Threshold Calibration: Making All Large Predicted Probabilities Trustworthy

Alexandru Lopotenco
Stanford University

Edgar Dobriban
University of Pennsylvania

Abstract

Modern classifiers increasingly output full predictive distributions (class probabilities). In downstream decision-making, the *largest* predicted probabilities often trigger actions (e.g., medical follow-up when risk exceeds a threshold), so failures of calibration at the high end are particularly costly. At the same time, *full* multiclass calibration becomes statistically intractable for many classes. Motivated by these considerations, we propose *threshold calibration*: calibration of *all* predicted class probabilities above a fixed threshold. We define a corresponding miscalibration functional, give a partition-based debiased estimator with linear-time complexity, and prove a distribution-free consistency result. On the COVTYPE benchmark, we empirically evaluate threshold calibration across several predictors and post-hoc recalibrators, finding that more expressive recalibrators can substantially reduce threshold miscalibration.

1 Motivation

Probabilistic predictions are routinely used beyond accuracy-centric evaluation. A model that outputs a vector of predicted probabilities $z = f(x) \in \Delta_{K-1}$ for K classes given an input x is *calibrated* if the actual classes occur with frequency z in the long run given predictions z. Calibration is critical in domains where decisions depend on probabilities rather than hard labels: clinical risk stratification, selective prediction, autonomous systems, and any pipeline that thresholds on confidence (1–16).

In such pipelines, the *largest predicted probabilities* are often the ones acted upon (17, 18). For example, an alert might be triggered if some class probability exceeds $a = 0.1$ or $a = 0.3$. A practitioner may not care whether the remaining low-probability classes are perfectly calibrated, but does care that every probability above a is trustworthy.

1.1 Why existing notions can be inadequate

Full multiclass calibration is statistically too hard. Full multiclass calibration can be cast as a nonparametric problem on the $(K-1)$-simplex. The T-Cal framework (7) analyzes minimax testing rates for multiclass calibration under Hölder smoothness and shows a curse of dimensionality: the optimal separation deteriorates exponentially in K. This intractability forces us to consider weaker, application-driven notions.

Confidence-only calibration can hide systematic failures. A popular notion in deep learning is *confidence calibration*, which conditions only on the value of the top predicted probability, ignoring which class attains it (9). Because it aggregates over labels, it can mask per-class miscalibration: one class may be systematically overconfident while another is underconfident, yet the mixture looks calibrated when binned by confidence (19).

Top-label calibration is too narrow for thresholded decisions. Other notions such as top-label calibration (calibration of the probability assigned to the predicted class) track label identity (19). However, this is not enough to capture does not any pipelines act on *all* probabilities exceeding a threshold, not only the maximum. In multi-label screening, for instance, several classes can simultaneously exceed a, and we want *joint* trustworthiness of all such large entries (20, 21).

2 Threshold calibration

We consider multiclass classification with $K \geq 2$ classes. Let $X \in \mathcal{X}$ be features and $Y \in \{e_1, \ldots, e_K\}$ be a one-hot label. A probabilistic classifier $f : \mathcal{X} \to \Delta_{K-1}$

Workshop "Towards Trustworthy Predictions: Theory and Applications of Calibration for Modern AI" at AISTATS 2026, Tangier, Morocco.

outputs $Z = f(X) \in \Delta_{K-1}$, which is a probability vector in the simplex Δ_{K-1}. Let $[K] := \{1, \dots, K\}$. We denote by $[v]_k$ the k-th coordinate of a vector v, and for $S \subset [K]$, by $[v]_S$ the subvector of S formed by the coordinates in S.

Standard multiclass calibration requires that, given that the predicted probabilities are $Z = f(X)$, the jth class occurs (i.e., $Y_j = 1$) with probability z_j, for all $j \in [K]$. Thus, for all $z \in \Delta_{K-1}$ and for each $j \in [K]$,

$$\mathbb{P}([Y]_j = 1 | Z = z) = [z]_j. \tag{1}$$

Given z, the predicted class probabilities match the true probabilities.

Threshold selector. Let $a \in (0, 1)$ denote the threshold of interest. Define the *threshold selector* T_a as follows. For a vector of predicted probabilities $z \in \Delta_{K-1}$, $T_a(z)$ is the tuple of all coordinates $j \in [K]$, with $[z]_j \geq a$, sorted in an increasing order of the coordinates. Equivalently, $T_a(z) = (t_1, \dots, t_J)$ is the unique tuple such that $t_1 < \cdots < t_J$ and $\{t_1, \dots, t_J\} = \{k \in [K] : [z]_k \geq a\}$, where $J = \big|\{k \in [K] : [z]_k \geq a\}\big|$.

For instance, the threshold-0.3 selector $T_{0.3}$ acts as follows on two four-dimensional example inputs:

$$T_{0.3}\left(\begin{bmatrix}0.45\\0.4\\0.10\\0.05\end{bmatrix}\right) = T_{0.3}\left(\begin{bmatrix}0.4\\0.45\\0.10\\0.05\end{bmatrix}\right) = (1, 2).$$

This highlights that the threshold selector does not retain the ordering of the predicted probabilities; which is consistent with our motivation of selecting only the set of the large predictions. We then define threshold calibration as the condition that the large probabilities are calibrated.

Definition 1 (Threshold calibration)**.** *A classifier f is* threshold-a calibrated *if for all $z \in \Delta_{K-1}$ and associated set $S = T_a(z)$ of coordinates of size at least a, for each $j \in S$,*

$$\mathbb{P}\big([Y]_j = 1 \big| [Z]_{T_a(Z)} = [z]_S, T_a(Z) = S\big) = [z]_j. \tag{2}$$

In words, given that exactly the coordinates in $T_a(z)$ exceed the threshold a, the predicted class probabilities match the true probabilities. Compared to the standard notion of calibration from (1), the key difference is that only the probabilities larger than a are required to be calibrated. Moreover, for reasons similar to those in (19), we also condition on the specific coordinates $T_a(Z)$, as this provides a more granular notion of calibration. Not conditioning on $T_a(Z)$ is potentially also reasonable, and would be of interest to investigate in future work.

3 Measuring and estimating threshold miscalibration

3.1 Calibration error

Let P denote the joint distribution of (Z, Y) induced by (X, Y) and f. For a realized selection $Q = T_a(Z)$ of coordinates, define the (tuple-specific) threshold regression function as $\mathrm{reg}_P^{(Q)} : z \mapsto \mathbb{E}([Y]_Q \mid [Z]_Q = [z]_Q, T_a(Z) = Q)$, and the associated residual $\mathrm{res}_P^{(Q)}(z) = \mathrm{reg}_P^{(Q)}(z) - [z]_Q$. Threshold calibration is equivalent to $\mathrm{res}_P^{(Q)}(z) = 0$ almost surely for every Q in the range of T_a.

We measure deviation from calibration using the *threshold expected calibration error* (ECE), defined as

$$\ell_2\text{-ECE}_P^{(T_a)} = \left(\mathbb{E}\left[\left\|\mathrm{res}_P^{(T_a(Z))}(Z)\right\|_2^2\right]\right)^{1/2}. \tag{3}$$

3.2 A debiased partition-based estimator

We observe i.i.d. calibration data $(Z_i, Y_i)_{i=1}^n \sim P$. To estimate $\ell_2\text{-ECE}_P^{(T_a)}$, we partition the space of selected probabilities and use a local U-statistic. Fix binning parameters $m, M \in \mathbb{N}_+$. Let $\{B_{m,i}\}_{i=1}^M$ be a partition of the Δ_{K-1} such that for all $i = 1 \dots M$ and any $z, z' \in B_{m,i}$ we have that

$$T_a(z) = T_a(z') \text{ and } \|[z]_{T_a(z)} - [z']_{T_a(z')}\|_\infty \leq m^{-1}.$$

Let $\mathcal{I}_{m,i} = \{l \in [n] : Z_l \in B_{m,i}\}$ and $N_{m,i} = |\mathcal{I}_{m,i}|$. Then, we define the statistic $T_{m,n}$ via

$$\sum_{i : N_{m,i} \geq 2} \frac{N_{m,i}}{n(N_{m,i} - 1)^2} \sum_{l \neq j \in \mathcal{I}_{m,i}} [U_l]_{T_a(Z_l)}^\top [U_j]_{T_a(Z_j)}, \tag{4}$$

where $U_l := Y_l - Z_l$. This is a debiased plug-in estimator of $(\ell_2\text{-ECE}_P^{(T_a)})^2$ and can be computed in $O(nK)$ time by a single pass over the data (to compute the per-bin sums). We summarize the procedure of building the estimator $T_{m,n}$ in Algorithm 1.

3.3 A basic consistency guarantee

We now show that this estimator asymptotically converges to the true calibration error, which provides a basic level of justification for the estimator. This result requires only that the partition becomes finer in a way that the piecewise constant approximation over the partition converges to the true calibration error. That such a condition is required is reasonable, because our estimator is based on a piecewise U-statistic.

Assumption 3.1. *As $n, m, M \to \infty$, the partition becomes finer, such that the piecewise approximation*

Algorithm 1 One-pass computation of $T_{m,n}$ for threshold calibration

Input: data $\{(Z_i, Y_i)\}_{i=1}^n$, threshold a, partition $\{B_{m,i}\}_{i=1}^M$.
Initialize: for each bin i, set $N_i \leftarrow 0$, $S_i \leftarrow 0$ (vector), $Q_i \leftarrow 0$ (scalar).
for $\ell = 1$ to n **do**
  find bin index $i(\ell)$ such that $Z_\ell \in B_{m,i(\ell)}$.
  $u_\ell \leftarrow [Y_\ell - Z_\ell]_{T_a(Z_\ell)}$.
  $N_{i(\ell)} \leftarrow N_{i(\ell)} + 1$; $S_{i(\ell)} \leftarrow S_{i(\ell)} + u_\ell$; $Q_{i(\ell)} \leftarrow Q_{i(\ell)} + \|u_\ell\|_2^2$.
end for
Aggregate:

$$T_{m,n} \leftarrow \sum_{i:N_i \geq 2} \frac{N_i}{n(N_i-1)^2}\Big(\|S_i\|_2^2 - Q_i\Big).$$

Output: $T_{m,n}$.

of the calibration error converges to the true threshold ECE:

$$\sum_i \mathbb{P}(Z \in B_{m,i}) \, \|\mu_{m,i}\|^2 \to (\ell_2\text{-}ECE_P^{(T_a)})^2,$$

where $\mu_{m,i} = \mathbb{E}\left[\left\|\mathbb{E}[[Y-Z]_{T_a(Z)} | Z \in B_{i,m}]\right\|^2\right]$.

Under this condition we show the following consistency results. First observe that no more than $1/a$ coordinates can be selected, and so $J_a := \min\{K, \lfloor 1/a \rfloor\}$ is an Upper bound on the effective dimensionality of the problem.

Proposition 1 (Mean-square consistency)**.** *For any $a \in (0,1)$ and distribution P of (Z,Y), if $m = m(n) \to \infty$ and $m(n) = o(n^{1/J_a})$, then $T_{m(n),n}$ is mean-square consistent for $(\ell_2\text{-}ECE_P^{(T_a)})^2$, i.e., $\mathbb{E}\left[\left\{T_{m(n),n} - (\ell_2\text{-}ECE_P^{(T_a)})^2\right\}^2\right] \to 0$ as $n \to \infty$.*

The condition $m = o(n^{1/J_a})$ mirrors that a higher effective dimensionality must use a relatively smaller number of bins. The proof is contained in the supplementary material, and a sketch is provided below.

4 Proof sketch of Proposition 1

Let $U_j := Y_j - Z_j \in \mathbb{R}^K$. For each bin $B_{m,i}$, all points share the same selected tuple $Q = T_a(z)$ by construction, so we write $U_{m,j} := [U_j]_Q \in \mathbb{R}^{|Q|}$ for $Z_j \in B_{m,i}$. Let $\xi^2 = (\ell_2\text{-ECE}_P^{(T_a)})^2$. We use the bias–variance decomposition

$$\mathbb{E}\Big[\big(T_{m,n} - \xi^2\big)^2\Big] = \mathrm{Var}(T_{m,n}) + \big(\mathbb{E}[T_{m,n}] - \xi^2\big)^2.$$

A distribution-free variance bound. Apply the Efron–Stein inequality by replacing one observation with an independent copy. Changing one sample affects at most two bins (the original bin and the new bin). Within a bin with selected tuple size $\leq J_a$, every inner product satisfies $|u^\top v| \leq \|u\|_2 \|v\|_2 \leq J_a$, since entries of U lie in $[-1,1]$. A direct add/remove-one-point argument shows that each affected bin contribution changes by at most $C\,J_a/n$ for a universal constant C, hence $|T_{m,n} - T_{m,n}^{(s)}| \leq C' J_a/n$ for another constant C'. Efron–Stein then yields $\mathrm{Var}(T_{m,n}) \leq C'' J_a^2/n = O(1/n)$ uniformly over P and m.

Identify $\mathbb{E}[T_{m,n}]$ as a plug-in quantity plus a vanishing remainder. Let $p_{m,i} = \mathbb{P}(Z \in B_{m,i})$ and denote the within-bin mean residual vector $v_{m,i}^2 := \|\mu_{m,i}\|_2^2$. Conditioning on bin memberships, for $l_1 \neq l_2$ in the same bin, independence gives $\mathbb{E}[U_{m,l_1}^\top U_{m,l_2} \mid Z_{l_1}, Z_{l_2} \in B_{m,i}] = \mu_{m,i}^\top \mu_{m,i} = v_{m,i}^2$. Therefore,

$$\begin{aligned}\mathbb{E}[T_{m,n}] &= \sum_{i:\, p_{m,i}>0} \frac{1}{n}\,\mathbb{E}\left[\frac{N_{m,i}^2}{N_{m,i}-1}\mathbf{1}\{N_{m,i} \geq 2\}\right] v_{m,i}^2 \\ &= \sum_i p_{m,i} v_{m,i}^2 + R_{m,n},\end{aligned}$$

where the remainder satisfies $|R_{m,n}| \lesssim (J_a/n)\, M_m$ and M_m is the number of bins. In our construction $M_m \lesssim m^{J_a}$, so $m^{J_a} = o(n)$ implies $R_{m,n} \to 0$. Then, Assumption 3.1 gives us the desired result.

5 Experiments: threshold calibration on synthetic data

We evaluate the performance of our estimator introduced in (4), aiming to support the consistency statement introduced in Proposition 1 through simulations of a synthetic dataset.

Dataset and protocol. In particular, we consider the case when $K = 4$, i.e. we have data $(Z,Y) \in \Delta_{K-1} \times \{0,1\}^K$. We analyze the empirical convergence of threshold-a mis-calibration, where $a = 0.3$. To generate mis-calibrated data, we set the regression curve that determines mis-calibration as

$$[g_{\varepsilon_0}(z)]_c = \begin{cases} [z]_c + \frac{\varepsilon_0}{\sqrt{|T_a(z)|}} & \text{if } c \in T_a(z) \\ [z]_c \left(1 - \frac{\varepsilon_0 \sqrt{|T_a(z)|}}{\sum_{q \notin T_a(z)} [z]_q}\right) & \text{otherwise.} \end{cases}$$

In each of the three panels from Figure 1, we fix a target miscalibration level by choosing ε_0 so that the ground-truth value of $(\ell_2\text{-ECE}_P^{(T_a)})^2$ matches a prescribed value (estimated via 10^7 Monte Carlo draws).

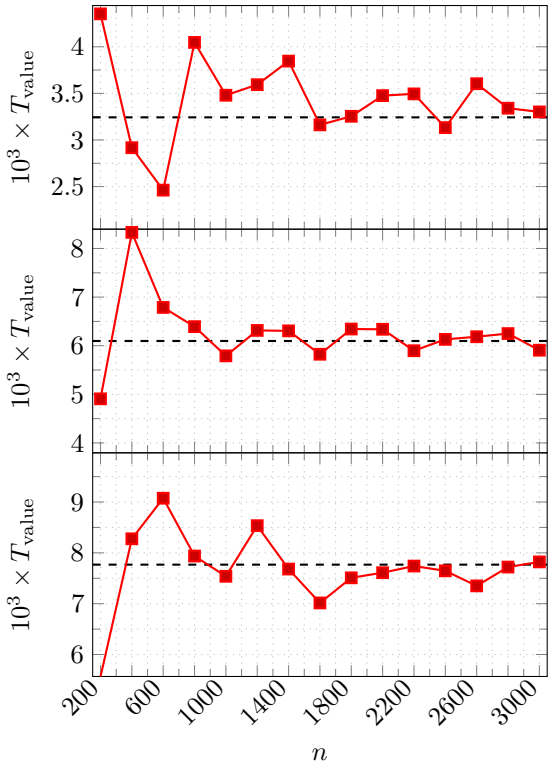

Figure 1: Convergence of T_{value} (shown as $10^3 \times T_{\text{value}}$) for threshold-$a$ calibration with $a = 0.3$ on synthetic data, for three fixed ground-truth levels. In each panel, the dashed line denotes the corresponding $\ell_2\text{-ECE}_P^{(T_a)}$. The dashed lines, i.e., the true $\ell_2\text{-ECE}_P^{(T_a)}$ are $10^{-3} \times 3.24, 10^{-3} \times 6.10$ and $10^{-3} \times 7.77$, from top to bottom, respectively.

To construct $T_{m,n}$, we pick our partition scheme hyperparameter $m = \lfloor n^{\frac{1}{4}} \rfloor \leq n^{\frac{1}{J_a}} = n^{\frac{1}{3}}$. We then generate i.i.d. samples of size n from the resulting model and compute the plug-in statistic $T_{m,n}$ from (4). We report the empirical mean of $T_{m,n}$ over 100 independent repetitions.

Across all three settings, the empirical mean $T_{m,n}$ approaches the ground-truth $(\ell_2\text{-ECE}_P^{(T_a)})^2$ as n increases, which is aligned with the consistency guarantee in Proposition 1. We also observe larger dispersion at smaller n and at higher miscalibration levels, consistent with the fact that $T_{m,n}$ aggregates within-bin second-order terms and is sensitive to bin counts when n is small.

Overall, these simulations indicate that threshold calibration error can be estimated reliably at moderate sample sizes even when $K > 2$. In fact, we want to point out that due to the fact that $a = 3$ the effective dimension of the problem is the same as full multicalibration (as we can select up to 3 points, so $J_a = 3$), yet we need as little as a few thousand points to see a reliable empirical convergence to the true squared mis-calibration error. This supports the premise that threshold calibration can be a practical alternative to full multiclass calibration in large-K problems where the latter is statistically intractable.

6 Discussion

In this work we introduced threshold calibration, wich isolates the part of the predictive distribution that is most actionable: all entries above a fixed threshold a. It is strictly weaker than full multiclass calibration, but it is tractable and aligns with many downstream decision rules that threshold on probabilities.

Several future directions are immediate. First, one can aim to develop methods for additional statistical problems such as hypothesis testing or confidence intervals, as in Lee et al. (7), Sun et al. (8) but with an effective dimension tied to $J_a = \min\{K, \lfloor 1/a \rfloor\}$ instead of K. Second, one can study post-hoc recalibration methods that directly target threshold calibration or general multi-calibration. Third, the threshold selector T_a is only one example of a data-dependent selection rule. A natural next step is to develop a general theory of calibration for selectors that choose a small subset of coordinates in a structured way, together with estimation and testing procedures whose complexity depends on the size of the tuples in the selector's range.

References

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

[2] Françoise Seillier-Moiseiwitsch and AP Dawid. On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, 88(421):355–359, 1993.

[3] Francis X Diebold and Roberto S Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 1995.

[4] Francis X Diebold, Todd A Gunther, and Anthony S Tay. Evaluating density forecasts with applications to financial risk management. *International Economic Review*, pages 863–883, 1998.

[5] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: a unifying framework. *Advances in Neural Information Processing Systems*, 32:12257–12267, 2019.

[6] David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests beyond classification. *arXiv preprint arXiv:2210.13355*, 2022.

[7] Donghwan Lee, Xinmeng Huang, Hamed Hassani, and Edgar Dobriban. T-cal: An optimal test for the calibration of predictive models. *arXiv preprint arXiv:2203.01850*, 2022.

[8] Yan Sun, Pratik Chaudhari, Ian J. Barnett, and Edgar Dobriban. A confidence interval for the ℓ_2 expected calibration error, 2024. URL https://arxiv.org/abs/2408.08998.

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[10] Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pages 2805–2814. PMLR, 2018.

[11] Sunil Thulasidasan, Gopinath Chennupati, Jeff 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.

[12] 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*, pages 11117–11128. PMLR, 2020.

[13] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33, 2020.

[14] 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.

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

[16] Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration. In *Advances in Neural Information Processing Systems*, pages 3792–3803, 2019.

[17] Sasank Chilamkurthy, Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G. Campeau, Vasantha Kumar Venugopal, Vidur Mahajan, Pooja Rao, and Prashant Warier. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *The Lancet*, 392 (10162):2388–2396, December 2018. doi: 10.1016/S0140-6736(18)31645-3. Epub 2018 Oct 11. PMID: 30318264.

[18] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8): 1819–1837, 2014. doi: 10.1109/TKDE.2013.39.

[19] Chirag Gupta and Aaditya K Ramdas. Top-label calibration and multiclass-to-binary reductions. *arXiv preprint arXiv:2107.08353*, 2021.

[20] Ivo Baltruschat, Leonhard Steinmeister, Hannes Nickisch, Axel Saalbach, Michael Grass, Gerhard Adam, Tobias Knopp, and Harald Ittrich. Smart chest x-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *European Radiology*, 31:3837–3845, 2021. doi: 10.1007/s00330-020-07480-7.

[21] Joy T. Wu, Ken C. L. Wong, Yaniv Gur, and et al. Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents. *JAMA Network Open*, 3(10):e2022779, 2020. doi: 10.1001/jamanetworkopen.2020.22779.

Supplementary Materials for Threshold Calibration

1 Proof of Proposition 1

Write $U_j = Y_j - Z_j \in \mathbb{R}^K$ and for any tuple Q that can be selected by T_a let $U_{j,Q} = [U_j]_Q \in \mathbb{R}^{|Q|}$,. In particular, note that for any specific bin $B_{m,i}$ we will have that the selected entries of the points $Z_j \in B_{m,i}$ are the same tuple by construction. Hence, for any $j \in \mathcal{I}_{m,i}$ for some i, we will denote $U_{m,j} := U_{j,Q} = [U_j]_Q \in \mathbb{R}^{|Q|}$, where Q is the unique tuple that can be selected by T_a for any datapoint within $B_{m,i}$. For a fixed m, recall the induced partition $(B_{m,i})_{m,i}$ of Δ_{K-1}, the index sets $\mathcal{I}_{m,i} = \{l \in [n] : Z_l \in B_{m,i}\}$, and counts $N_{m,i} = |\mathcal{I}_{m,i}|$. Using $\|\sum_l u_l\|_2^2 - \sum_l \|u_l\|_2^2 = \sum_{l_1 \neq l_2} u_{l_1}^\top u_{l_2}$, we can rewrite the statistic as

$$T_{m,n}^{(S)} = \sum_{i,Q:\, N_{m,i} \geq 2} \frac{N_{m,i}}{n(N_{m,i}-1)^2} \sum_{l_1 \neq l_2 \in \mathcal{I}_{m,i,Q}} U_{m,l_1}^\top U_{m,l_2}.$$

We prove mean-square consistency via bias–variance decomposition:

$$\mathbb{E}\Big[\big(T_{m,n}^{(S)} - \xi^2\big)^2\Big] = \mathrm{Var}(T_{m,n}^{(S)}) + \big(\mathbb{E}[T_{m,n}^{(S)}] - \xi^2\big)^2. \tag{1}$$

We show $\mathrm{Var}(T_{m,n}^{(S)}) = O(1/n)$ uniformly in P, m.

Distribution-free $O(1/n)$ variance bound. We use the Efron–Stein inequality. Let $\mathcal{D} = ((Z_1, Y_1), \ldots, (Z_n, Y_n))$, and for each $s \in [n]$ let $\mathcal{D}^{(s)}$ be the dataset where (Z_s, Y_s) is replaced by an independent copy (Z_s', Y_s'), independent of everything else. Let $T = T_{m,n}^{(S)}(\mathcal{D})$ and $T^{(s)} = T_{m,n}^{(S)}(\mathcal{D}^{(s)})$. Efron–Stein gives

$$\mathrm{Var}(T) \leq \frac{1}{2} \sum_{s=1}^{n} \mathbb{E}\big[(T - T^{(s)})^2\big].$$

It therefore suffices to bound $|T - T^{(s)}|$ uniformly by $C(J)/n$.

Changing one sample can only affect the contribution of at most two bins: the bin containing Z_s and the bin containing Z_s' (possibly the same). Thus it is enough to bound the change in a single bin functional under adding/removing one element.

Fix a bin and write I for its index set with $|I| = N$ and Q for the associated tuple that is selected by T_a within this bin. Define the bin contribution functional

$$\Psi(I) = \begin{cases} \frac{N}{n(N-1)^2} \sum_{l_1 \neq l_2 \in I} U_{m,l_1}^\top U_{m,l_2}, & N \geq 2, \\ 0, & N \leq 1. \end{cases}$$

Note that for any $u, v \in [-1,1]^{|Q|}$ we have $|u^\top v| \leq \|u\|_2 \|v\|_2 \leq \sqrt{|Q|}\sqrt{|Q|} \leq J$. Hence for any $N \geq 2$,

$$\Big| \sum_{l_1 \neq l_2 \in I} U_{m,l_1}^\top U_{m,l_2} \Big| \leq N(N-1)\, J.$$

Let I_- be obtained by removing one index from I (so $|I_-| = N-1$). Write $P(I) = \sum_{l_1 \neq l_2 \in I} U_{m,l_1}^\top U_{m,l_2}$. Then $|P(I) - P(I_-)|$ is bounded by the magnitude of the deleted ordered pairs, which is at most $2(N-1)J$. Therefore, for $N \geq 3$,

$$\begin{aligned} |\Psi(I) - \Psi(I_-)| &\leq \Big| \frac{N}{n(N-1)^2} \big(P(I) - P(I_-)\big) \Big| + \Big| \Big(\frac{N}{n(N-1)^2} - \frac{N-1}{n(N-2)^2} \Big) P(I_-) \Big| \\ &\leq \frac{N}{n(N-1)^2} \cdot 2(N-1)J + \frac{1}{n} \Big| \frac{N}{(N-1)^2} - \frac{N-1}{(N-2)^2} \Big| \cdot (N-1)(N-2)J. \end{aligned}$$

The first term is at most $\frac{2J}{n} \cdot \frac{N}{N-1} \le \frac{4J}{n}$. For the second term, consider $b(x) = \frac{x}{(x-1)^2}$ for $x \ge 2$. A direct derivative computation gives $|b'(x)| = \frac{x+1}{(x-1)^3}$, so for $x \ge 2$, $|b'(x)| \le 16/x^2$ (using $(x-1) \ge x/2$). By the mean value theorem, $\left|\frac{N}{(N-1)^2} - \frac{N-1}{(N-2)^2}\right| = |b(N) - b(N-1)| \le \sup_{x \in [N-1,N]} |b'(x)| \le 16/(N-1)^2$. Thus the second term is at most $\frac{1}{n} \cdot 16/(N-1)^2 \cdot (N-1)(N-2)J \le \frac{16J}{n}$. Hence $|\Psi(I) - \Psi(I_-)| \le \frac{20J}{n}$ for $N \ge 3$. The case $N = 2$ is even simpler: $\Psi(I_-) = 0$ and $|\Psi(I)| \le \frac{2}{n} \cdot 2J = \frac{4J}{n}$. Therefore, for all N, removing one point changes one bin contribution by at most $20J/n$. Adding one point is handled symmetrically, so modifying one observation changes at most two bin contributions by at most $2 \cdot 2 \cdot 20J/n = 80J/n$. Consequently, $|T - T^{(s)}| \le 80J/n$ almost surely, and thus

$$\mathrm{Var}(T) \le \frac{1}{2} \sum_{s=1}^{n} \mathbb{E}[(T - T^{(s)})^2] \le \frac{1}{2} \cdot n \cdot (80J/n)^2 = \frac{3200\, J^2}{n}.$$

In particular, $\mathrm{Var}(T_{m,n}^{(S)}) = O(1/n)$ uniformly over all P and all m.

Compute $\mathbb{E}[T_{m,n}^{(S)}]$ up to a vanishing $O(\tilde{m}/n)$ term. For each bin define the cell probability $p_{m,i} = \mathbb{P}(Z \in B_{m,i})$. If $p_{m,i} = 0$ the cell is null and contributes nothing; henceforth fix cells with $p_{m,i} > 0$. Recall the bin-specific residual function and define the within-cell mean residual vector

$$\mu_{m,i} := r(B_{i,m}) = \mathbb{E}\left[\left\|\mathbb{E}[[Y - Z]_{T_a(Z)} | Z \in B_{m,i}]\right\|^2\right], \qquad v_{m,i}^2 := \|\mu_{m,i}\|_2^2.$$

Note $\|\mathrm{res}_P^{(Q)}(z)\|_2^2 \le |Q| \le J$, so $0 \le v_{m,i}^2 \le J$.

Condition on the bin memberships and on the event $\{N_{m,i} = N \ge 2\}$. Within this cell, the i.i.d. structure implies that the ordered pairs (U_{m,l_1}, U_{m,l_2}) are independent for distinct indices, and $\mathbb{E}[U_{m,l} \mid Z_l \in B_{m,i}] = \mu_{m,i}$. Therefore, for $l_1 \ne l_2$ in the same cell,

$$\mathbb{E}[U_{m,l_1}^\top U_{m,l_2} \mid Z_{l_1}, Z_{l_2} \in B_{m,i}] = \mu_{m,i}^\top \mu_{m,i} = v_{m,i}^2.$$

Hence, conditional on $N_{m,i} = N \ge 2$, the expected ordered-pair sum is $N(N-1)v_{m,i}^2$. Thus the conditional expectation of the i-summand equals

$$\mathbb{E}\Big[\frac{N}{n(N-1)^2} \sum_{l_1 \ne l_2 \in \mathcal{I}_{m,i,Q}} U_{m,l_1}^\top U_{m,l_2} \,\Big|\, N_{m,i} = N\Big] = \frac{N}{n(N-1)^2} \cdot N(N-1)\, v_{m,i}^2 = \frac{N^2}{n(N-1)} v_{m,i}^2.$$

Since the summand is zero when $N \le 1$, we obtain

$$\mathbb{E}[T_{m,n}^{(S)}] = \sum_{i:\, p_{m,i} > 0} \frac{1}{n}\, \mathbb{E}\Big[\frac{N_{m,i}^2}{N_{m,i} - 1} I(N_{m,i} \ge 2)\Big]\, v_{m,i}^2. \tag{2}$$

Now decompose $\frac{N^2}{N-1} I(N \ge 2) = \big(N - I(N=1)\big) + \frac{N}{N-1} I(N \ge 2)$. Since $\mathbb{E}[N_{m,i}] = n p_{m,i}$ and $\frac{N}{N-1} I(N \ge 2) \le 2I(N \ge 2) \le 2$, we can write (2) as

$$\mathbb{E}[T_{m,n}^{(S)}] = \sum_{i,Q:\, p_{m,i} > 0} p_{m,i}\, v_{m,i}^2 + R_{m,n}, \tag{3}$$

where the remainder is

$$R_{m,n} = \frac{1}{n} \sum_{i:\, p_{m,i} > 0} \Big(- \mathbb{P}(N_{m,i} = 1) + \mathbb{E}\big[\tfrac{N_{m,i}}{N_{m,i}-1} I(N_{m,i} \ge 2)\big]\Big) v_{m,i}^2.$$

The bracket is bounded in absolute value by $1 + 2 = 3$, and $v_{m,i}^2 \le J$, so

$$|R_{m,n}| \le \frac{3J}{n} \cdot M_m,$$

where M_m is the number of (potentially) non-null cells $B_{m,i}$. In our construction $M_m \le |\mathcal{S}|\tilde{m}$ with $\tilde{m} = m + m^2 + \cdots + m^J \le J m^J$. Since $m^J = o(n)$ by assumption, $M_m/n \to 0$, hence $R_{m,n} \to 0$.

Then the result follows by assumption.

2 Example partition construction

See Figure 1 for an example of the induced partition $\{B_{m,i}\}$ for threshold calibration with $K = 3$ and threshold $a = 0.4$. The solid lines divide the simplex into disjoint selection regions $T_a^{-1}(Q)$, where Q records which coordinates satisfy $z_k \geq a$; the central gray region (no coordinate exceeds a) is never selected and therefore plays no role. Within each nonempty region, the dashed lines further subdivide $T_a^{-1}(Q)$ into bins $B_{m,i}$ induced by an equal-radius partition of the $|Q|$-dimensional simplex obtained after coordinate augmentation. Consequently, the bin boundaries depend only on the selected coordinates: for $|Q| = 1$ they are constant along the unselected directions (e.g., horizontal splits in the $T_a(z) = (3)$ region depend only on z_3), while for $|Q| = 2$ they arise from intersecting a 2D simplex partition with the corresponding triangular region.

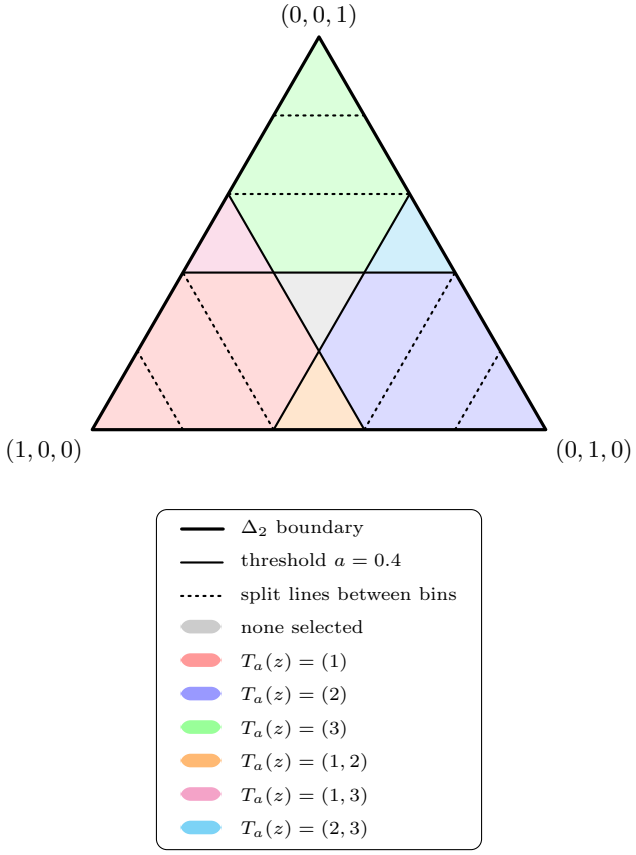

Figure 1: Induced bins $\{B_{m,i}\}$ for threshold calibration on Δ_2 with $a = 0.4$: solid lines show selection regions $T_a^{-1}(Q)$; dashed lines show within-region bins that depend only on the selected coordinates.