcal{E}(D; \varphi; \beta) \leq L_\varphi \sum_{b=1}^{m_B} \sum_{c=1}^{C} w_b \cdot \pi_c\Big(\varepsilon + 2\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2\, n_b^{\text{eff}}}} + \mathcal{O}(\gamma_n^2)\Big)\right) \geq 1 - \delta. \tag{30}$$

The additional $\mathcal{O}(\gamma_n^2)$ term vanishes as $n \to \infty$ under the kernel consistency assumptions.

B.3 Proof of Corollary 1.

Statement. 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:

$$\limsup_{\rho\to 0} \mathcal{E}(D;\varphi;\beta) \leq L_\varphi \sum_{b=1}^{m_B}\sum_{c=1}^{C} w_{b,c} \cdot \pi_c \sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|\Psi(b,c)|}} \tag{31}$$

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

$$\left|\mathrm{conf}_{b,c} - \hat{\theta}_{b,c}\right| \leq \frac{1}{|B_{b,c}|}\sum_{i\in I_{b,c}} |\hat{\mathbf{p}}_{i,c} - \hat{\theta}_c(\mathbf{y}_i \mid \mathbf{x}_i)| \leq \varepsilon.$$

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:

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

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:

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

Substituting back into (6) we obtain:

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

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

$$\boxed{\limsup_{\rho\to 0}\Pr\left(\mathcal{E}(D;\varphi;\beta) \leq L_\varphi \sum_{b=1}^{m_B}\sum_{c=1}^{C} w_{b,c}\cdot\pi_c\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|\Psi(b,c)|}}\right) \geq 1-\delta}. \tag{7}$$

□

B.4 Proof of Theorem 3

Statement. Let f be composed of a feature extractor ϕ and classification layer g. Let f be locally calibrated with error ε relative to kernel estimates $\hat{\theta}(\mathbf{y}_i \mid \mathbf{x}_i)$ computed on an independent arbitrarily large auxiliary dataset. Let the LCE be computed on $\phi(\mathbf{x}_i)$. Then, conditioning on the evaluation inputs $\{\mathbf{x}_i\}_{i=1}^n$ and on the corresponding consistent kernel estimates, for any $\delta \in [0,1]$, with at least probability $1-\delta$:

$$\mathrm{LCE} \leq \varepsilon + \frac{2}{n}\sum_{b=1}^{m_B}\sum_{i\in I_b}\sqrt{\frac{\log(\frac{2nC}{\delta})}{2n_i^{\mathrm{eff}}}} + \frac{L}{n}\sum_{b=1}^{m_B}\mathbb{E}\Big[\sum_{i\in I_b}\sum_{j\in I_b} w_{i,j}\|\phi(\mathbf{x}_j)-\phi(\mathbf{x}_i)\|_1\Big]s$$

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

Proof. 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 $\mathbf{x}_i$:

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

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

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

In the following, we bound the deviation coordinate-wise and obtain a uniform statement over all classes $c \in [C]$. Fix a bin b and consider one coordinate c. By triangle inequality,

$$\frac{1}{|B_b|} \sum_{i \in I_b} \big|\hat{\theta}(\hat{p}_{i,c} \mid \mathbf{x}_i) - \hat{\theta}(y_{i,c} \mid \mathbf{x}_i)\big| \le \frac{1}{|B_b|} \sum_{i \in I_b} \big|\hat{\theta}(\hat{p}_{i,c} \mid \mathbf{x}_i) - \hat{p}_{i,c}\big| + \frac{1}{|B_b|} \sum_{i \in I_b} \big|\hat{p}_{i,c} - \hat{\theta}(y_{i,c} \mid \mathbf{x}_i)\big|$$

By local calibration with respect to the (infinite-sample) kernel target $\theta(\cdot \mid \mathbf{x}_i)$, $\|\hat{\mathbf{p}}_i - \theta(\mathbf{y} \mid \mathbf{x}_i)\|_1 \le \varepsilon$, hence for every c, $|\hat{p}_{i,c} - \theta(y_{i,c} \mid \mathbf{x}_i)| \le \varepsilon$. Therefore, for each $i \in I_b$,

$$\big|\hat{\theta}(y_{i,c} \mid \mathbf{x}_i) - \hat{p}_{i,c}\big| \le \big|\hat{\theta}(y_{i,c} \mid \mathbf{x}_i) - \theta(y_{i,c} \mid \mathbf{x}_i)\big| + \varepsilon \tag{32}$$

Moreover, conditioning on the inputs, $\hat{\theta}(y_{i,c} \mid \mathbf{x}_i)$ is a weighted average of bounded independent label indicators. Thus, we apply Hoeffding's inequality and for each fixed (i, c) and with probability at least $1 - \delta'$,

$$\big|\hat{\theta}(y_{i,c} \mid \mathbf{x}_i) - \theta(y_{i,c} \mid \mathbf{x}_i)\big| \le \sqrt{\frac{\log(\frac{2}{\delta'})}{2\, n_i^{\mathrm{eff}}}}, \qquad n_i^{\mathrm{eff}} := \Big(\sum_{j \in I_b} w_{ij}^2\Big)^{-1} \tag{33}$$

Setting $\delta' = \delta/(nC)$ and applying a union bound over all anchors i and classes c implies that with probability at least $1 - \delta$,

$$\forall i, c: \quad \big|\hat{\theta}(y_{i,c} \mid \mathbf{x}_i) - \theta(y_{i,c} \mid \mathbf{x}_i)\big| \le \sqrt{\frac{\log(\frac{2nC}{\delta})}{2\, n_i^{\mathrm{eff}}}} \tag{34}$$

Averaging over $i \in I_b$ yields

$$\frac{1}{|B_b|} \sum_{i \in I_b} \big|\hat{\theta}(y_{i,c} \mid \mathbf{x}_i) - \hat{p}_{i,c}\big| \le \varepsilon + \frac{1}{|B_b|} \sum_{i \in I_b} \sqrt{\frac{\log(\frac{2nC}{\delta})}{2 n_i^{\mathrm{eff}}}} \tag{35}$$

Before proceeding let us rewrite:

$$\big|\, \hat{\theta}(\hat{p}_{i,c} \mid \mathbf{x}_i) - \hat{p}_{i,c} \,\big| = \Big|\sum_{j \in I_b} w_{ij} \hat{p}_{j,c} - \hat{p}_{i,c}\Big| = \big|\sum_{j \in I_b} w_{ij}(\hat{p}_{j,c} - \hat{p}_{i,c}) + \sum_{j \in I_b} w_{ij} \hat{p}_{i,c} - \hat{p}_{i,c}\,\big| =$$
$$\big|\sum_{j \in I_b} w_{ij}(\hat{p}_{j,c} - \hat{p}_{i,c}) + (\underbrace{\sum_{j \in I_b} w_{ij}}_{=1} - 1)\hat{p}_{i,c}\,\big| = \big|\sum_{j \in I_b} w_{ij}(\hat{p}_{j,c} - \hat{p}_{i,c})\,\big|$$

In addition note that:

$$\frac{1}{|B_b|} \sum_{i \in I_b} \big|\sum_{j \in I_b} w_{ij}(\hat{p}_{j,c} - \hat{p}_{i,c})\,\big| \le \frac{1}{|B_b|} \sum_{i \in I_b} \underbrace{\sum_{j \in I_b} w_{i,j} \big|\, \hat{p}_{j,c} - \hat{p}_{i,c}\,\big|}_{Z_i}$$

Again we follow the **conditioning assumptions** of Appendix A.2. Please note that each coordinate-wise distance satisfies $\hat{p}_{j,c} - \hat{p}_{i,c} \in [-1, 1]$, hence $Z_i \in [0, 1]$. 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:

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

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

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

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

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

Averaging for all anchors in the bin:

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

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 (Gao and Pavel, 2017),

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

More precisely, $L \leq 1$ if the kernel estimates are obtain 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 $1 - \delta$,

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

Conclusion: as $\sum_{c=1}^{C} 1/C = 1$, averaging over bins (weight $|B_b|/n$) yields the final bound:

$$\boxed{\Pr\left(\text{LCE} \leq \ \varepsilon \ + \ \frac{L}{n} \sum_{b=1}^{m_B} \mathbb{E}\Big[\sum_{i \in I_b} \sum_{j \in I_b} w_{i,j} \|\phi(\mathbf{x}_j) - \phi(\mathbf{x}_i)\|_1 \Big] \ + \ \frac{2}{n} \sum_{b=1}^{m_B} \sum_{i \in I_b} \sqrt{\frac{\log\left(\frac{2nC}{\delta}\right)}{2n_i^{\text{eff}}}} \right) \geq 1 - \delta} \tag{36}$$

□

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 $1/\sqrt{2n_i^{\text{eff}}}$. 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.

Class-wise metric. In the context of class-wise *LCE* the only difference is the introduction of a dependence between bins' cardinalities and the classes. Then, with probability at least $1-\delta$, the bound applies with the same exact form under the following minor changes:

$$\text{cwLCE} \;\leq\; \varepsilon \;+\; \frac{L}{nC}\sum_{c=1}^{C}\sum_{b=1}^{m_B}\mathbb{E}\Bigg[\sum_{i\in I_{b,c}}\sum_{j\in I_{b,c}} w_{i,j}\,\|\phi(\mathbf{x}_j)-\phi(\mathbf{x}_i)\|_1\Bigg] \;+\; \frac{2}{nC}\sum_{c=1}^{C}\sum_{b=1}^{m_B}\sum_{i\in I_{b,c}}\sqrt{\frac{\log\!\left(\frac{2\,n\,C}{\delta}\right)}{2\,n_{i,c}^{\text{eff}}}} \tag{37}$$

B.5 Proof of Theorem 4

Statement. Let f be a probabilistic classifier and let $\hat{\theta}(\cdot \mid \mathbf{x})$ be a kernel estimator of $p(\cdot \mid \mathbf{x})$ that is consistent in mean squared error. Then, under standard regularity conditions:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}_{\mathrm{JSD}}\left(\hat{\mathbf{p}}_i,\hat{\theta}(\mathbf{y}\mid\mathbf{x}_i\right) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}_{\mathrm{JSD}}\left(\hat{\mathbf{p}}_i,\mathbf{p}_i\right) \tag{38}$$

Proof. 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 $Q_i = \Pr(\mathbf{y}_i \mid \mathbf{x}_i)$, meaning:

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

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

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}_{\mathrm{JSD}}(\hat{P}_i\|\hat{Q}_i) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}_{\mathrm{JSD}}(\hat{P}_i\|Q_i), \tag{39}$$

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

Since the Jensen-Shannon distance $\mathrm{d}_{\mathrm{JSD}}$ is a metric, it satisfies the triangle inequality:

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

Averaging over i, we obtain:

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

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:

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

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:

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

By the consistency assumption of kernel estimator,

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

and therefore,

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

Combining we obtain:

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

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

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

and therefore:

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

As before, by symmetry:

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

Combining we obtain:

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

Which concludes our proof:

$$\boxed{\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}_{\mathrm{JSD}}(\hat{P}_i\|\hat{Q}_i) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{d}_{\mathrm{JSD}}(\hat{P}_i\|Q_i)}. \tag{40}$$

□

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, with slight discrepancies due to the different procedure of iteration over bins being applied. 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 $\mathcal{Y}=\{1,\dots,C\}$, 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 $f\colon\mathcal{X}\to\Delta^C$, 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 $\beta:\Delta^C\to\{1,\dots,m_B\}$ partition the probability simplex Δ^C into m_B multidimensional (Appendix A) 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 $S_b=\sum_{i\le b}|B_i|$. The $ECCE$ is:

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

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

$$\Pr\left(ECCE \le \sum_{c=1}^{C}\sum_{b=1}^{m_B}\pi_c\cdot w_b\left(\varepsilon+\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|\Psi(b,c)|}}\right)\right)\ge 1-\delta \tag{41}$$

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

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

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

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

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 $\frac{1}{|B_i|}\sum_{j\in I_i}\left(\mathbf{1}\{y_j=c\}-\hat{\theta}_c(\mathbf{y}_j\mid\mathbf{x}_j)\right)$ is the label variables $\{y_j\}_{j\in I_i}$. For each j, the summand satisfies:

1. **Zero mean:**
$$\mathbb{E}\Big[\mathbf{1}\{y_j=c\}-\hat{\theta}_c(\mathbf{y}_j\mid\mathbf{x}_j)\;\Big|\;\mathbf{x}_j\Big]=P(y_j=c\mid\mathbf{x}_j)-\hat{\theta}_c(\mathbf{y}_j\mid\mathbf{x}_j)\approx 0,$$
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 $\tau_b=\sqrt{\frac{\log(\frac{2Cm_B}{\delta})}{2|B_i|}}$ and applying the union bound over the $m_B\cdot C$ bin–class pairs yields: with probability at least $1-\delta$,

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

Then with high probability:

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

$$\leq\sum_{c=1}^{C}\sum_{b=1}^{m_B}\pi_c\frac{S_b}{n}\left[\sum_{i=1}^{b}\frac{|B_i|}{S_b}\sqrt{\frac{\log(\frac{Cm_B}{\delta})}{2|B_i|}}+\sum_{i=1}^{b}\frac{|B_i|}{S_b}\varepsilon\right]$$

Since $\sum_{i=1}^{b}|B_i|/S_b=1$, the bound then simplifies as follows:

$$\Pr\left(ECCE\leq\sum_{c=1}^{C}\sum_{b=1}^{m_B}\pi_c\cdot w_b\left(\varepsilon+\sqrt{\frac{\log(\frac{Cm_B}{\delta})}{2|\Psi(b,c)|}}\right)\right)\geq 1-\delta$$

with $\Psi(b,c)=B_{i^\star}$ with $i^\star=\arg\min_{i\leq b}|B_i|$ and $w_b={}^{S_b}/_{n}$.

Class-wise metric. 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: $|B_b|\to|B_{b,c}|$. Then, with probability at least $1-\delta$, the bound applies with the same exact form under the following minor changes:

$$\textit{class-wise } ECCE\leq\sum_{c=1}^{C}\sum_{b=1}^{m_B}\pi_c\frac{S_{b,c}}{n}\left(\varepsilon+\sqrt{\frac{\log(\frac{Cm_B}{\delta})}{2\min_i|B_{i,c}|}}\right)\tag{42}$$

and setting $\Psi(b,c)=B_{i^\star,c}$ with $i^\star=\arg\min_i|B_{i,c}|$ and $w_{b,c}={}^{S_{b,c}}/_{n}$ we obtain the bound of Theorem 2. □

C.2 Local Calibration and Proximity Bias

Proximity bias is a well-documented phenomenon in probabilistic classifiers (Xiong et al., 2023), 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:*

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

Let f be locally calibrated with error ε with respect to kernel estimates on an arbitrarily large auxiliary dataset, then conditioning on the evaluation samples $\{\mathbf{x}_i\}_{i=1}^n$ and the corresponding consistent kernel estimates, with probability at least $1-\delta\in[0,1]$ the difference in class frequencies between the two sub-bins is bounded as follows:

$$\Pr\left(\|\textit{freq}(S_1)-\textit{freq}(S_2)\|_1\leq 2\varepsilon+\sqrt{\frac{2\log(\frac{4C}{\delta})}{\min(|S_1|,|S_2|)}}+\|\textit{conf}(S_1)-\textit{conf}(S_2)\|_1\right)\geq 1-\delta.\tag{43}$$

Proof. 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 $\mathbf{x}_{(i,1)}, \ldots, \mathbf{x}_{(i,k)}$ are the k nearest neighbors of $\mathbf{x}_i$ in feature space (excluding $\mathbf{x}_i$ itself).

We begin by applying the triangle inequality:

$$\begin{aligned}
\|\mathrm{freq}(S_1) - \mathrm{freq}(S_2)\|_1 &= \left\| \Big(\mathrm{freq}(S_1) - \mathrm{conf}(S_1)\Big) + \Big(\mathrm{conf}(S_1) - \mathrm{conf}(S_2)\Big) + \Big(\mathrm{conf}(S_2) - \mathrm{freq}(S_2)\Big) \right\|_1 \\
&\leq \left\| \Big(\mathrm{freq}(S_1) - \mathrm{conf}(S_1)\Big) + \Big(\mathrm{conf}(S_1) - \mathrm{conf}(S_2)\Big) \right\|_1 + \|\mathrm{conf}(S_2) - \mathrm{freq}(S_2)\|_1 \\
&\leq \|\mathrm{freq}(S_1) - \mathrm{conf}(S_1)\|_1 + \|\mathrm{conf}(S_1) - \mathrm{conf}(S_2)\|_1 + \|\mathrm{conf}(S_2) - \mathrm{freq}(S_2)\|_1 . \qquad (44)
\end{aligned}$$

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:

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

Now define the following events:

$$A := \{\|\mathrm{freq}(S_1) - \mathrm{conf}(S_1)\|_1 > \varepsilon + \eta_{\delta_1}\}, \quad \text{where } \eta_{\delta_1} := \sqrt{\frac{\log(\frac{2C}{\delta'})}{2|S_1|}},$$

$$B := \{\|\mathrm{freq}(S_2) - \mathrm{conf}(S_2)\|_1 > \varepsilon + \eta_{\delta_2}\}, \quad \text{where } \eta_{\delta_2} := \sqrt{\frac{\log(\frac{2C}{\delta'})}{2|S_2|}}.$$

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:

$$\Pr(\neg A \cap \neg B) \geq 1 - 2\delta'.$$

Under this event, we can bound Eq. (44). More precisely, both the first and third terms is bounded by $\varepsilon + \sqrt{\frac{\log(\frac{2C}{\delta'})}{2n_{S_s}}}$, and by choosing $\delta' = \frac{\delta}{2}$ we can conclude:

$$\Pr\left(\|\mathrm{freq}(S_1) - \mathrm{freq}(S_2)\|_1 \leq 2\varepsilon + \sqrt{\frac{2\log(\frac{4C}{\delta})}{\min(|S_1|, |S_2|)}} + \|\mathrm{conf}(S_1) - \mathrm{conf}(S_2)\|_1 \right) \geq 1 - \delta. \qquad (45)$$

□

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 allows to conclude that, under *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 $(\mathbf{x}_i, \hat{\mathbf{p}}_\mathbf{i})$, with $h(\mathbf{x}_i \mid \hat{\mathbf{p}}_\mathbf{i}) > 0$ and $q_f(\hat{\mathbf{p}}_\mathbf{i}) > 0$ for some $i \in S_s$. Then, the bound on the proximity bias asymptotically satisfies:*

$$\forall\, \xi > 0 : \lim_{n\to\infty} \Pr\Big(\|\mathrm{freq}(S_1) - \mathrm{freq}(S_2)\|_1 \le 2\varepsilon + \xi \Big) = 1 \tag{46}$$

Thus, the empirical difference in class frequencies between proximity sub-bins is controlled by ε up to vanishing stochastic fluctuations. This theoretical result allows to infer that *proximity bias* is directly controlled by the *local calibration* property of a model.

Proof. 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:

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

with radius $w_n \to 0$. 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 $(\mathbf{x}_i, \hat{\mathbf{p}}_\mathbf{i})$, with $h(\mathbf{x}_i \mid \hat{\mathbf{p}}_\mathbf{i}) > 0$ and $q_f(\hat{\mathbf{p}}_\mathbf{i}) > 0$ for some $i \in S_s$.

By the triangle inequality (as in (44)),

$$\|\mathrm{freq}(S_1) - \mathrm{freq}(S_2)\|_1 \le \|\mathrm{freq}(S_1) - \mathrm{conf}(S_1)\|_1 + \|\mathrm{conf}(S_1) - \mathrm{conf}(S_2)\|_1 + \|\mathrm{conf}(S_2) - \mathrm{freq}(S_2)\|_1.$$

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

$$\Pr\left(\|\mathrm{freq}(S_s) - \mathrm{conf}(S_s)\|_1 \le \varepsilon + \sqrt{\frac{\log(\frac{2C}{\delta'})}{2\min(|S_1|, |S_2|)}} \right) \;\ge\; 1 - \delta'$$

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

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

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

$$\forall\, \xi > 0 : \lim_{n\to\infty} \Pr\Big(\|\mathrm{freq}(S_1) - \mathrm{freq}(S_2)\|_1 \le 2\varepsilon + \xi \Big) = 1 \tag{47}$$

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 $S_1, S_2 \subseteq \mathcal{I}_n(\hat{\mathbf{p}}_0)$,

$$\|\mathrm{conf}(S_1) - \mathrm{conf}(S_2)\|_1 \le 2Cw_n.$$

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

$$\|\mathrm{conf}(S_1) - \mathrm{conf}(S_2)\|_1 \xrightarrow{n\to\infty} 0. \tag{48}$$

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 $\mathrm{vol}(B(0, r_n))$ 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:

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

Then, by the Lebesgue differentiation theorem (see e.g. Theorem 1.3 in (Stein and Shakarchi, 2009)), for sufficiently small r_n and w_i,

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

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 (Weiserstrass-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:

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

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:

$$c_h q_{\min}\, \mathrm{vol}(B(0, r_n))\, \mathrm{vol}(B(0, w_i)) \;\le\; \Pr\big(\mathbf{X} \in B(\mathbf{x}_i, r_n),\, f(\mathbf{X}) \in B_n(\hat{\mathbf{p}}_0)\big) \;\le\; C_h q_{\max}\, \mathrm{vol}(B(0, r_n))\, \mathrm{vol}(B(0, w_i)).$$

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 $E[|S_s|]$ 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:

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

we obtain:

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

while both $w_i, r_n \to 0$. 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 $|S_s| \sim \mathrm{Binomial}(n, p_n)$ with success probability $p_n = \Pr\big((\mathbf{x}, f(\mathbf{x})) \in B(\mathbf{x}_i, r_n) \times B_n(\hat{\mathbf{p}}_i, w_i)\big)$. As a consequence, for any $\eta \in (0, 1)$, the Chernoff bound gives:

$$\Pr\big(|S_s| \ge (1 - \eta)\, \mathbb{E}[|S_s|]\big) \ge 1 - \exp\Big(- \tfrac{\eta^2}{2}\, \mathbb{E}[|S_s|]\Big).$$

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

$$|S_s| \ge (1 - \eta)\, \mathbb{E}[|S_s|] \to \infty \tag{49}$$

Which allows us to conclude:

$$\forall\, \xi > 0 : \lim_{n \to \infty} \Pr\Big(\|\mathrm{freq}(S_1) - \mathrm{freq}(S_2)\|_1 \le 2\varepsilon + \xi \Big) = 1 \tag{50}$$

□

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 up to vanishing terms 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).

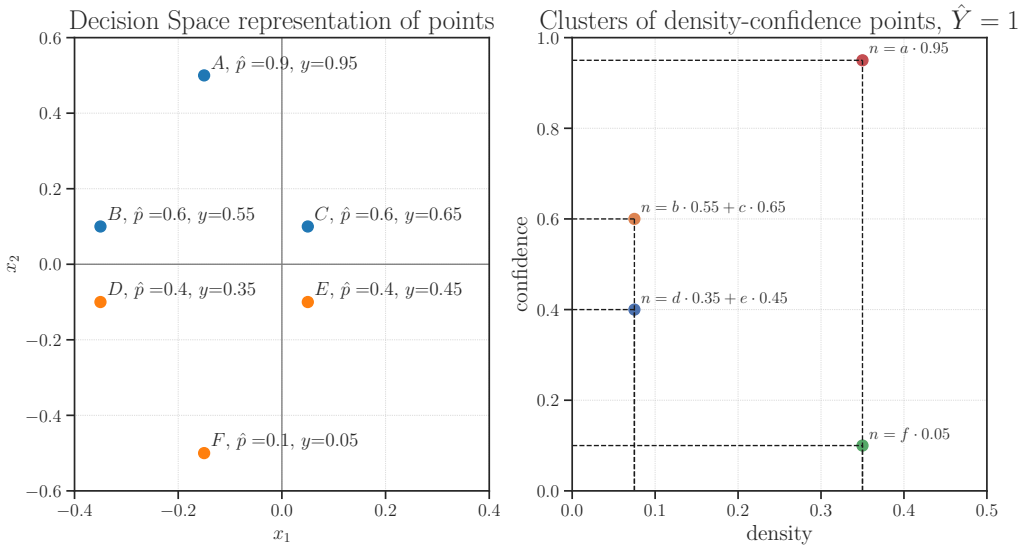


Figure 4: Points in the decision space (right) and their mapping to density-confidence space for calibration (left).

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**	**p**	**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:

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

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 (Guo et al., 2017; Kull et al., 2019). 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 (Kingma and Ba, 2015) 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 LN in the calibration experiments.

Residual Modelling. The LN 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 LN 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:

$$\phi'_{\text{PCA}}(\mathbf{x}) = \tilde{\phi}_{\text{PCA}}(\mathbf{x}) + w_\phi \cdot \phi_{\text{PCA}}(\mathbf{x}) + b_\phi, \quad g'(\mathbf{x}) = \tilde{g}(\mathbf{x}) + w_g \cdot g(\mathbf{x}) + b_g, \tag{51}$$

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 LN 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 LN 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 LN 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 LN 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 LN.

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.

E Detailed Results

Dataset	Method	$LCE \downarrow$	$MLCE \downarrow$	$ECCE \downarrow$	$ECE \downarrow$	$ACC \uparrow$	$NLL \downarrow$
cifar10	LN	**.0079 ± .0003**	**.6279 ± .0157**	.0017 ± .0001	.0048 ± .0003	.888 ± .002	.347 ± .002
	DC	.0099 ± .0003	.8133 ± .0195	**.0011 ± .0001**	**.0037 ± .0001**	.890 ± .002	**.332 ± .007**
	KO	.0086 ± .0003	.8190 ± .0183	.0012 ± .0001	.0045 ± .0002	.893 ± .001	.340 ± .002
	KC	.0095 ± .0005	.7923 ± .0269	.0015 ± .0003	.0054 ± .0004	**.893 ± .001**	.340 ± .003
	PS	.0180 ± .0005	.8542 ± .0224	.0014 ± .0001	.0085 ± .0003	.884 ± .001	.466 ± .004
	IR	.0115 ± .0003	.7978 ± .0162	.0014 ± .0002	.0047 ± .0001	.884 ± .001	.364 ± .008
	TS	.0127 ± .0006	.8200 ± .0191	.0036 ± .0008	.0072 ± .0010	.884 ± .001	.362 ± .008
	NC	.0154 ± .0005	.8738 ± .0238	.0065 ± .0004	.0153 ± .0006	.884 ± .001	.494 ± .018
cifar100	LN	**.0024 ± .0001**	**.7022 ± .0070**	.0007 ± .0001	**.0018 ± .0001**	.688 ± .001	**1.125 ± .002**
	DC	.0027 ± .0001	.8117 ± .0121	**.0007 ± .0001**	.0019 ± .0001	**.690 ± .002**	1.154 ± .009
	KO	.0043 ± .0001	.7575 ± .0101	.0013 ± .0001	.0044 ± .0001	.679 ± .003	1.352 ± .009
	KC	.0043 ± .0001	.7594 ± .0055	.0013 ± .0001	.0044 ± .0002	.679 ± .004	1.351 ± .007
	PS	.0053 ± .0001	.7031 ± .0409	.0011 ± .0001	.0033 ± .0001	.670 ± .002	1.618 ± .007
	IR	.0031 ± .0001	.8049 ± .0116	.0007 ± .0001	.0020 ± .0001	.670 ± .002	1.437 ± .029
	TS	.0034 ± .0001	.8239 ± .0173	.0013 ± .0001	.0024 ± .0001	.670 ± .002	1.277 ± .009
	NC	.0032 ± .0001	.8250 ± .0176	.0017 ± .0001	.0039 ± .0002	.670 ± .002	1.502 ± .036
tissue	LN	**.0144 ± .0012**	.7293 ± .0359	.0050 ± .0011	.0100 ± .0015	.630 ± .001	**1.012 ± .003**
	DC	.0276 ± .0013	.9638 ± .0104	.0021 ± .0005	**.0062 ± .0007**	.617 ± .003	1.052 ± .009
	KO	.0177 ± .0006	**.6913 ± .0349**	.0047 ± .0006	.0144 ± .0006	.632 ± .001	1.027 ± .002
	KC	.0165 ± .0008	.7179 ± .0250	.0043 ± .0006	.0126 ± .0010	**.632 ± .001**	1.021 ± .002
	PS	.0418 ± .0018	.9521 ± .0043	**.0013 ± .0001**	.0135 ± .0012	.603 ± .008	1.180 ± .008
	IR	.0360 ± .0024	.9605 ± .0101	.0026 ± .0004	.0095 ± .0005	.603 ± .008	1.096 ± .016
	TS	.0413 ± .0034	.9620 ± .0107	.0127 ± .0025	.0196 ± .0036	.603 ± .008	1.112 ± .023
	NC	.0768 ± .0019	.9695 ± .0127	.0308 ± .0016	.0725 ± .0020	.603 ± .008	2.100 ± .101

Table 3: Results for all experiments over five seeds.

Table 3 reports all the results for **Q1**, **Q2** and **Q3**.