Journal of Machine Learning Research 24 (2023) 1-72 Submitted 3/22; Revised 11/23; Published 11/23

# T-Cal: An Optimal Test for the Calibration of Predictive Models

**Donghwan Lee*** DH7401@SAS.UPENN.EDU
*Graduate Group in Applied Mathematics and Computational Science*
*University of Pennsylvania*
*Philadelphia, PA 19104-6340, USA*

**Xinmeng Huang*** XINMENGH@SAS.UPENN.EDU
*Graduate Group in Applied Mathematics and Computational Science*
*University of Pennsylvania*
*Philadelphia, PA 19104-6340, USA*

**Hamed Hassani** HASSANI@SEAS.UPENN.EDU
*Department of Electrical and Systems Engineering*
*University of Pennsylvania*
*Philadelphia, PA 19104-6340, USA*

**Edgar Dobriban** DOBRIBAN@WHARTON.UPENN.EDU
*Department of Statistics and Data Science*
*University of Pennsylvania*
*Philadelphia, PA 19104-6340, USA*

**Editor:** Ambuj Tewari

## Abstract

The prediction accuracy of machine learning methods is steadily increasing, but the calibration of their uncertainty predictions poses a significant challenge. Numerous works focus on obtaining well-calibrated predictive models, but less is known about reliably assessing model calibration. This limits our ability to know when algorithms for improving calibration have a real effect, and when their improvements are merely artifacts due to random noise in finite datasets. In this work, we consider detecting mis-calibration of predictive models using a finite validation dataset as a hypothesis testing problem. The null hypothesis is that the predictive model is calibrated, while the alternative hypothesis is that the deviation from calibration is sufficiently large.

We find that detecting mis-calibration is only possible when the conditional probabilities of the classes are sufficiently smooth functions of the predictions. When the conditional class probabilities are Hölder continuous, we propose *T-Cal*, a minimax optimal test for calibration based on a debiased plug-in estimator of the $\ell_2$-Expected Calibration Error (ECE). We further propose *adaptive T-Cal*, a version that is adaptive to unknown smoothness. We verify our theoretical findings with a broad range of experiments, including with several popular deep neural net architectures and several standard post-hoc calibration methods. T-Cal is a practical general-purpose tool, which—combined with classical tests for discrete-valued predictors—can be used to test the calibration of virtually any probabilistic classification method. T-Cal is available at `https://github.com/dh7401/T-Cal`.

∗. Equal Contribution.

**Keywords:** uncertainty quantification, calibration, nonparametric statistics, hypothesis testing, minimax optimality

## 1 Introduction

The prediction accuracy of contemporary machine learning methods such as deep neural networks is steadily increasing, leading to adoption in more and more safety-critical fields such as medical diagnosis (Esteva et al., 2017), self-driving vehicles (Bojarski et al., 2016), and recidivism forecasting (Berk, 2017). In these applications and beyond, machine learning models are required not only to be accurate but also to be well-calibrated: giving precise probability estimates for the correctness of their predictions.

To be concrete, consider a classification problem where the goal is to classify features $\mathbf{x}$ (such as images) into one of several classes $\mathbf{y}$ (such as a building, vehicle, etc.). A probabilistic classifier (or, probability predictor) $f$ assigns to each input $\mathbf{x}$ a probability distribution $f(\mathbf{x})$ over the classes. For a given input $\mathbf{x}$, the entries of $f(\mathbf{x})$ represent the probabilities assigned by the classifier to the event that the outcome belongs to the $k$-th class, for any $k = 1, \ldots, K$. This classifier is *calibrated* if for any value $\mathbf{z}$ taken by $f(\mathbf{x})$, and for all classes $k$, the probability that the outcome belongs to the $k$-th class, i.e., $[\mathbf{y}]_k = 1$, equals the predicted probability, i.e., the $k$-th coordinate $[\mathbf{z}]_k$ of $\mathbf{z}$:

$$P([\mathbf{y}]_k = 1 | f(\mathbf{x}) = \mathbf{z}) = [\mathbf{z}]_k.$$

This form of calibration is an important part of uncertainty quantification, decision science, analytics, and forecasting (see e.g., Hilden et al., 1978; Miller et al., 1991, 1993; Steyerberg et al., 2010; Hand, 1997; Jolliffe and Stephenson, 2012; Van Calster and Vickers, 2015; Harrell, 2015; Tetlock and Gardner, 2016; Shah et al., 2018; Steyerberg, 2019, etc). Unfortunately, however, recent works starting from at least Guo et al. (2017) have reported that modern machine learning methods are often poorly calibrated despite their high accuracy; which can lead to harmful consequences (e.g., Van Calster and Vickers, 2015; Steyerberg, 2019).

To address this problem, there has been a surge of works aimed at improving the calibration of machine learning models. These methods seek to achieve calibration either by modifying the training procedure (Harrell, 2015; Lakshminarayanan et al., 2017; Kumar et al., 2018; Thulasidasan et al., 2019; Zhang et al., 2020; Mukhoti et al., 2020) or by learning a re-calibration function that transforms, in a post-hoc way, the predictions to well-calibrated ones (Cox, 1958; Mincer and Zarnowitz, 1969; Steyerberg et al., 2010; Platt, 1999; Zadrozny and Elkan, 2001, 2002; Guo et al., 2017; Kumar et al., 2019; Kisamori et al., 2020).

In this regard, a key challenge is to rigorously assess and compare the performance of calibration methods. Without such assessments, we have limited ability to know when algorithms for improving calibration have a real effect, and when their improvements are merely artifacts due to random noise in finite-size datasets. As it turns out, existing works do not offer a satisfactory solution to this challenge.

In more detail, in this work, we consider the problem of detecting mis-calibration of predictive models using a finite validation dataset. We focus on models whose probability predictions are continuously distributed—which is generally reasonable for many modern

machine learning methods, including deep neural nets. We develop efficient and provably optimal algorithms to test their calibration.

Detecting mis-calibration has been studied from the perspective of statistical hypothesis testing. The seminal work of Cox (1958) formulated a test of calibration for a collection of binary (yes-no) predictions, and proposed using a score test for a logistic regression model. This has been widely used and further developed, leading to various tests for the so-called calibration slope and calibration intercept, which can validate various qualitative versions of model calibration, see e.g., Hosmer and Lemesbow (1980); Miller et al. (1991); Steyerberg (2019) and references therein. In pioneering work, Miller (1962) suggested a chi-squared test for testing calibration of multiple series of binary predictions. To deal with the challenging problem of setting critical values (i.e., how large of an empirical mis-calibration is statistically significant?) for testing calibration, bootstrap methods have become common, see e.g., Harrell (2015). We refer to Section 1.1 for more details and for a discussion of other related works.

In contrast to the above works that aim to test calibration slopes and intercepts, we aim to develop a *nonparametric hypothesis test* for calibration, which does not assume a specific functional form (such as a logistic regression model), for the deviations to be detected from perfect calibration. A nonparametric approach has the advantage that it can detect subtle forms of mis-calibration even after re-calibration by parametric methods. However, existing approaches for nonparametric testing often rely on ad hoc techniques for binning the probability predictions, which is a limitation because the results can depend on the way that the binning has been performed (Harrell, 2015; Steyerberg, 2019). In contrast, our *adaptive tests* automatically select an optimal binning scheme. Finally, as a new development in the area of testing calibration, T-Cal has theoretically guaranteed *minimax optimality properties* for detecting certain reasonable types of smooth mis-calibration. These properties make T-Cal both practically and theoretically appealing.

We consider a given multi-class probabilistic classifier, and are interested in testing if it is calibrated. We make the following contributions:

- As a candidate test statistic, we consider the plug-in estimator of $\ell_2$-expected calibration error (ECE), which is the expectation of the squared distance between the probability predictions and class probabilities given these predictions. This is also known as the mean calibration error (e.g., Harrell, 2015, p. 105). While the plug-in estimator is biased (i.e., its expectation is not zero even under perfect calibration), we show how to construct a *debiased plug-in estimator* (DPE).

  We consider detecting mis-calibration when the deviation between predicted class probabilities and their true values—the "mis-calibration curve"—satisfies a classical smoothness condition known as Hölder continuity. We later show that a smoothness condition is essentially unavoidable. Under this condition, we show that T-Cal can detect mis-calibration if the ECE is sufficiently large and the number of bins is chosen appropriately, depending on the smoothness (Theorem 3).

- To make T-Cal practical, we present a version that is adaptive to the unknown smoothness parameter (Theorem 5). This makes T-Cal fully tuning-free and practically useful. From a theoretical perspective, adaptivity only requires a minor additional increase in the level of mis-calibration that can be detected; by a $\log n$ factor.

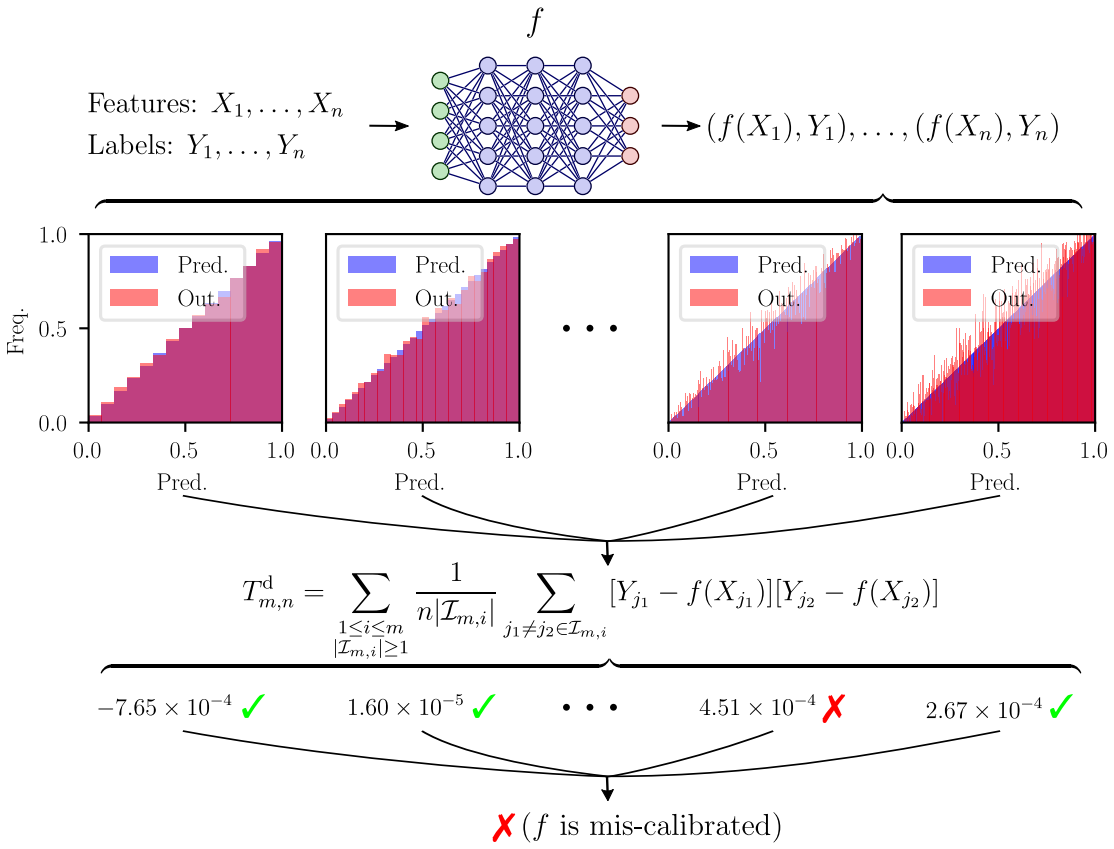



Figure 1: An overview of adaptive T-Cal. For a given probability predictor $f$, we compute $T^{\mathrm{d}}_{m;n}$, the debiased plug-in estimator (DPE), binned over several scales (See (6) for the definition). We then compare each value with the hypothetical distribution of DPE that we would get if the model were perfectly calibrated. The hypothesis of perfect calibration is rejected if at least one of the scales is detected to be mis-calibrated. This multi-scale approach ensures that T-Cal adaptively detects mis-calibration.

- We support our theoretical results with a broad range of experiments. We provide simulations, which support our theoretical optimality results. We also provide experiments with several popular deep neural net architectures (ResNet-50, VGG-19, DenseNet-121, etc), on benchmark datasets (CIFAR 10 and 100, ImageNet) and several standard post-hoc calibration methods (Platt scaling, histogram binning, isotonic regression, etc).

- To complement these results, we argue that T-Cal is optimal, by providing a number of fundamental lower bounds. We prove that detecting mis-calibration from a finite dataset is only possible when the mis-calibration curve is sufficiently smooth, and it is not possible when the curve is just continuous (Proposition 9).

  When the mis-calibration curves are Hölder smooth, we show that the calibration error required for reliable detection of mis-calibration has to be appropriately large (Theorem 10). This minimax result relies on Ingster's (or the chi-squared) method. Combined with our previous results, this shows that T-Cal is minimax optimal.

- To further put our problem in context, we show that testing calibration can be reduced to a well-known problem in statistical inference—the two-sample goodness-of-fit problem—by a novel randomization technique. Based on this insight, and building on the results of Arias-Castro et al. (2018); Kim et al. (2022) on goodness-of-fit testing, we present another asymptotically minimax optimal test for mis-calibration that matches the lower bound (Theorem 12). While this method is theoretically optimal, it relies on sample splitting and is not as sample-efficient as our previous method in experiments.

- In the proofs, we have the following innovations:

  1. We introduce an *equal-volume binning scheme* for the probability simplex $\Delta_{K-1}$ (Appendix B.3). We decompose the probability simplex into hypersimplices by taking intersections with smaller hypercubes composing the unit hypercube $[0,1]^K$. Then we further decompose the hypersimplices into equal-volume simplices using results in polyhedral combinatorics. This construction enables us to extend proof techniques from the nonparametric hypothesis testing literature to our setting.
  2. To analyze our plug-in estimator, we need to deal with terms involving probability scores of inputs, which are continuous random variables. This is different from the structure of chi-squared statistics such as that of Ingster (1987). Thus, computing the mean and variance of the DPE requires a different analysis.
  3. While densities on the probability simplex can take arbitrary positive values, the conditional expectation of probability predictions has to lie in the probability simplex. This requires a careful construction of alternative distributions to use Ingster's method.

Our numerical results can be reproduced with code available at `https://github.com/dh7401/T-Cal`.

We now summarize some key takeaways:

- **The need for statistical significance to claim calibration.** It is crucial to perform rigorous statistical tests to assess the calibration of machine learning methods. While models with smaller empirical ECE generally tend to be better calibrated, these values can be highly influenced by noise and randomness inherent in finite datasets. Hence it is crucial to develop and use tools to assess statistical significance—such as the hypothesis tests of calibration that we develop—when claiming improved calibration.
- **Potential suboptimality of popular approaches.** The currently prevalent usage of popular metrics, such as the empirical ECE, may be suboptimal. The current standard is to evaluate mis-calibration metrics using a fixed number of bins (such as 15) of the probability scores, for all prediction models (ResNet, VGG, etc), and all datasets. Our results show theoretically that the optimal number of bins increases with the level of oscillations and non-smoothness expected in the probability predictor. Modern machine learning methods are becoming more and more over-parametrized and data-adaptive. This suggests that it is ever more important to use a careful model- and data-adaptive test (and number of bins) when testing calibration.

### 1.1 Related Works

There is a great body of related work on evaluating the calibration of prediction methods, on improving calibration accuracy, and on nonparametric hypothesis testing techniques. We review the most closely related works.

**Broader Context.** Broadly speaking, the study of calibration is an important part of the study of classification, prediction, analytics, and forecasting (e.g., Hilden et al., 1978; Miller et al., 1991, 1993; Steyerberg et al., 2010; Hand, 1997; Jolliffe and Stephenson, 2012; Gneiting and Katzfuss, 2014; Van Calster and Vickers, 2015; Harrell, 2015; Tetlock and Gardner, 2016; Shah et al., 2018; Steyerberg, 2019, etc).

**Calibration.** As recounted in Lichtenstein et al. (1977), research on calibration dates back at least to the early 1900s, when meteorologists suggested expressing predictions as probabilities and comparing them to observed empirical frequencies. Calibration has since been studied in a variety of areas, including meteorology, statistics, medicine, computer science, and social science; and under a variety of names, such as realism or realism of confidence, appropriateness of confidence, validity, external validity, secondary validity, and reliability (Lichtenstein et al., 1977). A general finding in this area is that human forecasters are often overconfident and thus mis-calibrated (e.g., Keren, 1991, etc), as codified for instance in Tversky and Kahneman's celebrated work on prospect theory (Kahneman and Tversky, 2013).

Beyond our hypothesis testing perspective, approaches to study calibration include Bayesian perspectives (e.g., Dawid, 1982; Kadane and Lichtenstein, 1982, etc.) and on-line settings (e.g., Foster and Vohra, 1998; Vovk and Shafer, 2005, etc.). See also Section 10.9 of Harrell (2015), Section 15.3 of Steyerberg (2019), and Hastie and Tibshirani (1998); Ivanov et al. (1999); Garczarek (2002); Buja et al. (2005); Toll et al. (2008); Gebel (2009); Serrano (2012); Van Calster et al. (2019); Huang et al. (2020), among others.

**Calibration Measures.** Proper scoring rules (Good, 1952; De Finetti, 1962; Savage, 1971; Winkler et al., 1996; DeGroot and Fienberg, 1983; Gneiting et al., 2007) such as

the Brier score (Brier, 1950) and negative log-likelihood (e.g., Winkler et al., 1996, etc) are objective functions of two probability distributions (a true distribution and a predicted distribution). They are minimized when the predicted distribution equals the true distribution; see also Bickel (2007).

As discussed in Sections 4.5 and 10.9 of Harrell (2015), some of the standard techniques in the area include plotting calibration curves, also known as reliability diagrams (estimated probabilities against predicted ones); which can be bias-corrected using the bootstrap; and re-calibration by fitting statistical models to these curves (e.g., Austin and Steyerberg, 2014, etc). More recently, the notion of ECE, also known as mean absolute calibration error (e.g., Harrell, 2015, p. 105) is popularized in Naeini et al. (2015) and later generalized to multi-class settings in Vaicenavicius et al. (2019). Gupta et al. (2021) develop a binning-free calibration measure based on the Kolmogorov-Smirnov test. Arrieta-Ibarra et al. (2022) introduce calibration metrics building on Komogorov-Smirnov and Kuiper statistics.

**Calibration in Modern Machine Learning.** Guo et al. (2017) draw attention to the mis-calibration of modern neural networks and compare different recalibration methods based on ECE. Many other works (Milios et al., 2018; Kull et al., 2019; Zhang et al., 2020) also evaluate their methods using ECE or its variants. In the following works of Vaicenavicius et al. (2019); Kumar et al. (2019), it has been recognized that ECE evaluated on a fixed binning scheme can underestimate the calibration error. The limitation of fixed binning has been known for the analogous problems of testing probability distributions and densities, see e.g., Mann and Wald (1942), or page 19 of Ingster (2012). Kumar et al. (2019) proposes a debiased ECE, but only for probability predictors with a finite number of outputs. Nixon et al. (2019) empirically study various versions of ECE obtained by adjusting hyperparameters involved in the estimator of ECE (such as norm, binning scheme, and class conditionality), and find that the choice of calibration measure is crucial when comparing different calibration methods. **?** propose a heuristic for choosing an optimal number of bins when computing ECE. Zhang et al. (2020) use kernel density estimation to estimate ECE without relying on a binning scheme. Zhao et al. (2020) show that individual calibration is possible by randomized predictions and propose a training objective to enforce individual calibration. See also Niculescu-Mizil and Caruana (2005); Kull et al. (2017); Bai et al. (2021), among others.

There has been interest in a variety of forms of calibration. We study the strongest form, multi-class calibration, which is stronger than other definitions such as marginal calibration and confidence (top) calibration (Vaicenavicius et al., 2019; Widmann et al., 2019).

**Nonparametric Hypothesis Testing.** Ingster (1986) derives the minimax testing rate for two-sample testing where $s$-Hölder continuous densities on $[0,1]$ are separated in an $L^2$ sense, and shows that the chi-squared test achieves the minimax optimal rate $n^{-2s/(4s+1)}$. Ingster (1987) extends this result to $L^p$ metrics and derives the minimax optimal rate $n^{-s/(2s+1-\max\{2,p\}^{-1})}$ for $1 \quad p < 1$ and $(n/\log n)^{s/(2s+1)}$ for $p = 1$. Ingster (2000) proposes an adaptive version of the test at the cost of $(\log\log n)^{s/(4s+1)}$ factor in the minimax rate. Arias-Castro et al. (2018) extend these results to densities on $[0,1]^d$ and show the minimax rate $n^{-2s/(4s+d)}$. Kim et al. (2022) prove that a permutation test can also achieve the same optimal rate. Butucea and Tribouley (2006) study two-sample testing for one-dimensional densities in Besov spaces; they also prove adaptivity. See also Balakrishnan

and Wasserman (2018); Donoho and Jin (2015); Jin and Ke (2016); Chhor and Carpentier (2021); Dubois et al. (2021); Berrett et al. (2021) for reviews and further related works.

**Nonparametric Functional Estimation.** Bickel and Ritov (1988) study the problem of estimating the quadratic integral functional of the $k$-th derivative of $s$-Hölder probability densities on $\mathsf{R}$, and prove that the optimal convergence rate is $n^{-[4(s-k)=(4s+1)\wedge 1=2]}$. Donoho and Nussbaum (1990); Brown and Low (1996) show an analogous result for regression functions in the Gaussian white noise model. Birgé and Massart (1995) generalize these results to smooth integral functionals of Hölder densities and their derivatives, and prove the same convergence rate. Kerkyacharian and Picard (1996) provide optimal Haar wavelet-based estimators of cubic functionals of densities over the broader class of Besov spaces; and also discuss estimating integrals of other powers of the density. Laurent (1996) studies estimation of functionals of the form $\int_{\mathsf{R}} \phi(f(x), x) d\mu(x)$ of densities $f$, where $\phi$ is a sufficiently smooth function and $\mu$ is a measure. This work constructs estimators attaining the optimal parametric rate using orthogonal projections, including showing semiparametric efficiency, when the smoothness $s > d/4$ in dimension $d$. Robins et al. (2008); Giné and Nickl (2008); Tchetgen et al. (2008) introduce estimation method using higher-order U-statistics. Efromovich and Low (1996); Cai and Low (2006); Giné and Nickl (2008); Mukherjee et al. (2015) propose estimation methods adaptive to unknown smoothness $s$ based on Lepski's method (Lepski, 1991; Lepski and Spokoiny, 1997). See Giné and Nickl (2021) for a more thorough review of related literature.

**Hypothesis Testing for Calibration.** Cox (1958) formulates a test of calibration for a collection of Bernoulli random variables, as a test that their success probabilities are equal to some given values; and proposed using a score test for a logistic regression model. These tests are referred to as testing the calibration slope and intercept, and they are part of a broader hierarchy of calibration (Van Calster et al., 2016). See also Miller et al. (1991); Steyerberg (2019) and references therein. Miller (1962), Section 5, suggests a chi-squared test for testing calibration of a collection of sequences of Bernoulli random variables. Spiegelhalter (1986) proposes a test of calibration based on the Brier score, for discrete-valued probability predictors. The Hosmer-Lemeshow test (Hosmer and Lemesbow, 1980) is a goodness-of-fit test for logistic regression models. The test is based on a chi-squared statistic that measures differences between expected and observed numbers of events in subgroups, and thus has, on the surface, a similarity to the types of test statistics we consider. There are also related tests for comparing predictors (Schervish, 1989; Diebold and Mariano, 1995).

Seillier-Moiseiwitsch and Dawid (1993) study testing the calibration of sequential probability forecasts. Bröcker and Smith (2007) study the bootstrap-based procedure they call consistency resampling to produce standard error bars in reliability diagrams; without focusing on its optimality. For testing the calibration of forecasted densities, Dawid (1984); Diebold et al. (1998) propose the probability integral transform (PIT). Held et al. (2010) propose a score-based approach for testing calibration. Vaicenavicius et al. (2019) use consistency resampling to test a hypothesis of perfect calibration; again without studying its optimality. Widmann et al. (2019) propose kernel-based mis-calibration measures together with their estimators, and argue that the estimators can be viewed as calibration test statistics. Tamás and Csáji (2021) suggest distribution-free hypothesis tests for the null $H_0 : \mathsf{E}[Y \mid X] = X$ based on conditional kernel mean embedding.

**Note on Terminology.** The term calibration sometimes has a different meaning in a variety of areas of human activity, including measurement technology, engineering, economics, and even statistics, etc., see e.g., Franklin (1999); Dawkins et al. (2001); Kodovskỳ and Fridrich (2009); Osborne (1991); Vovk et al. (2020); Angelopoulos et al. (2021). These generally mean adjusting a measurement to agree with a desired standard, within a specified accuracy. However, in our work, we focus on the notion of probabilistic calibration described so far.

### 1.2 Notations

For an integer $d \geq 1$, and a vector $\mathbf{v} \in \mathbb{R}^d$, we refer to the coordinates of $\mathbf{v}$ as both $[\mathbf{v}]_1, \ldots, [\mathbf{v}]_d$ and $v_1, \ldots, v_d$. For any $p \geq 1$, and for an integer $K \geq 2$, we denote the $\ell_p$-norm of $\mathbf{x} = (x_1, \ldots, x_K)^\top \in \mathbb{R}^K$ by $\|\mathbf{x}\|_p := (\sum_{i=1}^K |x_i|^p)^{1/p}$. When $p$ is unspecified, $\|\cdot\|$ stands for $\|\cdot\|_2$. For an event $A$, we denote by $I(A)$ its indicator random variable, where $I(A) = 1$ if event $A$ happens, and $I(A) = 0$ otherwise. For two real numbers $a, b$, we denote $a \wedge b := \min(a, b)$. For two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ with $b_n \neq 0$, we write $a_n \asymp b_n$ if $0 < \liminf_n a_n/b_n \leq \limsup_n a_n/b_n < \infty$. When the index $n$ is self-evident, we may omit it above. We use the Bachmann-Landau asymptotic notations $\Omega(\cdot), \Theta(\cdot)$ to hide constant factors in inequalities and use $\tilde{\Omega}(\cdot), \tilde{\Theta}(\cdot)$ to also hide logarithmic factors. For a Lebesgue measurable set $A \subseteq \mathbb{R}^d$, we denote by $\mathbf{1}_A : \mathbb{R}^d \to \{0, 1\}$ its indicator function where $\mathbf{1}_A(\mathbf{x}) = 1$ if $\mathbf{x} \in A$ and $\mathbf{1}_A(\mathbf{x}) = 0$ otherwise. For a real number $s \in \mathbb{R}$, we denote the largest integer less than or equal to $s$ by $\lfloor s \rfloor$. Also, the smallest integer greater than or equal to $s$ is denoted by $\lceil s \rceil$.

## 2 Definitions and Setup

For $K \geq 2$, consider a $K$-class classification problem where $X \in \mathcal{X}$ is the input feature vector (for instance, an image) and $Y \in \mathcal{Y} := \{\mathbf{y} = (y_1, \ldots, y_K)^\top \in \{0, 1\}^K : \sum_{i=1}^K y_i = 1\}$ is the one-hot encoded output label (for instance, the indicator of the class of the image: building, vehicle, etc).

We consider a probabilistic classifier $f$ mapping the feature space to probability distributions over $K$ classes. Formally, the output space is the $(K-1)$-dimensional probability simplex $\Delta_{K-1}$,

$$\Delta_{K-1} := \{\mathbf{z} = (z_1, \ldots, z_K)^\top \in [0, 1]^K : z_1 + \cdots + z_K = 1\},$$

i.e., $f : \mathcal{X} \to \Delta_{K-1}$. For any $k \in \{1, \ldots, K\}$, the individual component $[f(X)]_k$ denotes the predicted probability of the $k$-th class. Thus, $f$ is also referred to as a probability predictor. The probability predictor $f$ is assumed to be pre-trained on data that are independent of our calibration data at hand.

We assume that the feature-label pair $(X, Y)$ has an unknown joint probability distribution $P$ on $\mathcal{X} \times \mathcal{Y}$. Calibration requires that the predicted probabilities of correctness are equal to the true probabilities. Thus, given that we predicted the probabilities $f(X) = \mathbf{z}$, and thus $[f(X)]_k = [\mathbf{z}]_k$, the true probability that $[Y]_k = 1$ should be equal to $[f(X)]_k = [\mathbf{z}]_k$. Thus, for almost every $\mathbf{z}$, calibration requires that for all $k = 1, \ldots, K$,

$$P[[Y]_k = 1 \mid f(X) = \mathbf{z}] = [\mathbf{z}]_k.$$

We can reformulate this in a way that is more convenient to study. The map $(\mathbf{x}, \mathbf{y}) \mapsto (f(\mathbf{x}), \mathbf{y})$ induces a probability distribution on $\Delta_{K-1} \times Y$; where we can think of $(\mathbf{x}, \mathbf{y})$ as a realization of $(X, Y)$. As will be discussed shortly, calibration only depends on the joint distribution of $(f(X), Y)$. For this reason, we also denote the joint distribution of $(f(X), Y)$ by $P$ when there is no confusion. We write $Z := f(X)$ for the predicted probabilities corresponding to $X$.

We define the *regression function* $\mathrm{reg}_f : \Delta_{K-1} \to \Delta_{K-1}$ as

$$\mathrm{reg}_f(\mathbf{z}) := \mathbb{E}[Y \mid f(X) = \mathbf{z}] = \mathbb{E}[Y \mid Z = \mathbf{z}],$$

where the expectation is conditioned on the score $Z$ with $(Z, Y) \sim P$. Note that each component, for $k = 1, \ldots, K$, has the form $\mathbb{E}[[Y]_k \mid f(X) = \mathbf{z}] = P([Y]_k = 1 \mid f(X) = \mathbf{z})$. Especially for binary classification, this is also referred to as the *calibration curve* of the probabilistic classifier $f$ (Harrell, 2015). Since we are particularly interested in continuous probability predictors, we assume that the marginal distribution $P_Z$ of $Z$ has a density with respect to the uniform measure on $\Delta_{K-1}$. Then, this expectation is well-defined almost everywhere.

In this language, the probabilistic classifier $f$ is *perfectly calibrated* if $\mathrm{reg}_f(Z) = Z$ almost everywhere.[1] Further, it turns out that it is important to study *the deviations from calibration*. For this reason, we define the *residual function* $\mathrm{res}_f : \Delta_{K-1} \to \mathbb{R}^K$ as

$$\mathrm{res}_f(\mathbf{z}) := \mathrm{reg}_f(\mathbf{z}) - \mathbf{z},$$

so that perfect calibration amounts to $\mathrm{res}_f(Z) = 0$ almost everywhere. When $(Z, Y)$ have a joint distribution $P$, we sometimes write $\mathrm{res}_f = \mathrm{res}_{f;P}$ to display the dependence of the mis-calibration curve on $P$. As we will see, the structure of the residual function crucially determines our ability to detect mis-calibration. In analogy to the notion of calibration curves mentioned above, we may also call $\mathrm{res}_f$ the *mis-calibration curve* of the probabilistic classifier $f$.

We observe calibration data $(Z_i, Y_i) \in \Delta_{K-1} \times Y$, $i \in \{1, \ldots, n\}$, sampled i.i.d. from $P$, and denote their joint product distribution as $P^n$. Our goal is to rigorously test if $f$ is perfectly calibrated based on this finite calibration dataset. The calibration properties of the probabilistic classifier $f$ can be expressed equivalently in terms of the distribution $P$ of $(f(X), Y) = (Z, Y)$. Therefore, we will sometimes refer to testing the calibration of the distribution $P$, and the probabilistic classifier will be implicit.

**Expected Calibration Error.** The $\ell_p$-ECE (Expected Calibration Error) for the distribution $P$, also known as the mean calibration error (e.g., Harrell, 2015, p. 105), is

$$\ell_p\text{-ECE}(f) = \ell_p\text{-ECE}_P(f) = \mathbb{E}_{Z \sim P_Z}\left[\|\mathrm{reg}_f(Z) - Z\|_p^p\right]^{\frac{1}{p}} = \mathbb{E}_{Z \sim P_Z}\left[\sum_{k=1}^{K} |[\mathrm{res}_f(Z)]_k|^p\right]^{\frac{1}{p}}. \tag{1}$$

1. In the binary case ($K = 2$), we identify $\Delta_{K-1}$ with $[0, 1]$ via the map $(z, 1-z)^\top \mapsto z$ and use $Y = \{0, 1\}$ instead of the one-hot encoded output space. We say $f$ is perfectly calibrated if $\mathrm{reg}_f(z) := P(Y = 1 \mid f(X) = z) = z$ almost everywhere.

In words, this quantity measures the average over all classes $k = 1, \dots, K$ and over the data distribution $X \sim P_X$ of the per-class error $[\mathrm{res}_f(\mathbf{z})]_k = \mathsf{E}[[Y]_k \mid f(X) = \mathbf{z}] - [\mathbf{z}]_k$ between the predicted probability of class $k$ for input $X$—i.e., $[\mathbf{z}]_k = [f(X)]_k$—and the actual probability $\mathsf{E}[[Y]_k \mid f(X) = \mathbf{z}] = P[[Y]_k = 1 \mid f(X) = \mathbf{z}]$ of that class. For instance, when the number of classes is $K = 2$, and the power is $p = 1$, we have $\ell_1\text{-ECE}(f) = \mathsf{E}_{X \sim P_X} \sum_{k=1}^{2} |P[[Y]_k = 1 \mid f(X)] - [f(X)]_k| = 2\mathsf{E}_{X \sim P_X} |P[[Y]_1 = 1 \mid f(X)] - [f(X)]_1|$.

**Hölder Continuity.** We describe the notion of Hölder continuity for functions defined on $\Delta_{K-1}$. For simplicity, we only provide the definition for $K = 2$. See Appendix B.1 for the complete definition for general $K \geq 2$.

Identifying $\Delta_1$ with $[0, 1]$ via the map $(z, 1-z)^\top \mapsto z$, a function $g : \Delta_1 \to \mathsf{R}$ can be equivalently understood as a function $g : [0, 1] \to \mathsf{R}$. For an integer $d \geq 0$ and a function $g : [0, 1] \to \mathsf{R}$, let $g^{(d)}$ be the $d$-th derivative of the function $g$. For a real number $s$, we denote the smallest integer greater than or equal to $s$ by $\lceil s \rceil$.

For a Hölder smoothness parameter $s > 0$ and a Hölder constant $L > 0$, let $H_K(s, L)$ be the class of $(s, L)$-Hölder continuous functions $g : [0, 1] \to \mathsf{R}$ satisfying, for all $x_1, x_2 \in [0, 1]$

$$|g^{(\lceil s \rceil - 1)}(x_1) - g^{(\lceil s \rceil - 1)}(x_2)| \leq L|x_1 - x_2|^{s - \lceil s \rceil + 1}. \tag{2}$$

In particular, $H_K(1, L)$ denotes all $L$-Lipschitz functions. We consider $L > 0$ as an arbitrary fixed constant, and we do not display the dependence of our results on its value. For instance, when the Lipschitz constant is $L = 1$, and the Hölder smoothness parameter is $s = 1.5$, this is the set of real-valued functions $g$ defined on $[0, 1]$ such that for all $x_1, x_2 \in [0, 1]$, $|g'(x_1) - g'(x_2)| \leq L|x_1 - x_2|^{0.5}$.

**Goal.** Our goal is to test the null hypothesis of perfect calibration, i.e., $\mathrm{res}_f = 0$, against the alternative hypothesis that the model is mis-calibrated. To quantify mis-calibration, we use the notion of the $\ell_p$-ECE$(f)$ from (1). We study the signal strength needed so that reliable mis-calibration detection is possible. Further, we assume that the mis-calibration curves are Hölder continuous because we will show that by only assuming continuity, reliable detection of mis-calibration is impossible. In Remark 11, We will also discuss what happens when the mis-calibration function is not Hölder smooth.

Let $\mathcal{P}$ be the family of all distributions $P$ over $(Z, Y) \in \Delta_{K-1} \times Y$ such that the marginal distribution $P_Z$ of $Z$ has a density with respect to the uniform measure on $\Delta_{K-1}$. Define the collection $\mathcal{P}_0$ of joint distributions $P$ of $(Z, Y)$ under which the probability predictor $f$ is perfectly calibrated:

$$\mathcal{P}_0 := \{P \in \mathcal{P} : \mathrm{res}_{f;P}(Z) = 0, \ P_Z\text{-a.s.}\}.$$

For a Hölder smoothness parameter $s$ and a Hölder constant $L$, let $\mathcal{P}_{s;L;K}$ be the family of probability distributions $P \in \mathcal{P}$ over the predictions and labels $(f(X), Y) = (Z, Y) \in \Delta_{K-1} \times Y$ under which the residual map $\mathbf{z} \mapsto [\mathrm{res}_{f;P}(\mathbf{z})]_k$ (i.e., the map $\mathrm{res}_f$ under the distribution $(Z, Y) \sim P$) belongs to the class of $(s, L)$-Hölder continuous functions $H_K(s, L)$ for every $k \in \{1, \dots, K\}$. For a separation rate $\varepsilon > 0$, define the collection $\mathcal{P}_1(\varepsilon, p, s)$ of joint distributions $P \in \mathcal{P}_{s;L;K}$ under which the $\ell_p$-ECE of $f$ is at least $\varepsilon$:

$$\mathcal{P}_1(\varepsilon, p, s) := \{P \in \mathcal{P}_{s;L;K} : \ell_p\text{-ECE}_P(f) \geq \varepsilon\}. \tag{3}$$

We will also refer to these distributions as *$\varepsilon$-mis-calibrated*. Our goal is to test the *null hypothesis* of calibration against the *alternative* of an $\varepsilon$-calibration error:

$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_1(\varepsilon, p, s). \tag{4}$$

Although we consider the null hypothesis of perfect calibration, we generally do not expect a model trained on finite data to be perfectly calibrated. In this regard, the purpose of testing (4) is to check if there is statistically significant evidence of mis-calibration, and *not* to check whether the predictor $f$ is perfectly calibrated. As usual in hypothesis testing, not rejecting the null hypothesis does not mean that we accept that $f$ is perfectly calibrated but means that there is no statistically significant evidence of mis-calibration. In this case, to gain more confidence that the model is calibrated, one may consider testing other hypotheses about calibration—such as top-$k$ calibration, (Guo et al., 2017)—or collecting more data; of course, this may require dealing with multiple testing problems. Meanwhile, since the null of calibration is not rejected, one may use the classifier as if it was calibrated until evidence to the contrary is presented.

Moreover, in Remark 4, we also provide results for the null hypothesis of a small enough calibration error.

**Hypothesis Testing.** We recall some notions from hypothesis testing (e.g., Lehmann and Romano, 2005; Ingster, 2012, etc) that we use to formulate our problem. A *test* $\xi$ is a function[2] $\xi : (\Delta_{K-1} \times Y)^n \to \{0, 1\}$ of the data, given a dataset $S = \{(X_i, Y_i)\}_{i=1}^n \in (\Delta_{K-1} \times Y)^n$, the decision $\xi(S)$ of rejecting the null hypothesis. In other words, for a given dataset $S$, $\xi(S) = 1$ means that we detect mis-calibration, and $\xi(S) = 0$ means that we do not detect mis-calibration.

Denote the set of all *level* $\alpha \in (0, 1)$ tests, which have a *false detection rate (or, false positive rate; type I error)* bounded by $\alpha$, as

$$\Phi_n(\alpha) := \{\xi : \sup_{P \in \mathcal{P}_0} P(\xi = 1) \leq \alpha\}.$$

The probability $P(\xi = 1)$ is taken with respect to the distribution of the sample. For $\varepsilon > 0$ and $P \in \mathcal{P}_1(\varepsilon, p, s)$ from (3), we want to minimize the *false negative rate (type II error)* $P(\xi = 0)$, the probability of not detecting mis-calibration. We consider the worst possible value (maximum or rather supremum) $\sup_{P \in \mathcal{P}_1(\varepsilon;p;s)} P(\xi = 0)$ of the type II error, over all distributions $P \in \mathcal{P}_1(\varepsilon, p, s)$. We then want to minimize this over all tests $\xi \in \Phi_n(\alpha)$ that appropriately control the level, leading to the *minimax risk* (minimax type II error)

$$R_n(\varepsilon, p, s) := \inf_{\xi \in \Phi_n(\alpha)} \sup_{P \in \mathcal{P}_1(\varepsilon;p;s)} P(\xi = 0).$$

In words, among all tests that have a false detection rate of $\alpha < 1$ using a sample of size $n$, we want to find the one with the best possible (smallest) mis-detection rate over all $\varepsilon$-mis-calibrated distributions.

We consider $\alpha \in (0, 1)$ as a fixed constant, and we do not display the dependence of our results on its value. We want to understand how large the $\ell_p$-ECE (as measured by

2. To be rigorous, a Borel measurable function.

$\varepsilon$ in $P_1(\varepsilon, p, s)$) needs to be to ensure reliable detection of mis-calibration. This amounts to finding $\varepsilon'$ such that the best possible worst-case risk $R_n(\varepsilon', p, s)$ is small. For a fixed $\beta \in (0, 1 - \alpha)$, the minimum separation (signal strength) for $s$-Hölder functions, in the $\ell_p$-norm, needed for a minimax type II error of at most $\beta$ is defined as

$$\varepsilon_n(\beta; p, s) := \varepsilon_n(p, s) = \inf\{\varepsilon' : R_n(\varepsilon', p, s) \leq \beta\}.$$

Since $\beta \in (0, 1 - \alpha)$ is fixed, we usually omit the dependence of $\varepsilon_n$ on this value.

**Remark 1 (Comparison with classical nonparametric hypothesis testing)** *As we summarize in Section 1.1, prior works such as Ingster (1987, 2000, 2012); Berman et al. (2014) have studied the problem of testing that the $L^p$ norm of a function is zero against the alternative that it is nonzero, where the function is either a probability density or a regression function in the Gaussian white noise model. Our task here is different from the classical problem since $\mathrm{reg}_f$ is not a probability density, and we are not provided independent observations of the function $\mathrm{reg}_f$ or $\mathrm{res}_f$ in the Gaussian white noise model. Rather, our observation model is closer to multinomial regression; which is heteroskedastic and differs from the above models. While our proposed test shares ideas with the chi-squared test of Ingster (1987, 2000), it requires a different analysis for the above-mentioned reasons.*

## 3 An Adaptive Debiased Calibration Test

Here we describe our main test for calibration. This relies on a debiased plug-in estimator for $\ell_2$-ECE$(f)^2$. We prove that the test is minimax optimal and discuss why debiasing is necessary. We also provide an adaptive plug-in test, which can adapt to an unknown Hölder smoothness parameter $s$.

### 3.1 Debiased Plug-in Estimator

The calibration error of a continuous probability predictor $f$ is often estimated by a discretized plug-in estimator associated with a partition (or binning) of the probability simplex $\Delta_{K-1}$ (e.g., Cox, 1958; Harrell, 2015). The early work of Cox (1958) already recommended grouping together similar probability forecasts. More recently, Guo et al. (2017) divide the interval $[0, 1]$ into bins of equal width and compute the (top-1) ECE by averaging the difference between confidence and accuracy in each bin. Vaicenavicius et al. (2019) generalize this idea to $K$-class classification and data-dependent partitions.

In this work, we use an equal-volume partition $B_m$ of the probability simplex $\Delta_{K-1}$, which is parametrized by a binning scheme parameter $m \in \mathbb{N}_+$. The partition $B_m$ consists of $m^{K-1}$ simplices with equal volumes and diameters proportional to $m^{-1}$. To construct a such partition, we first divide the simplex $\Delta_{K-1}$ into $K - 1$ hypersimplices—generalizations of the standard probability simplex that can have more vertices and edges—by taking intersections with $m^{-1}$-scaled and translated $K$-dimensional hypercubes. The hypersimplices are further divided into unit volume simplices using the result of Stanley (1977); Sturmfels (1996). The construction of $B_m$ is elaborated in Appendix B.3. The purpose of using an equal-volume partition $B_m$ is only for a simpler description of our results, and any partition with $\Theta(m^{-K+1})$ volumes and $\Theta(m^{-1})$ diameters can be used.

Let us denote the sets comprising the partition as $\mathcal{B}_m = \{B_1, \ldots, B_{m^{K-1}}\}$. For each $i \in \{1, \ldots, m^{K-1}\}$, define the indices of data points falling into the bin $B_i$ as $I_{m,i} := \{j : Z_j \in B_i, 1 \le j \le n\}$. Then, for each $i \in \{1, \ldots, m^{K-1}\}$, the averaged difference between probability predictions $Z_j = f(X_j)$ and true labels $Y_j$ for the probability predictions in $B_i$ is $|I_{m,i}|^{-1} \sum_{j \in I_{m,i}} (Y_j - Z_j)$. This estimates $\mathbb{E}[Y - Z \mid Z \in B_i] = \mathbb{E}[\mathrm{res}_f(Z) \mid Z \in B_i]$. Now, the quantity $\ell_2\text{-ECE}(f)^2 = \mathbb{E}[\|\mathrm{res}_f(Z)\|_k^2]$ can be approximated by piecewise averaging as $\sum_{1 \le i \le m^{K-1}} P_Z(B_i) \|\mathbb{E}[\mathrm{res}_f(Z) \mid Z \in B_i]\|_k^2$. Plugging in the estimate $\|I_{m,i}|^{-1} \sum_{j \in I_{m,i}} (Y_j - Z_j)\|_k^2$ of $\|\mathbb{E}[\mathrm{res}_f(Z) \mid Z \in B_i]\|_k^2$, we can define a plug-in estimator of $\ell_2\text{-ECE}(f)^2$ as follows:

$$T^b_{m,n} := \sum_{\substack{1 \le i \le m^{K-1} \\ |I_{m,i}| \ge 1}} \frac{|I_{m,i}|}{n} \left\| \frac{1}{|I_{m,i}|} \sum_{j \in I_{m,i}} (Y_j - Z_j) \right\|_k^2. \tag{5}$$

Above, the sum is taken over bins $B_i$ containing at least one datapoint. As will be discussed in Section 3.3, the plug-in estimator is biased in the sense that its expectation is not zero under perfectly calibrated distributions. Moreover, it does not lead to an optimal test statistic. Informally, this happens because we are estimating both $\mathbb{E}[Y \mid Z \in B_i]$ and $\mathbb{E}[Z \mid Z \in B_i]$ with the same sample $(Z_i, Y_i), i \in \{1, \ldots, n\}$. We hence define the Debiased Plug-in Estimator (DPE):

$$T^d_{m,n} := \sum_{\substack{1 \le i \le m^{K-1} \\ |I_{m,i}| \ge 1}} \frac{|I_{m,i}|}{n} \left[ \left\| \frac{1}{|I_{m,i}|} \sum_{j \in I_{m,i}} (Y_j - Z_j) \right\|_k^2 - \frac{1}{|I_{m,i}|^2} \sum_{j \in I_{m,i}} \|Y_j - Z_j\|_k^2 \right]. \tag{6}$$

The debiasing term in (6) ensures that $T^d_{m,n}$ has mean zero under a distribution $P \in \mathcal{P}_0$ under which $f$ is a calibrated probability predictor. Due to the discretization, the mean of $T^d_{m,n}$ is not exactly $\ell_2\text{-ECE}(f)^2$ under $P \in \mathcal{P}_1(\varepsilon; p, s)$, but the debiasing makes it comparable to $\ell_2\text{-ECE}(f)^2$. This will be a crucial step when proving the optimality of $T^d_{m,n}$.

Remark 2 (Connection to nonparametric functional estimation) The definition of $T^d_{m,n}$ is closely related to the U-statistic for estimating the quadratic integral functional of a probability density (Kerkyacharian and Picard, 1996; Laurent, 1996). To see this, let $\{\phi_i(x) = P_Z(B_i)^{-1/2} 1_{B_i}(x) : i = 1, \ldots, m^{K-1}\}$ be the Haar scaling functions associated to the partition $\mathcal{B}_m$. For each $1 \le k \le K$, the U-statistic

$$\frac{1}{n(n-1)} \sum_{1 \le i \le m^{K-1}} \sum_{1 \le j_1 \ne j_2 \le n} [Y_{j_1} - Z_{j_1}]_k [Y_{j_2} - Z_{j_2}]_k \phi_i(Z_{j_1}) \phi_i(Z_{j_2})$$

is an unbiased estimate of $\sum_{1 \le i \le m^{K-1}} P_Z(B_i)^{-1} \left( \int_{B_i} [\mathrm{res}_f(z)]_k \, dP_Z(z) \right)^2$. Summing over $1 \le k \le K$ and plugging in $P_Z(B_i) \approx |I_{m,i}|/n$, we recover (6) with the minor modification of changing $n \to n-1$ in the scaling.

However, as noted in Remark 1, our problem differs from those studied in classical nonparametric statistic literature. Specifically, our definition of $T^d_{m,n}$ additionally requires an estimation of $P_Z(B_i)$ by $|I_{m,i}|/n$. Therefore, prior results on nonparametric functional

**Algorithm 1** T-Cal: an optimal test for calibration (based on debiased plug-in estimation of the calibration error)

Input: Probability predictor $f : X \to \Delta_{K-1}$, i.i.d. sample $\{(X_i, Y_i) \in X \times Y : i \in \{1, \ldots, n\}\}$, false detection rate $\alpha \in (0, 1)$, true detection rate $\beta \in (0, 1-\alpha)$, Hölder smoothness $s$

Initialize: $m \leftarrow b\, n^{2/(4s+K-1)}c$, $T^d_{m,n} \leftarrow 0$, $Z_i \leftarrow f(X_i)$ for $1 \le i \le n$, define $\{B_1, \ldots, B_{m^{K-1}}\}$ as in Appendix B.3

for $i = 1$ to $m^{K-1}$ do

$I_{m,i} \leftarrow \{j : Z_j \in B_i, 1 \le j \le n\}$

$T^d_{m,n} \leftarrow T^d_{m,n} + \frac{|I_{m,i}|}{n} \left[ \frac{1}{|I_{m,i}|} \sum_{j \in I_{m,i}} (Y_j - Z_j) \right]^2 - \frac{1}{|I_{m,i}|^2} \sum_{j \in I_{m,i}} \|Y_j - Z_j\|^2$

end for

$\phi_{m,n} \leftarrow \mathbb{I}\left[ T^d_{m,n} > \frac{\sqrt{2}K}{\alpha} m^{\frac{K-1}{2}} n^{-1} + m^{-\frac{K-1}{2}} \right]$

Output: Reject $H_0$ if $\phi_{m,n} = 1$

estimation (Bickel and Ritov, 1988; Donoho and Nussbaum, 1990; Birgé and Massart, 1995) cannot be directly applied to $\ell_2$-ECE$(f)^2$.

Wang et al. (2008); Shen et al. (2020) consider quadratic functional estimation for an unknown distribution of covariates and show that the minimax rate also depends on the Hölder smoothness of the covariate density function.

In the following theorem, we prove that $T^d_{m,n}$ leads to a minimax optimal test when the number of bins is chosen in a specific way, namely $m \asymp n^{2/(4s+K-1)}$. Crucially, the number of bins required decreases with the smoothness parameter $s$. In this sense, our result parallels the well-known results on the optimal choice of the number of bins for testing probability distributions and densities (Mann and Wald, 1942; Ingster, 2012).

The guarantee on the power (or, Type II error control) requires the following mild condition, stated in Assumption 1. This ensures that the probability of each bin is proportional to the inverse of the number of bins up to some absolute constant. In particular, this holds if the density of the probabilities predicted is close to uniform. This assumption is necessary when extending the results of Arias-Castro et al. (2018); Kim et al. (2022) to a general base probability measure $\mu$ of the probability predictions over the probability simplex. See Appendix B.2 for more discussion.

**Assumption 1** (Bounded marginal density) Let $\mu$ be the uniform probability measure on the probability simplex $\Delta_{K-1}$. There exist constants $\ell_\mu, u_\mu > 0$ such that $\ell_\mu \le dP_Z/d\mu \le u_\mu$ almost everywhere.

**Theorem 3** (Calibration test via debiased plug-in estimation) Suppose $p \ge 2$ and assume that the Hölder smoothness parameter $s$ is known. For a binning scheme parameter $m \in \mathbb{N}_+$, let

$$\phi_{m,n}(\cdot) = \phi_{m,n} := \mathbb{I}\left[ T^d_{m,n} > \frac{\sqrt{2}K}{\alpha} m^{\frac{K-1}{2}} n^{-1} + m^{-\frac{K-1}{2}} \right].$$

Under Assumption 1 and for $m = \lfloor n^{2/(4s+K-1)} c \rfloor$, we have

1. False detection rate control. For every $P$ for which $f$ is perfectly calibrated, i.e., for $P \in P_0$, the probability of falsely claiming mis-calibration is at most $\alpha$, i.e., $P(\phi_{m,n} = 1) \leq \alpha$.

2. True detection rate control. There exists $c > 0$ depending only on $(s, L, K, \beta_l, \beta_u, \alpha, \gamma)$ such that when

$$\varepsilon \geq c n^{-2s/(4s+K-1)},$$

then for every $P \in P_1(\varepsilon, p, s)$|i.e., when $f$ is mis-calibrated with an $\ell_p$-ECE of $\varepsilon$|the power (true positive rate) is bounded as $P(\phi_{m,n} = 1) \geq 1 - \beta$.

The proof can be found in Appendix A.1. The proof follows the classical structure of upper bound arguments in nonparametric hypothesis testing, see e.g., Arias-Castro et al. (2018); Kim et al. (2022) for recent examples. We compute the mean and variance of $T^d_{m,n}$ under null distributions $P_0 \in P_0$ and alternative distributions $P_1 \in P_{s,L,K}$ with a large ECE. Using Lemma 13, we can find a lower bound on $E_{P_1}[T^d_{m,n}] - E_{P_0}[T^d_{m,n}]$. The variances $\mathrm{Var}_{P_0}(T^d_{m,n})$ and $\mathrm{Var}_{P_1}(T^d_{m,n})$ can be also upper bounded. We argue that the mean difference $E_{P_1}[T^d_{m,n}] - E_{P_0}[T^d_{m,n}]$ is significantly larger than the square root of the variances $\mathrm{Var}_{P_0}(T^d_{m,n})$ and $\mathrm{Var}_{P_1}(T^d_{m,n})$. The conclusion follows from Chebyshev's inequality.

Combined with our lower bound in Theorem 10, this result shows the desired property that our test is minimax optimal. This holds for all $p \geq 2$, so that the test is minimax optimal even when the mis-calibration is measured in the $\ell_p$ norm with $p < 2$. This is consistent with experimental findings such as those of Nixon et al. (2019), where the empirical $\ell_2$-ECE performs better than the empirical $\ell_1$-ECE as a measure of calibration error. Also see Section 4.1 for a comparison of the empirical $\ell_1$-ECE and $\ell_2$-ECE as a test statistic.

Although we present explicit critical values in Theorem 3, they can be conservative in practice, as in other works in nonparametric testing (Ingster, 1987; Arias-Castro et al., 2018; Kim et al., 2022). Therefore, we recommend choosing the critical values via a version of bootstrap: consistency resampling (Brocker and Smith, 2007; Vaicenavicius et al., 2019). See Appendix C.3 for further details on choosing critical values.

Remark 4 So far, we considered the null hypothesis of perfect calibration. However, since the predictor $f$ is trained on a finite dataset, we cannot expect it to be perfectly calibrated. We can extend Theorem 3 to the null hypothesis of "small enough" mis-calibration, namely, for any given constant $c_0 > 0$, an $\ell_p$-ECE of at most $c_0 n^{-2s/(4s+K-1)}$. Then, the true and false positive rates of the test

$$\phi^{\mathrm{comp}}_{m,n} := \mathbb{I}\left\{ T^d_{m,n} > \sqrt{2\, m^{K-1} n^{-2} \wedge m^{(K-1)} + 5 c_0^2 n^{\frac{4s}{4s+K-1}} n^{-1} \wedge m^{(K-1)}} \right\}$$

can be controlled as in Theorem 3. See Appendix A.3 for the proof.

### 3.2 An Adaptive Test

The binning scheme used in our plug-in test requires knowing the smoothness parameter $s$ to be minimax optimal. However, in practice, this parameter is usually unknown. Can we design an adaptive test that does not require knowing this parameter? Here we answer this question in the affirmative. As in prior works in nonparametric hypothesis testing, e.g., Ingster (2000); Arias-Castro et al. (2018); Kim et al. (2022), we propose an adaptive test that can adapt to an unknown Hölder smoothness parameters. The idea is to evaluate the plug-in test over a variety of partitions, and thus be able to detect mis-calibration at various different scales.

In more detail, we evaluate the test with a number of bins ranging over a dyadic grid $2, 2^2, \ldots, 2^B$. In addition, to make sure that we control the false detection rate, we need to divide the level $\alpha$ by the number of tests performed. Thus, for a number $B = \lceil \frac{2}{K-1} \log_2(n/\sqrt{\log n}) \rceil$ of tests performed, we let the adaptive test

$$\mathrm{ad}_n := \max_{1 \le b \le B} \mathrm{ad}_{2^b;n}^{\alpha/B} \tag{7}$$

detect mis-calibration if any of the debiased plug-in tests $\mathrm{ad}_{2^b;n}^{\alpha/B}$, with the number of bins $2^b$, $b \in \{1, \ldots, B\}$, detects mis-calibration at level $\alpha/B$. We summarize the procedure in Algorithm 2.

**Theorem 5 (Adaptive plug-in test)** Suppose $p \ge 2$. Under Assumption 1, the adaptive test from (7) enjoys

1. **False detection rate control.** For every $P$ for which $f$ is perfectly calibrated, i.e., for $P \in \mathcal{P}_0$, the probability of falsely claiming mis-calibration is at most $\alpha$, i.e., $P(\mathrm{ad}_n = 1) \le \alpha$.
2. **True detection rate control.** There exists $c_{\mathrm{ad}} > 0$ depending on $(s, L, K, \alpha_l, \alpha_u, \alpha, \beta)$ such that the power (true positive rate) is lower bounded as $P(\mathrm{ad}_n = 1) \ge 1 - \beta$ for every $P \in \mathcal{P}_1(\varepsilon, p, s)$ i.e., when $f$ is mis-calibrated with an $\ell_p$-ECE of at least $\varepsilon \ge c_{\mathrm{ad}}(n/\sqrt{\log n})^{-2s/(4s+K-1)}$.

See Appendix A.4 for the proof. Compared to the non-adaptive test, this test requires a mild additional factor of $(\log n)^{s/(4s+K-1)}$ in the separation rate $\varepsilon$ to guarantee detection. It is well understood in the area of nonparametric hypothesis testing that some adaptation cost is unavoidable, see for instance Spokoiny (1996); Ingster (2000). For more discussion, see Remark 14.

**Remark 6** We remark that the false detection rate control of Theorem 3 and 5 does not require a Hölder smoothness assumption.

### 3.3 Necessity of Debiasing

Recall from (5) that $T^b_{m;n}$ is the plug-in estimator of $\ell_2\text{-ECE}(f)^2$ without the debiasing term in (6). We argue that this biased estimator is not an optimal test statistic, even for $m = m' = \lfloor n^{2/(4s+K-1)} \rfloor$ from Theorem 3 (which is optimal for the debiased test), by presenting a failure case in the following example.

**Algorithm 2** Adaptive T-Cal: an adaptive test for calibration

Input: Probability predictor $f : X \to \Delta_{K-1}$; i.i.d. sample $\{(X_i, Y_i) \in X \times Y : i \in \{1, \dots, n\}\}$; false detection rate $\alpha \in (0, 1)$; true detection rate $\beta \in (0, 1-\alpha)$
Initialize: $B \leftarrow \lceil \frac{2}{K-1} \log_2(n/\sqrt{\log n}) \rceil$; $Z_i \leftarrow f(X_i)$ for $1 \le i \le n$; $\phi^{ad}_n \leftarrow 0$
for $b = 1$ to $B$ do
  Compute $\phi_{2^b;n}(\frac{\alpha}{B})$ as in Algorithm 1
  if $\phi_{2^b;n}(\frac{\alpha}{B}) = 1$ then
    $\phi^{ad}_n \leftarrow 1$
    break
  end if
end for
Output: Reject $H_0$ if $\phi^{ad}_n = 1$

Example 1 (Failure of naive plug-in) Consider binary classification problem with $K = 2$, $m = \lfloor n^{2/(4s+1)} \rfloor$ (assumed to be divisible by four), and the partition

$$B_m = \{B_1, \dots, B_m\} = \left\{0, \frac{1}{m}\right), \dots, \left[\frac{m-1}{m}, 1\right].$$

Let $P_0$ be the distribution over $(Z, Y) \in [0,1] \times \{0,1\}$ given by $Z \sim P_0$ Unif$([0,1])$ and $Y \mid Z = z \sim P_0$ Ber$(z)$ for all $z \in [0,1]$. Under $P_0$, the probability predictor $f$ is perfectly calibrated, i.e., $P_0 \in \mathcal{P}_0$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be the function defined by

$$\psi(x) := e^{-\frac{1}{x(1-x)}} 1_{(0,1)}(x). \tag{8}$$

Let $g : [0,1] \to [0,1]$ be the function (corresponding to the calibration curve of the probability predictor $f$)

$$g(z) := z - m^{-s} \sum_{j=0}^{\frac{m}{4}-1} \psi\left(m\left(z - \frac{j}{m}\right)\right) + m^{-s} \sum_{j=\frac{m}{4}}^{\frac{m}{2}-1} \psi\left(m\left(z - \frac{j}{m}\right)\right) \tag{9}$$

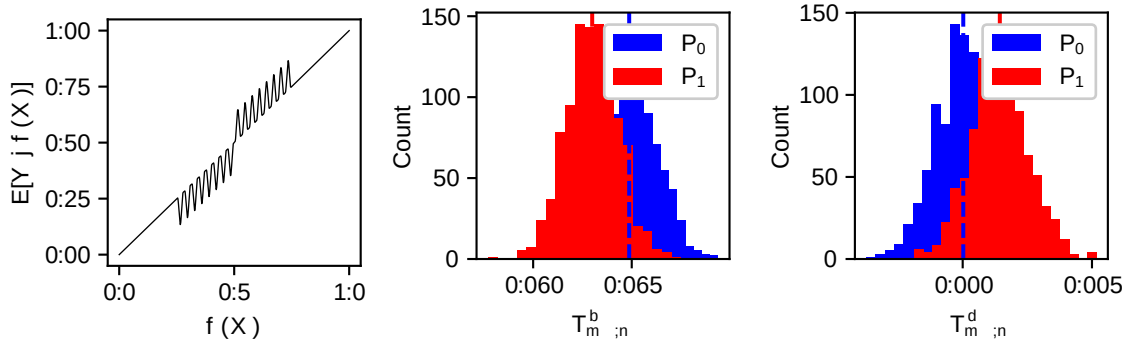
for $s \in (\frac{1}{4}, \frac{1}{2})$, $\varepsilon > 0$, and $\psi$ defined in (8). Define the distribution $P_1$ over $(Z, Y)$ by

$$Z \sim P_1 \text{ Unif}([0,1]) \text{ and } Y \mid Z = z \sim P_1 \text{ Ber}(g(z))$$

for all $z \in [0,1]$. As we will show in the proof of Theorem 10, (1) the mis-calibration curve $g(z) - z = \text{res}_f(z)$ is $s$-Hölder and (2) $\ell_2\text{-ECE}_{P_1}(f) = \Theta(n^{-2s/(4s+1)})$.

As can be seen in Figure 2a, the probability predictor $f$ under $P_1$ is an example of a mis-calibrated predictor, as $f(X)$ is smaller than $E[Y \mid f(X)]$ when $f(X)$ is above 0.5; and vice versa. However, the mean of $T^b_{m,n}$ under the mis-calibrated distribution $P_1$ is surprisingly smaller than the mean under the calibrated distribution $P_0$ when $n$ is large enough (Proposition 7).

That is, the statistic $T^b_{m,n}$ does not capture the amount of mis-calibration, and therefore the calibration test based on it will not perform well. Figure 2b confirms this finding, and Figure 2c displays that this effect can be removed by using the debiased statistic $T^d_{m,n}$.



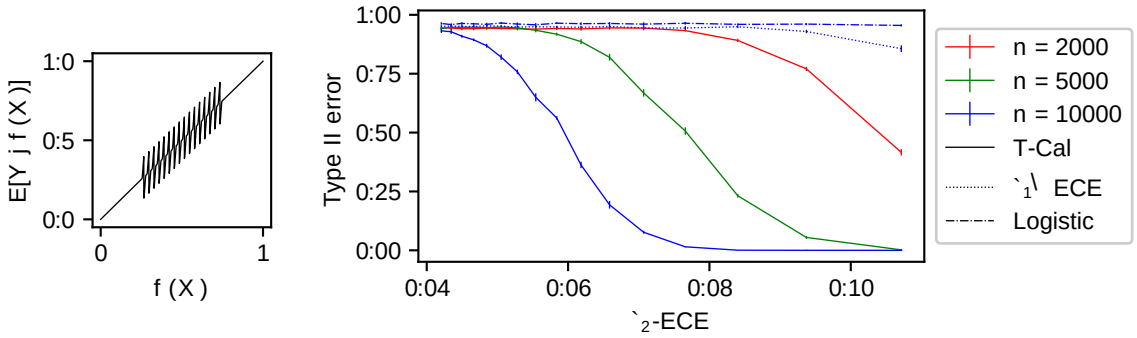

(a) Calibration curve under $P_1$ (b) Histograms of $T^b_{m;n}$ (c) Histograms of $T^d_{m;n}$

Figure 2: (a) A graph of the calibration curve $z \mapsto g(z) = E_{P_1}[Y \mid f(X) = z]$ defined in (9). When the true label probability is above/below 0.5, the model outputs a smaller/larger score. Hence $f$ is a mis-calibrated probability predictor under $P_1$. (b) Histograms of $T^b_{m;n}$ and $T^d_{m;n}$ under $P_0$ and $P_1$ are obtained from 1,000 independent observations. We use the parameters $n = 10{,}000$, $s = 0.3$, and $\alpha = 100$. The dashed line indicates the empirical mean of each distribution. Note that the biased estimator $T^b_{m;n}$ has a smaller mean under $P_1$, which aligns with Proposition 7. (c) We see this effect disappears after debiasing and that the mean of $T^d_{m;n}$ becomes zero.

Proposition 7 (Failure of naive plug-in test) Let $P_0$ and $P_1$ be the distributions defined in Example 1, and $m = \lfloor n^{2/(4s+1)} \rfloor$. Then $E_{P_0}[T^b_{m;n}] < E_{P_1}[T^b_{m;n}]$ for all large enough $n \in \mathbb{N}_+$.

See Appendix A.5 for the proof. We remark that it is possible to avoid the phenomenon in Proposition 7, by choosing a different $m$. Proposition 7 aims only to highlight that the effect of the bias in $T^b_{m;n}$ can be extreme in certain cases, and we do not claim that $m = m^*$ is also the optimal choice for the biased statistic.

We finally comment on the related results of Brocker (2012); Ferro and Fricker (2012); Kumar et al. (2019). In Brocker (2012); Ferro and Fricker (2012), the plug-in estimator of the squared $\ell_2$-ECE is decomposed into terms related to reliability and resolution. Based on this observation, Kumar et al. (2019) propose a debiased estimator for the squared $\ell_2$-ECE and show an improved sample complexity for estimation. However, their analysis is restricted to the binary classification case and probability predictors with only finitely many output values. It is not clear how to adapt their method to predictors with continuous outputs, because this would require discretizing the outputs. Our debiased plug-in estimator $T^d_{m;n}$ is more general, as it can be used for multi-class problems and continuous probability predictors $f$. Also, our reason to introduce $T^d_{m;n}$ (testing) differs from theirs (estimation).




(a) Calibration curve under $P_{1;m}$

(b) Comparison of calibration tests

Figure 3: (a) A graph of the calibration curve $z \mapsto g_m(z) = \mathbb{E}_{P_{1;m}}[Y \mid f(X) = z]$ defined in (10). The mis-calibration curve alternates between negative and positive values, making detection challenging. (b) We compare our test $_{m;n}$ with other commonly used calibration tests. Since our test optimally adjusts the number of bins $m = \lfloor n^{2/(4s+K-1)} \rfloor$ according to the sample size $n$, it can detect mis-calibration over smaller and smaller intervals as $n$ grows. On the other hand, the plug-in test $_{m;n}$, with a fixed-in-$n$ binning scheme parameter $m$, fails to detect mis-calibration over intervals smaller than the bin width. This issue remains when the sample size $n$ increases. The test based on the calibration slope and intercept also suffers from the same issue. Standard error bars are plotted over 10 repetitions.

## 4 Experiments

We perform experiments on both synthetic and empirical datasets to support our theoretical results. These experiments suggest that T-Cal is in general superior to state-of-the-art methods.

### 4.1 Synthetic Data: Power Analysis

Let $P_0 \in \mathcal{P}_0$ be the distribution defined in Example 1—a distribution under which $f$ is perfectly calibrated. For $m \in \mathbb{N}_+$, $s > 0$, $\cdot > 0$, and $\cdot : \mathbb{R} \to \mathbb{R}$ from (8), define $g_m : [0;1] \to [0;1]$ by

$$g_m(z) := z + \cdot m^{-s} \sum_{j=0}^{m-1} (-1)^j \cdot 2mz - \frac{m}{2} - j \cdot : \tag{10}$$

This function oscillates strongly, as shown in Figure 3a. Let $P_{1;m}$ be the distribution over $(Z;Y) \in [0;1] \times \{0;1\}$ given by $Z \sim P_{1;m}$ Unif$([0;1])$ and $Y \mid Z = z \sim P_{1;m}$ Ber$(g_m(z))$ for all $z \in [0;1]$. Under $P_{1;m}$, the probability predictor $f$ is mis-calibrated with an $\ell_p$-ECE of at least $\varepsilon = \cdot \cdot \|\cdot\|_{L^p} m^{-s}$. However, since the mis-calibration curve $g_m$ oscillates strongly, mis-calibration can be challenging to detect.

We study the type II error of tests against the alternative where the mis-calibration is specified as $H_1 : (Z;Y) \sim P_{1;m}$. This gives a lower bound on the worst-case type II error

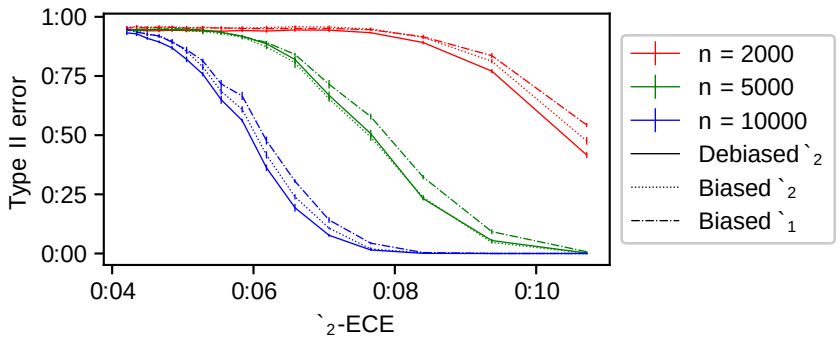



Figure 4: Type II error comparison for $T^{d}_{m\ ;n}$ (T-Cal), $T^{b}_{m\ ;n}$, and $T^{`1}_{m\ ;n}$. Using $`_2$ is better than $`_1$, and debiased$`_2$ (T-Cal) is better than biased $`_2$. Standard error bars are plotted over 10 repetitions.

over the alternative hypothesis $P_1(";p;s)$. We repeat the experiment for di erent values of m to obtain a plot of $`_2$-ECE versus type II error.

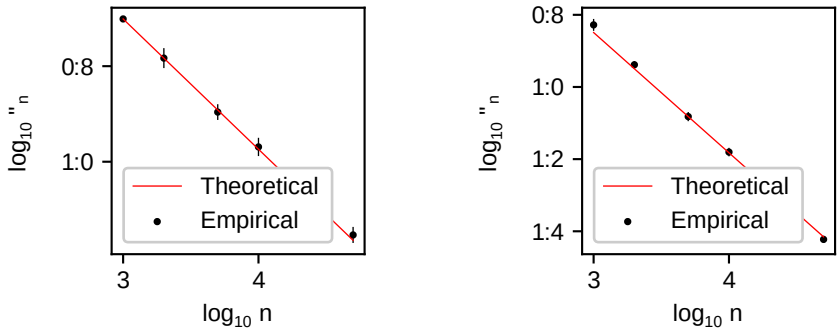
Comparison of Tests. We compare the test $_{m\ ;n}$ with classical calibration tests dating back to Cox (1958), and discussed in Harrell (2015); Vaicenavicius et al. (2019). Harrell (2015) re ts a logistic model

$$P(Y = 1 \mid Z) = \frac{1}{1 + \exp\left( {}_0 + {}_1 \log \frac{Z}{1 - Z} \right)}$$

on the sample$f(Z_i;Y_i) : i \in f1;:::;ngg$ and tests the null hypothesis of $_0 = 0$ and $_1 = 1$. Speci cally, we perform the score test (Rao, 1948; Silvey, 1959), with the test statistic derived from the gradient of log-likelihood with respect to the tested parameters. There are several approaches to set the critical values, including by using the asymptotic distribution theory of sampling statistics under the null hypothesis, or by data reuse methods such as the bootstrap. We estimate the critical values via 1000 Monte Carlo simulations.

Vaicenavicius et al. (2019) use$`\hat{1}$-ECE, the plug-in estimator for $`_1$-ECE, as their test statistic. They approximate the distribution of $`\hat{1}$-ECE by a bootstrapping procedure called consistency resampling (in which both the probability predictions and the labels are resampled) and compute a p-value based on this approximation. This test also uses a plug-in estimator as the test statistic but di ers from T-Cal as it is neither debiased nor adaptive. Since the data-generating distribution is known in this synthetic experiment, we set the critical value via 1000 Monte Carlo simulations.

We control the false detection rate at a level $= 0:05$ and run experiments for $n = 2;000;5;000$, and $10;000$. We nd that our proposed test achieves the lowest type II error, see Figure 3b. We also nd that other tests do not leverage the growing sample size. For this reason, we only display $n = 10;000$ for the other two tests. As can be seen in Figure 3b, T-Cal outperforms other testing methods in true detection rate by a large margin.




(a) $s = 0:3, \quad = 25$ (b) $s = 0:5, \quad = 50$

Figure 5: The dots are $\log \varepsilon_n$ computed for different sample sizes $n$. The red line has a slope $\frac{2s}{4s+1}$. Standard error bars are plotted over 10 repetitions. See the text for more details.

$\ell_1$-ECE versus $\ell_2$-ECE. To confirm the effectiveness of using the $\ell_2$-ECE estimator $T^{d}_{m,n}$, we compare it with a plug-in $\ell_1$-ECE estimator defined as

$$T^{\ell_1}_{m;n} := \sum_{\substack{1 \le i \le m^{K-1} \\ |I_{m;i}| \ge 1}} \frac{|I_{m;i}|}{n} \left| \frac{1}{|I_{m;i}|} \sum_{j \in I_{m;i}} (Y_j - Z_j) \right| :$$

At the moment, it is unknown how to debias this estimator. We use the optimal binning parameter $m = m^* = \lfloor n^{2/(4s+K-1)} \rfloor$ for both $\ell_1$ and $\ell_2$ estimators; because it is unknown what the $\ell_1$-optimal binning scheme is. Also, to isolate the effect of debiasing, we compare the biased $\ell_2$ estimator $T^{b}_{m,n}$ as well, with the same number of bins. In Figure 4, we see the $\ell_2$ estimators consistently outperform the $\ell_1$ estimator, regardless of debiasing. While it is a common practice to use a plug-in estimator of $\ell_1$-ECE, our result suggests T-Cal compares favorably to it.

Minimum Detection Rate. We perform an experiment to support the result on the minimum detection rate of T-Cal, presented in Theorem 3. For each $n$, we find the largest integer, denoted $m(n)$, such that the type II error against $H_1 : (Z;Y) \sim P_{1;m(n)}$ is less than 0:05. We compute $\varepsilon_n := \ell_2\text{-ECE}_{P_{1;m(n)}}(f) = \|f\|_{L^2} m(n)^{-s}$ (a lower bound on the minimum detection rate) and plot $\log \varepsilon_n$ versus $\log n$ in Figure 5. We see that the logarithm decreases as $n$ grows, with the slope $\frac{2s}{4s+1}$ predicted by Theorem 3.

### 4.2 Results on Empirical Datasets

To verify the performance of adaptive T-Cal empirically, we apply it to the probability predictions output by deep neural networks trained on several datasets. Since our goal is to test calibration, we calculate the probabilities predicted by pre-trained models on the test sets. As in (Guo et al., 2017; Kumar et al., 2019; Nixon et al., 2019, etc), we binarize the test labels by taking the top-1 confidence as the new probability prediction,

| | DenseNet 121 | | ResNet 50 | | VGG-19 | |
|---|---|---|---|---|---|---|
| | $\ell_1$-ECE | Calibrated? | $\ell_1$-ECE | Calibrated? | $\ell_1$-ECE | Calibrated? |
| No Calibration | 2.02% | reject | 2.23% | reject | 2.13% | reject |
| Platt Scaling | 2.32% | reject | 1.78% | reject | 1.71% | reject |
| Poly. Scaling | 1.71% | reject | 1.29% | reject | 0.90% | accept |
| Isot. Regression | 1.16% | reject | 0.62% | reject | 1.13% | accept |
| Hist. Binning | 0.97% | reject | 1.12% | reject | 1.28% | reject |
| Scal. Binning | 1.94% | reject | 1.21% | reject | 1.67% | reject |

Table 1: The values of the empirical $\ell_1$-ECE (Guo et al., 2017) and the testing results, via adaptive T-Cal and multiple binomial testing, of models trained on CIFAR-10.

and the labels as the results of the top-1 classification, i.e., $Z = \max_{1 \le k \le K} [Z]_k$ and $Y = \mathbb{I}$(correctly classified by the top-1 prediction). This changes the problem of detecting the full-class mis-calibration to testing the mis-calibration of a binary classifier. Hence, we choose $K = 2$ for adaptive T-Cal in the experiments below. We refer readers to (Gupta and Ramdas, 2022) for more details about binarization via the top-1 prediction.

CIFAR-10. For the CIFAR-10 dataset, the models are DenseNet 121, ResNet 50, and VGG-19. We first apply the adaptive test directly to the $10,000$ uncalibrated probability predictions output by each model, with the false detection rate controlled at the level $\alpha = 0.05$.

For every choice of the number of bins $m$, we estimate the critical value by taking the upper 5% quantile of the values of the test statistic over $3000$ bootstrap re-samples of the probability predictions. The labels are also chosen randomly, following Bernoulli distributions with the probability prediction as the success probability. We also provide the values of the standard empirical $\ell_1$-ECE calculated with Guo et al. (2017)'s approach for the reader's reference, and with 15 equal-width bins.

We then test the probability predictions of these three models calibrated by several post-calibration methods: Platt scaling (Platt, 1999), polynomial scaling, isotonic regression (Zadrozny and Elkan, 2002), histogram binning (Zadrozny and Elkan, 2001), and scaling-binning (Kumar et al., 2019). To this end, we split the original dataset of $10,000$ images into 2 sets of sizes $2,000$ and $8,000$. The first set is used to calibrate the model, and the second is used to perform adaptive T-Cal and calculate the empirical $\ell_1$-ECE. In polynomial scaling, we use polynomials of order 3 to do regression on all the prediction-label pairs $(Z_i, Y_i)$, and truncate the calibrated prediction values into the interval $[0, 1]$. We set the binning scheme in both histogram binning and scaling binning as 15 equal-mass bins. Our implementation is adapted from Kumar et al. (2019).

Since the recalibrated probability predictions output by the latter two methods belong to a finite set, we use a test based on the binomial distribution. See Appendix B.4 for details. For completeness, we also provide the debiased empirical $\ell_2$-ECE values (Kumar et al., 2019) for models calibrated by the two discrete methods, see the details in Table 4, Appendix C.1.

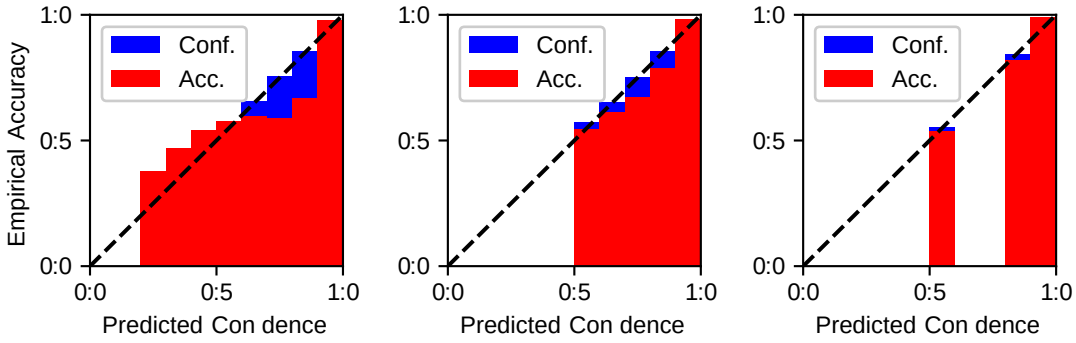



Figure 6: The reliability diagrams for VGG-19, trained on CIFAR-10, calibrated by Platt scaling (left), polynomial scaling (middle), and histogram binning (right). The bins (bars) containing less than 10 data points, where the sample noise dominates, are omitted for clarity. The dashed lines correspond to perfect calibration.

The results are listed in Table 1, where we use \accept" to denote that the test does not reject. The models with smaller empirical $\ell_1$-ECE are more likely to be accepted, by adaptive T-Cal and by multiple binomial testing, as perfectly calibrated. This can be further illustrated by the three empirical reliability diagrams given in Figure 6, where the model's predictions calibrated by Platt scaling (left) are visually more \mis-calibrated" than those calibrated by polynomial scaling (middle) and histogram binning (right).

CIFAR-100. We perform the same experimental procedure for three models pre-trained on the CIFAR-100 dataset: MobileNet-v2, ResNet 56, and Shu eNet-v2 (Chen, 2021). The test set provided by CIFAR-100 is split into two parts, containing $2,000$ and $8,000$ images, respectively. Since the regression functions $r_i$ of models trained on the larger CIFAR-100 dataset can be more complicated than those of models trained on CIFAR-10, we set the polynomial degree as $\ve$ in polynomial scaling.

The results are listed in Table 2. The values of the debiased empirical $\ell_2$-ECE (Kumar et al., 2019) for the two discrete calibration methods are provided in Table 5, Appendix C.1. The results roughly align with the magnitude of the empirical ECE value.

However, as can be observed in the column corresponding to ResNet 56, this trend is certainly not monotone. The calibrated ResNet 56 with the empirical ECE $1.84\%$ is accepted while the calibrated ResNet 56 with a smaller value $1.57\%$ is rejected. Furthermore, the test results reveal that models with relatively large (or small) empirical $\ell_1$-ECE values may not necessarily be poorly (or well) calibrated since the $\ell_1$-ECE values measured can be highly dominated by the sample noise.

ImageNet. We repeat the above experiments on models pre-trained on the ImageNet dataset. We examine three pre-trained models provided in the torchvision package in PyTorch: DenseNet 161, ResNet 152, and E cientNet-b7. We split the validation set of $50,000$

| | MobileNet-v2 | | ResNet 56 | | Shu eNet-v2 | |
|---|---|---|---|---|---|---|
| | $\ell_1$-ECE | Calibrated? | $\ell_1$-ECE | Calibrated? | $\ell_1$-ECE | Calibrated? |
| No Calibration | 11.87% | reject | 15.2% | reject | 9.08% | reject |
| Platt Scaling | 1.40% | accept | 1.84% | accept | 1.34% | accept |
| Poly. Scaling | 1.69% | reject | 1.91% | reject | 1.81% | accept |
| Isot. Regression | 1.76% | accept | 2.33% | reject | 1.38% | accept |
| Hist. Binning | 1.66% | reject | 2.44% | reject | 2.77% | reject |
| Scal. Binning | 1.85% | reject | 1.57% | reject | 1.65% | accept |

Table 2: The values of the empirical $\ell_1$-ECE (Guo et al., 2017) and the testing results, via adaptive T-Cal and multiple binomial testing, of models trained on CIFAR-100.

images into a calibration set and a test set—of sizes 10,000 and 40,000, respectively. We use polynomials of degree 5 in polynomial scaling.

The results are listed in Table 3. The values of the debiased empirical $\ell_2$-ECE (Kumar et al., 2019) are provided in Table 6, in Appendix C.1. As can be seen, the test results here generally align with the empirical ECE values.

Remark 8 While it is hard to verify the Hölder smoothness of res in these empirical datasets, we believe that some level of justification would be possible assuming (1) enough differentiability on $f$ (which we think to be true for common neural net architectures and activation functions), (2) $Y$ being deterministic given $X$ (which is reasonable for low-noise datasets such as ImageNet), and (3) the set of inputs $C_k = \{x \in \mathcal{X} : (Y|X = x) = e_k\}$ corresponding to each class being a Lipschitz domain, locally the graph of a Lipschitz continuous function. Given these assumptions, each coordinate of the regression function $\mathrm{reg}_f(z)$ can be written as $[\mathrm{reg}_f(z)]_k = \mathbb{E}[[Y]_k | f(X) = z] = P_Z(C_k \cap \{f(x) = z\})/P_Z(\{f(x) = z\})$; assuming the denominator is strictly positive. Then, Hölder continuity may follow from the inverse function theorem and the coarea formula (Federer, 2014, Theorem 3.2.3), which expresses the measure of a level set as an integral of the Jacobian.

However, this argument still requires making the essentially unverifiable Assumption (3) from the above paragraph. In some cases, this assumption can be viewed as reasonable: for instance, one may reasonably think that image manifolds for classes in ImageNet are locally Lipschitz; and thus Assumption (3) may hold. Under such conditions, this rough argument may provide an idea of why Hölder smoothness could be reasonable for certain predictive models such as neural net architectures.

When the Hölder condition does not hold, we still have the type I error guarantee, but may not have type II error control. We expect that the Hölder condition might be drastically violated when there is obvious discontinuity of the predictors/regression functions (e.g., a decision tree/random forest trained on data having discrete features).

| | DenseNet 161 | | ResNet 152 | | EfficientNet-b7 | |
|---|---|---|---|---|---|---|
| | $\ell_1$-ECE | Calibrated? | $\ell_1$-ECE | Calibrated? | $\ell_1$-ECE | Calibrated? |
| No Calibration | 5.67% | reject | 4.99% | reject | 2.82% | reject |
| Platt Scaling | 1.58% | reject | 1.41% | reject | 1.90% | reject |
| Poly. Scaling | 0.62% | accept | 0.64% | accept | 0.71% | accept |
| Isot. Regression | 0.63% | reject | 0.80% | reject | 1.06% | reject |
| Hist. Binning | 0.46% | reject | 1.26% | reject | 0.88% | reject |
| Scal. Binning | 1.55% | reject | 1.40% | reject | 1.97% | reject |

Table 3: The values of the empirical $\ell_1$-ECE (Guo et al., 2017) and the testing results, via adaptive T-Cal and multiple binomial testing, of models trained on ImageNet.

## 5 Lower Bounds for Detecting Mis-calibration

To complement our results on the performance of the plug-in tests proposed earlier, we now show some fundamental lower bounds for detecting mis-calibration. We also provide a reduction that allows us to test calibration via two-sample tests and sample splitting. We show that this has a minimax optimal performance, but empirically does not perform as well as our previous test.

### 5.1 Impossibility for General Continuous Mis-calibration Curves

In Proposition 9, we show that detecting mis-calibration is impossible, even when the sample size $n$ is arbitrarily large unless the mis-calibration curve $\mathrm{res}_{\cdot}$ has some level of smoothness. Intuitively, if the mis-calibration curve can be arbitrarily non-smooth, then it can oscillate between positive and negative values with arbitrarily high frequency, and these oscillations cannot be detected from a finite sample.

In this regard, one needs to be careful when concluding the quality of calibration from a finite sample. If we only assume that the mis-calibration curve is a continuous function of the probability predictions, then it is impossible to tell apart calibrated and mis-calibrated models. Further, for more complex models such as deep neural networks, one expects the predicted probabilities to be able to capture larger and larger classes of functions; thus this result is even more relevant for modern large-scale machine learning.

Let $P_1^{\mathrm{cont}}(\varepsilon; p)$ be the family of probability distributions $P$ over $(Z, Y)$ such that $\ell_p\text{-ECE}_P(f) \geq \varepsilon$ and every entry of the mis-calibration curve $\mathrm{res}_{\cdot;P}$ is continuous. This is a larger set of distributions than $P_1(\varepsilon; p; s)$ in (3), because we only assume continuity, not Hölder smoothness. Denote the corresponding minimax type II error by $R_n^{\mathrm{cont}}(\varepsilon; p)$, namely

$$R_n^{\mathrm{cont}}(\varepsilon; p) := \inf_{\phi \in \Phi_n(\alpha)} \sup_{P \in P_1^{\mathrm{cont}}(\varepsilon; p)} P(\phi = 0).$$

This has the same interpretation as before, namely, it is the best possible false negative rate for detecting mis-calibration for data distributions belonging to $P_1^{\mathrm{cont}}(\varepsilon; p)$, in a worst-case sense.

Proposition 9 (Impossibility of detecting mis-calibration) Let $\varepsilon_0 = 0.1$. For any level $\alpha \in (0,1)$, the minimax type II error $R_n^{\text{cont}}(\varepsilon_0, p)$ for testing the null hypothesis of calibration at level $\alpha$ against the hypothesis $P \in P_1^{\text{cont}}(\varepsilon_0, p)$ of general continuous mis-calibration curves satisfies $R_n^{\text{cont}}(\varepsilon_0, p) \geq 1 - \alpha$ for all $n$.

In words, this result shows that for a certain fixed $\ell_p$ calibration error $\varepsilon_0$, and for a fixed false positive rate $\alpha > 0$, the false negative rate is at least $1 - \alpha$. Thus, it is not possible to detect mis-calibration in this setting. The choice of the constant $\varepsilon_0 = 0.1$ is arbitrary and can be replaced by any other constant; the result holds with minor modifications to the proof.

The proof can be found in Appendix A.7. We make a few remarks on related results. While Example 3.2 of Kumar et al. (2019) demonstrates that—related to earlier results on probability distribution and density estimation (Mann and Wald, 1942)—using a binned estimator of ECE can arbitrarily underestimate the calibration error, we show a fundamental failure not due to binning, but instead due to the finite sample size. Also, our result echoes Theorem 3 of Gupta et al. (2020) which states that asymptotically perfect calibration is only possible for probability predictors with a countable support; but this does not overlap with Proposition 9 as our conclusion is about the impossibility of detecting mis-calibration.

### 5.2 Hölder Alternatives

As is customary in nonparametric statistics (Ingster, 1987; Low, 1997; Gyorfi et al., 2002; Ingster, 2012), we consider testing against Hölder continuous alternatives; or, differently put, detecting mis-calibration when the mis-calibration curves are Hölder continuous. This excludes the pathological examples where the mis-calibration curves oscillate widely that were discussed in Section 5.1; but still allow a very rich class of possible mis-calibration curves, including non-smooth ones.

Theorem 10 states that, for a $K$-class classification problem and for alternatives with a Hölder smoothness parameters, the mis-calibration of a model can be detected only when the calibration error is of order $(n^{-2s/(4s+K-1)})$. In other words, the smallest possible calibration error that can be detected using a sample of size $n$ is of order $n^{-2s/(4s+K-1)}$.

Testing calibration of a probability predictor in our nonparametric model leads to rates that are slower than the parametric case $n^{-1/2}$. This is because $2s/(4s+K-1) < 1/2$ for $s > 0$ and $K \geq 2$. The rate becomes even slower as the number of classes $K$ grows. This indicates that evaluating model calibration on a small-sized dataset can be problematic. Further, it suggests that multi-class calibration may be even harder to achieve.

This rate is what one may expect based on results for similar problems in nonparametric hypothesis testing (Ingster, 2012), with $K - 1$ interpreted as the dimension. Specifically, the rate is equal to the minimum separation rate in two-sample goodness-of-fit testing for densities on $[0,1]^{K-1}$. This connection to two-sample testing will be made clear in Section 6.

Theorem 10 (Lower bound for detecting mis-calibration) Given a level $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, consider the hypothesis testing problem (4), in which we test the calibration of the $K$-class probability predictor $f$ assuming $(s, L)$-Hölder continuity of mis-calibration curves as defined in (62). There exists $c_{\text{lower}} > 0$ depending only on

$(p; s; L; K; ; )$ such that, for any $p > 0$, the minimum $\ell_p$-ECE of $f$, i.e. $\varepsilon_n(p; s)$, required to have a test with a false positive rate (type I error) at most $$ and with a true positive rate (power) at least $1$ satisfies $\varepsilon_n(p; s) \geq c_{\text{lower}} n^{-2s/(4s+K-1)}$ for all $n$.

See Appendix A.8 for the proof. The proofs of both Proposition 9 and Theorem 10 are based on Ingster's method, also known as the chi-squared or Ingster-Suslina method (Ingster, 1987, 2012). Informally, Ingster's method states that if we can select alternative distributions with an average likelihood ratio to a null distribution close to unity, then no test with a fixed level can control the minimax type II error below a certain threshold.

Remark 11 (Hölder smoothness assumption) Since the residual function $\text{res}_{f;P}$ depends on the unknown joint distribution $P$ of $(Z; Y)$, the Hölder continuity of the map $z \mapsto [\text{res}_f(z)]_k$ for each $k \in \{1; \ldots; K\}$ is in general an assumption that we need to make. When the residual map $\text{res}_f$ does not satisfy the Hölder assumption, we still have the false detection rate control in Theorem 3 and 5, but we cannot guarantee the true detection rate control and the lower bound in Theorem 10. Extending our approach beyond the Hölder assumption may be possible in future work, inspired by works in nonparametric hypothesis testing that study|for instance|Besov spaces of functions (Ingster, 2012).

## 6 Reduction to Two-sample Goodness-of-fit Testing

To further put our work in context in the literature on nonparametric hypothesis testing, in this section we carefully examine the connections between the problem of testing calibration, and a well-known problem in that area. Specifically, we describe a novel randomization scheme that allows us to reduce the null hypothesis of perfect calibration to a hypothesis of equality of two distributions|making a strong connection to the problem of two-sample goodness of fit testing In other words, we can use a calibrated probabilistic classifier and randomization to generate two samples from an identical distribution. For a mis-calibrated classifier the same scheme will generally result in two samples from two different distributions.

If a classifier is perfectly calibrated, then its class probability predictions will match the true prediction-conditional class probabilities. Therefore, randomly sampling labels according to the classifier's probability predictions will yield a sample from the empirical distribution. We rely on sample splitting to obtain two samples: the empirical and the generated one. Then we can use any classical test to check if the two samples are generated from the same distribution. As we will show, the resulting test has a theoretically optimal detection rate. However, due to the sample splitting step, its empirical performance is inferior to the test based on the debiased plug-in estimator from Section 3.

We split our sample into two parts. For $i \in \{\lfloor n/2 \rfloor + 1; \ldots; n\}$, we generate random variables $\tilde{Y}_i$ following the categorical distribution $\text{Cat}(Z_i)$ over classes $\mathcal{Y} = \{1; \ldots; K\}$, with a $K$-class probability distribution $Z_i = f(X_i)$ predicted by the classifier $f$. These $\tilde{Y}_i$ are independent of each other and of $Y_1; \ldots; Y_{\lfloor n/2 \rfloor}$, due to the sample splitting step. For each $k \in \{1; \ldots; K\}$, define

$$V_k := \sum_{i=\lfloor \frac{n}{2} \rfloor + 1}^{n} Z_i \cdot [Y_i]_k = 1; 1 \quad i$$

and

$$W_k := \left\{ Z_i : [\hat{Y}_i]_k = 1,\ \left\lfloor \frac{n}{2} \right\rfloor + 1 \quad i \quad n \right\} :$$

By construction, $V_k$ is an i.i.d. sample from the distribution on the probability simplex $_{K\ 1}$ with a density[3]

$$^{V}_{k}(z) := \frac{[\mathrm{reg}_f(z)]_k}{\int_{K\ 1} [\mathrm{reg}_f(z)]_k \, dP_Z(z)} = \frac{[\mathrm{reg}_f(z)]_k}{E[Y]_k}$$

with respect to $P_Z$. Similarly, $W_k$ is an i.i.d. sample from the distribution on $_{K\ 1}$ with a density

$$^{W}_{k}(z) := \frac{[z]_k}{\int_{K\ 1} [z]_k \, dP_Z(z)} = \frac{[z]_k}{E[Z]_k} :$$

Now we consider testing the null hypothesis

$$H_0 : \ ^{V}_{k} = \ ^{W}_{k} \text{ for all } k \in \{1, \ldots, K\}$$

against the complement of $H_0$. We claim that if we use an appropriate test for this null hypothesis, with an additional procedure to rule out "easily detectable" alternatives, then we can obtain a test that attains the optimal rate specified in Theorem 10.

We describe the main idea of this reduction. A formal result can be found in Theorem 12. Let $P \in \mathcal{P}_1(\varepsilon; p; s)$ and assume $\varepsilon = \tilde{\Theta}(n^{-2s/(4s+K\ 1)})$.[4] The squared distance between $^{V}_{k}$ and $^{W}_{k}$ in $L^2(P_Z)$ is

$$\int_{K\ 1} \left( \frac{[\mathrm{reg}_f(z)]_k}{E[Y]_k} - \frac{[z]_k}{E[Z]_k} \right)^2 dP_Z(z) = \int_{K\ 1} \left( \frac{[\mathrm{res}_f(z)]_k}{E[Y]_k} + \frac{[z]_k}{E[Y]_k} - \frac{[z]_k}{E[Z]_k} \right)^2 dP_Z(z)$$
$$\frac{1}{(E[Y]_k)^2} \int_{K\ 1} [\mathrm{res}_f(z)]_k^2 \, dP_Z(z) + \frac{2E[Z \quad Y]_k}{(E[Y]_k)^2 E[Z]_k} \int_{K\ 1} [z]_k [\mathrm{res}_f(z)]_k \, dP_Z(z) : \qquad (11)$$

Further,

$$\int_{K\ 1} [z]_k [\mathrm{res}_f(z)]_k \, dP_Z(z) = E_P[[Z]_k E[Y \quad Z \mid Z]_k] = E_P[[Z]_k [Y \quad Z]_k] : \qquad (12)$$

Since $E[Y \quad Z]_k = E[[Z]_k [Y \quad Z]_k] = 0$ and $\mathrm{Var}([Y \quad Z]_k); \mathrm{Var}([Z]_k [Y \quad Z]_k) \quad 1$ under $H_0$, we can detect mis-calibration for the alternatives $P \in \mathcal{P}_1(\varepsilon; p; s)$ such that $E_P[Y \quad Z]_k = (n^{-1/2})$ or $E_P[[Z]_k [Y \quad Z]_k] = (n^{-1/2})$ by rejecting the null hypothesis $H_0$ of calibration if $\frac{1}{n} \sum_{i=1}^{n} [Y_i \quad Z_i]_k \quad cn^{-1/2}$ or $\frac{1}{n} \sum_{i=1}^{n} [Z_i]_k [Y_i \quad Z_i]_k \quad cn^{-1/2}$ for some $c > 0$. For the remaining alternatives, choose $k_0 \in \{1, \ldots, K\}$ such that

$$\frac{1}{(E[Y]_{k_0})^2} \int_{K\ 1} [\mathrm{res}_f(z)]_k^2 \, dP_Z(z) \quad \frac{\varepsilon^2}{K (E[Y]_{k_0})^2} = \tilde{\Theta}(n^{-\frac{4s}{4s+K\ 1}}) :$$

3. We assume that the densities $^{V}_{k}$ and $^{W}_{k}$ are well defined, and in particular that $E[Y]_k > 0$ and $E[Z]_k > 0$ for every $k \in \{1, \ldots, K\}$. This follows from Assumption 2, which will be introduced later in the section.
4. Here we use the notation $\tilde{\Theta}(\ )$ to include the adaptive case. See Corollary 13 and Remark 14 for further details.

Then, $k \overset{V}{k} \overset{W}{k} k^2_{L^2(P_Z)}$ is at least

$$\tilde{\Omega}(n^{-\frac{4s}{4s+K-1}}) + \frac{2E[Z - Y]_{k_0} E[[Z]_{k_0}[Y - Z]_{k_0}]}{(E[Y]_{k_0})^2 E[Z]_{k_0}} = \tilde{\Omega}(n^{-\frac{4s}{4s+K-1}}).$$

Since $|V_{k_0}|, |W_{k_0}| = \Omega(n)$ with high probability, the power of the test can be controlled using standard results on two sample testing. The full procedure is described in Algorithm 3.

In general, for a positive integer $d > 0$, we allow using an arbitrary deterministic two-sample testing procedure $TS_{\alpha,\beta} : ([0,1]^d)^{n_1} \times ([0,1]^d)^{n_2} \to \{0,1\}$, which takes in two $d$-dimensional samples $\{V_1, \ldots, V_{n_1}\}$, $\{W_1, \ldots, W_{n_2}\}$ and outputs "1" if and only if the null hypothesis is rejected. The two samples $\{V_1, \ldots, V_{n_1}\}$ and $\{W_1, \ldots, W_{n_2}\}$ are sampled i.i.d. from distributions with densities $f_1$ and $f_2$, respectively, with respect to an appropriate probability measure $\mu$ on $[0,1]^d$. Further, it is assumed that $f_1 - f_2$ is $(s, L)$-Hölder continuous for a Hölder smoothness parameter $s > 0$ and a Hölder constant $L > 0$. Given $\alpha \in (0,1)$ and $\beta \in (0, 1-\alpha)$, the two-sample test is required to satisfy, for some $c_{ts} > 0$ depending on $(s, L, d, \alpha, \beta)$ and on $\ell, u$ from Assumption 1 to be introduced next,

$$\begin{aligned} &P(TS_{\alpha,\beta}(V_1, \ldots, V_{n_1}, W_1, \ldots, W_{n_2}) = 1) && \text{if } f_1 = f_2, \\ &P(TS_{\alpha,\beta}(V_1, \ldots, V_{n_1}, W_1, \ldots, W_{n_2}) = 0) && \text{if } \|f_1 - f_2\|_{L^2(\mu)} \geq c_{ts}(n_1 \wedge n_2)^{-\frac{2s}{4s+d}}. \end{aligned} \quad (13)$$

There are a number of such tests proposed in prior work, see e.g., Ingster (2012); Arias-Castro et al. (2018); Kim et al. (2022) and Appendix B.2. Our general approach allows using any of these. It is also known that there are *adaptive* tests $TS^{ad}$ that do not require knowing the Hölder smoothness parameters. In the adaptive setting, the best-known minimum required separation in this general dimensional situation is

$$\|f_1 - f_2\|_{L^2(\mu)} \geq c_{ad} \left(\frac{n_1 \wedge n_2}{\log\log(n_1 \wedge n_2)}\right)^{-\frac{2s}{4s+d}} \quad (14)$$

for some $c_{ad} > 0$. See Appendix B.2 for examples of $TS$ and $TS^{ad}$.

Next, we state an additional assumption required in our theorem. See Appendix B.2 for more discussion. Assumption 2 guarantees that every class appears in the dataset. This is reasonable in many practical settings, as classes that do not appear can be omitted.

Assumption 2 (Lower bounded class probability) There exists a constant $d_c > 0$ such that $E[Y]_k > d_c$ for all $k \in \{1, \ldots, K\}$.

Our result is as follows.

Theorem 12 (Optimal calibration test via sample splitting) Suppose $p \geq 2$ and let $\overset{split}{n}$ be the test described in Algorithm 3. Assume the Hölder smoothness parameter $s$ is known. Under Assumption 1 and 2, we have

1. False detection rate control. For every $P$ for which $f$ is perfectly calibrated, i.e., for $P \in \mathcal{P}_0$, the probability of falsely claiming mis-calibration is at most $\alpha$, i.e., $P(\overset{split}{n} = 1) \leq \alpha$.

**Algorithm 3** Sample splitting calibration test $\phi_n^{\text{split}}$

Input: Probability predictor $f : \mathcal{X} \to \Delta_{K-1}$; i.i.d. sample $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i \in \{1, \dots, n\}\}$; false detection rate $\alpha \in (0, 1)$; true detection rate $\beta \in (0, 1-\alpha)$; Hölder smoothness $s$; minimax optimal two-sample density test TS

Procedure: $Z_i \sim f(X_i)$ for $i \in \{1, \dots, n\}$; independently sample $\tilde{Y}_i \sim \mathrm{Cat}(Z_i)$ for $i \in \{\lfloor \frac{n}{2} \rfloor + 1, \dots, n\}$

for $k = 1$ to $K$ do

$T_{1,k} \leftarrow \frac{1}{n} \sum_{i=1}^{n} [Y_i - Z_i]_k$, $T_{2,k} \leftarrow \frac{1}{n} \sum_{i=1}^{n} [Z_i]_k [Y_i - Z_i]_k$

$V_k \leftarrow \{Z_i : [Y_i]_k = 1, 1 \le i \le \frac{n}{2}\}$, $W_k \leftarrow \{Z_i : [\tilde{Y}_i]_k = 1, \frac{n}{2} + 1 \le i \le n\}$

$b_k \leftarrow \mathbb{I}\{|T_{1,k}| \ge \sqrt{\frac{3K}{n}}\} \cdot \mathbb{I}\{|T_{2,k}| \ge \sqrt{\frac{3K}{n}}\} \cdot \mathrm{TS}_{\frac{\alpha}{3K}, \frac{\beta}{2}}(V_k, W_k)$

end for

Output: Reject $H_0$ if $\phi_n^{\text{split}} := \max\{b_k : k \in \{1, \dots, K\}\} = 1$

2. True detection rate control. There exists $c_{\text{split}} > 0$ depending on $(s, L, K, \|\cdot\|_l, \|\cdot\|_u, d_c, \alpha, \beta)$ such that the power (true positive rate) is bounded as $P(\phi_n^{\text{split}} = 1) \ge 1 - \beta$ for every $P \in \mathcal{P}_1(\varepsilon, p, s)$|i.e., when $f$ is mis-calibrated with an $\ell_p$-ECE of at least $\varepsilon \ge c_{\text{split}} n^{-2s/(4s+K-1)}$.

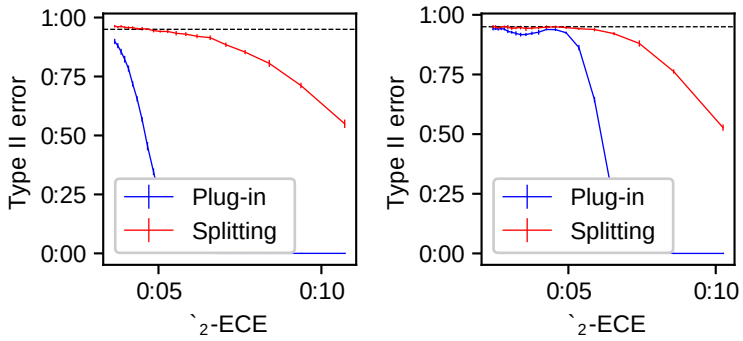
The proof is in Appendix A.9. Theorem 10 and Theorem 12 together imply that the minimax optimal detection rate for calibration is $\varepsilon_n(p, s) \asymp n^{-2s/(4s+K-1)}$. By replacing TS with an adaptive test $\mathrm{TS}^{\text{ad}}$, we obtain an adaptive version of the test $\phi_n^{\text{split}}$.

Corollary 13 (Adaptive test via sample splitting) Suppose $p \ge 2$ and let $\phi_n^{\text{ad-s}}$ be the test described in Algorithm 3 with TS replaced by an adaptive two-sample test $\mathrm{TS}^{\text{ad}}$. Under Assumption 1 and 2, we have

1. False detection rate control. For every $P$ for which $f$ is perfectly calibrated, i.e., for $P \in \mathcal{P}_0$, the probability of falsely claiming mis-calibration is at most $\alpha$, i.e., $P(\phi_n^{\text{ad-s}} = 1) \le \alpha$.

2. True detection rate control. There exists $c_{\text{ad-s}} > 0$ depending on $(s, L, K, \|\cdot\|_l, \|\cdot\|_u, d_c, \alpha, \beta)$ such that the power (true positive rate) is bounded as $P(\phi_n^{\text{ad-s}} = 1) \ge 1 - \beta$ for every $P \in \mathcal{P}_1(\varepsilon, p, s)$|i.e., when $f$ is mis-calibrated with an $\ell_p$-ECE of at least $\varepsilon \ge c_{\text{ad-s}}(n/\log\log n)^{-2s/(4s+K-1)}$.

Remark 14 (Adaptation cost and optimality) Spokoiny (1996); Ingster (2000) develop an adaptive chi-squared test for one-dimensional goodness-of-fit testing which can adapt to an unknown Hölder smoothness parameter $s$ while only losing a $(\log\log n)^{s/(4s+1)}$ factor in the separation rate. The test was proven to be minimax optimal in the adaptive setting. Arias-Castro et al. (2018) extend the adaptive test to a general dimension $d$ and attain an adaptive test at the cost of $\sqrt{\log n}$ factor. Kim et al. (2022) provide a stronger analysis for their permutation test and reduce the adaptation cost to $(\log\log n)^{2s/(4s+d)}$. To our knowledge, the minimax optimality of these adaptive tests in general dimensions is so far not established.




(a) $s = 0:6, \quad = 100$ (b) $s = 0:8, \quad = 200$

Figure 7: Type II error comparison for $_{m;n}$ and $^{split}_{n}$. The horizontal dashed line indicates a type II error of $1 \quad = 0:95$. Since $^{split}_{n}$ relies on sampling splitting, its e ective sample size is much smaller than that of the plug-in test. This results in higher type II errors as can be seen in the gure. Standard error bars are plotted over 10 repetitions.

While the adaptive test in Corollary 13 requires an additional factor of $(\log\log n)^{2s=(4s+K \quad 1)}$ in the separation rate, Theorem 5 requires a factor of $(\log n)^{s=(4s+K \quad 1)}$. This gap comes from the requirement(14) which we borrow from Kim et al. (2022). Theorem 6.1 and Lemma C.1 of Kim et al. (2022) develops combinatorial concentration inequalities to improve a polynomial dependency on in the separation rate to a logarithmic dependency. This results in the$(\log\log n)^{2s=(4s+K \quad 1)}$ factor in their adaptive test. Since our proof of Theorem 3 uses a quadratic tail bound from Chebyshev's inequality, the adaptation cost in Theorem 5 is$(\log n)^{s=(4s+K \quad 1)}$. Currently, it appears challenging to improve the polynomial dependence for$T^{d}_{m;n}$, due to its complicated conditional structure.

### 6.1 Comparison with the Debiased Plug-in Test

We compare the empirical performances of the debiased plug-in test$_{m;n}$ and the sample splitting test $^{split}_{n}$. As described in Section 4.1, we study the type II error against the xed alternative where the mis-calibration is speci ed as$H_1 : (Z;Y) \quad P_{1;m}$, for various values of m. We use a sample size of$n = 20;000$ and pairs of Holder smoothness and scaling parameter indicated in Figure 7. The critical value for $= 0:05$ and the corresponding type II error are estimated via 1,000 Monte Carlo simulations. For the sample splitting test $^{split}_{n}$, we use the chi-squared two-sample test of Arias-Castro et al. (2018).

Since $^{split}_{n}$ relies on sample splitting and discards some of the observations, its e ective sample size is smaller than that of the debiased plug-in test. For this reason, we nd that T-Cal outperforms the sample splitting test by a large margin. While the sample splitting test reveals a theoretically interesting connection to two-sample density testing, it appears empirically suboptimal.

## 7 Conclusion

This paper studied the problem of testing model calibration from a nite sample. We analyzed the plug-in estimator of $\ell_2$-ECE$(f)^2$ as a test statistic for calibration testing. We discovered that the estimator needs debiasing and becomes minimax optimal when the number of bins is chosen appropriately. We also provided an adaptive version of the test, which can be used without knowing the Holder smoothness parameter $s$. We tested T-Cal with a broad range of experiments, including several neural net architectures and post-hoc calibration methods.

On the theoretical side, we provided an impossibility result for testing calibration against general continuous alternatives. Assuming that the calibration curve is $s$-Holder-smooth, we derived a lower bound of $(n^{-2s/(4s+K-1)})$ on the calibration error required for a model to be distinguished from a perfectly calibrated one. We also discussed a reduction to two-sample testing and showed that the resulting test also matches the lower bound.

Interesting future directions include (1) developing a testing framework for comparing calibration of predictive models, (2) extending the theoretical result to $\ell_p$-ECE with $p > 2$ and other calibration concepts such as top$(k)$, within-$k$, and marginal calibration, (3) developing a minimax estimation theory for calibration error, and (4) establishing local rates for testing calibration.

## Acknowledgments and Disclosure of Funding

This work was supported in part by the NSF TRIPODS 1934960, NSF DMS 2046874 (CAREER), ARO W911NF-20-1-0080, DCIST, and NSF CAREER award CIF-1943064, and Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) #FA9550-20-1-0111 award. We thank the editors and associate editor for their work in handling our manuscript; and the reviewers for their very thorough reading and many helpful suggestions that have significantly improved our work. We are grateful to a number of individuals for their comments, discussion, and feedback; in particular, we would like to thank Sivaraman Balakrishnan, Chao Gao, Chirag Gupta, Lucas Janson, Adel Javanmard, Edward Kennedy, Ananya Kumar, Yuval Kluger, Mohammed Mehrabi, Alexander Podkopaev, Yury Polyanskiy, Mark Tygert, and Larry Wasserman.

## Appendix A. Proofs

Notations. For completeness and to help the reader, we present (and in some cases recall) some notations. We will use the symbols := or =: to define quantities in equations. We will occasionally use bold font for vectors. For an integer $d \geq 1$, we denote $[d] := \{1, \ldots, d\}$ and $1_d := (1, 1, \ldots, 1)^\top \in \mathbb{R}^d$. For a vector $v \in \mathbb{R}^d$, we will sometimes write $[v]_i$ for the $i$-th coordinate of $v$, for any $i \in [d]$. The minimum of two scalars $a, b \in \mathbb{R}$ is denoted by $\min(a, b)$ or $a \wedge b$; their maximum is denoted by $\max(a, b)$ or $a \vee b$. We denote the $d$-dimensional Lebesgue measure by $\mathrm{Leb}_d$. For a function $h : \mathbb{R}^d \to \mathbb{R}$, $1 \leq p < \infty$, and a measure $\mu$ on $\mathbb{R}^d$, we let $\|h\|_{L^p(\mu)} := (\int |h|^p d\mu)^{1/p}$. When $\mu = \mathrm{Leb}_d$, we omit $\mu$ and write $\|h\|_{L^p}$. If $p = \infty$, then $\|h\|_{L^p} := \operatorname{ess\,sup}_{x \in \mathbb{R}^d} |h(x)|$. We denote the $\ell_p$-norm of $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$ by $\|x\|_p := (\sum_{i=1}^d |x_i|^p)^{1/p}$. When $p$ is unspecified, $\|\cdot\|$ stands for $\|\cdot\|_2$.

For two sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$ with $b_n \neq 0$, we write $a_n \asymp b_n$ if $0 < \liminf_n a_n/b_n \leq \limsup_n a_n/b_n < \infty$. When the index $n$ is self-evident, we may omit it above. We use the Bachmann-Landau asymptotic notations $O(\cdot), \Omega(\cdot)$ to hide constant factors in inequalities and use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$ to also hide logarithmic factors. For a Lebesgue measurable set $A \subseteq \mathbb{R}^d$, we denote by $1_A : \mathbb{R}^d \to \{0, 1\}$ its indicator function where $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ otherwise. For a real number $s \in \mathbb{R}$, we denote the largest integer less than or equal to $s$ by $\lfloor s \rfloor$. Also, the smallest integer greater than or equal to $s$ is denoted by $\lceil s \rceil$.

For an integer $d \geq 1$, a vector $\alpha = (\alpha_1, \ldots, \alpha_d)^\top \in \mathbb{N}^d$ is called a multi-index. We write $|\alpha| := \alpha_1 + \cdots + \alpha_d$. For a vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)^\top \in \mathbb{N}^d$, we write $x^\alpha := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$. For a sufficiently smooth function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its partial derivative of order $\alpha = (\alpha_1, \ldots, \alpha_d)^\top$ by $f^{(\alpha)} := \partial^{\alpha_1} \cdots \partial^{\alpha_d} f$. A pure partial derivative with respect to an individual coordinate $i \in [d]$ is also denoted as $\partial_i f$. For two sets $S, T$, a map $f : S \to T$, and a subset $S' \subseteq S$, we denote by $f(S')$ the image of $S'$ under $f$. The support of a function $f : S \to \mathbb{R}$ is the set of points $\mathrm{supp}(f) := \{a \in S : f(a) \neq 0\}$.

The uniform distribution on a compact set $S \subseteq \mathbb{R}^d$ is denoted by $\mathrm{Unif}(S)$. The binomial distribution with $n \in \mathbb{N}$ trials and success probability $p \in [0, 1]$ is denoted by $\mathrm{Bin}(n, p)$, and write $\mathrm{Ber}(p) := \mathrm{Bin}(1, p)$. For an integer $d \geq 2$, we let $\Delta_{d-1} := \{z = (z_1, \ldots, z_d)^\top \in [0, 1]^d : z_1 + \cdots + z_d = 1\}$ be the $(d-1)$-dimensional probability simplex. We denote the multinomial distribution with $n \in \mathbb{N}$ trials and class probability vector $p \in \Delta_{d-1}$ by $\mathrm{Multi}(n, p)$, and write $\mathrm{Cat}(p) := \mathrm{Multi}(1, p)$. For a joint distribution $(X, Y) \sim P$, we will write $P_X, P_Y$ for the marginal distributions of $X, Y$, respectively. For a distribution $Q$ and a random variable $Z \sim Q$, we will denote expectations of functions of $Z$ with respect to $Q$ as $\mathbb{E}f(Z)$, $\mathbb{E}_Z f(Z)$, $\mathbb{E}_Q f(Z)$, or $\mathbb{E}_{Z \sim Q} f(Z)$. We abbreviate almost surely by "a.s.", and almost everywhere by "a.e."

### A.1 Proof of Theorem 3

We first state a Lemma used in the proof. This lemma generalizes Lemma 3 in Arias-Castro et al. (2018) from the uniform measure on the cube to a general probability measure on the probability simplex. See Section 3.2.2 of Ingster (2012) for a discussion of how such results connect to geometric notions like Kolmogorov diameters.

Lemma 15 For $m \geq N_+$, let $B_m = \{B_1, \ldots, B_{m^{K-1}}\}$ be the partition of $\mathcal{I}^{K-1}$ defined in Appendix B.3, and $\mu$ be a probability measure on $\mathcal{I}^{K-1}$ such that $\mu(B_i) > 0$ for all $i \in [m^{K-1}]$. For any continuous function $h : \mathcal{I}^{K-1} \to \mathbb{R}$, define

$$W_m[h] := \sum_{i=1}^{m^{K-1}} \frac{\int_{B_i} h(z) d\mu(z)}{\mu(B_i)} 1_{B_i}.$$

There are $b_1, b_2 > 0$ depending on $(K, s, L)$ such that for every $h \in H_K(s, L)$,

$$\|W_m[h]\|_{L^2(\mu)} \leq b_1 \|h\|_{L^2(\mu)} + b_2 m^{-s}.$$

In other words,

$$\left( \sum_{i=1}^{m^{K-1}} \frac{\mathbb{E}[h(Z) \mathbb{I}(Z \in B_i)]^2}{\mu(B_i)} \right)^{\frac{1}{2}} \leq b_1 \mathbb{E}[h(Z)^2]^{\frac{1}{2}} + b_2 m^{-s}.$$

The proof can be found in Appendix A.2.

Overview of the proof. The proof follows the classical structure of upper bound arguments in nonparametric hypothesis testing, see e.g., Arias-Castro et al. (2018); Kim et al. (2022) for recent examples. We compute or bound the mean and variance of $\overline{T}^d_{m,n}$ under null distributions $P_0 \in \mathcal{P}_0$ and alternative distributions $P_1 \in \mathcal{P}_{s;L;K}$ with a large ECE. Using Lemma 15, we can find a lower bound on $\mathbb{E}_{P_1}[T^d_{m,n}] - \mathbb{E}_{P_0}[T^d_{m,n}]$. The variances $\mathrm{Var}_{P_0}(T^d_{m,n})$ and $\mathrm{Var}_{P_1}(T^d_{m,n})$ can be also upper bounded. We argue that the mean difference $\mathbb{E}_{P_1}[T^d_{m,n}] - \mathbb{E}_{P_0}[T^d_{m,n}]$ is significantly larger than the square root of the variances $\mathrm{Var}_{P_0}(T^d_{m,n})$ and $\mathrm{Var}_{P_1}(T^d_{m,n})$. The conclusion follows from Chebyshev's inequality.

Proof Let $N_i := |I_{m,i}|$ for each $i \in [m^{K-1}]$ and $I := \{i \in [m^{K-1}] : N_i \geq 1\}$. Also write $N := (N_1, \ldots, N_{m^{K-1}})^\top$, $Z := (Z_1, \ldots, Z_n)^\top$, $\overline{Y} := Y - Z$, and $\overline{Y}_j := Y_j - Z_j$ for all $j \in [n]$. By Assumption 1,

$$\int_{\mathcal{I}^{K-1}} \mathrm{kres}_i(z)^2 dP_Z(z) \leq u \int_{\mathcal{I}^{K-1}} \mathrm{kres}_i(z)^2 dz,$$

where the latter integral is with respect to the uniform measure $\mathrm{Unif}(\mathcal{I}^{K-1})$ on $\mathcal{I}^{K-1}$. Therefore, we may assume $P_Z = \mathrm{Unif}(\mathcal{I}^{K-1})$ by merging $u$ with $c$. We prove the theorem for $p = 2$. Then, the general case follows since $\mathcal{P}_1(\varepsilon; p; s) \subseteq \mathcal{P}_1(\varepsilon; 2; s)$ for all $p \geq 2$.

Let $P_0 \in \mathcal{P}_0$ and $P_1 \in \mathcal{P}_{s;L;K}$ be a null and an alternative distribution over $(Z, Y)$, respectively. Write $\varepsilon := \ell_p\text{-ECE}_{P_1}(f)$. Under $P_0$ and conditioned on $Z$, recalling $T^d_{m;n}$ from (6),

$$\mathbb{E}_{P_0}[T^d_{m,n} \mid Z] = \frac{1}{n} \sum_{i \in I} \frac{1}{N_i} \sum_{j_1 \neq j_2 \in I_{m,i}} \mathbb{E}_{P_0}\left[ \overline{Y}_{j_1}^\top \overline{Y}_{j_2} \mid Z \right] = 0$$

because $\mathbb{E}_{P_0}[\overline{Y}_{j_1}^\top \overline{Y}_{j_2} \mid Z] = \mathbb{E}_{P_0}[\overline{Y}_{j_1} \mid Z]^\top \mathbb{E}_{P_0}[\overline{Y}_{j_2} \mid Z] = 0$ for all $j_1 \neq j_2 \in I_{m,i}$. Therefore,

$$\mathbb{E}_{P_0}[T^d_{m,n}] = \mathbb{E}_{P_0}[\mathbb{E}_{P_0}[T^d_{m,n} \mid Z]] = 0. \tag{15}$$

Also,

$$\mathrm{Var}_{P_0}(T^d_{m,n} \mid Z) = \frac{1}{n^2}\sum_{i\in I} \frac{1}{N_i^2}\mathrm{Var}_{P_0}\Big(2\sum_{j_1<j_2\in I_{m,i}} \overline{Y}^{>}_{j_1}\overline{Y}_{j_2} \mid Z\Big) = \frac{1}{n^2}\sum_{i\in I} \frac{4}{N_i^2}\sum_{j_1<j_2\in I_{m,i}} \mathrm{Var}_{P_0}\left(\overline{Y}^{>}_{j_1}\overline{Y}_{j_2} \mid Z\right).$$

Here we used that the cross terms in the expansion of $\mathrm{Var}_{P_0}(\sum_{j_1<j_2\in I_{m,i}} \overline{Y}_{j_1}\overline{Y}_{j_2} \mid Z)$ vanish since for $j_1, j_2, j_3 \in I_{m,i}$ that are pairwise different,

$$\mathrm{Cov}_{P_0}\left(\overline{Y}^{>}_{j_1}\overline{Y}_{j_2}, \overline{Y}^{>}_{j_1}\overline{Y}_{j_3} \mid Z\right) = E_{P_0}\left[\overline{Y}^{>}_{j_2}\overline{Y}_{j_1}\overline{Y}^{>}_{j_1}\overline{Y}_{j_3} \mid Z\right] - E_{P_0}\left[\overline{Y}^{>}_{j_1}\overline{Y}_{j_2} \mid Z\right] E_{P_0}\left[\overline{Y}^{>}_{j_1}\overline{Y}_{j_3} \mid Z\right]$$
$$= E_{P_0}\left[\overline{Y}_{j_2} \mid Z\right]^{>} E_{P_0}\left[\overline{Y}_{j_1}\overline{Y}^{>}_{j_1} \mid Z\right] E_{P_0}\left[\overline{Y}_{j_3} \mid Z\right] = 0,$$

and for $j_1, j_2, j_3, j_4 \in I_{m,i}$ that are pairwise different, $\mathrm{Cov}_{P_0}\left(\overline{Y}^{>}_{j_1}\overline{Y}_{j_2}, \overline{Y}^{>}_{j_3}\overline{Y}_{j_4} \mid Z\right) = 0$ by the independence of $\overline{Y}_{j_1}, \overline{Y}_{j_2}, \overline{Y}_{j_3}, \overline{Y}_{j_4}$ given $Z$. Further, since $\mathrm{Var}_{P_0}(\overline{Y}^{>}_{j_1}\overline{Y}_{j_2} \mid Z) \le K^2$ for all $j_1 < j_2 \in I_{m,i}$,

$$\mathrm{Var}_{P_0}(T^d_{m,n} \mid Z) \le \frac{1}{n^2}\sum_{i\in I} \frac{2K^2 N_i(N_i - 1)}{N_i^2} \le 2K^2 n^{-2}\sum_{i=1}^{m^{K-1}} I(N_i \ge 2).$$

Thus, by the law of total variance,

$$\mathrm{Var}_{P_0}(T^d_{m,n}) = E_{P_0}[\mathrm{Var}_{P_0}(T^d_{m,n} \mid Z)] + \mathrm{Var}_{P_0}(E_{P_0}[T^d_{m,n} \mid Z]) \le 2K^2 n^{-2}\sum_{i=1}^{m^{K-1}} P_0(N_i \ge 2)$$
$$= 2K^2 m^{K-1} n^{-2} P_0(N_1 \ge 2).$$

Since $N_1$ follows a Binomial distribution with $n$ trials and success probability $m^{-K+1}$, and as $(1-x)^{n-1} \ge 1 - (n-1)x$ for any $x \in [0,1]$, we see that

$$P_0(N_1 \ge 2) = 1 - \left(1 - \frac{1}{m^{K-1}}\right)^{n-1} - \frac{n-1}{m^{K-1}} \le 1 \wedge \frac{n^2}{m^{2(K-1)}}. \tag{16}$$

Therefore, defining $\sigma^2$ below,

$$\mathrm{Var}_{P_0}(T^d_{m,n}) \le 2K^2 \left(m^{K-1} n^{-2} \wedge m^{-(K-1)}\right) =: \sigma^2. \tag{17}$$

Under $P_1$, we have

$$E_{P_1}[T^d_{m,n} \mid Z] = \frac{1}{n}\sum_{i\in I} \frac{1}{N_i}\sum_{j_1\neq j_2\in I_{m,i}} E_{P_1}\left[\overline{Y}^{>}_{j_1}\overline{Y}_{j_2} \mid Z\right]$$
$$= \frac{1}{n}\sum_{i\in I} \frac{1}{N_i}\sum_{j_1\neq j_2\in I_{m,i}} \mathrm{res}_i(Z_{j_1})^{>}\mathrm{res}_i(Z_{j_2});$$

since for each $i \in I$, $E_{P_1}[\overline{Y}_{j_1} > \overline{Y}_{j_2} \mid Z] = E_{P_1}[\overline{Y}_{j_1} \mid Z] > E_{P_1}[\overline{Y}_{j_2} \mid Z] = \mathrm{res}_f(Z_{j_1}) > \mathrm{res}_f(Z_{j_2})$ for all $j_1 \neq j_2 \in I_{m,i}$. Moreover,

$$\begin{aligned} E_{P_1}[T^d_{m,n} \mid N] &= E_{P_1}[E_{P_1}[T^d_{m,n} \mid Z] \mid N] \\ &= \frac{1}{n} \sum_{i \in I} \frac{N_i(N_i - 1)}{N_i} E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i] > E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i] \\ &= \frac{1}{n} \sum_{i \in I} (N_i - 1)\, \| E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i] \|^2 \end{aligned} \tag{18}$$

and

$$\begin{aligned} E_{P_1}[T^d_{m,n}] &= E_{P_1}[E_{P_1}[T^d_{m,n} \mid N]] \\ &= \frac{1}{n} \sum_{i=1}^{m^{K-1}} E_{P_1}[I(N_i - 1)(N_i - 1)]\, \| E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i] \|^2 . \end{aligned}$$

For any $x \in [0,1]$, we have $1 - nx + \frac{n}{2}x^2 - \frac{n}{3}x^3 \leq (1-x)^n \leq 1 - nx + \frac{n}{2}x^2$. Applying the inequality for $x = m^{-(K-1)}$, we derive

$$\frac{1}{4}\left( \frac{n}{m^{K-1}} \wedge \frac{n^2}{m^{2(K-1)}} \right) \leq \frac{n}{2 m^{2(K-1)}} - \frac{n}{3 m^{3(K-1)}} - \frac{n}{m^{K-1}} \leq 1$$

$$E_{P_1}[I(N_i - 1)(N_i - 1)] = \frac{n}{m^{K-1}} \left( 1 + \left( 1 - \frac{1}{m^{K-1}} \right)^n - \frac{n}{2} \frac{1}{m^{2(K-1)}} - \frac{n^2}{m^{2(K-1)}} \right).$$

Also, since $(1-x)^n \geq 1$ for $x \in [0,1]$, we have

$$E_{P_1}[I(N_i - 1)(N_i - 1)] \leq \frac{n}{m^{K-1}}.$$

Therefore,

$$\frac{1}{4}\left( \frac{n}{m^{K-1}} \wedge \frac{n^2}{m^{2(K-1)}} \right) \leq E_{P_1}[I(N_i - 1)(N_i - 1)] \leq \left( \frac{n}{m^{K-1}} \wedge \frac{n^2}{m^{2(K-1)}} \right). \tag{19}$$

By Lemma 15, and as" $= \left( \sum_{k=1}^K E_{P_1}[[\mathrm{res}_f(Z)]_k^2] \right)^{1/2} \leq K^{1/2} \sum_{k=1}^K \left( E_{P_1}[[\mathrm{res}_f(Z)]_k^2] \right)^{1/2}$ by the Cauchy-Schwarz inequality,

$$\begin{aligned} &\sum_{i=1}^{m^{K-1}} m^{-(K-1)} \| E_{P_1}[\mathrm{res}_f(Z) I(Z \in B_i)] \|^2 = \sum_{i=1}^{m^{K-1}} \sum_{k=1}^K m^{-(K-1)} E_{P_1}\left[ [\mathrm{res}_f(Z)]_k^2 I(Z \in B_i) \right] \\ &= \sum_{k=1}^K W_m^2\left[ [\mathrm{res}_f(Z)]_k \right] \leq \sum_{k=1}^K \left( b_1 E_{P_1}\left[ [\mathrm{res}_f(Z)]_k^2 \right]^{\frac{1}{2}} + b_2 m^{-s} \right)^2 \\ &\leq \sum_{k=1}^K b_1^2 E_{P_1}\left[ [\mathrm{res}_f(Z)]_k^2 \right] + 2 b_1 b_2 E_{P_1}\left[ [\mathrm{res}_f(Z)]_k^2 \right]^{\frac{1}{2}} m^{-s} + b_1^2 "^2 + 2 \sqrt{K} b_1 b_2 m^{-s} ". \end{aligned} \tag{20}$$

By (19) and (20), de ning below,

$$E_{P_1}[T^d_{m;n}] \quad \frac{1}{4}(b_1^{2"2} \quad 2^p \overline{K} b_1 b_2 m^{s"}) \quad 1 \wedge \frac{n}{m^{K\ 1}} =: \ : \tag{21}$$

Moreover, we nd

$$\mathrm{Var}_{P_1}(T^d_{m;n} \mid N) = \frac{1}{n^2} \sum_{i \in I} \frac{1}{N_i^2} \mathrm{Var}_{P_1}\Big(2 \sum_{j_1<j_2 \in I_{m;i}} \overline{Y}^{>}_{j_1} \overline{Y}_{j_2} \mid N\Big)$$

$$= \frac{1}{n^2} \sum_{i \in I} \frac{1}{N_i^2} \Big[ 2N_i(N_i \quad 1)\mathrm{Var}_{P_1}\big(\overline{Y}^{>}_1 \overline{Y}_2 \mid Z_1; Z_2 \in B_i\big) + 4N_i(N_i \quad 1)(N_i \quad 2)\mathrm{Cov}_{P_1}\big(\overline{Y}^{>}_1 \overline{Y}_2; \overline{Y}^{>}_1 \overline{Y}_3 \mid Z_1; Z_2; Z_3 \in B_i\big)\Big]$$

$$\frac{1}{n^2} \sum_{i \in I} \frac{2K^2 N_i(N_i \quad 1)}{N_i^2} + \frac{4K^2 N_i(N_i \quad 1)(N_i \quad 2)}{N_i^2} k E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i]k^2 \ :$$

Further, by equations (16), (19) and since $kE_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i]k^2 \quad E_{P_1}[k\mathrm{res}_f(Z)k^2 \mid Z \in B_i]$,

$$E_{P_1}[\mathrm{Var}_{P_1}(T^d_{m;n} \mid N)]$$

$$\frac{1}{n^2} \sum_{i=1}^{m^{K\ 1}} 2K^2 P_1(N_i \quad 2) + 4K^2 E_{P_1}[(N_i \quad 1)I(N_i \quad 1)] k E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i]k^2$$

$$^2 + 4K^2 m^{K\ 1} n^{\ 2} E_{P_1}[(N_1 \quad 1)I(N_i \quad 1)] \sum_{i=1}^{m^{K\ 1}} E_{P_1}[k\mathrm{res}_f(Z)k^2 I(Z \in B_i)]$$

$$^2 + 4K^{2"2}(n^{\ 1} \wedge m^{\ (K\ 1)}): \tag{22}$$

Also, from (18),

$$\mathrm{Var}_{P_1}(E_{P_1}[T^d_{m;n} \mid N]) \quad \frac{1}{n^2}\Big( \sum_{i=1}^{m^{K\ 1}} \sqrt{\mathrm{Var}_{P_1}[I(N_i \quad 1)(N_i \quad 1)]}\, k E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i]k^2 \Big)^2$$

$$= \frac{1}{n^2}\mathrm{Var}_{P_1}[I(N_1 \quad 1)(N_1 \quad 1)] m^{2(K\ 1)"4}:$$

Writing $x = m^{\ (K\ 1)}$, we have

$$\mathrm{Var}_{P_1}[I(N_1 \quad 1)(N_1 \quad 1)] = nx(1 \quad x) + (1 \quad x)^n \quad (1 \quad x)^{2n} \quad 2nx(1 \quad x)^n:$$

Using that $1 \quad nx \quad (1 \quad x)^n$, we nd

$$\mathrm{Var}_{P_1}[I(N_i \quad 1)(N_i \quad 1)] = nx(1 \quad x) + (1 \quad x)^n[1 \quad (1 \quad x)^n \quad 2nx]$$

$$nx(1 \quad x) \quad nx = \frac{n}{m^{K\ 1}}:$$

Similarly,

$$\mathrm{Var}_{P_1}[I(N_i \quad 1)(N_i \quad 1)] \quad nx + (1 \quad x)^{n \quad 1}[1 \quad (1 \quad x)^n \quad 2nx]$$
$$nx + (1 \quad (n \quad 1)x)(\quad nx) \quad n^2x^2 = \frac{n^2}{m^{2(K \quad 1)}}:$$

Therefore,

$$\mathrm{Var}_{P_1}(E_{P_1}[T^d_{m\ ;n} \mid N]) \qquad \frac{m^{K \quad 1}}{n} \wedge 1 \quad "^4: \tag{23}$$

By equations (22), (23), and the law of total variance, de ning $^2$ below,

$$\mathrm{Var}_{P_1}(T^d_{m\ ;n}) = \mathrm{Var}_{P_1}(E_{P_1}[T^d_{m\ ;n} \mid N]) + E_{P_1}[\mathrm{Var}_{P_1}(T^d_{m\ ;n} \mid N)]$$
$$^2 + 4K^2"^2(n^{\ 1} \wedge m^{\ (K \quad 1)}) + \frac{m^{K \quad 1}}{n} \wedge 1 \quad "^4 =: \quad ^2: \tag{24}$$

Recalling that $m = \lfloor n^{2=(4s+K \quad 1)} c$, we choose $c > 0$ such that $" \quad cn^{\ 2s=(4s+K \quad 1)}$ implies

$$\left(\frac{\sqrt{2}}{} + \frac{\sqrt{2}}{}\right) K(m^{\frac{K \quad 1}{2}} n^{\ 1} \wedge m^{\frac{K \quad 1}{2}}) + \frac{\sqrt{4}}{}K"^{\quad} n^{\frac{1}{2}} \wedge m^{\frac{K \quad 1}{2}} + \frac{\sqrt{1}}{}\left(\frac{m^{K \quad 1}}{n} \wedge 1\right)^{\frac{1}{2}} "^2$$
$$\frac{1}{4}(b_1^2"^2 \quad 2\sqrt{K}b_1b_2m^{\ s}") \quad 1 \wedge \frac{n}{m^{K \quad 1}}$$

for all large enough $n$. For , , and from (17), (21), and (24), this gives

$$\overline{p}_= + \overline{p}_= \quad : \tag{25}$$

By equations (15), (17), the de nition of $_{m\ ;n}$ from Algorithm (1), and Chebyshev's inequality,

$$P_0(\ _{m\ ;n} = 1) \quad P_0 \quad T^d_{m\ ;n} \quad \overline{p}_= \quad \frac{\mathrm{Var}_{P_0}(T^d_{m\ ;n})}{^2_=} \quad : \tag{26}$$

By equations (21), (24), (25), and Chebyshev's inequality,

$$P_1 \quad T^d_{m\ ;n} < \overline{p}_= \quad P_1 \quad T^d_{m\ ;n} \quad E_{P_1}[T^d_{m\ ;n}] \quad \overline{p}_=$$
$$P_1 \quad T^d_{m\ ;n} \quad E_{P_1}[T^d_{m\ ;n}] \quad \overline{p}_= \quad \frac{\mathrm{Var}_{P_1}(T^d_{m\ ;n})}{^2_=} \quad : \tag{27}$$

By the above arguments, Theorem 3 holds for all $n \quad N$, where $N \ 2 \ N_+$ depends on $(s; L; K; \ _l; \ _u; \ ; \ )$: If we require $c$ to further satisfy $c \quad N^{\ 2s=(4s+K \quad 1)}$, then the family $P_1(";p;s)$ is empty for $n < N$ given $" \quad cn^{\ 2s=(4s+K \quad 1)} > 1$. Therefore, Theorem 3 becomes vacuously true for $n < N$, and thereby true for all $n \ 2 \ N$. This nishes the proof.

■

A.2 Proof of Lemma 15

We state and prove Lemma 16 and 17 which we use in the proof of Lemma 15.

Lemma 16 Fix $h \in H_K(s, L)$ and $z_0 \in \mathcal{X}_{K-1}$. Let $u$ be the $(\lceil s \rceil - 1)$-th order Taylor series of $h$ at $z_0$. There is $L'$ depending on $(K, s, L)$ such that

$$|h(z) - u(z)| \le L' \|z - z_0\|^s \tag{28}$$

for all $z \in \mathcal{X}_{K-1}$.

Proof Let $\Phi = \Phi_K$ as in (62). Recall that for a multi-index $\alpha = (\alpha_1, \ldots, \alpha_{K-1})^T \in \mathbb{N}^{K-1}$, we write $\Phi^\alpha(z) - \Phi^\alpha(z_0)) = \prod_{j \in [K-1]} (\Phi_j(z) - \Phi_j(z_0))^{\alpha_j}$. By a Taylor series expansion, there exists $t \in [0, 1]$ such that

$$\begin{aligned} h(z) - (h \circ \Phi^{-1})(\Phi(z)) = & \sum_{\substack{\alpha \in \mathbb{N}^{K-1} \\ 0 \le |\alpha| \le \lceil s \rceil - 2}} \frac{(h \circ \Phi^{-1})^{(\alpha)}(\Phi(z_0))}{\alpha!} (\Phi(z) - \Phi(z_0))^\alpha \\ & + \sum_{\substack{\alpha \in \mathbb{N}^{K-1} \\ |\alpha| = \lceil s \rceil - 1}} \frac{(h \circ \Phi^{-1})^{(\alpha)}(t\Phi(z) + (1-t)\Phi(z_0))}{\alpha!} (\Phi(z) - \Phi(z_0))^\alpha. \end{aligned}$$

Then, $h(z) - u(z)$ equals

$$\sum_{\substack{\alpha \in \mathbb{N}^{K-1} \\ |\alpha| = \lceil s \rceil - 1}} \frac{(h \circ \Phi^{-1})^{(\alpha)}(t\Phi(z) + (1-t)\Phi(z_0)) - (h \circ \Phi^{-1})^{(\alpha)}(\Phi(z_0))}{\alpha!} (\Phi(z) - \Phi(z_0))^\alpha.$$

By the triangle inequality and the $s$-Hölder continuity of $h$,

$$\begin{aligned} & |h(z) - u(z)| \\ & \le \sum_{\substack{\alpha \in \mathbb{N}^{K-1} \\ |\alpha| = \lceil s \rceil - 1}} \frac{|(h \circ \Phi^{-1})^{(\alpha)}(t\Phi(z) + (1-t)\Phi(z_0)) - (h \circ \Phi^{-1})^{(\alpha)}(\Phi(z_0))|}{\alpha!} |(\Phi(z) - \Phi(z_0))^\alpha| \\ & \le \sum_{\substack{\alpha \in \mathbb{N}^{K-1} \\ |\alpha| = \lceil s \rceil - 1}} \frac{L(t\|\Phi(z) - \Phi(z_0)\|)^{s - \lceil s \rceil + 1}}{\alpha!} \|\Phi(z) - \Phi(z_0)\|^{\lceil s \rceil - 1} \\ & \le L' \|\Phi(z) - \Phi(z_0)\|^s \le L' \|z - z_0\|^s. \end{aligned}$$

■

Lemma 17 Let $P_q^K$ be the class of polynomials on $\mathbb{R}^K$ of degree at most $q$. There are $a_1, a_2 > 0$ depending on $(K, q)$ such that

$$\|W_m[v]\|_{L^2(\mu)} \ge a_1 \|v\|_{L^2(\mu)} \tag{29}$$

for every $v \in P_q^K$ and $m \ge a_2$.

Proof If (29) does not hold, then we can find a sequence $\{m_l\}_{l=1}^{\infty}$ increasing to infinity and a sequence of polynomials $\{v_l\}_{l=1}^{\infty} \subset P_q^K$ such that $\|W_{m_l}[v_l]\|_{L^2(\mu)} < \frac{1}{l}\|v_l\|_{L^2(\mu)}$. Dividing $v_l$ by $\|v_l\|_{L^2(\mu)}$, we may assume $\|v_l\|_{L^2(\mu)} = 1$. Now $\{v \in P_q^K : \|v\|_{L^2(\mu)} = 1\}$ is compact in the topology induced by the norm $\|\cdot\|_{L^2(\mu)}$, due to the Heine-Borel theorem because it is closed and bounded. Thus, we can find a convergent subsequence $\{v_{l_k}\}_{k=1}^{\infty}$. Denote the limit by $v_1$. On one hand,

$$\|W_{m_{l_k}}[v_1]\|_{L^2(\mu)} \le \|W_{m_{l_k}}[v_1 - v_{l_k}]\|_{L^2(\mu)} + \|W_{m_{l_k}}[v_{l_k}]\|_{L^2(\mu)} \le \|v_1 - v_{l_k}\|_{L^2(\mu)} + \frac{1}{l_k} \to 0.$$

Here, we used that

$$\|W_m[v]\|_{L^2(\mu)}^2 = \sum_{i=1}^{m^{K-1}} \frac{\left(\int_{B_i} v(z)\,d\mu(z)\right)^2}{\mu(B_i)} \le \sum_{i=1}^{m^{K-1}} \int_{B_i} v(z)^2 d\mu(z) = \|v\|_{L^2(\mu)}^2. \tag{30}$$

On the other hand, $\|W_{m_{l_k}}[v_1]\|_{L^2(\mu)} \to \|v_1\|_{L^2(\mu)} = 1$ since $W_{m_{l_k}}[v_1] \to v_1$ a.e. and $\|W_{m_{l_k}}[v_1]\|_{L^1} \le \|v_1\|_{L^1}$. The conclusion follows due to the contradiction. ∎

Now we proceed with the proof of Lemma 15. Since $a_1$ does not depend on $m$, for sufficiently large $m$, we can choose $r$ such that $m/r$ is an integer and $m/r \ge a_2$. Partition $[0,1]^{K-1}$ into $M := (m/r)^{K-1}$ simplices as described in Appendix B.3 and call them $\tilde{B}_1, \ldots, \tilde{B}_M$. By this construction, we can ensure that each $\tilde{B}_j$, $j \in [M]$, consists of $r^{K-1}$ different simplices $B_i$ (also defined in Appendix B.3). For $j \in [M]$, let $u_j$ be the $(d \vee 1)$-th order Taylor expansion of $h$ at an arbitrary vertex of $\tilde{B}_j$. Define $u := \sum_{j=1}^{M} u_j 1_{\tilde{B}_j}$. By equation (28),

$$|h(z) - u(z)| \le L^0 \operatorname{diam}(\tilde{B}_1)^s = L^0 \operatorname{diam}([0,1]^{K-1})^s \left(\frac{r}{m}\right)^s =: b m^{-s} \tag{31}$$

for all $z \in [0,1]^{K-1}$. Therefore, by (30) and Lemma 16,

$$\begin{aligned} \|W_m[h]\|_{L^2(\mu)} &\le \|W_m[u]\|_{L^2(\mu)} + \|W_m[u-h]\|_{L^2(\mu)} \\ &\le \|W_m[u]\|_{L^2(\mu)} + \|u-h\|_{L^2(\mu)} \le \|W_m[u]\|_{L^2(\mu)} + b m^{-s}. \end{aligned}$$

Note that

$$\|W_m[u]\|_{L^2(\mu)}^2 = \sum_{j=1}^{M} \|W_m[u_j 1_{\tilde{B}_j}]\|_{L^2(\mu)}^2,$$

and that Lemma 17 with $q = d \vee 1$ can be applied to $\tilde{B}_j$ and its $r^{K-1}$ sub-simplices to get

$$\|W_m[u_j 1_{\tilde{B}_j}]\|_{L^2(\mu)}^2 \le a_1^2 \|u_j 1_{\tilde{B}_j}\|_{L^2(\mu)}^2.$$

Thus,

$$\|W_m[u]\|_{L^2(\mu)}^2 \le a_1^2 \sum_{j=1}^{M} \|u_j 1_{\tilde{B}_j}\|_{L^2(\mu)}^2 = a_1^2 \|u\|_{L^2(\mu)}^2.$$

In conclusion, combining the above inequalities, and by (31)

$$\|W_m[h]\|_{L^2(\Pi)} \le a_1\|u\|_{L^2(\Pi)} + bm^{-s} \le a_1(\|h\|_{L^2(\Pi)} + bm^{-s}) + bm^{-s} =: b_1\|h\|_{L^2(\Pi)} + b_2 m^{-s}.$$

This finishes the proof of Lemma 15.

### A.3 Proof of Remark 4

We follow the same strategy in Appendix A.1. For null distributions $P_0$ and alternative distributions $P_1$ such that $\ell_p\text{-ECE}_{P_0}(f) \ge c_0 n^{-2s/(4s+K-1)}$ and $\varepsilon := \ell_p\text{-ECE}_{P_1}(f) \ge c_1 n^{-2s/(4s+K-1)}$, we show the mean difference $E_{P_1}[T^d_{m,n}] - E_{P_0}[T^d_{m,n}]$ is larger than $\mathrm{Var}_{P_0}(T^d_{m,n})^{1/2}$ and $\mathrm{Var}_{P_1}(T^d_{m,n})^{1/2}$.

While $E_{P_0}[T^d_{m,n}] = 0$ under the null hypothesis of perfect calibration, we now have

$$\begin{aligned} E_{P_0}[T^d_{m,n}] &= \frac{1}{n}\sum_{i=1}^{m^{K-1}} E_{P_0}[|(N_i - 1)(N_i - 2)|]\,|E_{P_0}[\mathrm{res}_f(Z) \mid Z \in B_i]|^2 \\ &\ge (m^{-(K-1)} \wedge m^{-2(K-1)}n)\sum_{i=1}^{m^{K-1}} E_{P_0}[|\mathrm{res}_f(Z)|^2 \mid Z \in B_i] \\ &\ge c_0^2 n^{-\frac{4s}{4s+K-1}}(1 \wedge m^{-(K-1)}n). \end{aligned} \tag{32}$$

Therefore,

$$E_{P_1}[T^d_{m,n}] - E_{P_0}[T^d_{m,n}] \ge c_0^2 n^{-\frac{4s}{4s+K-1}}(1 \wedge m^{-(K-1)}n) =: \Delta_0.$$

By the equation (24),

$$\mathrm{Var}_{P_0}(T^d_{m,n}) \le 2 + 5K^2 c_0^2 n^{-\frac{4s}{4s+K-1}}(n^{-1} \wedge m^{-(K-1)}) =: (\sigma_0)^2.$$

Similar to (25), we can choose large enough $b_1 > 0$ such that

$$\Phi\Big(-\frac{\Delta_0}{\sigma_0} + \Phi^{-1}(\alpha)\Big) \le \alpha.$$

The conclusion follows from Chebyshev's inequality as in (26) and (27).

### A.4 Proof of Theorem 5

By the union bound, for $P \in \mathcal{P}_0$,

$$P(\phi^{\mathrm{ad}}_n = 1) \le \sum_{b=1}^{B} P\Big(T_{2^b,n} \ge \frac{\alpha}{B}\Big) = \sum_{b=1}^{B} \frac{\alpha}{B} = \alpha.$$

There exists $b_0 \in \{1, \dots, B\}$ such that $2^{b_0-1} < (n/\log n)^{2/(4s+K-1)} \le 2^{b_0}$. Let $m_0 = 2^{b_0}$ and repeat the argument in the proof of Theorem 3. The condition (25) for type II error control is now changed to

$$\sqrt{\frac{2}{\alpha}}K m_0^{\frac{K-1}{2}} n^{-1} \sqrt{\log\tfrac{1}{\alpha}} + \sqrt{\frac{7}{\alpha}}K m_0^{\frac{K-1}{2}} n^{-1} \le \frac{1}{4}(b_1\varepsilon/2 - 2\sqrt{K}b_1 b_2 m_0^{-s}\varepsilon);$$

which is satisfied when $\varepsilon \geq c_{ad}(n=\sqrt{\log n})^{-2s/(4s+K-1)}$ for a sufficiently large $c_{ad} > 0$. Assuming $\varepsilon \geq c_{ad}(n=\sqrt{\log n})^{-2s/(4s+K-1)}$ and $P \in P_1(\varepsilon, p, s)$, we have

$$P(\hat{\alpha}^{ad}_n = 1) \geq P(\alpha_{m_0,n} = 1) \geq 1 - \cdots.$$

This finishes the proof.

## A.5 Proof of Proposition 7

Overview of the proof. We repeat the computation in Appendix A.1. However, due to the bias term, now the mean difference $E_{P_1}[T^b_{m,n}] - E_{P_0}[T^b_{m,n}]$ cannot be lower bounded by a positive number. Instead, we prove that $E_{P_0}[T^b_{m,n}] \leq E_{P_1}[T^b_{m,n}]$ holds for all large enough $n$.

Proof We use the same notations as in Appendix A.1. Since

$$E_{P_0}[T^b_{m,n} \mid Z] = \frac{1}{n}\sum_{i \in I} \left[\frac{1}{N_i}\left[\sum_{j \in I_{m,i}} E_{P_0}\left[\overline{Y}_j^2 \mid Z\right] + \sum_{j_1 \neq j_2 \in I_{m,i}} E_{P_0}\left[\overline{Y}_{j_1}\overline{Y}_{j_2} \mid Z\right]\right]\right]$$
$$= \frac{1}{n}\sum_{i \in I}\left[\frac{1}{N_i}\sum_{j \in I_{m,i}} Z_j - Z_j^2\right],$$

we have

$$E_{P_0}[T^b_{m,n} \mid N] = E_{P_0}[E_{P_0}[T^b_{m,n} \mid Z] \mid N] = \frac{1}{n}\sum_{i \in I} E_{P_0}\left[Z - Z^2 \mid Z \in B_i\right]. \tag{33}$$

Similarly,

$$E_{P_1}[T^b_{m,n} \mid Z] = \frac{1}{n}\sum_{i \in I}\left[\frac{1}{N_i}\left[\sum_{j \in I_{m,i}} E_{P_1}\left[\overline{Y}_j^2 \mid Z\right] + \sum_{j_1 \neq j_2 \in I_{m,i}} E_{P_1}\left[\overline{Y}_{j_1}\overline{Y}_{j_2} \mid Z\right]\right]\right]$$
$$= \frac{1}{n}\sum_{i \in I}\left[\frac{1}{N_i}\left[\sum_{j \in I_{m,i}} (\mathrm{reg}_f(Z_j) - \mathrm{reg}_f(Z_j)^2 + \mathrm{res}_f(Z_j)^2) + \sum_{j_1 \neq j_2 \in I_{m,i}} \mathrm{res}_f(Z_{j_1})\mathrm{res}_f(Z_{j_2})\right]\right];$$

and thus

$$E_{P_1}[T^b_{m,n} \mid N] = E_{P_1}[E_{P_1}[T^b_{m,n} \mid Z] \mid N] \tag{34}$$
$$= \frac{1}{n}\sum_{i \in I} E_{P_1}\left[\mathrm{reg}_f(Z) - \mathrm{reg}_f(Z)^2 + \mathrm{res}_f(Z)^2 \mid Z \in B_i\right] + (N_i - 1)E_{P_1}[\mathrm{res}_f(Z) \mid Z \in B_i]^2.$$

Since $Z \sim \mathrm{Unif}([0,1])$ under both $P_0$ and $P_1$, the equations (33) and (34) imply

$$E_{P_0}[T^b_{m,n} \mid N] - E_{P_1}[T^b_{m,n} \mid N]$$
$$= \frac{1}{n}\sum_{i \in I} E[\mathrm{res}_f(Z)(2Z-1) \mid Z \in B_i] - (N_i - 1)E[\mathrm{res}_f(Z) \mid Z \in B_i]^2$$
$$\leq \frac{1}{n}\sum_{i \in I} E[\mathrm{res}_f(Z)(2Z-1) \mid Z \in B_i] \leq 2k\left(k^2_{L_1} m\right)^{-2s}.$$

Here we used that $E[\mathrm{res}_i(Z) \mid Z \in B_i]^2 \le 2k\, k_{L-1} m^{-2s}$ for all $i \in [m]$ and $\sum_{i \in I} (N_i - 1) \le n$. Taking total expectation,

$$E_{P_0}[T^b_{m,n}] - E_{P_1}[T^b_{m,n}] \le \frac{1}{n} E\Big[\sum_{i \in I} E[\mathrm{res}_i(Z)(2Z-1) \mid Z \in B_i]\Big] \le 2k\, k_{L-1} m^{-2s}$$

$$= \frac{1}{n}\sum_{i=1}^{m} P(N_i - 1) E[\mathrm{res}_i(Z)(2Z-1) \mid Z \in B_i] \le 2k\, k_{L-1} m^{-2s}$$

$$= \frac{1}{n} P(N_1 - 1) \sum_{i=1}^{m} E[\mathrm{res}_i(Z)(2Z-1) \mid Z \in B_i] \le 2k\, k_{L-1} m^{-2s}. \tag{35}$$

From (9), we see that $E[\mathrm{res}_i(Z)(2Z-1) \mid Z \in B_i] = 0$ for all $i \in [m]$. Thus,

$$\sum_{i=1}^{m} E[\mathrm{res}_i(Z)(2Z-1) \mid Z \in B_i] \ge \sum_{i=\frac{m}{4}+1}^{\frac{m}{2}} E[\mathrm{res}_i(Z)(2Z-1) \mid Z \in B_i]$$

$$\frac{1}{4} \sum_{i=\frac{m}{4}+1}^{\frac{m}{2}} E[|\mathrm{res}_i(Z)| \mid Z \in B_i] \ge \frac{1}{32} k\, k_{L-1} m^{1-s}. \tag{36}$$

Combining (35) and (36), we find

$$E_{P_0}[T^b_{m,n}] - E_{P_1}[T^b_{m,n}] \ge \frac{1}{32} P(N_1 - 1)\, k\, k_{L-1} m^{1-s} n^{-1} \ge 2k\, k_{L-1} m^{-2s}. \tag{37}$$

Since $m = \lfloor n^{2/(4s+1)} \rfloor$ and $\frac{2}{4s+1} < 1$, we find

$$\lim_{n \to \infty} P(N_1 - 1) = \lim_{n \to \infty} 1 - \Big(1 - \frac{1}{m}\Big)^n = 1.$$

Also, we have $m^{1-s} n^{-1} \ge n^{(1-6s)/(4s+1)}$ and $m^{-2s} \ge n^{-4s/(4s+1)}$ with $\frac{1-6s}{4s+1} > \frac{-4s}{4s+1}$. In conclusion, the RHS of (37) is positive for all large enough $n$.

■

### A.6 Ingster's method

**Lemma 18 (Ingster's method for the lower bound)** Let $P_0 \in \mathcal{P}_0$ and $P_1, \ldots, P_M \in \mathcal{P}_1(\varepsilon; p, s)$ be probability distributions on $\mathcal{X}$, and suppose that $P_1, \ldots, P_M$ are absolutely continuous with respect to $P_0$. For an i.i.d. sample $\{(Z_i, Y_i) : i \in \{1, \ldots, n\}\}$ from $P_0$, define the average likelihood ratio between $P_1, \ldots, P_M$ and $P_0$ as

$$L_n := \frac{1}{M} \sum_{i=1}^{M} \prod_{j=1}^{n} \frac{dP_i}{dP_0}(Z_j, Y_j).$$

If $E_{P_0}[L_n^2] \le 1 + (1+\gamma)^2$, then the minimax type II error (false negative rate) for testing $H_0 : P \in \mathcal{P}_0$ against $H_1 : P \in \mathcal{P}_1(\varepsilon; p, s)$ at level $\alpha$ satisfies $R_n(\varepsilon; p, s) \ge \gamma$ and the minimum separation rate to ensure type II error at most $\gamma$ obeys $\varepsilon_n(\gamma; p, s) \ge \varepsilon$.

The proof follows from the results of Ingster (1987, 2012); see also Lemma G.1 in Kim et al. (2022) for a very clear statement. By de nition, it holds that

$$\begin{aligned} R_n(\varepsilon; p, s) &= \inf_{\phi \in \Phi_n(\alpha)} \sup_{P \in \mathcal{P}_1(\varepsilon;p,s)} E_P[1 - \phi] \geq \inf_{\phi \in \Phi_n(\alpha)} \frac{1}{M} \sum_{i=1}^{M} E_{P_i}[1 - \phi] \\ &= \inf_{\phi \in \Phi_n(\alpha)} E_{P_0}[1 - \phi] + \frac{1}{M} \sum_{i=1}^{M} \left( E_{P_i}[1 - \phi] - E_{P_0}[1 - \phi] \right) \\ &\geq 1 - \alpha + \inf_{\phi \in \Phi_n(\alpha)} \frac{1}{M} \sum_{i=1}^{M} \left( E_{P_i}[1 - \phi] - E_{P_0}[1 - \phi] \right), \end{aligned}$$

where the last inequality holds because $E_{P_0}[\phi] \leq \alpha$. Further,

$$\frac{1}{M} \sum_{i=1}^{M} E_{P_i}[1 - \phi] - E_{P_0}[1 - \phi] = E_{P_0}[\phi] - \frac{1}{M} \sum_{i=1}^{M} E_{P_i}[\phi] = |E_{P_0}[\phi] - E_{P_0}[\phi L_n]| \leq \sqrt{E_{P_0}[|L_n - 1|]} \sqrt{E_{P_0}[L_n^2] - 1},$$

by a change of variables and the Cauchy-Schwarz inequality. Therefore, we have $R_n(\varepsilon; p, s) \geq 1 - \alpha - (1 - \alpha) = \beta$. Finally, since $\varepsilon \mapsto R_n(\varepsilon; p, s)$ is non-increasing, we find $\varepsilon_n(p, s) \geq \varepsilon$.

## A.7 Proof of Proposition 9

**Overview of the proof.** We construct distributions $P_1, \ldots, P_M$ over $(Z, Y)$ under which the predictor $f$ has an $\ell_p$-ECE of at least $\varepsilon_0 = 0.1$. We can choose the mis-calibration curves of $P_1, \ldots, P_M$ to be orthogonal in $L^2$, so that the cross terms in the expansion of $E_{P_0}[L_n^2]$ cancel out. By choosing $M$ sufficiently large, we can ensure that $E_{P_0}[L_n^2]$ is at most $1 + (1 - \alpha)^2$. The conclusion follows from Lemma 18.

**Proof** We prove Proposition 9 for the binary case. The generalization to the multi-class case follows the same argument and is omitted. The construction in this proof is inspired by Ingster (1987, 2000); Burnashev (1979). Let $P_0$ be a null distribution over $(Z, Y) \in [0, 1] \times \{0, 1\}$ de ned as follows: the distribution of the predicted probabilities follows $Z \overset{P_0}{\sim} \mathrm{Unif}([0, 1])$ and $P_0(Y = 1 \mid Z = z) = z$ for all $z \in [0, 1]$. Under $P_0$, the probability predictor $f$ is perfectly calibrated. For each $i \in [M]$, let

$$g_i(u) := \begin{cases} u + \sqrt{\frac{u(1-u)}{3}} \sin(2i\pi(u - \frac{1}{4})) & u \in [\frac{1}{4}, \frac{3}{4}], \\ u & u \notin [\frac{1}{4}, \frac{3}{4}], \end{cases}$$

and de ne $P_i$ as follows: $Z \overset{P_i}{\sim} \mathrm{Unif}([0, 1])$ and $P_i(Y = 1 \mid Z = z) = g_i(z)$ for all $z \in [0, 1]$.

It can be verified that $0 \le g_i(u) \le 1$ for all $u \in [0,1]$. Since $p \ge 1$, for all $i \in [M]$, the $\ell_p$-ECE of the probability predictor $f$ under $P_i$ is lower bounded as

$$\begin{aligned}\ell_p\text{-ECE}_{P_i}(f) &\ge \ell_1\text{-ECE}_{P_i}(f) = 2\int_0^1 |g_i(u) - u|\,du \\ &= 2\int_{\frac{1}{4}}^{\frac{3}{4}} \sqrt{\frac{u(1-u)}{3}}\,\sin\left(2i\left(u - \frac{1}{4}\right)\right)du \le 0.1.\end{aligned}$$

Thus we know that $P_i \in P_1^{\text{cont}}(\varepsilon_0, p)$ for all $i \in [M]$. Now, observe that

$$L_n = \frac{1}{M}\sum_{i=1}^{M}\prod_{j=1}^{n}\frac{dP_i}{dP_0}(Z_j, Y_j) = \frac{1}{M}\sum_{i=1}^{M}\prod_{j=1}^{n}\frac{1 - Y_j + (2Y_j - 1)g_i(Z_j)}{1 - Y_j + (2Y_j - 1)Z_j}$$

and thus, for a random variable $(Z, Y) \sim P_0$, and defining $A_{a,b}$ below

$$\begin{aligned}E_{P_0}[L_n^2] &= \frac{1}{M^2}\sum_{a,b\in[M]} E_{P_0}\left[\prod_{j=1}^{n}\frac{1 - Y_j + (2Y_j - 1)g_a(Z_j)}{1 - Y_j + (2Y_j - 1)Z_j}\cdot\frac{1 - Y_j + (2Y_j - 1)g_b(Z_j)}{1 - Y_j + (2Y_j - 1)Z_j}\right] \\ &= \frac{1}{M^2}\sum_{a,b\in[M]} E_{P_0}\left[\frac{1 - Y + (2Y - 1)g_a(Z)}{1 - Y + (2Y - 1)Z}\cdot\frac{1 - Y + (2Y - 1)g_b(Z)}{1 - Y + (2Y - 1)Z}\right]^n \\ &=: \frac{1}{M^2}\sum_{a,b\in[M]} E_{P_0}[A_{a,b}]^n.\end{aligned}$$

In the second line, we have used the independence of the observations. If $a = b$, then

$$\begin{aligned}E_{P_0}[A_{a,b}] &= \int_0^1 u\frac{g_a(u)^2}{u^2} + (1-u)\frac{(1 - g_a(u))^2}{(1-u)^2}du = 1 + \int_0^1 \frac{(g_a(u) - u)^2}{u(1-u)}du \\ &= 1 + \int_{\frac{1}{4}}^{\frac{3}{4}} \frac{1}{3}\sin^2\left(2a\left(u - \frac{1}{4}\right)\right)du = \frac{13}{12}.\end{aligned}$$

If $a \ne b$, then

$$\begin{aligned}E_{P_0}[A_{a,b}] &= \int_0^1 u\frac{g_a(u)g_b(u)}{u^2} + (1-u)\frac{(1 - g_a(u))(1 - g_b(u))}{(1-u)^2}du \\ &= 1 + \int_0^1 \frac{(g_a(u) - u)(g_b(u) - u)}{u(1-u)}du \\ &= 1 + \int_{\frac{1}{4}}^{\frac{3}{4}} \frac{1}{3}\sin\left(2a\left(u - \frac{1}{4}\right)\right)\sin\left(2b\left(u - \frac{1}{4}\right)\right)du = 1.\end{aligned}$$

Therefore,

$$E_{P_0}[L_n^2] = \frac{1}{M^2}\left(M - \left(\frac{13}{12}\right)^n + (M^2 - M)\right).$$

Choose a large enough $M \in \mathbb{N}_+$ such that $M \ge [(13/12)^n - 1]/(1 - \varepsilon)^2$. Then,

$$\frac{1}{M^2}\left(M - \left(\frac{13}{12}\right)^n + (M^2 - M)\right) \le 1 - (1 - \varepsilon)^2,$$

and the result follows by Lemma 18.

■

### A.8 Proof of Theorem 10

**Overview of the proof.** We construct $m^{K-1}$ distributions under which the predictor $f$ has an $\ell_p$-ECE of $(n^{-2s/(4s+K-1)})$. The mis-calibration curves are constructed by linearly combining bump functions with disjoint supports. By properly scaling them, we can guarantee Hölder continuity. Also, the mis-calibration curves are chosen to be "almost" orthogonal in $L^2$, so that the cross terms in the expansion of $\mathbb{E}_{P_0}[L_n^2]$ are small. We use Lemma 18 to conclude.

**Proof** The proof is inspired by the lower bound arguments in Arias-Castro et al. (2018). For $m := \lceil n^{2/(4s+K-1)} \rceil$ and $\omega \in \{-1,1\}^{[m]^{K-1}}$, we define alternative distributions $P_\omega \in \mathcal{P}_1(\varepsilon, p, s)$ with $\varepsilon := c_{\text{lower}} n^{-2s/(4s+K-1)}$ and use Lemma 18 to prove $\mathcal{C}_n(p,s) \geq c_{\text{lower}} n^{-2s/(4s+K-1)}$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be the function from (8). It can be verified that $\psi$ is infinitely differentiable and its derivatives of every order are bounded. For $x = (x_1, \ldots, x_K)^\top \in (z_1, \ldots, z_{K-1})^\top$, we see that $[\frac{1}{2K}, \frac{1}{K}]^{K-1} \subset (x_1, \ldots, x_{K-1}) \cap [\frac{1}{2K}, 1]^K)$. For each $j = (j_1, \ldots, j_{K-1})^\top \in [m]^{K-1}$, define $\psi_j : \mathbb{R}^{K-1} \to \mathbb{R}$ by

$$\psi_j(x_1, \ldots, x_{K-1}) := m^{-s} \prod_{k=1}^{K-1} \psi(2Kmx_k - 1) - j_k + 1) \,. \tag{38}$$

Then, each $\psi_j$ is supported on the cube

$$\operatorname{supp}(\psi_j) = \prod_{k=1}^{K-1} \left[ \frac{j_k - 1 + m}{2Km}, \frac{j_k + m}{2Km} \right] \,.$$

The sets $\operatorname{supp}(\psi_j)$ are disjoint for different indices $j \in [m]^{K-1}$, and we have

$$\bigcup_{j \in [m]^{K-1}} \operatorname{supp}(\psi_j) \subset \left[\frac{1}{2K}, \frac{1}{K}\right]^{K-1} \subset (x_1, \ldots, x_{K-1}) \cap \left[\frac{1}{2K}, 1\right]^K \Big) \,.$$

Let $c_{\psi,\alpha} := \log\left(1 + (1 - 2^{-\alpha})^{2^{-1/4}}\right)$ and

$$\gamma := \max_{t \in \{0,\ldots,d_{\text{se}}\}} \|\psi^{(t)}\|_{L^1}^{-1} \wedge \frac{1}{2K} \wedge \frac{L(2K)^{-d_{\text{se}}}}{2\sqrt{K-1}} \wedge \frac{L(2K)^{-d_{\text{se}}+1}}{4} \wedge \frac{c_{\psi,\alpha}(2K)^{-\frac{K-1}{2}}}{2\sqrt{K!}} \,. \tag{39}$$

By the definition of $\gamma$ in (39), we see that

$$\|\psi\|_{L^1}^{K-1} \leq \frac{1}{2K} \,; \tag{40}$$

$$\sqrt{K-1}(2K)^{-d_{\text{se}}} \max_{t \in \{0,\ldots,d_{\text{se}}\}} \|\psi^{(t)}\|_{L^1}^{K-1} \leq \frac{L}{2} \,; \tag{41}$$

$$2\,(2K)^{dse\ 1}\ \max_{t2f\ 0;:::;dse\ 1g}\ (t)^{K\ 1}\ _{L^1}\ \frac{L}{2}; \tag{42}$$

and

$$4\ ^2K!(2K)^{\ K+1}\ k\ k_{L^2}^{2(K\ 1)}\ c_{;}^2\ \ 1: \tag{43}$$

For each $\ 2\ f\ 1g^{[m]^{K\ 1}}$, de ne $g\ :\ _{K\ 1}\ !\ R^K$ by

$$g\ (z) := z + \Big(\sum_{j2[m]^{K\ 1}}\ _j(\ _j\ )(z)\Big)(1;\ \ 1;0;:::;0)^{>}:$$

Then we have $g(\ _{K\ 1})\ \ _{K\ 1}$ for all $\ 2\ f\ 1g^{[m]^{K\ 1}}$. This is because

$$\sum_{k=1}^K [g\ (z)]_k = \sum_{k=1}^K [z]_k + \Big(\sum_{j2[m]^{K\ 1}}\ _j(\ _j\ )(z)\Big) - \Big(\sum_{j2[m]^{K\ 1}}\ _j(\ _j\ )(z)\Big) = 1$$

for all $z\ 2\ _{K\ 1}$ and

$$[g\ (z)]_k = \begin{cases} [z]_k\ \ \sum_{j2[m]^{K\ 1}}\ _j(\ _j\ )(z)\ \ \frac{1}{2K}\ \ m^{\ s}\,k\ k_{L^1}^{K\ 1}; & \text{if } z\ 2\ [\frac{1}{2K};1]^K; k\ 2\ f\ 1;2g; \\ [z]_k; & \text{otherwise.} \end{cases}$$

for all $z\ 2\ _{K\ 1}$ and $k\ 2\ f\ 1;:::;K\ g$ by (40).

Next, we claim that each coordinate function of the mapping $z\ 7!\ g\ (z)\ \ z$ belongs to $H_K(s;L)$, i.e., for all multi-indices $\ 2\ N^{K\ 1}$ with $j\ j = dse\ \ 1$ and $x_1;x_2\ 2\ (\ _{K\ 1})$,

$$\sum_{j2[m]^{K\ 1}}\ _j\ ^{(\ )}_j(x_1)\ \ \ ^{(\ )}_j(x_2)\ \ \ L\,kx_1\ \ x_2k^{s\ d\ se+1}: \tag{44}$$

If $m\,kx_1\ \ x_2k\ \ 1$, then from the mean value theorem

$$^{(\ )}_j(x_1)\ \ \ ^{(\ )}_j(x_2)\ \ \ \sqrt{K\ \ 1}\ \max_{\substack{2N^{K\ 1}\\ j\ ^0j=dse}}\ \ ^{(\ ^0)}_j\ _{L^1}\ kx_1\ \ x_2k: \tag{45}$$

By the de nition of $_j$ in (38), for any $\ ^0=(\ ^0_1;:::;\ ^0_{K\ 1})^{>}\ 2\ N^{K\ 1}$ with $j\ ^0j = dse$,

$$\Big\|\ ^{(\ ^0)}_j\Big\|_{L^1} = m^{j\ ^0j\ s}(2K)^{j\ ^0j}\prod_{k=1}^{K\ 1}\ \|\ ^{(\ ^0_k)}\|_{L^1}\ \ m^{dse\ s}(2K)^{dse}\ \max_{t2f\ 0;:::;dseg}\ (t)^{K\ 1}\ _{L^1}: \tag{46}$$

Plugging (46) into (45), and using (41), we reach

$$^{(\ )}_j(x_1)\ \ \ ^{(\ )}_j(x_2)\ \ \ \sqrt{K\ \ 1}m^{dse\ s}(2K)^{dse}\ \max_{t2f\ 0;:::;dseg}\ (t)^{K\ 1}\ _{L^1}\ kx_1\ \ x_2k$$
$$\frac{L}{2}m^{dse\ s}kx_1\ \ x_2k\ \ \frac{L}{2}kx_1\ \ x_2k^{s\ d\ se+1}:$$

If $m\|x_1 - x_2\| > 1$, using (42), we similarly get

$$\left|\phi_j^{(\alpha)}(x_1) - \phi_j^{(\alpha)}(x_2)\right| \le 2 \max_{j' : j'_{\mathrm{dse}-1} = j_{\mathrm{dse}-1}} \left\|\phi_{j'}^{(\alpha)}\right\|_{L^1}$$

$$\le 2\, m^{\mathrm{dse}-s-1} (2K)^{\mathrm{dse}-1} \max_{t \in \{0,\dots,\mathrm{dse}-1\}} \left\|\phi^{(t)}\right\|_{L^1} \le \frac{L}{2} m^{\mathrm{dse}-s-1} \le \frac{L}{2} \|x_1 - x_2\|^{s-\mathrm{dse}+1}.$$

Thus, for any $j \in [m]^{K-1}$ and $x_1, x_2 \in \Delta_{K-1}$, it holds that

$$\left|\phi_j^{(\alpha)}(x_1) - \phi_j^{(\alpha)}(x_2)\right| \le \frac{L}{2} \|x_1 - x_2\|^{s-\mathrm{dse}+1}. \tag{47}$$

Given $x_1, x_2 \in \Delta_{K-1}$, there can be two cases: (1) there exists $j_1 \in [m]^{K-1}$ such that

$$\left(\sum_{j \in [m]^{K-1}} \beta_j \phi_j^{(\alpha)}(x_1) - \phi_j^{(\alpha)}(x_2)\right) = \beta_{j_1} \left|\phi_{j_1}^{(\alpha)}(x_1) - \phi_{j_1}^{(\alpha)}(x_2)\right|,$$

or (2) there exist distinct $j_1, j_2 \in [m]^{K-1}$ such that $x_1 \in \mathrm{supp}(\phi_{j_1})$ and $x_2 \in \mathrm{supp}(\phi_{j_2})$. In the first case, (44) directly follows from (47). In the second case, choose a point $x_3$ on the line segment connecting $x_1$ and $x_2$ such that $x_3 \in \mathrm{supp}(\phi_{j_1}) \cap \mathrm{supp}(\phi_{j_2})$. Such a point exists since $\mathrm{supp}(\phi_{j_1})$, $\mathrm{supp}(\phi_{j_2})$ are open and $\mathrm{supp}(\phi_{j_1}) \cap \mathrm{supp}(\phi_{j_2}) = \emptyset$. For any $\beta \in \mathbb{N}^{K-1}$ with $|\beta| = \mathrm{dse} - 1$, we have

$$\left|\sum_{j \in [m]^{K-1}} \beta_j \phi_j^{(\alpha)}(x_1) - \phi_j^{(\alpha)}(x_2)\right| = \left|\beta_{j_1} \phi_{j_1}^{(\alpha)}(x_1) - \beta_{j_2} \phi_{j_2}^{(\alpha)}(x_2)\right|$$

$$= \left|\beta_{j_1} \phi_{j_1}^{(\alpha)}(x_1) - \beta_{j_1} \phi_{j_1}^{(\alpha)}(x_3) + \beta_{j_2} \phi_{j_2}^{(\alpha)}(x_3) - \beta_{j_2} \phi_{j_2}^{(\alpha)}(z_2)\right|$$

$$\le \left|\phi_{j_1}^{(\alpha)}(x_1) - \phi_{j_1}^{(\alpha)}(x_3)\right| + \left|\phi_{j_2}^{(\alpha)}(x_3) - \phi_{j_2}^{(\alpha)}(x_2)\right|$$

$$\le \frac{L}{2} \|x_1 - x_3\|^{s-\mathrm{dse}+1} + \frac{L}{2} \|x_3 - x_2\|^{s-\mathrm{dse}+1} \le L \|x_1 - x_2\|^{s-\mathrm{dse}+1}.$$

The second inequality holds because of (47), and the last inequality holds because $\|x_1 - x_3\|, \|x_3 - x_2\| \le \|x_1 - x_2\|$ and $s - \mathrm{dse} + 1 > 0$. This finishes the proof of (44).

Now, let $P_0$ and $P_\sigma$, $\sigma \in \{-1, 1\}^{[m]^{K-1}}$, be the distributions of $(Z, Y) \in \Delta_{K-1} \times Y$ characterized by

$$Z \sim P_0 \ \mathrm{Unif}(\Delta_{K-1}), \quad Y \mid Z = z \sim P_0 \ \mathrm{Cat}(z) \text{ for all } z \in \Delta_{K-1},$$

and

$$Z \sim P_\sigma \ \mathrm{Unif}(\Delta_{K-1}), \quad Y \mid Z = z \sim P_\sigma \ \mathrm{Cat}(g_\sigma(z)) \text{ for all } z \in \Delta_{K-1}.$$

We have $P_0 \in \mathcal{P}_0$ by definition. For $Z \sim \mathrm{Unif}(\Delta_{K-1})$ and $j_0 := 1_{K-1} \in [m]^{K-1}$,

$$\ell_p\text{-ECE}_{P_\sigma}(f)^p = \mathbb{E}\left[\sum_{k=1}^{K} |[g_\sigma(Z) - Z]_k|^p\right] = 2^p \mathbb{E}\left[\left|\sum_{j \in [m]^{K-1}} \sigma_j (\phi_j - \phi_{j_0})(Z)\right|^p\right]$$

$$= 2(K-1)!^p m^{K-1} \int_{R^{K-1}} |\sigma_{j_0}(x)|^p dx.$$

Further,

$$m^{K-1}\int_{R^{K-1}} |\beta_0(x)|^p dx = m^{ps}\prod_{k=1}^{K-1} m\int_R |(m(2Kx_k - 1))|^p dx_k$$
$$= m^{ps}(2K)^{K+1}\|k\|_{L_p}^{(K-1)p}. \quad (48)$$

Thus, we have

$$\ell_p\text{-ECE}_P(f) = (2(2K)^{K+1}(K-1)!)^{\frac{1}{p}} m^{-s}\|k\|_{L_p}^{K-1} c_{\text{lower}} n^{-\frac{2s}{4s+K-1}} \quad (49)$$

for some $c_{\text{lower}} > 0$ because

$$\lim_{n\to\infty} \frac{m^{-s}}{n^{-\frac{2s}{4s+K-1}}} = \lim_{n\to\infty} \lceil n^{\frac{2}{4s+K-1}}\rceil^{-s} n^{-\frac{2s}{4s+K-1}} = 1.$$

From (44) and (49), we see that $P_\beta \in \mathcal{P}_1(\varepsilon; p; s)$ with $\varepsilon = c_{\text{lower}} n^{-2s/(4s+K-1)}$ for all $\beta \in \{-1\}^{[m]^{K-1}}$.

The final step is to apply Lemma 18. Given $n$ i.i.d. observations $\{(Z_i, Y_i) : i \in [n]\}$, the average likelihood ratio between $P_\beta$, $\beta \in \{-1\}^{[m]^{K-1}}$, and $P_0$ is

$$L_n = \frac{1}{2^{m^{K-1}}} \sum_{\beta\in\{-1\}^{[m]^{K-1}}} \prod_{i=1}^n \frac{[g_\beta(Z_i)]_{\text{argmax}_k [Y_i]_k}}{[Z_i]_{\text{argmax}_k [Y_i]_k}}.$$

Let $\beta^1$, $\beta^2$ be independent random variables uniformly drawn from $\{-1\}^{[m]^{K-1}}$, and $(Z, Y) \sim P_0$. Then,

$$E_{P_0}[L_n^2] = E_{\beta^1,\beta^2} E_{P_0}\left[\prod_{i=1}^n \frac{[g_{\beta^1}(Z_i)]_{\text{argmax}_k [Y_i]_k}}{[Z_i]_{\text{argmax}_k [Y_i]_k}} \frac{[g_{\beta^2}(Z_i)]_{\text{argmax}_k [Y_i]_k}}{[Z_i]_{\text{argmax}_k [Y_i]_k}}\right]$$
$$= E_{\beta^1,\beta^2}\left[E_{P_0}\frac{[g_{\beta^1}(Z)]_{\text{argmax}_k [Y]_k}[g_{\beta^2}(Z)]_{\text{argmax}_k [Y]_k}}{[Z]^2_{\text{argmax}_k [Y]_k}}\right]^n.$$

Moreover,

$$E_{P_0}\frac{[g_{\beta^1}(Z)]_{\text{argmax}_k [Y]_k}[g_{\beta^2}(Z)]_{\text{argmax}_k [Y]_k}}{[Z]^2_{\text{argmax}_k [Y]_k}} = \int_{[-1,1]^{K-1}} \left(\sum_{k=1}^K [z]_k \frac{[g_{\beta^1}(z)]_k[g_{\beta^2}(z)]_k}{[z]_k^2}\right) dz$$
$$= 1 + \int_{[-1,1]^{K-1}} \sum_{k=1}^K \frac{[g_{\beta^1}(z) - z]_k[g_{\beta^2}(z) - z]_k}{[z]_k} dz$$
$$= 1 + \int_{[-1,1]^{K-1}} \left(\sum_{j\in[m]^{K-1}} \beta^1_j(\mathbb{1}_j)(z)\right)\left(\sum_{j\in[m]^{K-1}} \beta^2_j(\mathbb{1}_j)(z)\right)\left(\frac{1}{[z]_0} + \frac{1}{[z]_1}\right) dz$$

where the integral $\int_{K-1}$ is over Unif($_{K-1}$). Since $\{\text{supp}(\ _j) : j \in [m]^{K-1}\}$ is a collection of pairwise disjoint sets,

$$\int \Big(\sum_{j\in[m]^{K-1}} {}^1_j(\ _j)(z)\Big)\Big(\sum_{j\in[m]^{K-1}} {}^2_j(\ _j)(z)\Big)\Big(\frac{1}{[z]_0}+\frac{1}{[z]_1}\Big)\,dz$$
$$= \int \Big(\sum_{j\in[m]^{K-1}} {}^1_j\,{}^2_j(\ _j)(z)^2\Big)\Big(\frac{1}{[z]_0}+\frac{1}{[z]_1}\Big)\,dz.$$

The random variable $(\ ^1_j\ ^2_j)_{j\in[m]^{K-1}}$ is also uniformly distributed on $\{-1,1\}^{[m]^{K-1}}$. Hence,

$$E_{P_0}[L_n^2] = E_{\ 1}\Big[\Big(1+\ ^2\sum_{j\in[m]^{K-1}}\ ^1_j\int_{\ K-1}(\ _j)(z)^2\Big(\frac{1}{[z]_0}+\frac{1}{[z]_1}\Big)\,dz\Big)^n\Big]. \qquad (50)$$

Since $\bigcup_{j\in[m]^{K-1}}\text{supp}(\ _j) \subseteq (\ _{K-1}\setminus[\frac{1}{2K};1]^K)$, we have $\frac{1}{[z]_0}+\frac{1}{[z]_1} \le 4K$ for every $z \in\ _{K-1}$ such that $(\ _j)(z) \neq 0$ for at least one $j \in [m]^{K-1}$. Thus,

$$\ ^2\sum_{j\in[m]^{K-1}}\ _j\int_{\ K-1}(\ _j)(z)^2\Big(\frac{1}{[z]_0}+\frac{1}{[z]_1}\Big)\,dz$$
$$\le\ ^2\sum_{j\in[m]^{K-1}}\int_{\ K-1}(\ _j)(z)^2\Big(\frac{1}{[z]_0}+\frac{1}{[z]_1}\Big)\,dz$$
$$\le\ ^2\sum_{j\in[m]^{K-1}}4K\int_{\ K-1}(\ _j)(z)^2dz = 4\ ^2K!m^{K-1}\int_{R^{K-1}}{}_{j0}(x)^2dx. \qquad (51)$$

The last equality is because Unif($_{K-1}$) has density $(K-1)!$ with respect to $\text{Leb}_{K-1}$, when projected to $R^{K-1}$. Also, by (43) and (48),

$$4\ ^2K!m^{K-1}\int_{R^{K-1}}{}_{j0}(x)^2dx = 4\ ^2K!m^{-2s}(2K)^{K+1}\,k\,k_{L^2}^{2(K-1)}\ c_{;}^2\ m^{-2s} \le 1.$$

By (50) and that $(1+x)^n \le \exp(nx)$ for all $x \in (-1;1]$,

$$E_{P_0}[L_n^2] \le E_{\ 1}\Big[\exp\Big(n\ ^2\sum_{j\in[m]^{K-1}}\ ^1_j\int_{\ K-1}(\ _j)(z)^2\Big(\frac{1}{[z]_0}+\frac{1}{[z]_1}\Big)\,dz\Big)\Big]. \qquad (52)$$

Since $\{\epsilon_j^1 : j \in [m]^{K-1}\}$ is a set of i.i.d. random variables drawn from Unif$(\{\pm 1\})$, we have, with $\cosh(x) := [\exp(x) + \exp(-x)]/2$,

$$\begin{aligned}
&E_1 \exp\Big(n^2 \sum_{j\in[m]^{K-1}} \int_0^1 (\epsilon_j^1 - \epsilon_j^0)(z)^2 \Big(\frac{1}{[z]_0} + \frac{1}{[z]_1}\Big) dz\Big) \\
&= \prod_{j\in[m]^{K-1}} E_{\epsilon_j^1} \exp\Big(n^2 \int_0^1 (\epsilon_j^1 - \epsilon_j^0)(z)^2 \Big(\frac{1}{[z]_0} + \frac{1}{[z]_1}\Big) dz\Big) \\
&= \prod_{j\in[m]^{K-1}} \cosh\Big(n^2 \int_0^1 (\epsilon_j^1 - \epsilon_j^0)(z)^2 \Big(\frac{1}{[z]_0} + \frac{1}{[z]_1}\Big) dz\Big).
\end{aligned}$$

Similarly to (51) and (52), we have

$$n^2 \int_0^1 (\epsilon_j^1 - \epsilon_j^0)(z)^2 \Big(\frac{1}{[z]_0} + \frac{1}{[z]_1}\Big) dz \le c_{\epsilon,s}^2 m^{2s-K+1} n^{-1} \tag{53}$$

for each $j \in [m]^{K-1}$. Using that $\cosh(x) \le 1 + x^2 \le e^{x^2}$ for $x \in [-1, 1]$,

$$\begin{aligned}
&\prod_{j\in[m]^{K-1}} \cosh\Big(n^2 \int_0^1 (\epsilon_j^1 - \epsilon_j^0)(z)^2 \Big(\frac{1}{[z]_0} + \frac{1}{[z]_1}\Big) dz\Big) \\
&\le \exp\Big(\sum_{j\in[m]^{K-1}} n^2 \int_0^1 (\epsilon_j^1 - \epsilon_j^0)(z)^2 \Big(\frac{1}{[z]_0} + \frac{1}{[z]_1}\Big) dz\Big)^2.
\end{aligned}$$

Again from (53), it follows that

$$\begin{aligned}
&\exp\Big(\sum_{j\in[m]^{K-1}} n^2 \int_0^1 (\epsilon_j^1 - \epsilon_j^0)(z)^2 \Big(\frac{1}{[z]_0} + \frac{1}{[z]_1}\Big) dz\Big)^2 \\
&\le \exp(c_{\epsilon,s}^4 m^{4s-K+1} n^{-2}) \le 1 + (1+\epsilon)^2 \epsilon^2.
\end{aligned}$$

In conclusion, $\epsilon_n(p; s) \ge c_{\text{lower}} n^{-2s/(4s+K-1)}$ by Lemma 18.

■

## A.9 Proof of Theorem 12

Overview of the proof. Under $P \in P_0$, we prove that $T_{1;k}$ and $T_{2;k}$ have zero mean, and their variances are bounded by unity. By rejecting $H_0$ when $|T_{1;k}| \ge \sqrt{3K/n}$ or $|T_{2;k}| \ge \sqrt{3K/n}$, we can filter out distributions $P \in P_1(\epsilon; p; s)$ such that $E_P[T_{1;k}] = \Omega(n^{-1/2})$ or $E_P[T_{2;k}] = \Omega(n^{-1/2})$. For the remaining cases, we compute $\|V_k - W_k\|_{L^2(P_Z)}$ and show it is lower bounded by $\Omega(n^{-2s/(4s+K-1)})$. We conclude using the minimax optimality of the two-sample test $T_S$.

Proof We prove the theorem for $p = 2$. Then, the general case follows since $P_1(\varepsilon; p; s) \subseteq P_1(\varepsilon; 2; s)$ for all $p \geq 2$. Assume $P \in P_0$. By the union bound,

$$P(T_n^{\text{split}} = 1) \leq \sum_{k=1}^{K} \left[ P\left( |T_{1;k}| \geq \sqrt{\frac{3K}{n}} \right) + P\left( |T_{2;k}| \geq \sqrt{\frac{3K}{n}} \right) + P\left( TS_{\frac{1}{3K};\frac{1}{2}}(V_k; W_k) = 1 \right) \right].$$

Moreover, for all $k \in \{1; \ldots; K\}$,

$$E_P[Y - Z]_k = E_P[E_P[[Y - Z]_k | Z]] = E_P[[E_P[Y|Z] - Z]_k] = 0$$

and $\mathrm{Var}_P([Y - Z]_k) = E_P[[Y - Z]_k^2] \leq 1$. Thus, by Chebyshev's inequality

$$P\left( |T_{1;k}| \geq \sqrt{\frac{3K}{n}} \right) \leq \frac{n}{3K} \mathrm{Var}_P(T_{1;k}) = \frac{1}{3K} \mathrm{Var}_P([Y - Z]_k) \leq \frac{1}{3K}.$$

Similarly, we have

$$P\left( |T_{2;k}| \geq \sqrt{\frac{3K}{n}} \right) \leq \frac{n}{3K} \mathrm{Var}_P(T_{2;k}) = \frac{1}{3K} \mathrm{Var}_P([Z]_k[Y - Z]_k) \leq \frac{1}{3K}.$$

From (13), we know that $P(TS_{\frac{1}{3K};\frac{1}{2}}(V_k; W_k) = 1) \leq \frac{1}{3K}$. Therefore,

$$P(T_n^{\text{split}} = 1) \leq \sum_{k=1}^{K} \left( \frac{1}{3K} + \frac{1}{3K} + \frac{1}{3K} \right) = \varepsilon.$$

Let $P \in P_1(\varepsilon; p; s)$ and suppose that, for some $k \in \{1; \ldots; K\}$,

$$|E_P[T_{1;k}]| = |E_P[Y - Z]_k| \geq \frac{1}{\sqrt{n}} \left( \sqrt{3K} + \Phi^{-1}(p) \right). \tag{54}$$

By Chebyshev's inequality,

$$P\left( |T_{1;k} - E_P[T_{1;k}]| \geq \frac{1}{\sqrt{n}} \Phi^{-1}(p) \right) \leq \frac{1}{(\Phi^{-1}(p))^2} n \, \mathrm{Var}_P(T_{1;k}) \leq 1 - p.$$

Note that (54) and $|T_{1;k} - E_P[T_{1;k}]| \leq 1/\sqrt{n}$ imply

$$|T_{1;k}| \geq |E_P[T_{1;k}]| - |T_{1;k} - E_P[T_{1;k}]| \geq \sqrt{\frac{3K}{n}}.$$

Therefore,

$$P(T_n^{\text{split}} = 1) \geq P\left( |T_{1;k}| \geq \sqrt{\frac{3K}{n}} \right) \geq P\left( |T_{1;k} - E_P[T_{1;k}]| \geq \frac{1}{\sqrt{n}} \Phi^{-1}(p) \right) \geq 1 - p.$$

The same conclusion can be drawn when

$$|E_P[T_{2;k}]| = |E_P[[Z]_k[Y - Z]_k]| \geq \frac{1}{\sqrt{n}} \left( \sqrt{3K} + \Phi^{-1}(p) \right)$$

for some $k \in \{1, \ldots, K\}$.

Now it remains to prove the claim for $P \in H(\varepsilon; p; s)$ such that

$$|E_P[Y - Z]_k| - |E_P[[Z]_k[Y - Z]_k]| < \frac{1}{\sqrt{n}}\left(\sqrt{3K} + p^{1/4}\right) \tag{55}$$

for every $k \in \{1, \ldots, K\}$. Since

$$\ell_2\text{-ECE}_P(f)^2 = \sum_{k=1}^{K} \int_{\mathcal{I}_{K-1}} [\text{res}_f(z)]_k^2 \, dP_Z(z) \geq \varepsilon^2,$$

we can choose $k_0 \in \{1, \ldots, K\}$ such that $\int_{\mathcal{I}_{K-1}} [\text{res}_f(z)]_{k_0}^2 \, dP_Z(z) \geq \frac{\varepsilon^2}{K}$. Choose $c_{\text{split}}' > 0$ such that, for $d_c$ from Assumption 2 and $c_{ts}$ from (13),

$$\frac{(c_{\text{split}}')^2}{K} \geq \frac{4}{d_c^3}\left(\sqrt{3K} + p^{1/4}\right)^2 + c_{ts}^2 \left(\frac{d_c}{8}\right)^{\frac{4s}{4s+K-1}}. \tag{56}$$

There exists $N \in \mathbb{N}_+$ such that for all $n \geq N$,

$$\frac{1}{\sqrt{n}}\left(\sqrt{3K} + p^{1/4}\right) \leq \frac{d_c}{2}, \quad 2\exp\left(-\frac{d_c n}{8}\right) \leq \frac{1}{2}. \tag{57}$$

Let $c_{\text{split}} = c_{\text{split}}' \wedge N^{2s/(4s+K-1)}$. If $n < N$, then $P_1(\varepsilon; p; s)$ is empty since $\varepsilon \leq c_{\text{split}} n^{-2s/(4s+K-1)} > 1$, so the claim is vacuously true. Assume $n \geq N$. By (55), (57), and Assumption 2,

$$E_P[Z]_{k_0} \geq E_P[Y]_{k_0} - |E_P[Y - Z]_{k_0}| \geq \frac{d_c}{2}. \tag{58}$$

By (11), (12), (55), and (58), $\|\hat{v}_{k_0} - w_{k_0}\|_{L^2(P_Z)}^2$ is lower bounded by

$$\frac{1}{(E_P[Y]_{k_0})^2} \int_{\mathcal{I}_{K-1}} [\text{res}_f(z)]_{k_0}^2 \, dP_Z(z) + \frac{2E_P[Z - Y]_{k_0} E_P[[Z]_{k_0}[Y - Z]_{k_0}]}{(E_P[Y]_{k_0})^2 E_P[Z]_{k_0}}$$
$$\geq \frac{\varepsilon^2}{K} - \frac{4}{d_c^3}\left(\sqrt{3K} + p^{1/4}\right)^2 n^{-1}.$$

Further by $\varepsilon \geq c_{\text{split}}' n^{-2s/(4s+K-1)}$ and (56),

$$\frac{\varepsilon^2}{K} - \frac{4}{d_c^3}\left(\sqrt{3K} + p^{1/4}\right)^2 n^{-1} \geq \left(\frac{(c_{\text{split}}')^2}{K} - \frac{4}{d_c^3}\left(\sqrt{3K} + p^{1/4}\right)^2\right) n^{-\frac{4s}{4s+K-1}}$$
$$\geq c_{ts}^2 \left(\frac{d_c n}{8}\right)^{-\frac{4s}{4s+K-1}}. \tag{59}$$

In conclusion,

$$\|\hat{v}_{k_0} - w_{k_0}\|_{L^2(P_Z)} \geq c_{ts}\left(\frac{d_c n}{8}\right)^{-\frac{2s}{4s+K-1}}.$$

Note that $\overset{V}{k_0}$ $\overset{W}{k_0}$ is s-Hölder since it is a linear combination of two s-Hölder functions $z \mapsto [\mathrm{res}_{\cdot}(z)]_{k_0}$ and $z \mapsto [z]_{k_0}$, possibly with different Hölder constants. Thus by (13), we have

$$P\left[ TS_{\frac{3K}{}:\frac{}{2}}(V_{k_0}; W_{k_0}) = 1 \mid jV_{k_0}j = v; jW_{k_0}j = w \right] \quad 1 \quad \frac{}{2} \tag{60}$$

given that $v; w \quad \frac{d_c n}{8}$. For convenience, assume that $n$ is even (if required, drop an observation). Since $jV_{k_0}j \quad \mathrm{Bin}(\frac{n}{2}; E_P[Y]_{k_0})$ and $jW_{k_0}j \quad \mathrm{Bin}(\frac{n}{2}; E_P[Z]_{k_0})$, we find

$$P\left[ jV_{k_0}j < \frac{nE_P[Y]_{k_0}}{4} \right] \quad \frac{2}{e}^{\frac{nE_P[Y]_{k_0}}{4}}; \quad P\left[ jW_{k_0}j < \frac{nE_P[Z]_{k_0}}{4} \right] \quad \frac{2}{e}^{\frac{nE_P[Z]_{k_0}}{4}}$$

by Cherno 's inequality (Exercise 2.3.2 of Vershynin (2018)). Therefore,

$$\begin{aligned} P\left[ jV_{k_0}j < \frac{d_c n}{8} \text{ or } jW_{k_0}j < \frac{d_c n}{8} \right] \quad & P\left[ jV_{k_0}j < \frac{d_c n}{8} \right] + P\left[ jW_{k_0}j < \frac{d_c n}{8} \right] \\ & P\left[ jV_{k_0}j < \frac{nE_P[Y]_{k_0}}{4} \right] + P\left[ jW_{k_0}j < \frac{nE_P[Z]_{k_0}}{4} \right] \\ & \frac{2}{e}^{\frac{nE_P[Y]_{k_0}}{4}} + \frac{2}{e}^{\frac{nE_P[Z]_{k_0}}{4}} \quad 2 \frac{2}{e}^{\frac{d_c n}{8}} \quad \frac{}{2}; \end{aligned}$$

and thus

$$P\left[ jV_{k_0}j; jW_{k_0}j \quad \frac{d_c n}{8} \right] \quad 1 \quad \frac{}{2}: \tag{61}$$

The last inequality holds by (56). Finally by (60) and (61),

$$\begin{aligned} & P\left[ TS_{\frac{3K}{}:\frac{}{2}}(V_{k_0}; W_{k_0}) = 1 \right] \\ & \quad \int_{f(v;w) 2 z^2 : v;w \quad \frac{d_c n}{8} g} P\left[ TS_{\frac{3K}{}:\frac{}{2}}(V_{k_0}; W_{k_0}) = 1 \mid jV_{k_0}j = v; jW_{k_0}j = w \right] dP_{jV_{k_0}j;jW_{k_0}j}(v; w) \\ & \quad 1 \quad \frac{}{2} \quad P\left[ jV_{k_0}j; jW_{k_0}j \quad \frac{d_c n}{8} \right] \quad 1 \quad \frac{}{2}^2 \quad 1 \end{aligned}$$

where the integral is with respect to the joint distribution of $jV_{k_0}j$ and $jW_{k_0}j$. This proves that

$$P(\overset{split}{n} = 1) \quad P\left[ TS_{\frac{3K}{}:\frac{}{2}}(V_{k_0}; W_{k_0}) = 1 \right] \quad 1 \quad :$$

■

### A.10 Proof of Corollary 13

The proof of Theorem 12 can be repeated by replacing (13) with (14). The only difference is (59), where we used $n^{-4s=(4s+K-1)} \quad n^{-1}$. In the adaptive setting, we instead need $(n/\log\log n)^{-4s=(4s+K-1)} \quad n^{-1}$. This inequality holds for all large $n \in \mathbb{N}_+$, say $n \quad N^0$. Then, we can define $c_{\text{ad-s}}$ to be larger than $(N^0/\log\log N^0)^{2s=(4s+K-1)}$, so that $P_1(\cdot; p; s)$ becomes empty when $n < N^0$.

## Appendix B. Background

### B.1 Hölder Continuity on the Probability Simplex

To define derivatives—and thus the class of functions we study—on $\Delta_{K-1}$, a coordinate chart $\phi : \Delta_{K-1} \to \mathbb{R}^{K-1}$ has to be specified. For example, we can consider the canonical projection $\phi_k : (z_1, \ldots, z_K)^\top \mapsto (z_1, \ldots, z_{k-1}, z_{k+1}, \ldots, z_K)^\top$. The definition of Hölder smoothness below depends on the choice of $\phi$. We assume $\phi = \phi_K$, but all conclusions and proofs remain the same for any choice of the coordinate chart $\phi$.

For an integer $d \geq 1$, a vector $\alpha = (\alpha_1, \ldots, \alpha_d)^\top \in \mathbb{N}^d$ is called a multi-index. We write $|\alpha| := \alpha_1 + \cdots + \alpha_d$. For a sufficiently smooth function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its partial derivative of order $\alpha = (\alpha_1, \ldots, \alpha_d)^\top$ by $f^{(\alpha)} := \frac{\partial^{\alpha_1}}{\partial x_1} \cdots \frac{\partial^{\alpha_d}}{\partial x_d} f$. For a Hölder smoothness parameter $s > 0$ and a Hölder constant $L > 0$, let $H_K(s, L)$ be the class of $(s, L)$-Hölder continuous functions $g : \Delta_{K-1} \to \mathbb{R}$ satisfying, for all $x_1, x_2 \in \phi(\Delta_{K-1})$ and multi-indices $\alpha \in \mathbb{N}^{K-1}$ with $|\alpha| = \lfloor s \rfloor \geq 1$,

$$(g \circ \phi^{-1})^{(\alpha)}(x_1) - (g \circ \phi^{-1})^{(\alpha)}(x_2) \leq L \|x_1 - x_2\|^{s - \lfloor s \rfloor + 1}. \tag{62}$$

In particular, $H_K(1, L)$ denotes all $L$-Lipschitz functions. ]

### B.2 Two-sample Goodness-of-fit Tests

Here we state and slightly extend the results of Arias-Castro et al. (2018); Kim et al. (2022). For $d \in \mathbb{N}_+$, let $\mu$ be a measure on $[0,1]^d$ which is absolutely continuous with respect to $\mathrm{Leb}_d$ and satisfies

$$l \leq \frac{d\mu}{d\mathrm{Leb}_d} \leq u \tag{63}$$

almost everywhere for some constants $l, u > 0$. For $n_1, n_2 \in \mathbb{N}_+$, suppose we have two samples $\{V_1, \ldots, V_{n_1}\}$ and $\{W_1, \ldots, W_{n_2}\}$ i.i.d. sampled from the distributions on $[0,1]^d$ with densities $f_1$ and $f_2$, respectively, with respect to $\mu$. We also assume $f_1 - f_2$ is $(s, L)$-Hölder continuous for a Hölder smoothness parameter $s > 0$ and a Hölder constant $L > 0$.

For $m \in \mathbb{N}_+$ and $i = (i_1, \ldots, i_d)^\top \in [m]^d$, let $R_{m,i} := \prod_{k=1}^d \left(\frac{i_k - 1}{m}, \frac{i_k}{m}\right]$.

$$v_{i,m} := |\{j \in [n_1] : V_j \in R_{m,i}\}|,$$

$$w_{i,m} := |\{j \in [n_2] : W_j \in R_{m,i}\}|.$$

The unnormalized chi-squared statistic is defined by

$$T_{m,n_1,n_2} := \sum_{i \in [m]^d} (n_2 v_{i,m} - n_1 w_{i,m})^2.$$

Theorem 19 (Chi-squared test, Arias-Castro et al. (2018)) Consider the two-sample goodness-of-fit testing problem described above. Assume the Hölder smoothness parameter $s$ is known and let $m = \lfloor (n_1 \wedge n_2)^{2/(4s+d)} \rfloor$. For any $\alpha \in (0,1)$ and $\beta \in (0, 1 - \alpha)$, there

exist $c > 0$ depending on $(d; L)$ and $c_{ts} > 0$ depending on $(s; d; L; \ _l; \ _u; \ ; \ )$ such that for $:= n_1 n_2 (n_1 + n_2) + c n_1 n_2 m^{d=2}$,

$$P(\ _{m;n_1;n_2} \quad ) \qquad \text{if } f_1 = f_2;$$

$$P(\ _{m;n_1;n_2} \quad ) \quad 1 \qquad \text{if } \|f_1 \quad f_2\|_{L^2(\ )} \quad c_{ts}(n_1 \wedge n_2)^{\frac{2s}{4s+d}}:$$

For $m \ 2 \ \mathbb{N}_+$, let $k_m : ([0;1]^d)^4 \ ! \ \mathbb{R}$ be the kernel

$$k_m(v_1; v_2; w_1; w_2) = \sum_{i \ 2 \ [m]^d} 1_{R_{m;i}}(v_1) 1_{R_{m;i}}(v_2) + 1_{R_{m;i}}(w_1) 1_{R_{m;i}}(w_2)$$

$$1_{R_{m;i}}(v_1) 1_{R_{m;i}}(w_2) \quad 1_{R_{m;i}}(w_1) 1_{R_{m;i}}(v_2) \ :$$

For the two samples $\{V_1; : : : ; V_{n_1}\}$, $\{W_1; : : : ; W_{n_2}\}$ described above, de ne

$$U_{m;n_1;n_2} := \frac{1}{n_1(n_1 \quad 1) n_2(n_2 \quad 1)} \sum_{i_1 \neq i_2 \ 2 \ [n_1]} \sum_{j_1 \neq j_2 \ 2 \ [n_2]} k_m(V_{i_1}; V_{i_2}; W_{j_1}; W_{j_2}):$$

For any $\ 2 \ (0;1)$, the $1 \quad$ quantile $c_{1 \ ;m;n_1;n_2}$ of the U-statistic $U_{m;n_1;n_2}$ can be found by the permutation procedure described in Section 2.1 of Kim et al. (2022).

Theorem 20 (Permutation test, Kim et al. (2022)) Consider the two-sample goodness-of- t testing described above. Assume the Holder smoothness parameter $s$ is known and let $m = b(n_1 \wedge n_2)^{2=(4s+d)} c$. For any $\ 2 \ (0;1)$ and $\ 2 \ (0; 1 \quad )$, there exists $c_{ts} > 0$ depending on $(s; d; L; \ _l; \ _u; \ ; \ )$ such that

$$P(U_{m \ ;n_1;n_2} \quad c_{1 \ ;m \ ;n_1;n_2}) \qquad \text{if } f_1 = f_2;$$

$$P(U_{m \ ;n_1;n_2} \quad c_{1 \ ;m \ ;n_1;n_2}) \quad 1 \qquad \text{if } \|f_1 \quad f_2\|_{L^2(\ )} \quad c_{ts}(n_1 \wedge n_2)^{\frac{2s}{4s+d}}:$$

Corollary 21 (Multi-scale permutation test, Kim et al. (2022)) Consider the two-sample goodness-of- t testing problem described above. Let $B = d\frac{2}{d} \log_2(\frac{n_1 \wedge n_2}{\log\log(n_1 \wedge n_2)})e$ and de ne

$$^{perm}_{n_1;n_2} := \max_{b \ 2 \ \{1;:::;B\}} I(U_{2^b;n_1;n_2} \quad c_{1 \ =B; \ 2^b;n_1;n_2}):$$

For any $\ 2 \ (0;1)$ and $\ 2 \ (0; 1 \quad )$, there exists $c_{ad} > 0$ depending on $(d; L; \ _l; \ _u; \ ; \ )$ such that

$$P(\ ^{perm}_{n_1;n_2} = 1) \qquad \text{if } f_1 = f_2;$$

$$P(\ ^{perm}_{n_1;n_2} = 1) \quad 1 \qquad \text{if } \|f_1 \quad f_2\|_{L^2(\ )} \quad c_{ad} \ \frac{n_1 \wedge n_2}{\log\log(n_1 \wedge n_2)} \ ^{\frac{2s}{4s+d}}:$$

Remark 22 (Comment on the proofs) Theorem 19, Theorem 20, and Corollary 21 are generalization of Theorem 4 of Arias-Castro et al. (2018), Proposition 4.6 of Kim et al. (2022), and Proposition 7.1 of Kim et al. (2022), respectively. The original statements are for $= \text{Leb}_d$. The proofs in Arias-Castro et al. (2018); Kim et al. (2022) can be adapted with only minor di erences. For example, equation (93) of Arias-Castro et al. (2018) is still true assuming (63) above. Also, we proved Lemma 15 in this paper, which is a generalization of Lemma 3 of Arias-Castro et al. (2018) to a general measure . Other parts of the proofs in Arias-Castro et al. (2018); Kim et al. (2022) can be repeated without any modi cation.

B.3 Binning Scheme for the Probability Simplex

Here we describe a binning scheme for the probability simplex $\Delta_{K-1} \subset \mathbb{R}^K$ into equal volume simplices|i.e., a ne transforms of the standard probability simplex.

Hypersimplex. For $u \geq 2$ and $v \in [u-1]$, de ne the $(u-1)$-dimensional polytope hypersimplex $\Delta_{u-1;v}$|a generalization of the standard probability simplex that can have more vertices and edges|as

$$\Delta_{u-1;v} := \left\{ (x_1, \ldots, x_u)^\top \in [0,1)^u : x_1 + \cdots + x_u = v \right\}.$$

Let $A_{u-1;v-1}$ be the Eulerian number:

$$A_{u-1;v-1} := \sum_{i=0}^{v} (-1)^i \binom{u}{i} (v-i)^{u-1}.$$

It is known that the hypersimplex $\Delta_{u-1;v}$ can be partitioned into $A_{u-1;v-1}$ simplices (Stanley, 1977; Sturmfels, 1996) whose volumes are identical to that of the unit probability simplex $\Delta_{u-1}$.

Construction. Let $m \in \mathbb{N}_+$ and $R_{m;i} := \prod_{k=1}^{K} \left[ \frac{i_k - 1}{m}, \frac{i_k}{m} \right)$ for $i = (i_1, \ldots, i_K)^\top \in [m]^K$. The hypercube $R_{m;i}$ has a positive-volume intersection with $\Delta_{K-1}$ when

$$m + 1 \leq \sum_{k=1}^{K} i_k \leq m + K - 1. \tag{64}$$

Suppose that (64) holds and we write $|i| = \sum_{k=1}^{K} i_k$. Then, the intersection is

$$\begin{aligned} \Delta_{K-1} \cap R_{m;i} &= \left\{ (x_1, \ldots, x_K)^\top \in R_{m;i} : x_1 + \cdots + x_K = 1 \right\} \\ &= \frac{i - 1_K}{m} + \frac{1}{m} \left\{ (x_1, \ldots, x_K)^\top \in [0,1)^K : x_1 + \cdots + x_K = m + K - |i| \right\}, \end{aligned}$$

which is a $\frac{1}{m}$-scaled and translated version of the hypersimplex $\Delta_{K-1;m+K-|i|}$. Recall that this hypersimplex can be further partitioned into $A_{K-1;m+K-|i|-1}$ simplices with the volume $\mathrm{vol}(\Delta_{K-1})/m^{K-1}$. Let $S_i$ be the set of such $A_{K-1;m+K-|i|-1}$ simplices. Now, we de ne

$$B_m := \bigcup_{\substack{i \in [m]^K \\ m+1 \leq |i| \leq m+K-1}} \left\{ \frac{i - 1_K}{m} + \frac{1}{m} S : S \in S_i \right\},$$

which is the collection of all simplices obtained from decomposing $\Delta_{K-1} \cap R_{m;i}$ for each $i \in [m]^K$ satisfying (64). Since each simplex in $B_m$ has a volume $\mathrm{vol}(\Delta_{K-1})/m^{K-1}$, it follows that $|B_m| = m^{K-1}$. Noting that there are $\binom{|i|-1}{K-1}$ multi-indices $i \in [m]^K$ with $|i| = |i|$, it also directly follows from Worpitzky's identity (Equation 6.37 in Graham et al. (1994))

that

$$|B_m| = \sum_{j=m+1}^{m+K-1} \binom{j-1}{K-1} A_{K-1,m+K-j-1} = \sum_{j=0}^{K-2} \binom{m+j}{K-1} A_{K-1,K-j-2}$$
$$= \sum_{j=0}^{K-2} \binom{m+j}{K-1} A_{K-1,j} = m^{K-1}.$$

We finally index the partition as $B_m = \{B_1, \ldots, B_{m^{K-1}}\}$.

### B.4 Calibration Test for Discrete Predictions

Testing calibration of discrete probability predictions has been studied in Cox (1958); Miller (1962); Harrell (2015). Here we describe a test based on testing multiple binomial parameters. This test is related to the chi-squared test from Miller (1962), but does not use the asymptotic distribution of the test statistic when choosing critical values. Let $\{v_1, \ldots, v_t\} \subseteq [0,1]$ be the range of a discrete-valued probability predictor $f$. For each $i \in [t]$, we let $N_i = |\{j \in [n] : f(X_j) = v_i\}|$, $M_i = |\{j \in [n] : f(X_j) = v_i, Y_j = 1\}|$, and $p_i = P(Y = 1 \mid f(X) = v_i)$. In this setting, the random variable $M_i$, under the null hypothesis of perfect calibration, follows the binomial distribution $\text{Binom}(N_i, p_i)$ with $N_i$ trials and success probability $p_i$, given $N_i$. We use an exact binomial test to test the null hypothesis $H_{0,i} : p_i = v_i$ for each $i \in [t]$ and apply the Bonferroni correction to control the false detection rate under the null hypothesis $H_0 = \bigcap_{i=1}^{t} H_{0,i}$.

## Appendix C. More Experiments

### C.1 Debiased ECE from Kumar et al. (2019)

In this subsection, we provide additional experiments, to evaluate calibration methods whose outputs range over a finite set. When the predicted probabilities belong to a finite set, Kumar et al. (2019) propose a debiased version of the empirical $\ell_2$-ECE, whose sample complexity required is smaller than that of the plug-in estimator; see Section 3.3 for more discussion. We calculate Kumar et al. (2019)'s debiased estimator along with calibration testing results. We use models trained on CIFAR-10 (Table 4), CIFAR-100 (Table 5), and ImageNet (Table 6). The values of the debiased $\ell_2$-ECE estimator are typically very small, which is consistent with the experimental results in Kumar et al. (2019). As can be observed, there is no clear relation between the debiased empirical $\ell_2$-ECE values and test results.

| | DenseNet 121 | | ResNet 50 | | VGG-19 | |
|---|---|---|---|---|---|---|
| | $\ell_2$-ECE$^{db}$ | Calibrated? | $\ell_2$-ECE$^{db}$ | Calibrated? | $\ell_2$-ECE$^{db}$ | Calibrated? |
| Hist. Bin. | 0.02% | reject | 0.02% | reject | 0.05% | reject |
| Scal. Bin. | 0.11% | reject | 0.10% | reject | 0.20% | reject |

Table 4: The values of the debiased empirical $\ell_2$-ECE (Kumar et al., 2019) and the testing results, via multiple binomial testing, of models trained on CIFAR-10 with two discrete calibration methods.

| | MobileNet-v2 | | ResNet 56 | | Shu eNet-v2 | |
|---|---|---|---|---|---|---|
| | $\ell_2$-ECE$^{db}$ | Calibrated? | $\ell_2$-ECE$^{db}$ | Calibrated? | $\ell_2$-ECE$^{db}$ | Calibrated? |
| Hist. Bin. | 0.04% | reject | 0.09% | reject | 0.15% | reject |
| Scal. Bin. | 0.04% | reject | 0.03% | reject | 0.02% | accept |

Table 5: The values of the debiased empirical $\ell_2$-ECE (Kumar et al., 2019) and the testing results, via multiple binomial testing, of models trained on CIFAR-100 with two discrete calibration methods.

| | DenseNet 161 | | ResNet 152 | | E cientNet-b7 | |
|---|---|---|---|---|---|---|
| | $\ell_2$-ECE$^{db}$ | Calibrated? | $\ell_2$-ECE$^{db}$ | Calibrated? | $\ell_2$-ECE$^{db}$ | Calibrated? |
| Hist. Bin. | 0.01% | reject | 0.03% | reject | 0.02% | reject |
| Scal. Bin. | 0.05% | reject | 0.03% | reject | 0.06% | reject |

Table 6: The values of the debiased empirical $\ell_2$-ECE (Kumar et al., 2019) and the testing results, via multiple binomial testing, of models trained on ImageNet with two discrete calibration methods.

### C.2 Estimation of ECE via DPE

The DPE $T^d_{m;n}$ is used as a test statistic in the main text, but it can also be interpreted as an estimate of $\ell_2$-ECE$(f)^2$. While it builds on ideas from nonparametric functional estimation (Remark 2), the convergence rate $n^{-[4s/(4s+K-1)\wedge 1/2]}$ does not directly follow here, since our construction involves an additional estimation step $|P_Z(B_i) - I_{m;i}|/n$. If we use the same binning parameter $m = m^*_{n^{2/(4s+K-1)}}$, then the ratio between the standard deviation and the mean of the estimate $|I_{m;i}|/n$ is

$$\sqrt{\frac{(1 - P_Z(B_i))/nP_Z(B_i)}{}} \asymp n^{(K-1-4s)/[2(4s+K-1)]},$$

which is large for small $s$. For this reason, the DPE fails to achieve the parametric convergence rate when $s \leq \frac{K-1}{4}$. We experimentally support this claim in Figure 8, where we use

Figure 8: Estimation error of $T_{m\ ;n}^{d}$, varying $n$, $s$ over a log scale. The dashed line has slope $\frac{1}{2}$, which corresponds to the parametric convergence rate. The lines for $s = 0.4$ and $0.6$ fail to achieve the parametric convergence rate, even though $s$ is larger than the threshold value 0.25 where the parametric rate arises in standard nonparametric estimation.

$K = 2$ and $Z \sim \text{Unif}([0,1])$, with a deterministic choice of $Y = 1$. Optimal estimation of $\ell_2\text{-ECE}(f)^2$ remains an open problem.

### C.3 Comparison of Critical Values

We compare three choices of critical values, where we sample $\{(Z_i^*, Y_i^*)\}_{i=1}^n$ according to the following rules.

1. Oracle Monte Carlo: $Z_i^* \sim P_Z$, $Y_i^* \sim \text{Cat}(Z_i^*)$.
2. Full bootstrapping (consistency resampling): $Z_i^* \sim \text{Unif}(\{Z_i\}_{i=1}^n)$, $Y_i^* \sim \text{Cat}(Z_i^*)$.
3. $Y$-only bootstrapping: $Z_i^* = Z_i$, $Y_i^* \sim \text{Cat}(Z_i^*)$.

Here, $\{(Z_i, Y_i)\}_{i=1}^n$ is the original calibration sample, and all sampling is done independently. We repeat the above sampling procedures $N$ times to create $\{(Z_i^j, Y_i^j)\}_{i=1}^n$, $j \in [N]$ and compute the DPE $T_m^{d;j}{}_{;n}$ test statistics for each $j \in [N]$. Denote their order statistics by $T_m^{d;(1)}{}_{;n} \le \cdots \le T_m^{d;(N)}{}_{;n}$. For oracle Monte Carlo and $Y$-only bootstrapping, the DPEs $T_{m\ ;n}^{d}, T_m^{d;1}{}_{;n}, \ldots, T_m^{d;N}{}_{;n}$ are exchangeable under the null, so (assuming ties happen with zero probability)

$$P(T_{m\ ;n}^{d} \le T_m^{d;(j)}{}_{;n}) = \frac{N+1-j}{N+1},$$

Figure 9: Type I error of the full/ $Y$-only bootstrapping. We use $N = 19$; $= 0.05$ and compute the type I error from 10,000 oracle Monte Carlo trials. Standard error bars are plotted over 1,000 repetitions.

for any $j \in [N]$. In other words, the test that rejects when $I(T^{d}_{m,n} \quad T^{d;(j)}_{m,n})$ has type I error $\frac{j}{N+1}$. We do not have such a guarantee for consistency resampling.

Figure 9 compares the type I errors when using critical values based on full/ $Y$-only bootstrapping. We see that there is no significant difference between the two, and they stay relatively close to the nominal level $= 0.05$. Since consistency resampling and $Y$-only bootstrapping give randomized critical values, the true type I error is estimated by the sample mean of type I errors obtained from 10,000 independent trials.

### C.4 Cross-fitting

The idea of cross-fitting (see e.g., Hajek, 1962; Schick, 1986; Newey and Robins, 2018; Kennedy, 2020, for related ideas) can be applied to the sample splitting test in Section 6. Specifically, we compute the two-sample test statistic again by swapping the role of $f(Z_i, Y_i)g_{i=1}^{bn/2c}$ and $f(Z_i, Y_i)g_{i=bn/2c+1}^{n}$, and use the average of two test statistics. The critical value for $= 0.05$ and the corresponding type II error are estimated via 1,000 oracle Monte Carlo simulations. In Figure 10, we compare the type II error of the cross-fitted test statistics with other tests discussed in the main text. The experimental setup is identical to Figure 7. We observe that the cross-fitting procedure does not significantly improve the power.

(a) $s = 0.6, \quad = 100$ (b) $s = 0.8, \quad = 200$

Figure 10: Type II error comparison for plug-in, sample-splitting, and cross- tting test. The horizontal dashed line indicates a type II error of $1 \quad = 0.95$. Standard error bars are plotted over 10 repetitions.

## References

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv preprint arXiv:2110.01052, 2021.

Ery Arias-Castro, Bruno Pelletier, and Venkatesh Saligrama. Remember the curse of dimensionality: the case of goodness-of- t testing in arbitrary dimension. Journal of Nonparametric Statistics, 30(2):448{471, 2018.

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

Peter C Austin and Ewout W Steyerberg. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Statistics in Medicine, 33(3):517{535, 2014.

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-con dence: Theoretical analysis of calibration in binary classi cation. In International Conference on Machine Learning PMLR, 2021.

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: a selective review. The Annals of Applied Statistics, 12(2):727{749, 2018.

Richard Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. Journal of Experimental Criminology, 13(2):193{216, 2017.

Piotr Berman, Sofya Raskhodnikova, and Grigory Yaroslavtsev. $L_p$-testing. In Proceedings of the forty-sixth annual ACM symposium on Theory of computing, 2014.

Thomas B Berrett, Ioannis Kontoyiannis, and Richard J Samworth. Optimal rates for independence testing via u-statistic permutation tests. The Annals of Statistics, 49(5): 2457{2490, 2021.

J Eric Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. Decision Analysis, 4(2):49{65, 2007.

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: Sharp best order of convergence estimates. Sankhya: The Indian Journal of Statistics, Series A (1961-2002), 50(3):381{393, 1988.

Lucien Birge and Pascal Massart. Estimation of integral functionals of a density. The Annals of Statistics, 23(1):11{29, 1995.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316, 2016.

Glenn W Brier. Veri cation of forecasts expressed in terms of probability. Monthly Weather Review, 78(1):1{3, 1950.

Jochen Bröcker. Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. Climate Dynamics, 39(3):655{667, 2012.

Jochen Bröcker and Leonard A Smith. Increasing the reliability of reliability diagrams. Weather and Forecasting 22(3):651{661, 2007.

Lawrence D Brown and Mark G Low. Asymptotic equivalence of nonparametric regression and white noise. The Annals of Statistics, 24(6):2384{2398, 1996.

Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classi cation: structure and applications, 2005.

MV Burnashev. On the minimax detection of an inaccurately known signal in a white gaussian noise background. Theory of Probability & Its Applications, 24(1):107{119, 1979.

Cristina Butucea and Karine Tribouley. Nonparametric homogeneity tests. Journal of Statistical Planning and Inference, 136(3):597{639, 2006.

T Tony Cai and Mark G Low. Optimal adaptive estimation of a quadratic functional. The Annals of Statistics, 34(5):2298{2325, 2006.

Yaofo Chen. Pytorch cifar models, 2021. URL https://github.com/chenyaofo/pytorch-cifar-models.

Julien Chhor and Alexandra Carpentier. Goodness-of- t testing for Hölder-continuous densities: Sharp local minimax rates. arXiv preprint arXiv:2109.04346, 2021.

David R Cox. Two further applications of a model for binary regression. Biometrika, 45 (3/4):562{565, 1958.

A Philip Dawid. The well-calibrated bayesian. Journal of the American Statistical Association, 77(379):605{610, 1982.

A Philip Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. Journal of the Royal Statistical Society. Series A (General), 147(2):278{292, 1984.

Christina Dawkins, Thirukodikaval Nilakanta Srinivasan, and John Whalley. Calibration. In Handbook of econometrics, volume 5, pages 3653{3703. Elsevier, 2001.

Bruno De Finetti. Does it make sense to speak of \good probability appraisers". The Scientist Speculates: An Anthology of Partly-baked Ideas, pages 257{364, 1962.

Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. Journal of the Royal Statistical Society: Series D (The Statistician), 32(1-2):12{22, 1983.

Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3):134{144, 1995.

Francis X Diebold, Todd A Gunther, and Anthony S Tay. Evaluating density forecasts with applications to nancial risk management. International Economic Review, 39(4): 863{883, 1998.

David Donoho and Jiashun Jin. Higher criticism for large-scale inference, especially for rare and weak e ects. Statistical Science, 30(1):1{25, 2015.

David L Donoho and Michael Nussbaum. Minimax quadratic estimation of a quadratic functional. Journal of Complexity, 6(3):290{323, 1990.

Amandine Dubois, Thomas Berrett, and Cristina Butucea. Goodness-of- t testing for Hölder continuous densities under local di erential privacy. arXiv preprint arXiv:2107.02439, 2021.

Sam Efromovich and Mark Low. On optimal adaptive estimation of a quadratic functional. The Annals of Statistics, 24(3):1106{1125, 1996.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin M. Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classi cation of skin cancer with deep neural networks. Nature, 542:115{118, 2017.

Herbert Federer. Geometric measure theory Springer, 2014.

Christopher AT Ferro and Thomas E Fricker. A bias-corrected decomposition of the brier score. Quarterly Journal of the Royal Meteorological Society, 138(668):1954{1960, 2012.

Dean P Foster and Rakesh V Vohra. Asymptotic calibration. Biometrika, 85(2):379{390, 1998.

Allan Franklin. Calibration. In Can that be Right?, pages 237{272. Springer, 1999.

Ursula Garczarek. Classification rules in standardized partition spaces. PhD thesis, Universität Dortmund, 2002.

Martin Gebel. Multivariate calibration of classifier scores into the probability space. PhD thesis, Universität Dortmund, 2009.

Evarist Giné and Richard Nickl. A simple adaptive estimator of the integrated square of a density. Bernoulli, 14(1):47–61, 2008.

Evarist Giné and Richard Nickl. Mathematical foundations of infinite-dimensional statistical models. Cambridge University Press, 2021.

Tilmann Gneiting and Matthias Katzfuss. Probabilistic forecasting. Annual Review of Statistics and Its Application, 1:125–151, 2014.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):243–268, 2007.

Irving John Good. Rational decisions. Journal of the Royal Statistical Society: Series B (Methodological), 14(1):107–114, 1952.

Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. Concrete Mathematics: A Foundation for Computer Science. Addison-Wesley Longman Publishing Co., Inc., USA, 2nd edition, 1994. ISBN 0201558025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In International Conference on Machine Learning. PMLR, 2017.

Chirag Gupta and Aaditya Ramdas. Top-label calibration and multiclass-to-binary reductions. In International Conference on Learning Representations, 2022. URL https://openreview.net/forum?id=WqoBaaPHS-.

Chirag Gupta, Aleksandr Podkopaev, and Aaditya Ramdas. Distribution-free binary classification: prediction sets, confidence intervals and calibration. Advances in Neural Information Processing Systems, 33, 2020.

Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In International Conference on Learning Representations, 2021. URL https://openreview.net/forum?id=eQe8DEWNN2W

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. A distribution-free theory of nonparametric regression. Springer, 2002.

Jaroslav Hajek. Asymptotically Most Powerful Rank-Order Tests. The Annals of Mathematical Statistics, 33(3):1124–1147, 1962.

David J Hand. Construction and assessment of classification rules. Wiley, 1997.

Frank E Harrell. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis Springer, 2015.

Trevor Hastie and Robert Tibshirani. Classi cation by pairwise coupling. The Annals of Statistics, 26(2):451 { 471, 1998.

L Held, Kaspar Ru bach, and Fadoua Balabdaoui. A score regression approach to assess calibration of continuous probabilistic predictions. Biometrics, 66(4):1295{1305, 2010.

Jorgen Hilden, J Dik F Habbema, and Beth Bjerregaard. The measurement of performance in probabilistic diagnosis. Methods of Information in Medicine, 17(4):227{237, 1978.

David W Hosmer and Stanley Lemesbow. Goodness of t tests for the multiple logistic regression model. Communications in Statistics-Theory and Methods, 9(10):1043{1069, 1980.

Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. A tutorial on calibration measurements and calibration models for clinical prediction models. Journal of the American Medical Informatics Association, 27(4):621{633, 2020.

Irina A Ingster, Yuri Iand Suslina. Nonparametric goodness-of- t testing under Gaussian models Springer Science & Business Media, 2012.

Yuri I Ingster. An asymptotic minimax testing of nonparametric hypotheses on the density of the distribution of an independent sample. Journal of Soviet Mathematics, 33(1): 744{758, 1986.

Yuri I Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the $L_p$ metrics. Theory of Probability & Its Applications , 31(2):333{337, 1987.

Yuri I Ingster. Adaptive chi-square tests. Journal of Mathematical Sciences, 99(2):1110{1119, 2000.

Joan Ivanov, Jack V Tu, and C David Naylor. Ready-made, recalibrated, or remodeled? issues in the use of risk indexes for assessing mortality after coronary artery bypass graft surgery. Circulation , 99(16):2098{2104, 1999.

Jiashun Jin and Zheng Tracy Ke. Rare and weak e ects in large-scale inference: methods and phase diagrams.Statistica Sinica, pages 1{34, 2016.

Ian T Jolli e and David B Stephenson. Forecast veri cation: a practitioner's guide in atmospheric science John Wiley & Sons, 2012.

Joseph B Kadane and Sarah Lichtenstein. A subjectivist view of calibration. Technical report, Decision Research, Eugene, OR, 1982.

Daniel Kahneman and Amos Tversky. Prospect theory: an analysis of decision under risk. In Handbook of the Fundamentals of Financial Decision Making: Part I, pages 99{127. World Scienti c, 2013.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. arXiv preprint arXiv:2004.14497, 2020.

Gideon Keren. Calibration and probability judgements: conceptual and methodological issues. Acta Psychologica, 77(3):217–273, 1991.

Gérard Kerkyacharian and Dominique Picard. Estimating nonquadratic functionals of a density using haar wavelets. The Annals of Statistics, 24(2):485–507, 1996.

Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. The Annals of Statistics, 50(1):225–251, 2022.

Keiichi Kisamori, Motonobu Kanagawa, and Keisuke Yamazaki. Simulator calibration under covariate shift with kernels. In International Conference on Artificial Intelligence and Statistics, pages 1244–1253. PMLR, 2020.

Jan Kodovsky and Jessica Fridrich. Calibration revisited. In Proceedings of the 11th ACM workshop on Multimedia and security, 2009.

Meelis Kull, Telmo M Silva Filho, and Peter Flach. Beyond sigmoids: how to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electronic Journal of Statistics, 11(2):5052–5080, 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in Neural Information Processing Systems, 2019.

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In Advances in Neural Information Processing Systems, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In International Conference on Machine Learning. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in Neural Information Processing Systems, 2017.

Béatrice Laurent. Efficient estimation of integral functionals of a density. The Annals of Statistics, 24(2):659–681, 1996.

Erich L Lehmann and Joseph P Romano. Testing statistical hypotheses Springer Science & Business Media, 2005.

Oleg V Lepski. On a problem of adaptive estimation in gaussian white noise. Theory of Probability & Its Applications, 35(3):454–466, 1991.

Oleg V Lepski and Vladimir G Spokoiny. Optimal pointwise adaptive methods in nonparametric estimation. The Annals of Statistics, 25(6):2512–2546, 1997.

Sarah Lichtenstein, Baruch Fischho , and Lawrence D Phillips. Calibration of probabilities: the state of the art. In Decision Making and Change in Human A airs: Proceedings of the Fifth Research Conference on Subjective Probability, Utility, and Decision Making pages 275{324. Springer, 1977.

Mark G Low. On nonparametric con dence intervals. The Annals of Statistics, 25(6): 2547{2554, 1997.

HB Mann and Abraham Wald. On the choice of the number of class intervals in the application of the chi square test. The Annals of Mathematical Statistics, 13(3):306{317, 1942.

Dimitrios Milios, Ra aello Camoriano, Pietro Michiardi, Lorenzo Rosasco, and Maurizio Filippone. Dirichlet-based gaussian processes for large-scale calibrated classi cation. In Advances in Neural Information Processing Systems, 2018.

Michael E Miller, Siu L Hui, and William M Tierney. Validation techniques for logistic regression models. Statistics in Medicine, 10(8):1213{1226, 1991.

Michael E Miller, Carl D Langefeld, William M Tierney, Siu L Hui, and Clement J McDonald. Validation of probabilistic predictions. Medical Decision Making, 13(1):49{57, 1993.

Robert G Miller. Statistical prediction by discriminant analysis. Springer, 1962.

Jacob A Mincer and Victor Zarnowitz. The evaluation of economic forecasts. In Economic forecasts and expectations: Analysis of forecasting behavior and performance pages 3{46. NBER, 1969.

Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and James Robins. Lepski's method and adaptive estimation of nonlinear integral functionals of density. arXiv preprint arXiv:1508.00249, 2015.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. In Advances in Neural Information Processing Systems, 2020.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In AAAI Conference on Arti cial Intelligence , 2015.

Whitney K Newey and James R Robins. Cross- tting and fast remainder rates for semiparametric estimation. arXiv preprint arXiv:1801.09138 , 2018.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In International Conference on Machine learning, 2005.

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

Christine Osborne. Statistical calibration: A review. International Statistical Review / Revue Internationale de Statistique, 59(3):309{336, 1991.

John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classi ers, pages 61{74, 1999.

C Radhakrishna Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In Mathematical Proceedings of the Cambridge Philosophical Society volume 44, pages 50{57. Cambridge University Press, 1948.

James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order in uence functions and minimax estimation of nonlinear functionals. Probability and statistics: essays in honor of David A. Freedman, 2:335{421, 2008.

Leonard J Savage. Elicitation of personal probabilities and expectations. Journal of the American Statistical Association, 66(336):783{801, 1971.

Mark J Schervish. A general method for comparing probability assessors. The Annals of Statistics, 17(4):1856{1879, 1989.

Anton Schick. On Asymptotically E cient Estimation in Semiparametric Models. The Annals of Statistics, 14(3):1139 { 1151, 1986.

Francoise Seillier-Moiseiwitsch and A Philip Dawid. On testing the validity of sequential probability forecasts. Journal of the American Statistical Association, 88(421):355{359, 1993.

Nicolas Serrano. Calibration strategies to validate predictive models: is new always better? Intensive Care Medicine, 38(8):1246{1248, 2012.

Nilay D Shah, Ewout W Steyerberg, and David M Kent. Big data and predictive analytics: recalibrating expectations. JAMA , 320(1):27{28, 2018.

Yandi Shen, Chao Gao, Daniela Witten, and Fang Han. Optimal estimation of variance in nonparametric regression with random design. The Annals of Statistics, 48(6):3589{3618, 2020.

Samuel D Silvey. The lagrangian multiplier test. The Annals of Mathematical Statistics, 30 (2):389{407, 1959.

David J Spiegelhalter. Probabilistic prediction in patient management and clinical trials. Statistics in Medicine, 5(5):421{433, 1986.

Vladimir G Spokoiny. Adaptive hypothesis testing using wavelets. The Annals of Statistics, 24(6):2477{2498, 1996.

Richard Stanley. Eulerian partitions of a unit hypercube. Higher Combinatorics, 31:49, 1977.

Ewout W Steyerberg. Clinical prediction models. Springer, 2019.

Ewout W Steyerberg, Andrew J Vickers, Nancy R Cook, Thomas Gerds, Mithat Gonen, Nancy Obuchowski, Michael J Pencina, and Michael W Kattan. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1):128, 2010.

Bernd Sturmfels. Grobner Bases and Convex Polytopes. American Mathematical Soc., 1996.

Ambrus Tamas and Balazs Csanad Csaji. Exact distribution-free hypothesis tests for the regression function of binary classi cation via conditional kernel mean embeddings. IEEE Control Systems Letters, 2021.

Eric Tchetgen, Lingling Li, James Robins, and Aad van der Vaart. Minimax estimation of the integral of a power of a density. Statistics & probability letters, 78(18):3307{3311, 2008.

Philip E Tetlock and Dan Gardner. Superforecasting: The Art and Science of Prediction. Random House, 2016.

Sunil Thulasidasan, Gopinath Chennupati, Je A. Bilmes, Tanmoy Bhattacharya, and Sarah Ellen Michalak. On mixup training: improved calibration and predictive uncertainty for deep neural networks. In Advances in Neural Information Processing Systems, 2019.

DB Toll, KJM Janssen, Y Vergouwe, and KGM Moons. Validation, updating and impact of clinical prediction rules: a review. Journal of Clinical Epidemiology, 61(11):1085{1094, 2008.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schon. Evaluating model calibration in classi cation. In International Conference on Arti cial Intelligence and Statistics . PMLR, 2019.

Ben Van Calster and Andrew J Vickers. Calibration of risk prediction models: impact on decision-analytic performance. Medical Decision Making, 35(2):162{169, 2015.

Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was de ned: from utopia to empirical data. Journal of Clinical Epidemiology, 74:167{176, 2016.

Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. Calibration: the achilles heel of predictive analytics. BMC medicine, 17(1): 1{7, 2019.

Roman Vershynin. High-dimensional probability: An introduction with applications in data science. Cambridge University Press, 2018.

Vladimir Vovk and Glenn Shafer. Good randomized sequential probability forecasting is always possible. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(5):747{763, 2005.

Vladimir Vovk, Ivan Petej, Paolo Toccaceli, Alexander Gammerman, Ernst Ahlberg, and Lars Carlsson. Conformal calibrators. In *Conformal and Probabilistic Prediction and Applications*. PMLR, 2020.

Lie Wang, Lawrence D Brown, T Tony Cai, and Michael Levine. Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664, 2008.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: a unifying framework. *Advances in Neural Information Processing Systems*, 2019.

Robert L Winkler, Javier Munoz, José L Cervera, José M Bernardo, Gail Blattenberger, Joseph B Kadane, Dennis V Lindley, Allan H Murphy, Robert M Oliver, and David Ríos-Insua. Scoring rules and the evaluation of probabilities. *Test*, 5(1):1–60, 1996.

Bianca Zadrozny and Charles Peter Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference of Machine Learning*, 2001.

Bianca Zadrozny and Charles Peter Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, 2002.

Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. Mix-n-match: ensemble and compositional methods for uncertainty calibration in deep learning. In *International Conference on Machine Learning*. PMLR, 2020.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020.