Measuring multi-calibration

Ido Guy, Daniel Haimovich, Fridolin Linder, Nastaran Okati, Lorenzo Perini, Niek Tax, and Mark Tygert

Central Applied Science at Meta and Fundamental Artificial Intelligence Research at Meta

Abstract

A suitable scalar metric can help measure multi-calibration, defined as follows. When the expected values of observed responses are equal to corresponding predicted probabilities, the probabilistic predictions are known as "perfectly calibrated." When the predicted probabilities are perfectly calibrated simultaneously across several subpopulations, the probabilistic predictions are known as "perfectly multi-calibrated." In practice, predicted probabilities are seldom perfectly multi-calibrated, so a statistic measuring the distance from perfect multi-calibration is informative. A recently proposed metric for calibration, based on the classical Kuiper statistic, is a natural basis for a new metric of multi-calibration and avoids well-known problems of metrics based on binning or kernel density estimation. The newly proposed metric weights the contributions of different subpopulations in proportion to their signal-to-noise ratios; data analyses' ablations demonstrate that the metric becomes noisy when omitting the signal-to-noise ratios from the metric. Numerical examples on benchmark data sets illustrate the new metric.

1 Introduction

1.1 Informal definitions of calibration and multi-calibration

Predicted probabilities are known as "perfectly calibrated" when the expected values of corresponding observed responses are equal to the predicted probabilities. In perfectly calibrated predictions of weather, for example, it would rain on precisely 80% of the days predicted to have an 80% chance of rain (and similarly for any other percentage). Good calibration helps quantify uncertainty and assign levels of confidence to predictions. There are many metrics for measuring calibration of a given set of observations; see, for example, the work of Arrieta-Ibarra *et al.* (2022).

Good calibration of the entire population of observations need not imply good calibration in any particular subpopulation. The task of ensuring good calibration for sundry specified subpopulations simultaneously has been known as "multi-calibration" due to Hebert-Johnson *et al.* (2018). The present paper proposes metrics for measuring multi-calibration.

The present paper builds and follows up on theoretical advances of Haghtalab *et al.* (2023). The latter showed that accounting for statistical fluctuations is critical. Omitting normalization for signal-to-noise ratios (as in other extant works on multi-calibration) results in estimates from finite samples that are mainly noise in the setting of discrete classification; Subsection 3.5 illustrates this in detail via numerical experiments. The present paper proposes practical procedures for measuring multi-calibration, complete with the weighting by signal-to-noise ratios from the theorems of Haghtalab *et al.* (2023). For data sets containing only finitely many observations, extensive earlier work reviewed in Appendix D enables sample-efficient measurement of calibration, which the present paper extends to multi-calibration. The following subsection of this introduction reviews key insights of Haghtalab *et al.* (2023).

1.2 Purposes of metrics for multi-calibration

Any metric for multi-calibration merely complements without replacing traditional metrics of performance. This is key in constructing metrics for multi-calibration. Other metrics could include precision and recall, or sensitivity and specificity. That said, the information and signals in metrics of multi-calibration may help improve accuracy and other metrics, even when multi-calibration is of little direct interest. After all, multi-calibration inherently involves specification of subpopulations or covariates of interest, thus incorporating a-priori, domain-specific knowledge and insights.

Indeed, metrics of calibration for the full population alone (without consideration for any proper subpopulation) can obscure signals from different subpopulations that are suspected to impact accuracy; suspecting such impact on accuracy is common due to domain-specific knowledge. Breaking the full population into subpopulations corresponding to different geographical regions or different human attributes often reveals information that looking at the full population may obscure due to noise (noise can cause the different subpopulations to interfere with each other's estimates).

So, any metric for multi-calibration ideally should account for how the calibrations of the various subpopulations are likely to contribute to the accuracy for the full population. The metric proposed in the present paper takes this into account. The metric weights the contribution from each subpopulation under consideration in inverse proportion to the standard deviation of the corresponding metric of calibration for that single subpopulation. The standard deviation is calculated under the null hypothesis of perfect calibration. This results in smaller subpopulations contributing less, with the contribution to the overall metric being roughly in proportion to the square root of the effective number of observations in the subpopulation, as advocated earlier by Haghtalab *et al.* (2023). Other works omit this weighting that Haghtalab *et al.* (2023) proved to be essential.

To be precise, this noise level for estimates from a subpopulation is inversely proportional to the square root of the number of observations (or to the square root of the total weight of the observations for weighted samples) when drawing the predicted probabilities from a distribution that is independent of the number of observations. The present paper details how to handle any arbitrary distribution of the predicted probabilities, whether or not the underlying distribution depends on the number of observations.

1.3 Maximal benefits in metrics for multi-calibration

To combine estimates of calibration from many subpopulations, the metric for multi-calibration proposed below takes a maximum over all the given subpopulations. Specifically, the metric first measures calibration for every one of the specified subpopulations, then weights them by their signal-to-noise ratios, and finally returns the maximum over all subpopulations. Users typically specify as many subpopulations as they can conceive. Users are welcome to include as many potentially causal connections as they can conceive, forming subpopulations according to cuts along thresholds of covariates known or suspected a-priori to be causally important. With the normalization by the standard deviation discussed above, taking a maximum focuses power on the largest subpopulations when their calibration is poor, while reserving power for the smaller subpopulations when the largest subpopulations are already well-calibrated. Users generally address gross problems with calibration before drilling down into more refined problems; the metric proposed below measures progress made this way. And interpreting a maximum is trivial, focusing attention on the most problematic subpopulations. Taking a maximum to focus on the worst-case has been popular with others, such as Gohar & Cheng (2023), Hansen *et al.* (2024), and Shabat *et al.* (2020). This also aligns with Hebert-Johnson *et al.* (2018) and Haghtalab *et al.* (2023), who consider the full population to be well–multi-calibrated only when every single subpopulation considered is well-calibrated.

1.4 Benefits of good multi-calibration

Many benefits accrue from enhancements to multi-calibration (that is, from enhancing the calibration of each of many subpopulations simultaneously). One benefit could be called "in-distribution generalization." Indeed, such good multi-calibration enhances the weighted average accuracy across the population when re-weighted according to different distributions. With only good global calibration and average accuracy, averaged over the full population, nothing guarantees good performance across different weightings of the examples in the data set. The ability to attain good performance without re-training has various mathematical formulations, including the "universal adaptability" of Kim *et al.* (2022) and Ye & Li (2024), as well as the "out-of-distribution generalization" of Wu *et al.* (2024). Moreover, multi-calibration can automatically adjust for problematic outliers, as part of the "omniprediction" of Gopalan *et al.* (2023) and others.

1.5 Further related work

The work of Hebert-Johnson *et al.* (2018) introduced the term, "multi-calibration," along with an algorithm for improving multi-calibration that leveraged boosting. A software package, "`mcboost`," based on that algorithm was developed by Pfisterer *et al.* (2021). A related notion is "multi–accuracy-in-expectation," with the relation developed by Kim *et al.* (2019) and others. Multi–accuracy-in-expectation concerns the average accuracy for each of several subpopulations, while multi-calibration concerns calibration for the subpopulations. The present paper focuses on metrics of multi-calibration rather than algorithms for improving multi-calibration.

Appendix D reviews popular yet irreparably problematic metrics for calibration, namely the ICI ("integrated calibration index") of Austin & Steyerberg (2019) and the ECE ("empirical calibration error," "estimated calibration error," or "expected calibration error") and reliability diagrams reviewed by Bröcker & Smith (2007), Bröcker (2008), Wilks (2019), and others. As reviewed in Appendix D, the ECE and ICI are inherently limited for the case of assessing calibration from a finite sample of an infinite data set, so would also be inappropriate for assessing multi-calibration. Instead, the present paper uses an approach due to Kolmogorov (1933) and Smirnov (1939) and to Kuiper (1962), later extended to calibration by Arrieta-Ibarra *et al.* (2022), Tygert (2021), and others. The Kuiper metric avoids making an explicit trade-off between averaging away random noise and the statistical power to resolve variations as a function of the predicted probabilities. Lee *et al.* (2023) proved rigorously that such an undesirable trade-off is unavoidable in methods such as the ECE or ICI which leverage binning into histograms or kernel density estimation.

1.6 The presentation in the present paper

This paper adheres to the terminological convention of machine learning that "classification" refers to predictions into discrete classes, while "regression" refers to predictions that can take on a continuum of real values. In contrast, the convention among some statisticians is that "regression" refers to both classification into discrete classes as well as predictions that can take on a continuum of values. "Logistic regression" actually refers to classification; the terminology, "logistic regression," comes from statisticians. For simplicity and concreteness, the present paper focuses on classification into the two classes of Bernoulli variates, 0 ("failure") and 1 ("success"), mentioning the straightforward generalization to general regressions briefly in Appendix C.

The remainder of the present paper has the following structure: Section 2 introduces the main methodology and metric. Section 3 presents results from applying the methods to a couple bench-

mark data sets.* Section 4 summarizes the findings. Appendix A points to the related notion of "proportional" multi-calibration introduced by La Cava *et al.* (2023). Appendix B proposes an adjustment suitable for use in the case of significant imbalance between the numbers of observed instances of the different classes. Appendix C mentions a possible generalization to observed responses that can take real values other than the 0 (failure) and 1 (success) from Bernoulli trials. Appendix D reviews the ECE and ICI and their intrinsic limitations. Appendix E reviews schemes for improving multi-calibration based on boosting or augmentation. Appendix F constructs a synthetic data set relevant for unit testing of the associated software package. Appendix G supplements Section 3 with tables of further numerical results.

2 Methods

This section introduces the methodology of the present paper. Subsection 2.1 sets notation used throughout. Subsection 2.2 introduces the main metric for multi-calibration. Subsection 2.3 details an algorithm for generating subpopulations automatically based on given covariates.

2.1 Notation

This subsection sets notational conventions used throughout the present paper. All results mentioned in the present subsection are discussed, for example, by Tygert (2021).

Consider predicted probabilities $S_1 \le S_2 \le \cdots \le S_{n_0}$, known as "scores," and corresponding numbers R_1, R_2, ..., R_{n_0}, known as "responses," each of which can take either the value 0 or the value 1. Consider also corresponding positive real numbers W_1, W_2, ..., W_{n_0}, known as "weights" for a weighted sampling. (If the sampling is unweighted, then the weights are uniform, with $W_1 = W_2 = \cdots = W_{n_0}$.) The scores will be viewed as deterministic, not random, while the responses will be viewed as random variables that are independent. Consider ℓ subpopulations specified by indices $i_1^k < i_2^k < \cdots < i_{n_k}^k$ for $k = 0, 1, 2, \ldots, \ell$. "Subpopulation" 0 will be the full population, so $i_1^0 = 1$, and $i_2^0 = 2$, ..., and $i_{n_0}^0 = n_0$. The fact that every subpopulation is part of the full population yields that $n_k \le n_0$ for $k = 0, 1, \ldots, \ell$. Requiring n_k to be greater than, say, 9, for all $k = 0, 1, \ldots, \ell$, can help ensure that statistics estimated for the subpopulations are not mostly random noise, too.

Consider any $k = 0, 1, \ldots, \ell$. The cumulative differences for subpopulation k are

$$C_j^k = \frac{\sum_{m=1}^{j} \left(R_{i_m^k} - S_{i_m^k} \right) \cdot W_{i_m^k}}{\sum_{m=1}^{n_k} W_{i_m^k}} \tag{1}$$

for $j = 1, 2, \ldots, n_k$; obviously, $-1 \le C_j^k \le 1$ for $j = 1, 2, \ldots, n_k$. We also set $C_0^k = 0$. The Kuiper metric for subpopulation k is

$$D_k = \max_{0 \le j \le n_k} C_j^k - \min_{0 \le j \le n_k} C_j^k; \tag{2}$$

clearly, $0 \le D_k \le 1$, providing a universally normalized measure of the effect size for miscalibration. *The Kuiper metric is also equal to the absolute value of the total miscalibration, totaled over the interval of scores for which the absolute value of the total is greatest.* That is,

$$D_k = \max_{1 \le p \le q \le n_k} \left| \frac{\sum_{m=p}^{q} \left(R_{i_m^k} - S_{i_m^k} \right) \cdot W_{i_m^k}}{\sum_{m=1}^{n_k} W_{i_m^k}} \right|. \tag{3}$$

*An open-source software package that can automatically reproduce all results reported below is available at `https://github.com/facebookresearch/McMetric`

2.2 Metric for multi-calibration

This subsection introduces the main metric for multi-calibration that the present paper proposes. The appendices discuss possible alternatives that could be relevant for special circumstances.

The metric proposed here measures calibration for every one of the specified subpopulations, then normalizes them by their signal-to-noise ratios, and finally returns the worst case, that is, the maximum over all subpopulations. Focusing on the worst case identifies which subpopulation has the greatest potential for improvement, providing a readily interpretable and actionable diagnostic.

Thus, the multi-calibration metric which considers all subpopulations on an equal footing, in proportion to their signal-to-noise ratios, is

$$M = \max_{0 \le k \le \ell} \frac{D_k \cdot \left(\sum_{j=1}^{n_k} W_{i_j^k} \right) \sqrt{\sum_{j=1}^{n_0} [S_j (1 - S_j) (W_j)^2]}}{\left(\sum_{j=1}^{n_0} W_j \right) \sqrt{\sum_{j=1}^{n_k} \left[S_{i_j^k} \left(1 - S_{i_j^k} \right) \left(W_{i_j^k} \right)^2 \right]}}. \tag{4}$$

Indeed, under the null hypothesis that R_j is drawn independently from the Bernoulli distribution whose mean is S_j, for all $j = 1, 2, \ldots, n_0$, the standardized value of D_k is proportional to the factor in (4) that does not involve the full population $k = 0$. This normalization of D_k ensures that its distribution under the null hypothesis is similar to the distribution of the absolute value of a standard normal variate. Under the null hypothesis, after all, $C_0^k, C_1^k, C_2^k, \ldots, C_{n_k}^k$ is a driftless random walk and the standard deviation of $C_{n_k}^k$ is

$$\frac{\sqrt{\sum_{j=1}^{n_k} \left[S_{i_j^k} \left(1 - S_{i_j^k} \right) \left(W_{i_j^k} \right)^2 \right]}}{\sum_{j=1}^{n_k} W_{i_j^k}} \tag{5}$$

for $k = 0, 1, \ldots, \ell$. When interpreting (4) and (5), remember that the scores and weights are non-random while the responses are independent random variables.

The expected value of D_k calculated under the null hypothesis is at most $2\sqrt{2/\pi} \approx 1.6$ times the quantity in (5) and rapidly converges to that upper bound in the appropriate asymptotic limits, as proven in Section 3.1 of Tygert (2023). The dependence in (4) on the expression in (5) incorporates the lessons of Haghtalab *et al.* (2023).

The factor in (4) that does involve the full population $k = 0$ ensures that the weight of D_0 in the maximum from (4) is 1 (that is, the argument to the max for $k = 0$ is simply D_0). The most direct measure of effect size for the full population is D_0, so the metric M in (4) directly assesses effect size with reference to the same universal scale ranging from 0 to 1 as for any Kuiper metric.

When the denominator in (4) is 0 or nearly 0 and the numerator is also 0 or nearly 0, the resulting quotient should be taken to be 0. This consideration is critical for implementation in finite-precision arithmetic (such as floating-point or fixed-precision).

2.3 Algorithm for automatically generating subpopulations

This subsection proposes a randomized method for generating subpopulations which correspond to various ranges of the values of covariates. By construction, the subpopulations generated reflect information contained in the given covariates. This allows for the automatic generation of subpopulations given only variables that incorporate information of domain- or application-specific interest.

The algorithm aims to sample at random paths in a binary tree. The high-level strategy is to split recursively according to values of the covariates, building a binary tree for which each subpopulation generated corresponds to a path starting from the root. The generation includes all such subpopulations along the path from the root to a leaf; that is to say, the subpopulations generated include one for every subpath starting at the root.

The root of the tree corresponds to the full population. We then choose one of the covariates at random and retain only those members whose values of the covariate are less than the median of their unique values or instead are greater than or equal to the median (with the choice between less than and greater than or equal to the median made at random). We split subpopulations recursively, always choosing the next covariate at random (with replacement). The median of the "unique values" takes the median of the list of distinct values of the covariate for the members of the subpopulation prior to splitting. Again, all subpopulations along a single path from the root to a leaf are retained for the final set generated; the leaf corresponds to a subpopulation whose number of members is no less than a user-specified threshold (typically 10 or so, though the threshold can be as low as 1, if the user prefers).

For ordinal covariates, namely, variates for which their values come with a total ordering, the splits of the previous paragraph are uniquely, unambiguously defined. Ordinal variates include real-valued variates, integer-valued variates, and more generally any categorical variates for which the categories are ordered. For nominal variates, which are categorical variates for which the categories have no natural ordering, we re-order the categories at random every time we start building a path from the root to a leaf. We keep the same (randomized) order of the categories for each nominal variate until we reach a leaf, then re-randomize the orders for all nominal variates when starting a new path back at the root.

For implementations in practice, further optimizations are useful. There is no need to save in memory all subpopulations generated; instead, each subpopulation can be generated without saving any of the others. Furthermore, recording the splits used for subpopulations generated earlier can enable efficiently checking whether the same splits were used before (in which case generating the same subpopulation again would be redundant). Recording only the splits will typically require far less memory than storing all the indices in the subpopulations; after all, there can be only just so many levels in a binary tree.

Algorithm 1 provides pseudocode that omits the refinements mentioned in the previous paragraph. The Python codes in the associated GitHub repository include all the refinements.[†] The refinements are for efficiency only, not required for appreciation of the basic idea.

3 Results and discussion

This section presents and discusses the results of experiments on two standard benchmark data sets, namely the 1998 KDD Cup and the 2019 American Community Survey of the United States Census Bureau.[‡] Subsection 3.1 details the configurations of the experiments. Subsection 3.2 describes the machine-learned models considered. Subsection 3.3 describes the 2019 American Community Survey; Subsection 3.4 describes the 1998 KDD Cup. Subsection 3.5 discusses the results. Further examples are available in Appendices F and G.

The purpose of the present section is to illustrate the methods of the previous section, Section 2.

[†]Open-source Python scripts and libraries that can reproduce all results of the present paper are available at `https://github.com/facebookresearch/McMetric`

[‡]Open-source software that can automatically reproduce all results reported in the present paper is available at `https://github.com/facebookresearch/McMetric`

Algorithm 1: Randomized generation of subpopulations

Input: A positive integer m that will be the minimum possible number of members in a subpopulation, a positive integer ℓ that will be the number of subpopulations generated, and p sets of real numbers $a_1^{(1)}, a_2^{(1)}, \ldots, a_{n_0}^{(1)}$; $a_1^{(2)}, a_2^{(2)}, \ldots, a_{n_0}^{(2)}$; $\ldots$; $a_1^{(p)}, a_2^{(p)}, \ldots, a_{n_0}^{(p)}$ that are the values of "covariates" (also known as "features")

Output: A sequence of subpopulations $I_0, I_1, I_2, \ldots, I_\ell$, where each subpopulation refers to a set of indices specifying the subpopulation, namely, I_k is the set of indices $i_1^k, i_2^k, \ldots, i_{n_k}^k$

Set $k = 0$.
repeat
  For every nominal covariate, say the jth covariate, re-order at random the values taken by the full population's $a_1^{(j)}, a_2^{(j)}, \ldots, a_{n_0}^{(j)}$. (Maintain the original order of each ordinal covariate; randomize only the ordering of the categories for every nominal covariate.)
  Set I_k to be the set of all indices, 1, 2, $\ldots$, n_0 (recall that $i_1^0 = 1$, and $i_2^0 = 2$, $\ldots$, and $i_{n_0}^0 = n_0$).
  repeat
    Choose an integer j uniformly at random from 1, 2, $\ldots$, p, thus choosing one of the covariates at random.
    Calculate the median v of the distinct values that the values $a_{i_1^k}^{(j)}, a_{i_2^k}^{(j)}, \ldots, a_{i_{n_k}^k}^{(j)}$ of the jth covariate take.
    With probability 1/2, set I_{k+1} to be the members of I_k whose corresponding values $a_{i_1^k}^{(j)}, a_{i_2^k}^{(j)}, \ldots, a_{i_{n_k}^k}^{(j)}$ are less than the median v of their distinct values; while for the other probability 1/2, set I_{k+1} to be the members of I_k whose covariate values are greater than or equal to the median v.
    Increment k by 1 (that is, add 1 to k).
  until *there are strictly fewer than m members in I_k.*
  Decrement k by 1 (that is, subtract 1 from k).
until $k \geq \ell$.
return *the subpopulations I_0, I_1, I_2, $\ldots$, I_ℓ.*

In particular, Subsection 3.5 demonstrates the importance of normalizing by signal-to-noise ratios in the metrics for multi-calibration; omitting such normalization results in metrics that are mostly noise. The extensive tables of Appendix G refer to the metrics which omit such normalization as "multi-ablate," as appropriate for an ablation to demonstrate the necessity of normalization. The present section also discusses the results of calibrating via different methods and finds that isotonic regression is a strong baseline; Appendix G considers various methods designed specifically to improve multi-calibration, too. Finally, in most cases the metric for multi-calibration of (4) is significantly greater than the metric for calibration of (2) and (3), demonstrating that the metric for multi-calibration discovers miscalibration beyond that measured in the metric for calibration of the full population.

Appendix G includes a comprehensive battery of tests. The appendix considers all combinations of three base machine-learned models coupled with five different methods for calibration (plus the original base with no explicit calibration) coupled with the two data sets considered (with three different responses considered for the first data set) coupled with metrics of calibration and multi-calibration (both with and without weighting by signal-to-noise ratios). All together, this yields $3 \cdot 6 \cdot 4 = 72$ examples across every possible configuration, with each example evaluated across all metrics. Subsection 3.5 summarizes conclusions drawn from the results of the experiments.

3.1 Experimental set-up

This subsection provides details about the experiments conducted. The following subsection details the machine-learned models used.

All results reported are on a test set obtained by reserving a quarter of the data examples for the test set. We set the parameter ℓ in (4) and the implementation of Algorithm 1 (ℓ is the maximum number of subpopulations considered) to be $\ell = 1{,}000$. We set the parameter m in Algorithm 1 (m is the minimum size of a subpopulation) to be $m = 10$.

The cross-validations for Platt scaling and for isotonic regression leverage the Python package `scikit-learn` of Pedregosa *et al.* (2011), using `sklearn.calibration.CalibratedClassifierCV` with the default settings. The default settings perform five-fold cross-validation, holding out a fifth of the original training examples as a validation set for each of the five folds. The Platt scaling and isotonic regression use the validation set to calibrate the model that is fitted on the other four fifths of the original training set. The final calibrated estimator is the ensemble average over the five pairs of models and calibrations fit via the five folds.

3.2 Statistical models

This subsection details the machine-learned models considered in the numerical experiments of the present section. All values of metrics reported pertain to scores that are predictions made with these models.

We consider three models for classification: logistic regressions, decision trees (also known as "classification and regression trees" — CART), and the gradient-boosted decision-trees of LightGBM of Ke *et al.* (2017). For the first two, we use the implementations in the Python package `sklearn` of Pedregosa *et al.* (2011); for the last, we use the implementation in the Python package `lightgbm` of Ke *et al.* (2017).

We stick with the default parameters, aside from the following choices (which our experiments indicate are close to optimal for the data sets considered):

- `sklearn.linear_model.LogisticRegression`: `tol` $= 0.01$, `solver` $=$ saga, `max_iter` $= 1000$;

- `sklearn.tree.DecisionTreeClassifier`: `max_depth` = 4;
- `lightgbm.LGBMClassifier`: `max_depth` = 3, `num_leaves` = $2^{\texttt{max\_depth}}$, `n_estimators` = 10.

In general, none of the models was dramatically superior to the others in the empirical results reported by the tables. Decision trees or their boosted versions in LightGBM look to be safe choices.

3.3 American Community Survey of the U.S. Census Bureau

This subsection describes the first data set considered. Appendix G reports further numerical results from this data, via Tables 1–9. The present subsection includes two brief case studies for readers who prefer not to dive into the comprehensive battery of experiments in the appendix and its discussion in Subsection 3.5.

The 2019 American Community Survey of the U.S. Census Bureau samples households from across California; we focus on California's Orange County alone except in the last three paragraphs of the present subsection.[§] The sampling is weighted, so we retain only those households whose weights ("WGTP" in the microdata), personal income ("HINCP"), adjustment factor to income ("ADJINC"), and number of people ("NP") are strictly positive (eliminating vacant addresses and group quarters), leaving 10,680 households in the sample of Orange County. We included 233 covariates from the data set. We normalized every covariate by the maximum value that the retained households attain. We consider predicting one of three response variates: the presence of a tablet computer, the presence of a laptop computer, and the presence of a smartphone in the household (the response takes the value 1 when the associated computational device is present in the household and takes the value 0 otherwise). The covariates are also the features in Algorithm 1, so the parameter p of Algorithm 1 is $p = 233$ when analyzing Orange County.

With regard to all the different methods for calibrating the predicted probabilities considered in Appendix G, the empirical results reported in the tables of Appendix G indicate that isotonic regression is often the best or nearly the best among the methods considered. So the remainder of the present subsection focuses on isotonic regression when performing calibration.

For a brief case study, we train a LightGBM model and then calibrate via isotonic regression with cross-validation, when the response variate is the presence of a tablet computer in the household. Tables 1–3 show that this configuration coupling LightGBM with isotonic regression outperforms all other configurations. In this case, the Kuiper metric prior to calibration is 0.04085, while its expected value calculated under the null hypothesis of perfect calibration is about 0.01609, for a ratio of 2.539. The Kuiper metric after calibration is 0.01379, while its expected value under the null hypothesis is about 0.01442, for a ratio of 0.9562. Thus, isotonic regression greatly reduces both the absolute value of the Kuiper metric as well as its ratio relative to its expected value.

Similarly, an evaluation of the metric M for multi-calibration yields 0.07251 prior to calibration; dividing by the standard deviation from (5) for the full population (calculated under the null hypothesis) yields 7.192. This metric M of multi-calibration is 0.04085 after calibration; dividing by the standard deviation yields 4.520. Again, the isotonic regression drastically reduces both the absolute value as well as the ratio relative to the expected value under the null hypothesis. The argmax (worst-case) subpopulation prior to calibration arises from splitting along the variables for the adjusted household personal income and for one of the independent samples of the weights. The argmax subpopulation after calibration involves several more levels of refinement, splitting along the variables for the index in an ordered list of the language spoken, for the sexes of the

[§]The microdata of the American Community Survey is available from the United States Census Bureau at `https://www.census.gov/programs-surveys/acs/microdata.html`

householders (ordered to cover all cases), for an indicator that takes the value 1 when having a telephone is imputed and 0 otherwise, and for two independent samples of the weights. (The weights come with independent samples that help characterize the sampling design — the survey is not a full census but instead arises from random sampling whose design the independent samples of the weights help describe.) Thus, the isotonic regression results in the argmax subpopulation being much further down the binary tree of splits along covariates.

For completeness, we consider also all counties of California (rather than only Orange County), together with an extra nominal covariate that specifies the county for each individual household. The larger data set which includes all of California has a total of 134,094 households in its weighted sample. For this larger data set, we use ten other covariates (so eleven in all), instead of the hundreds considered earlier, in order to illustrate both settings. The full list of covariates is the county in which the members of the household reside, the number of people in the household, the number of related children in the household, the number of vehicles in the household, the number of rooms for the household, as well as binary indicators of whether the household got food stamps, has broadband access to the Internet, has high-speed access to the Internet, has satellite access to the Internet, has a laptop computer, and has a smartphone. We again consider predicting whether the household has a tablet computer. As in the earlier case study, we use LightGBM as the model and calibrate via isotonic regression with cross-validation (isotonic regression again outperformed across all metrics both Platt scaling and the three iterations of augmentation from Appendix E with any of the three models from Subsection 3.2).

For this larger example, the test error without including any covariates at all is 0.3471, the test error after training the model but before calibration reduces to 0.2485, and the test error reduces still further to 0.2434 after calibration with isotonic regression. The Kuiper metric is 0.05665 before calibration and 0.003738 after calibration with isotonic regression. The corresponding expected values calculated under the null hypothesis of perfect calibration — about 1.6 times the standard deviation from (5) — are 0.004970 and 0.004530. The corresponding numbers for an evaluation of the metric M of (4) for multi-calibration are 0.05665 and 0.01738. Thus, calibration via isotonic regression with cross-validation improves performance significantly across all metrics.

We could also consider ignoring the signal-to-noise ratio used for M in (4), reporting simply $\max_{0 \leq k \leq \ell} D_k$, where D_k is the Kuiper metric of (2) and (3) for subpopulation k. When ignoring the signal-to-noise ratio, the metric becomes .4069 prior to calibration and .5347 after calibration, while the expected values under the null hypothesis of perfect calibration for the argmax subpopulation would be 0.2041 and .1862 — around half the observed values. Given that 95% confidence intervals extend roughly two standard deviations with respect to the mean values, this shows that ignoring the signal-to-noise ratio results in metrics that are very noisy.

3.4 KDD Cup 1998

This subsection describes the other data considered. Tables 10–12 of Appendix G report numerical results on the data described in the present subsection. The following subsection, Subsection 3.5, discusses the results.

A national veterans organization in the U.S. conducted an experiment in 1994, mailing solicitations for donations. The 1998 Knowledge Discovery and Data-mining (KDD) Cup competition leveraged this data.[¶] We included 324 covariates from the data set for the 80,796 individuals who got mailings in both 1995 and 1996 and donated at least once. We used these covariates to predict whether the organization deems the recipient to be a "star donor" (the data set conveniently

[¶]The 1998 KDD Cup's data is available at `https://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html`

provides a response variate that takes the value 1 when an individual is a "star donor" and takes the value 0 otherwise). We normalized every covariate to range from 0 to 1, first subtracting off the minimum value of the covariate. The covariates are also the features in Algorithm 1, so the parameter p of Algorithm 1 is $p = 324$ in the present subsection.

As discussed in the following subsection, the empirical results reported in the tables of Appendix G demonstrate the importance of weighting by signal-to-noise ratios in the metric for multi-calibration. With regard to all the different methods for calibrating the predicted probabilities considered in Appendix G, the tables of Appendix G indicate that the thrice-augmented decision-trees are often the best or nearly the best among the methods considered. However, calibration via isotonic regression is competitive.

3.5 Discussion

This subsection discusses the results and implications for methodology.

Isotonic regression for calibration appears to be competitive in the numerical results reported by the tables for the data sets considered here. Thus, isotonic regression can be a good baseline for calibration, at the very least. Subsection 3.3 above considered this baseline in detail. In general, combining different methods appears to yield greater benefits than, say, using for the augmentations of Appendix E the same statistical model as used for training prior to any calibration.

Neglecting the signal-to-noise ratio when forming the metric of multi-calibration yields the "multi-ablate" metric reported in the tables of Appendix G, which is also the simple maximum ($\max_{0 \leq k \leq \ell} D_k$) whose discussion concludes Subsection 3.3 above. This metric is the same as the multi-calibration metric M defined in (4) except without any adjustment for the standard deviation in (5). This "multi-ablate" metric looks to be completely useless; its value is very large uniformly across all experiments and configurations. Omitting adjustment for the signal-to-noise ratio in (5) causes the metric to focus on noise from the smallest subpopulations.

In fact, the values of the "multi-ablate" metric are seldom more than a few times greater than the corresponding "expectation" values calculated under the null hypothesis of perfect calibration. This illustrates that neglecting the signal-to-noise ratio results in noisy metrics, given that 95% confidence intervals extend about two standard deviations with respect to the expected values. The conclusion of Subsection 3.3 discussed this; comparing the last two rows in each of Appendix G's Tables 1–12 gives 72 more examples. This strongly confirms the lessons of Haghtalab *et al.* (2023).

The metric M defined in (4) likely is most informative when the subpopulations are calibrated reasonably well, that is, when they are close to satisfying the null hypothesis of perfect calibration. Weighting subpopulations in inverse proportion to the standard deviations (under the null hypothesis) of their individual Kuiper metrics is most relevant when the subpopulations do not deviate too significantly from the null hypothesis. The metric M defined in (4) effects this weighting and so could be viewed as a regularized measure of calibration for the full population, regularized with consideration for the expected levels of noise on the metrics for the individual subpopulations.

When subpopulations are far from perfect calibration, the multiplicative weighting by the reciprocal of the expected level of noise in the metric M will tend to suppress the values of the Kuiper metric for smaller subpopulations that would otherwise have contributed strongly to the maximum over subpopulations that M reports. Thus, M focuses first on calibration for the full population and would pay more attention to miscalibration of the smaller subpopulations only as calibration improves — when calibration of the full population is poor, the maximum defining M will tend to report that miscalibration rather than any of the smaller subpopulations'.

Many users are likely to include subpopulations so small that noise becomes a real concern, requiring some sort of weighting according to statistical significance. Indeed, users may try to

squeeze the strongest possible statistical inferences out of the limited data that they have. The original proposition of Hebert-Johnson *et al.* (2018) addresses the resulting issues of statistical estimation by considering circuit complexity and oracles (as in the computational complexity theory of theoretical computer science). The present paper addresses statistical issues arising from statistics, instead. The statistical issues come from the responses being discrete when the task is classification (whereas the responses need not be discrete when the task is regression more generally). Even when the responses are discrete in a binary classification, the underlying probabilities need not be always either 0 or 1. The statistical issues arise from the noise inherent in the observed responses being only 0 or 1 while their expected values can take values strictly between 0 and 1, too.

4 Conclusion

A natural metric for measuring multi-calibration across a set of subpopulations is the maximum (over the set of subpopulations) of the subpopulation's Kuiper metric, each normalized by its standard deviation calculated under the null hypothesis of perfect calibration. The Kuiper metric here refers to that measuring miscalibration of the predicted probabilities for the corresponding subpopulation. There is also a straightforward algorithm for constructing subpopulations based on the values of covariates ("covariates" are also known as "features"), detailed in Subsection 2.3 and Algorithm 1. Taking the maximum in (4) and also multiplying by the signal-to-noise ratio (the reciprocal of the standard deviation) has many advantages, discussed in Section 1 and in Subsection 3.5; the metric proposed above secures these advantages. The following appendices suggest straightforward extensions required to address so-called "proportional multi-calibration," to address low-prevalence settings, and to address continuous regressions rather than only discrete classifications. The next appendix discusses advantages of the Kuiper metric over the ECE and ICI. The appendix after that describes a scheme for improving multi-calibration based on boosting or augmentation. The penultimate appendix provides a synthetic example with closed-form analytic expressions for the metrics. The final appendix provides further numerical illustrations on real data, complementing those in Section 3 above.

Acknowledgements

We would like to thank Léon Bottou, Kamalika Chaudhuri, and Adina Williams.

Appendices

Appendix A adapts the methodology of the present paper to what is known as "proportional" multi-calibration. Appendix B modifies the methods for cases of severe imbalance between the two classes (classes 0 and 1). Appendix C generalizes the methods from discrete classifications to continuous regressions. Appendix D reviews problems with popular metrics (the ECEs and ICIs) for calibration. Appendix E reviews procedures for improving multi-calibration based on boosting or augmentation. Appendix F calculates closed-form analytic expressions of the metrics for a synthetic data set, appropriate for unit testing of the software. Appendix G supplements the results reported in Subsection 3.3 with tables of further results.

A Proportional multi-calibration

"Proportional" calibration and its (entirely straightforward) generalization to multi-calibration was proposed by La Cava *et al.* (2023). Proportional calibration is the special case of weighted calibration in which the weight is inversely proportional to the score, that is, $W_j = 1/S_j$, where W_j is the weight and S_j is the score. A regularized variant was also proposed by La Cava *et al.* (2023); the variant introduces a regularization parameter — a positive real number ρ — and corresponds to the weighting $W_j = 1/\rho$ (when $0 \leq S_j \leq \rho$) and $W_j = 1/S_j$ (when $S_j > \rho$). Another way to regularize that is very similar is to take $W_j = 1/(S_j + \rho)$. This weighting highlights false positives more than false negatives, to the extent that the scores are predicting the classes correctly. The following appendix suggests another method for emphasizing false positives when the prevalence is low for the class for which responses take the value 1.

B Low prevalence

For simplicity, consider the case in which the original data is unweighted. The observed prevalence of the class for which responses take the value 1 is then simply the average response $A = \sum_{j=1}^{n_0} R_j/n_0$. Emphasizing false positives more than false negatives may be desirable when the average A is low. If the responses are assumed to be correct, then emphasizing false positives is possible by introducing weights, $W_j = 1/A$ for j such that $R_j = 1$, and $W_j = 1$ for j such that $R_j = 0$.

C Variance estimates for regressions

When the responses R_1, R_2, …, R_{n_0} can take real values other than only 0 and 1, the standard deviation of $C_{n_k}^k$ under the null hypothesis that $\mathbb{E}[R_j] = S_j$ for $j = 1, 2, \ldots, n_0$, is no longer the value from (5) but instead a good estimate becomes

$$\sqrt{\frac{\sum_{j=1}^{n_k-1} \left(R_{i_j^k} - S_{i_j^k} - R_{i_{j+1}^k} + S_{i_{j+1}^k}\right)^2 \left(W_{i_j^k} + W_{i_{j+1}^k}\right)^2}{4 \left(\sum_{j=1}^{n_k} W_{i_j^k}\right) \left(W_{i_1^k} + W_{i_{n_k}^k} + 2\sum_{j=2}^{n_k-1} W_{i_j^k}\right)}} \tag{6}$$

for $k = 0, 1, \ldots, \ell$. The factor of 4 in the denominator of (6) compensates for including in the numerator the square of twice the weights, as well as both the difference $\left(R_{i_j^k} - S_{i_j^k}\right)$ and the adjacent difference $\left(R_{i_{j+1}^k} - S_{i_{j+1}^k}\right)$ (these two differences are independent, so the variance of their difference is the sum of their variances). Including both differences cancels any linear trend that might occur when there is miscalibration, with the minor disadvantage of decreasing the resolution of the estimate slightly (due to the averaging across adjacent indices). The estimator in (6) is a special case of that proposed by Tygert (2024).

D ECEs and ICIs are popular yet poor metrics for calibration

The ECE ("empirical calibration error," "estimated calibration error," or "expected calibration error") and its other, less common name ("ICI" — "integrated calibration index") refer to long-standing metrics of calibration. Statistical estimators for them are inherently problematic when used for finite sample sizes (that is, for any finite data set), however. Such estimators involve

binning into histograms or use of kernel density estimation. The present paper uses the Kuiper metric of (2) and (3) instead, for the following reasons.

Arrieta-Ibarra *et al.* (2022) proved rigorously together with illustrative practical examples that the ECE or ICI require "observations whose density is infinitely higher than what the ECCEs [Kuiper metrics] require to attain the same consistency and power against any fixed alternative distribution, where 'infinitely higher' means asymptotically, in the limit of the sample size tending to infinity (or in the limit of the statistical confidence level tending to 100%).... The ECEs also exhibit an extreme dependence on the choice of bins, with different choices of bins yielding significantly different values for the ECE metrics; choosing among the possible binnings can be confusing, yet makes all the difference. In contrast, the ECCEs [Kuiper metrics] yield trustworthy results without needing such large numbers of observations and without needing to set any parameters."

Lee *et al.* (2023) proved rigorously that this failure of the ECE or ICI to meet even the most basic criteria for statistical acceptability (non-trivial power and/or consistency asymptotically) is unavoidable in any method based on histograms or kernel density estimation. Lists of references which point out problems with the ECE or ICI are available from Tygert (2021) and Arrieta-Ibarra *et al.* (2022), among others. The abstract of the latter concludes, "metrics based on binning or kernel density estimation unavoidably must trade-off statistical confidence for the ability to resolve variations as a function of the predicted probability or vice versa. Widening the bins or kernels averages away random noise while giving up some resolving power. Narrowing the bins or kernels enhances resolving power while not averaging away as much noise. The cumulative [Kuiper] methods do not impose such an explicit trade-off. Considering these results, practitioners probably should adopt the cumulative approach as a standard for best practices." The present paper follows the recommended best practices.

E Boosting for multi-calibration

This appendix reviews and amends an algorithm introduced by Hebert-Johnson *et al.* (2018) to improve calibration over many subpopulations simultaneously.

The original algorithm improved calibration by regressing the residuals (residual compared to perfect calibration) of successive improvements to calibration, repeatedly cycling through all subpopulations targeted. The original algorithm's improvements to calibration involved binning the scores and adjusting the predictions in each bin to make them perfectly calibrated on average.

The present paper considers a closely related technique, which enables much higher efficiency. We consider as input a model for classification (or regression) that has been trained to yield predictions for half the provided training set. Then, given any model for classification (or regression), we fit the new model to the other half of the provided training set, using not only the original covariates but also the predicted confidences from the already trained input model. That is, we fit the new model using the original covariates augmented with the predicted probabilities from the already trained model. Moreover, we can repeat this fitting several times (using the same half of the training data set), each time updating the scores used to augment the original variates with the confidences for the new predictions.

F A synthetic example with closed-form analytic expressions

This appendix synthesizes a data set for which closed-form analytic expressions are available for the Kuiper metrics, associated standard deviations, and metrics of multi-calibration.

We choose an odd positive integer q and set the number of observations in the full population to $n_0 = q(q+1)$.

We set the scores to be linear functions of the index, namely,

$$S_j = \frac{2j+q}{2(q+1)^2} \tag{7}$$

for $j = 1, 2, \ldots, n_0$.

We use uniform weights,

$$W_j = 1 \tag{8}$$

for $j = 1, 2, \ldots, n_0$.

We form q blocks of $q+1$ responses; for $j = 1, 2, \ldots, q$:

$$R_i = 1 \tag{9}$$

for $i = (j-1)(q+1)+1, (j-1)(q+1)+2, \ldots, (j-1)(q+1)+j$, and

$$R_i = 0 \tag{10}$$

for $i = (j-1)(q+1)+j+1, (j-1)(q+1)+j+2, \ldots, (j-1)(q+1)+q+1$.

Combining (9) and (10) yields that

$$\sum_{i=(j-1)(q+1)+1}^{j(q+1)} R_i = j \tag{11}$$

for $j = 1, 2, \ldots, q$. It follows from (7) that

$$\sum_{i=(j-1)(q+1)+1}^{j(q+1)} S_i = j \tag{12}$$

for $j = 1, 2, \ldots, q$, too.

The cumulative sum of the responses is therefore always greater than or equal to the cumulative sum of the scores, for every index. The maximum deviation will hence occur at one of the indices $(j-1)(q+1)+j$ for $j = 1, 2, \ldots, q$. The cumulative response at such an index is

$$\sum_{i=1}^{(j-1)(q+1)+j} R_i = \frac{j(j+1)}{2} \tag{13}$$

for $j = 1, 2, \ldots, q$. The corresponding cumulative score is

$$\sum_{i=1}^{(j-1)(q+1)+j} S_i = \frac{j(q+2)(j(q+2)-q-1)}{2(q+1)^2} \tag{14}$$

for $j = 1, 2, \ldots, q$.

The Kuiper metric depends on the maximum deviation between the cumulative response and the cumulative score. Combining (13) and (14) yields that the maximum deviation is

$$\max_{1 \le j \le q} \frac{j[(j+1)(q+1)^2 - (q+2)(j(q+2)-q-1)]}{2(q+1)^2}. \tag{15}$$

Differentiating the argument to the maximum in (15) with respect to j and setting to 0 yields that the argmax is $j = (q+1)/2$, so (15) becomes

$$\max_{1 \le j \le q} \frac{j[(j+1)(q+1)^2 - (q+2)(j(q+2) - q - 1)]}{2(q+1)^2} = \frac{2q+3}{8}. \tag{16}$$

Normalizing (16) by the size (n_0) of the full population yields the Kuiper metric

$$D_0 = \frac{2q+3}{8q(q+1)}. \tag{17}$$

For subpopulations, we select the middlemost blocks of indices

$$i_j^k = k(q+1) + j \tag{18}$$

for $j = 1, 2, \ldots, n_k$, with

$$n_k = n_0 - 2k(q+1) = (q-2k)(q+1) \tag{19}$$

for $k = 1, 2, \ldots, \ell$, with $\ell = (q-1)/2$.

For every subpopulation, $k = 1, 2, \ldots, \ell$, the maximum cumulative deviation again occurs at the index $i_j^k = (q-1)(q+1)/2 + (q+1)/2 = q(q+1)/2$, which determines the Kuiper metrics for the subpopulations. The sum of the responses at that point is

$$\sum_{j=k(q+1)+1}^{q(q+1)/2} R_j = \sum_{j=k}^{(q+1)/2} j = \frac{(q+1)(q+3)}{8} - \frac{k(k+1)}{2} \tag{20}$$

and the sum of the scores is

$$\sum_{j=k(q+1)+1}^{q(q+1)/2} S_j = \frac{(q(q+1)/2 + k(q+1) + 1 + q)(q(q+1)/2 - k(q+1))}{2(q+1)^2} = \frac{q(q+2) - 4k(k+1)}{8} \tag{21}$$

for $k = 1, 2, \ldots, \ell$, with $\ell = (q-1)/2$. Subtracting (21) from (20) yields the maximum cumulative deviation:

$$\frac{2q+3}{8} \tag{22}$$

for $k = 1, 2, \ldots, \ell$, with $\ell = (q-1)/2$. Normalizing (22) by the size (n_k) of the subpopulation yields the Kuiper metric

$$D_k = \frac{2q+3}{8(q-2k)(q+1)} \tag{23}$$

for $k = 1, 2, \ldots, \ell$, with $\ell = (q-1)/2$.

To calculate the standard deviation from (5), we need to compute the sum of the scores and the sum of the squares of the scores, for the full population and every subpopulation considered. The sum of the scores is

$$\sum_{j=k(q+1)+1}^{n_0 - k(q+1)} S_j = \frac{(k(q+1) + 1 + q(q+1) - k(q+1) + q)(q-2k)(q+1)}{2(q+1)^2} = \frac{(q-2k)(q+1)}{2} \tag{24}$$

for subpopulation k, for $k = 0, 1, \ldots, \ell$. The sum of the squares of the scores is

$$\sum_{j=k(q+1)+1}^{n_0-k(q+1)} (S_j)^2 = \frac{\sum_{j=k(q+1)+1}^{q(q+1)-k(q+1)} (4j^2 + 4jq + q^2)}{4(q+1)^4}$$

$$= \frac{2q(q(q+1) - k(q+1) + k(q+1) + 1)(q-2k)(q+1) + q^2(q-2k)(q+1) + 4\sum_{j=k(q+1)+1}^{q(q+1)-k(q+1)} j^2}{4(q+1)^4}$$

$$= \frac{1}{12(q+1)^4}\Big(4q^6 - 12kq^5 + 18q^5 + 12k^2q^4 - 48kq^4 + 33q^4 - 8k^3q^3 + 36k^2q^3 - 78kq^3 + 31q^3$$
$$- 24k^3q^2 + 36k^2q^2 - 66kq^2 + 14q^2 - 24k^3q + 12k^2q - 28kq + 2q - 8k^3 - 4k\Big) \quad (25)$$

for subpopulation k, for $k = 0, 1, \ldots, \ell$. Subtracting (25) from (24) yields the standard deviation from (5):

$$\frac{1}{n_k}\sqrt{\sum_{j=k(q+1)+1}^{n_0-k(q+1)} S_j(1-S_j)}$$

$$= \frac{1}{(q-2k)(q+1)^3\sqrt{12}}\sqrt{2q^6 + 12q^5 - 12k^2q^4 - 12kq^4 + 27q^4 + 8k^3q^3 - 36k^2q^3 - 42kq^3 + 29q^3}$$
$$\overline{+ 24k^3q^2 - 36k^2q^2 - 54kq^2 + 16q^2 + 24k^3q - 12k^2q - 32kq + 4q + 8k^3 - 8k} \quad (26)$$

for subpopulation k, for $k = 0, 1, \ldots, \ell$.

Combining (23) and (26) yields that the metric M of multi-calibration from (4) is

$$M = \max_{0\le k\le \ell} \frac{(2q+3)\sqrt{2q^5 + 12q^4 + 27q^3 + 29q^2 + 16q + 4}}{8(q+1)} \Big/ \sqrt{2q^7 + 12q^6}$$
$$\overline{- 12k^2q^5 - 12kq^5 + 27q^5 + 8k^3q^4 - 36k^2q^4 - 42kq^4 + 29q^4 + 24k^3q^3 - 36k^2q^3 - 54kq^3 + 16q^3}$$
$$\overline{+ 24k^3q^2 - 12k^2q^2 - 32kq^2 + 4q^2 + 8k^3q - 8kq}\,. \quad (27)$$

The partial derivative (with respect to k) of the square of the longest square-root in (27) is

$$-24kq^5 + 24k^2q^4 - 12q^5 - 72kq^4 + 72k^2q^3 - 42q^4 - 72kq^3 + 72k^2q^2 - 54q^3 - 24kq^2 + 24k^2q - 32q^2 - 8q, \quad (28)$$

which is negative for all k in the interval $[0, \ell]$, where $\ell = (q-1)/2$. Thus, the maximum in (27) occurs at $k = \ell$ and so evaluating (27) at $k = \ell$ yields

$$M = \frac{2q+3}{8(q+1)}\sqrt{\frac{2q^5 + 12q^4 + 27q^3 + 29q^2 + 16q + 4}{3q^6 + 15q^5 + 29q^4 + 27q^3 + 13q^2 + 3q}}. \quad (29)$$

The maximum of the Kuiper metrics $D_0, D_1, \ldots, D_\ell$ from (17) and (23) is

$$\max_{0\le k\le \ell} D_k = D_\ell = \frac{2q+3}{8(q+1)}, \quad (30)$$

which is larger than M in (29) by roughly $\sqrt{3q/2}$ when q is sufficiently large. The following appendix refers to (30) as the "multi-ablate" metric.

G Additional numerical results

This final appendix reports further results on the data sets from Subsections 3.3 and 3.4. Tables 1–9 pertain to the data set described in Subsection 3.3. Tables 10–12 pertain to the data set described in Subsection 3.4.

In the tables, the headings have the following meanings:

- "error" refers to the value obtained by subtracting the accuracy from 1; that is, the error is the fraction of classifications that are incorrect.
- "Kuiper" refers to the metric of calibration (for the full population) defined in (2) or, equivalently, in (3).
- "multi-cal." refers to the metric of multi-calibration defined in (4).
- "multi-ablate" refers to the same metric of multi-calibration defined in (4), but retaining only D_k and no other factors, thus reporting simply $\max_{0 \le k \le \ell} D_k$, with no adjustment for the standard deviation given by (5).
- "expectation" refers to the expected value of the Kuiper metric D_k calculated under the null hypothesis of perfect calibration for a subpopulation (the kth) which attains the maximum value of the "multi-ablate" metric $\max_{0 \le k \le \ell} D_k$, with the expected value being roughly 1.6 times the standard deviation calculated via (5). For discussion of the factor of 1.6 used here, see the paragraph following (5).
- "before calibration" refers to the performance of the model on the test set following only training on the training set.
- "Platt scaling" refers to further calibrating via cross-validation using Platt (temperature) scaling.
- "isotonic regression" refers to further calibrating via cross-validation using isotonic regression.
- "LightGBM" refers to three repetitions of augmentation as in Appendix E, using LightGBM (which performs boosting internally to itself).
- "decision trees" refers to three repetitions of augmentation as in Appendix E, using vanilla (unboosted) decision trees.
- "logistic regression" refers to three repetitions of augmentation as in Appendix E, using logistic regression.

Subsection 3.2 provides full details on the models, "LightGBM," "decision trees," and "logistic regression." The lines with "±" in the tables report twice the standard error of the mean for each metric on the previous line. Thus, the intervals of values reported in the tables are roughly 95% confidence intervals for the true mean values. The standard error is taken over three distinct random seeds for generating the subpopulations via Algorithm 1. The estimate of the standard error includes Bessel's correction (so divides by 2 rather than 3 in the estimate for the variance of the 3 observed values for the metric of multi-calibration, which makes the estimate unbiased and more conservative). The variation due to randomization in Algorithm 1 appears to be rather small in the experiments reported here.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.2148	.2273	**.2025**	.2033	.2108	.2650
Kuiper	.04085	.2599	**.01379**	.03886	.04161	.06345
multi-	.07205	.2997	**.03907**	.07085	.06734	.1296
cal.	±.00353	±.00669	±**.00336**	±.00495	±.00031	±.00000
multi-	.4189	.4778	.3765	.3731	.4605	.5983
ablate	±.02447	±.00509	±.00853	±.04365	±.12810	±.05048
expec-	.1965	.2432	.2092	.1966	.1966	.1966
tation	±.06411	±.04683	±.05236	±.06102	±.06102	±.06102

Table 1: LightGBM model on the test set of the U.S. Census Bureau's American Community Survey for tablet computers; the error when using no covariates at all is 0.2650.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.2075	.2294	.2074	**.2033**	.2108	.2650
Kuiper	.02280	.2560	**.01573**	.03886	.04161	.06345
multi-	**.05267**	.2949	.05269	.07085	.06734	.1296
cal.	±**.00325**	±.00398	±.00301	±.00495	±.00031	±.00000
multi-	.4534	.4765	.3912	.3731	.4605	.5983
ablate	±.05802	±.00121	±.03335	±.04365	±.12810	±.05048
expec-	.1965	.2432	.2092	.1966	.1966	.2269
tation	±.06411	±.04683	±.05236	±.06102	±.06102	±.02159

Table 2: Decision-tree model on the test set of the U.S. Census Bureau's American Community Survey for tablet computers; the error when using no covariates at all is 0.2650.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.2650	.2463	.2291	**.2033**	.2108	.2650
Kuiper	.06391	.2345	**.02007**	.03886	.04161	.06345
multi-	.1302	.2792	**.06360**	.07085	.06734	.1296
cal.	±.00000	±.00636	±**.00972**	±.00495	±.00031	±.00000
multi-	.6002	.4751	.5039	.3731	.4605	.5983
ablate	±.05013	±.00634	±.08941	±.04365	±.12810	±.05048
expec-	.2266	.2326	.2218	.1966	.1966	.2269
tation	±.02159	±.05165	±.02515	±.06102	±.06102	±.02159

Table 3: Logistic-regression model on the test set of the U.S. Census Bureau's American Community Survey for tablet computers; the error when using no covariates at all is 0.2650.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.07863	.08192	**.07331**	.08647	.09740	.1087
Kuiper	.03530	.3832	**.01273**	.02568	.03478	.09470
multi-	.04683	.4064	**.03213**	.04173	.07664	.1147
cal.	±.00198	±.00960	±**.00226**	±.00073	±.03826	±.00446
multi-	.4593	.4715	.4598	.4849	.5087	.6074
ablate	±.07141	±.00188	±.04422	±.03613	±.05003	±.04958
expec-	.1361	.2629	.1744	.1087	.1087	.1087
tation	±.08641	±.04788	±.03740	±.07011	±.07011	±.07011

Table 4: LightGBM model on the test set of the U.S. Census Bureau's American Community Survey for laptop computers; the error when using no covariates at all is 0.1087.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	**.07901**	.1087	.08031	.08647	.09740	.1087
Kuiper	**.01415**	.3736	.01609	.02568	.03478	.09470
multi-	.04531	.4003	**.03536**	.04173	.07664	.1147
cal.	±.01833	±.00842	±**.00248**	±.00073	±.03826	±.00446
multi-	.4937	.4660	.4727	.4849	.5087	.6074
ablate	±.04733	±.00170	±.06111	±.03613	±.05003	±.04958
expec-	.1361	.2629	.1744	.1087	.1087	.1818
tation	±.08641	±.04788	±.03740	±.07011	±.07011	±.03708

Table 5: Decision-tree model on the test set of the U.S. Census Bureau's American Community Survey for laptop computers; the error when using no covariates at all is 0.1087.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.1087	.1087	.08798	**.08647**	.09740	.1087
Kuiper	.09312	.3682	**.01039**	.02568	.03478	.09470
multi-	.1131	.3971	**.04089**	.04173	.07664	.1147
cal.	±.00452	±.00907	±**.00387**	±.00073	±.03826	±.00446
multi-	.6112	.4676	.5334	.4849	.5087	.6074
ablate	±.04863	±.00115	±.07509	±.03613	±.05003	±.04958
expec-	.1808	.2430	.1432	.1087	.1087	.1818
tation	±.03679	±.01243	±.02779	±.07011	±.07011	±.03708

Table 6: Logistic-regression model on the test set of the U.S. Census Bureau's American Community Survey for laptop computers; the error when using no covariates at all is 0.1087.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.06080	.06254	**.04585**	.04970	.06256	.06254
Kuiper	.02301	.4079	**.01168**	.01192	.02966	.08586
multi-	.03009	.4259	**.01778**	.01903	.04015	.09898
cal.	±.00415	±.00963	±**.00262**	±.00038	±.00522	±.00665
multi-	.3539	.4650	.2345	.2522	.5262	.4892
ablate	±.02678	±.00139	±.00873	±.03019	±.14432	±.13285
expec-	.1824	.2557	.2335	.1128	.1128	.1128
tation	±.00703	±.00588	±.02589	±.01410	±.01410	±.01410

Table 7: LightGBM model on the test set of the U.S. Census Bureau's American Community Survey for smartphones; the error when using no covariates at all is 0.06254.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.05895	.06254	**.04744**	.04970	.06256	.06254
Kuiper	**.008902**	.4068	.01770	.01192	.02966	.08586
multi-	.01918	.4236	.02125	**.01903**	.04015	.09898
cal.	±.00405	±.00921	±.00044	±**.00038**	±.00522	±.00665
multi-	.3083	.4654	.3103	.2522	.5262	.4892
ablate	±.04641	±.00116	±.07236	±.03019	±.14432	±.13285
expec-	.1824	.2557	.2335	.1128	.1128	.1692
tation	±.00703	±.00588	±.02589	±.01410	±.01410	±.02174

Table 8: Decision-tree model on the test set of the U.S. Census Bureau's American Community Survey for smartphones; the error when using no covariates at all is 0.06254.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.06254	.06254	.06768	**.05714**	.06256	.06254
Kuiper	.09031	.3987	**.006544**	.01161	.02966	.08586
multi-	.1032	.4147	.04495	**.02328**	.04015	.09898
cal.	±.00657	±.00855	±.00000	±**.00297**	±.00522	±.00665
multi-	.4838	.4631	.3015	.3187	.5262	.4892
ablate	±.13255	±.00040	±.03733	±.02098	±.14432	±.13285
expec-	.1715	.1923	.09129	.1128	.1128	.1692
tation	±.02193	±.04695	±.01025	±.01410	±.01410	±.02174

Table 9: Logistic-regression model on the test set of the U.S. Census Bureau's American Community Survey for smartphones; the error when using no covariates at all is 0.06254.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.3424	.3393	**.3388**	.3453	.3459	.4688
Kuiper	.08917	.1538	**.006591**	.06773	.008098	.02284
multi-	.08960	.1539	**.01508**	.06840	.01805	.03667
cal.	±.00000	±.00004	±**.00019**	±.00010	±.00161	±.00972
multi-	.3935	.4447	.4185	.3967	.4073	.4253
ablate	±.02546	±.04623	±.03513	±.03265	±.12692	±.04707
expec-	.2363	.2319	.2364	.2290	.2290	.2290
tation	±.00279	±.02261	±.00311	±.00151	±.00151	±.00151

Table 10: LightGBM model on the test set of the 1998 KDD Cup for star donors; the error when using no covariates at all is 0.4688.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.3514	.3496	.3509	**.3405**	.3459	.4688
Kuiper	**.004652**	.1434	.005189	.06780	.008098	.02284
multi-	.01383	.1436	**.01344**	.06827	.01805	.03667
cal.	±.00144	±.00009	±**.00105**	±.00028	±.00161	±.00972
multi-	.3897	.4681	.3851	.3914	.4073	.4253
ablate	±.04151	±.04725	±.04249	±.00828	±.12692	±.04707
expec-	.2363	.2319	.2364	.2290	.2290	.2317
tation	±.00279	±.02261	±.00311	±.00151	±.00151	±.02247

Table 11: Decision-tree model on the test set of the 1998 KDD Cup for star donors; the error when using no covariates at all is 0.4688.

		after calibration with cross-validation via:		after three augmentations via the specified method:		
metric	before calibration	Platt scaling	isotonic regression	LightGBM	decision trees	logistic regression
error	.4688	.4607	.4554	**.3405**	.3459	.4688
Kuiper	.02175	.03736	.009734	.06780	**.008098**	.02284
multi-	.03610	.04623	.02118	.06827	**.01805**	.03667
cal.	±.01000	±.00521	±.00160	±.00028	±**.00161**	±.00972
multi-	.4255	.4330	.4037	.3914	.4073	.4253
ablate	±.04490	±.06345	±.02877	±.00828	±.12692	±.04707
expec-	.2317	.2484	.2328	.2290	.2290	.2317
tation	±.02245	±.00783	±.00749	±.00151	±.00151	±.02247

Table 12: Logistic-regression model on the test set of the 1998 KDD Cup for star donors; the error when using no covariates at all is 0.4688.

References

Arrieta-Ibarra, Imanol, Gujral, Paman, Tannen, Jonathan, Tygert, Mark, & Xu, Cherie. 2022. Metrics of calibration for probabilistic predictions. *J. Mach. Learn. Res.*, **23**, 1–54.

Austin, Peter, & Steyerberg, Ewout. 2019. The integrated calibration index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat. Med.*, **38**(21), 4051–4065.

Bröcker, Jochen. 2008. Some remarks on the reliability of categorical probability forecasts. *Mon. Weather Rev.*, **136**(11), 4488–4502.

Bröcker, Jochen, & Smith, Leonard A. 2007. Increasing the reliability of reliability diagrams. *Weather Forecast.*, **22**(3), 651–661.

Gohar, Usman, & Cheng, Lu. 2023. A survey on intersectional fairness in machine learning: notions, mitigation, and challenges. *Pages 6619–6627 of:* Elkind, Edith (ed), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

Gopalan, Parikshit, Kim, Michael P., & Reingold, Omer. 2023. Swap agnostic learning, or characterizing omniprediction via multicalibration. *Pages 39936–39956 of:* Oh, Alice, Naumann, Tristan, Globerson, Amir, Saenko, Kate, Hardt, Moritz, & Levine, Sergey (eds), *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates. No. 1736.

Haghtalab, Nika, Jordan, Michael I., & Zhao, Eric. 2023. A unifying perspective on multicalibration: game dynamics for multi-objective learning. *Pages 72464–72506 of:* Oh, Alice, Naumann, Tristan, Globerson, Amir, Saenko, Kate, Hardt, Moritz, & Levine, Sergey (eds), *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates. No. 3169.

Hansen, Dutch, Devic, Siddartha, Nakkiran, Preetum, & Sharan, Vatsal. 2024. When is multicalibration post-processing necessary? *Pages 38383–38455 of:* Globerson, Amir, Mackey, Lester, Belgrave, Danielle, Fan, Angela, Paquet, Ulrich, Tomczak, Jakub, & Zhang, Chiyuan (eds), *Proceedings of the 38th International Conference on Neural Information Processing Systems*. Curran Associates. No. 1213.

Hebert-Johnson, Ursula, Kim, Michael, Reingold, Omer, & Rothblum, Guy. 2018. Multicalibration: calibration for the (computationally identifiable) masses. *Pages 1939–1948 of:* Dy, Jennifer, & Krause, Andreas (eds), *Proceedings of the 35th International Conference on Machine Learning (ICML)*. Proceedings of Machine Learning Research, vol. 80. Microtome Press.

Ke, Guolin, Meng, Qi, Finley, Thomas, Wang, Taifeng, Chen, Wei, Ma, Weidong, Ye, Qiwei, & Liu, Tie-Yan. 2017. LightGBM: a highly efficient gradient boosting decision tree. *Pages 3149–3157 of: Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates.

Kim, Michael P., Ghorbani, Amirata, & Zou, James. 2019. Multiaccuracy: black-box post-processing for fairness in classification. *Pages 247–254 of:* Conitzer, Vincent, Hadfield, Gillian, & Vallor, Shannon (eds), *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. Association for Computing Machinery.

Kim, Michael P., Kern, Christoph, Goldwasser, Shafi, Kreuter, Frauke, & Reingold, Omer. 2022. Universal adaptability: target-independent inference that competes with propensity scoring. *Proc. Natl. Acad. Sci. U.S.A.*, **119**(4), 1–6. No. e2108097119.

Kolmogorov, Andrei N. 1933. Sulla determinazione empirica di una legge di distribuzione (On the empirical determination of a distribution function). *Giorn. Ist. Ital. Attuar.*, **4**, 83–91.

Kuiper, Nicolaas H. 1962. Tests concerning random points on a circle. *Proc. Koninklijke Nederlandse Akademie van Wetenschappen Series A*, **63**, 38–47.

La Cava, William G., Lett, Elle, & Wan, Guangya. 2023. Fair admission risk prediction with proportional multicalibration. *Pages 350–378 of:* Mortazavi, Bobak J., Sarker, Tasmie, Beam, Andrew, & Ho, Joyce C. (eds), *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*. Proceedings of Machine Learning Research, vol. 209. Microtome Press.

Lee, Donghwan, Huang, Xinmeng, Hassani, Hamed, & Dobriban, Edgar. 2023. T-Cal: an optimal test for the calibration of predictive models. *J. Mach. Learn. Res.*, **24**, 1–72.

Pedregosa, Fabian, Varoquaux, Gaë, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, VanderPlas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, & Duchesnay, Edouard. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Pfisterer, Florian, Kern, Christoph, Dandl, Susanne, Sun, Matthew, Kim, Michael P., & Bischl, Bernd. 2021. mcboost: multi-calibration boosting for R. *Journal of Open Source Software*, **6**(64), 1–3. No. 3453.

Shabat, Eliran, Cohen, Lee, & Mansour, Yishay. 2020. Sample complexity of uniform convergence for multicalibration. *Pages 13331–13340 of:* Larochelle, Hugo, Ranzato, Marc-Aurelio, Hadsell, Raia, Balcan, Maria-Florina, & Lin, Hsuan-Tien (eds), *Proceedings of the 34th International Conference on Neural Information Processing Systems*. Curran Associates. No. 1118.

Smirnov, Nikolai. 1939. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscou*, **2**(2), 3–11.

Tygert, Mark. 2021. Cumulative deviation of a subpopulation from the full population. *J. Big Data*, **8**(117), 1–60. Available at `https://arxiv.org/abs/2008.01779`.

Tygert, Mark. 2023. Calibration of P-values for calibration and for deviation of a subpopulation from the full population. *Adv. Comput. Math.*, **49**(70), 1–22. Available at `https://arxiv.org/abs/2202.00100`.

Tygert, Mark. 2024. *Cumulative differences between subpopulations versus body mass index in the Behavioral Risk Factor Surveillance System data*. Tech. rept. 2411.07399. arXiv. Available at `https://arxiv.org/abs/2411.07399`.

Wilks, Daniel S. 2019. *Statistical Methods in the Atmospheric Sciences*. 4th edn. Elsevier.

Wu, Jiayun, Liu, Jiashuo, Cui, Peng, & Wu, Zhiwei Steven. 2024. Bridging multicalibration and out-of-distribution generalization beyond covariate shift. *Pages 73036–73078 of:* Globerson,

Amir, Mackey, Lester, Belgrave, Danielle, Fan, Angela, Paquet, Ulrich, Tomczak, Jakub, & Zhang, Chiyuan (eds), *Proceedings of the 38th International Conference on Neural Information Processing Systems*. Curran Associates. No. 2325.

Ye, Hanxuan, & Li, Hongzhe. 2024. *Multicalibration for modeling censored survival data with universal adaptability*. Tech. rept. 2405.15948. arXiv. Available at `https://arxiv.org/abs/2405.15948`.