When is Multicalibration Post-Processing Necessary?

Dutch Hansen (USC), Siddartha Devic (USC), Preetum Nakkiran (Apple), Vatsal Sharan (USC)

Background

- **Classifier Calibration**: Amongst all samples given score $p \in [0,1]$ by an ML algorithm, exactly a p-fraction of those samples have positive label.
- **Multicalibration**: Predictor must be *simultaneously* calibrated over all subgroups given as input.
  - Introduced in algorithmic fairness community [HKRR '18], alongside *multicalibration post-processing algorithms*.
  - Ensures our predictor does not discriminate against chosen subgroups of the data (e.g. minority populations).
- Surprising connections to loss minimization [GKRSW '21]. Some theoretical work suggests that multicalibration is a natural result of loss minimization over neural networks [BGHKN '23].

Q1: ***How calibrated are deep learning models on various groups within the data?***

Q2: *Additionally, when are multicalibration post-processing techniques practically effective?*

Experimental Methodology

Subgroups: We use a subgroup collection $G = \{g_1, \ldots, g_N\}$ which are defined by features and conjunctions of features in the data (e.g. "`Education = College AND Age > 40`").

Holdout data: We closely examine the effect of using different amounts of data for post-processing. To do this, we split data into disjoint training and calibration sets.

"Fair" comparison with ERM: On ERM runs without post-processing, we leave the calibration set empty. All data (except for the validation set) goes to ERM.

Multicalibration techniques: We use the multicalibration algorithms of [HKRR '18] and [HJZ '23]. We also release our implementations: `> pip install multicalibration`.

Scale: 40k experiments across 8 datasets and 10+ models. On each dataset, we test with multiple subgroup collections.

Tabular Data

Observation 1: Neural networks are well multi-calibrated. More generally, models which are *calibrated* are often *multi-calibrated* out of the box.

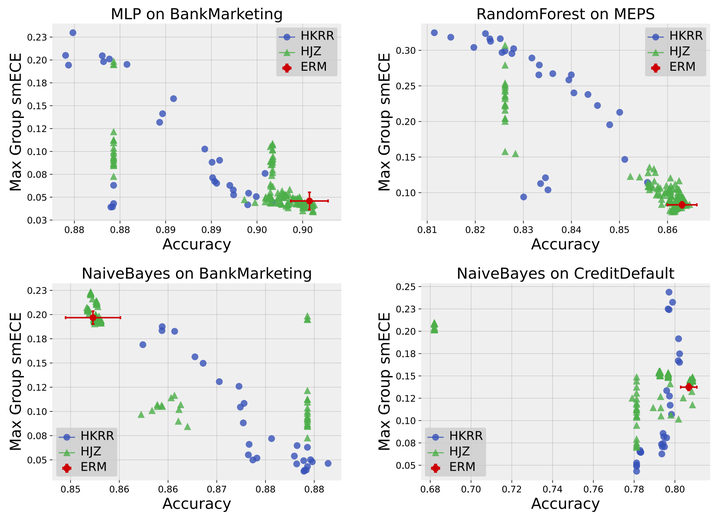


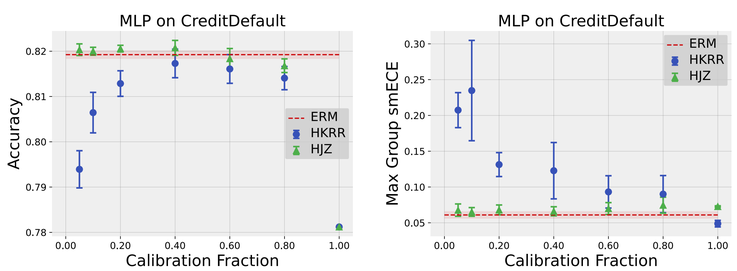
Figure 1: MLPs / random forests do not benefit from multicalibration post-processing (above), but naive Bayes does (below).

Observation 2: On tabular data, simple calibration methods like isotonic regression can improve worst group calibration error almost as well as multicalibration algorithms.

Training Procedure	Worst Group smECE	Accuracy
Naive Bayes ERM	0.287 ± 0.011	0.714 ± 0.018
NB + Multical. Post-Processing	**0.104 ± 0.002**	**0.835 ± 0.003**
NB + Isotonic Regression	0.122 ± 0.015	0.831 ± 0.006

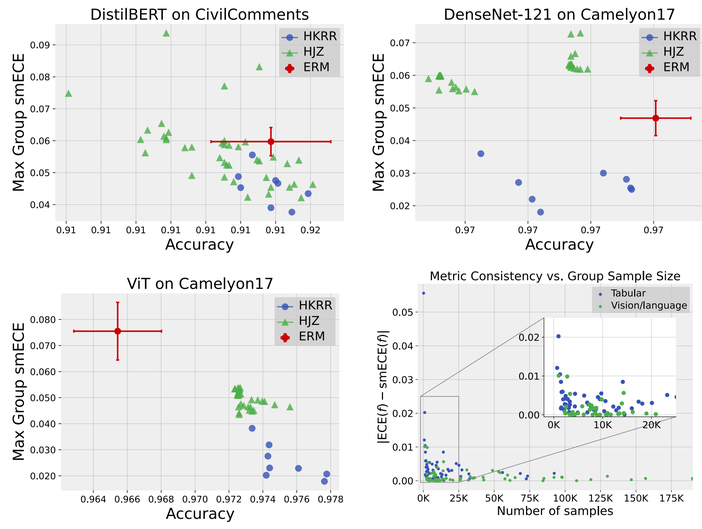
We measure worst group calibration error over all groups in G. We use Expected Calibration Error (ECE) and Smooth ECE (smECE) [BN '24].

Observation 3: A practitioner utilizing multicalibration post-processing can potentially face a trade-off between worst group calibration error and overall accuracy.

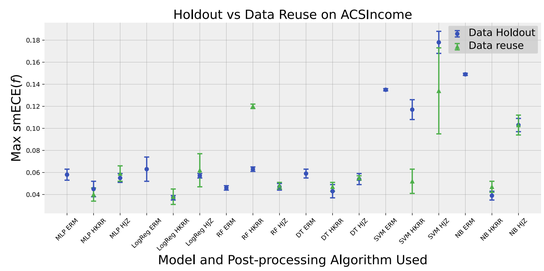


Language + Vision Data

Observation 4: On more complex data, multicalibration post-processing does improve upon ERM for most neural network architectures tested.



Observation 6: For best results, we recommend holding out data for multicalibration post-processing. Data *reuse* can provide varying results (at least on tabular data).



Conclusion: We initiate an empirical examination of the inherent multicalibration of neural networks relative to other ML models.

We also find barriers to the practical viability of existing multicalibration techniques. Such barriers suggest new avenues for the design of multicalibration algorithms.

Ask us about: measuring multicalibration in practice, best practices for using multicalibration algorithms, new directions for algorithm design.

References: [HKRR '18]: Hébert-Johnson, Kim, Reingold, and Rothblum. Calibration for the (Computationally-Identifiable) Masses. 2018. [GKRSW '21]: Gopalan, Kalai, Reingold, Sharan, Wieder. Omnipredictors. 2021. [BGHKN '23]: Blasiok, Gopalan, Hu, Kalai, Nakkiran. Loss Minimization Yields Multicalibration for Large Neural Networks. 2023. [HJZ' 23]: Haghtalab, Jordan, and Zhao. A Unifying Perspective on Multi-Calibration: Game Dynamics for Multi-Objective Learning. 2023. [BN '24]: Blasiok and Nakkiran. Smooth ECE: Principled Reliability Diagrams via Kernel Smoothing. 2024.