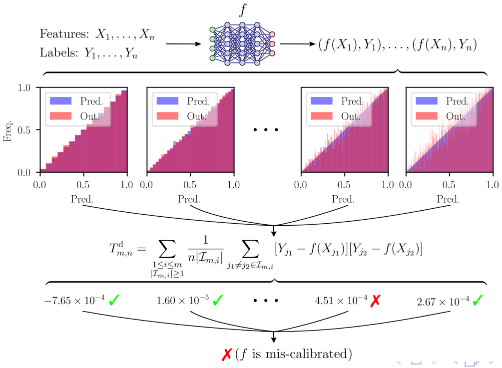
T-Cal: An optimal test for the calibration of predictive models

Edgar Dobriban, University of Pennsylvania

based on joint work with
Donghwan Lee, Xinmeng Huang, and Hamed Hassani



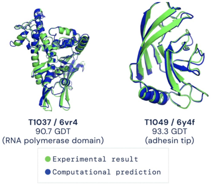
Overview

Context

- Prediction accuracy of deep learning is steadily increasing;

Context

- Prediction accuracy of deep learning is steadily increasing;
- Success stories: AlphaFold, cancer tissue image classification, computer vision, NLP ...

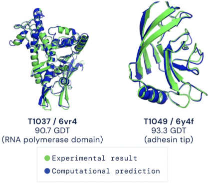


Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

Context

- Prediction accuracy of deep learning is steadily increasing;
- Success stories: AlphaFold, cancer tissue image classification, computer vision, NLP ...



Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.**

Figure 3.14: The GPT-3 generated news article that humans had the greatest difficulty distinguishing from a human written article (accuracy: 12%).

- Meanwhile, growing concerns: safety, ethics, energy- and sample-efficiency, uncertainty—due to overparametrization/use of big models

Uncertainty Quantification

- Given input x, the output y is often not uniquely determined

Uncertainty Quantification

- Given input x, the output y is often not uniquely determined
- Examples of uncertainty:
  - GPT-3: given text prompt, ...?

Uncertainty Quantification

- Given input x, the output y is often not uniquely determined
- Examples of uncertainty:
  - GPT-3: given text prompt, ...?
  - Skin cancer classification: given skin image, ...?

Uncertainty Quantification

- Given input x, the output y is often not uniquely determined
- Examples of uncertainty:
  - GPT-3: given text prompt, ...?
  - Skin cancer classification: given skin image, ...?
- Standard ML/DL pipeline does not provide a solution

Uncertainty Quantification

- Example problems:
  - Prediction Set: find mapping C of inputs to subsets of $\mathcal{Y}$: $P(y \in C(x)) \geqslant 1 - \alpha$, for some $\alpha \in (0, 1)$.

Uncertainty Quantification

- Example problems:
  - Prediction Set: find mapping C of inputs to subsets of $\mathcal{Y}$: $P(y \in C(x)) \geqslant 1 - \alpha$, for some $\alpha \in (0, 1)$.
  - Calibration: construct probability predictions that reflect true probabilities. For binary classification, for all appropriate z,

    $$P(y = 1 | f(x) = z) \approx z$$

    "When I say that the image is a car with probability 20%, it should really be 20%"

Calibration

- Input $x \in \mathbb{R}^d$; output: one-hot encoded label $y \in \{0,1\}^K$
- A probabilistic classifier $f : \mathbb{R}^d \to \Delta_{K-1}$ (simplex of probability distributions over $1, \ldots, K$) is calibrated if

$$P(\mathrm{y}_k = 1 | f(\mathrm{x}) = \mathrm{z}) = [f(x)]_k \quad \text{for any } k \in [K] = \{1, \ldots, K\} \text{ and } z \in \Delta_K$$

Calibration

- Input $x \in \mathbb{R}^d$; output: one-hot encoded label $y \in \{0,1\}^K$
- A probabilistic classifier $f : \mathbb{R}^d \to \Delta_{K-1}$ (simplex of probability distributions over $1, \ldots, K$) is calibrated if

$$P(\mathrm{y}_k = 1 | f(\mathrm{x}) = \mathrm{z}) = [f(x)]_k \quad \text{for any } k \in [K] = \{1, \ldots, K\} \text{ and } z \in \Delta_K$$

- Often focus on top-1 calibration: correctly predicting accuracy $f^+(x) = \max_k [f(x)]_k$

$$P(y = y_{\hat{k}(x)} | f^+(x)) = f^+(x)$$

Calibration in Modern ML

- Modern finding: powerful deep nets are *over-confident* and *mis-calibrated*; in contrast to smaller models

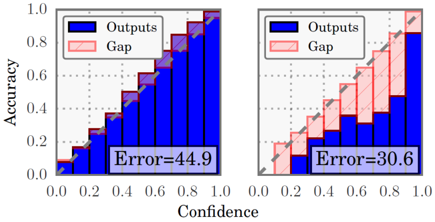


Figure: Guo et al, 2017: "On Calibration of Modern Neural Networks". Reliability diagram for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100.

Consequences of Mis-calibration

- Mis-calibration can have severe consequences in decision making

Consequences of Mis-calibration

- Mis-calibration can have severe consequences in decision making
- mHealth Example: Smartphone monitoring for timely intervention (w/ Ian Barnett, UPenn Biostats)
  - Monitor phone usage, sleep, physical activity, etc., of past mental health patients
  - Want to predict relapse, say in schizophrenia (using deep nets)
  - Mis-calibration can lead to poor risk evaluation and improper interventions

Figure: from colleaga.org

Improving Calibration

- Two main approaches to improve calibration: post-hoc calibration and penalization

Improving Calibration

- Two main approaches to improve calibration: post-hoc calibration and penalization
- In principle, possible to get nearly perfect calibration—think predicting a constant, the marginal class probabilities

Improving Calibration

- Two main approaches to improve calibration: post-hoc calibration and penalization
- In principle, possible to get nearly perfect calibration—think predicting a constant, the marginal class probabilities
- How well do current methods work? "... we improve from 3.1% to 1.9% ...".
  Inherent uncertainty due to limited data.
  Need rigorous tools to check/validate/test calibration

Testing Calibration

- Classical tests of calibration (Cox, Miller, 60s): Given $B_i \sim \text{Bernoulli}(q_i)$, $i = 1, \ldots, n$, test $q_i = p_i$, for a given probability predictions p_i.

Testing Calibration

- Classical tests of calibration (Cox, Miller, 60s): Given $B_i \sim \text{Bernoulli}(q_i)$, $i = 1, \ldots, n$, test $q_i = p_i$, for a given probability predictions p_i.
- Cox: $\log(q_i/(1 - q_i)) = \beta_0 + \beta_1 \log(p_i/(1 - p_i))$, test $\beta_0 = 0$, $\beta_1 = 1$

Testing Calibration

- Classical tests of calibration (Cox, Miller, 60s): Given $B_i \sim \text{Bernoulli}(q_i)$, $i = 1, \ldots, n$, test $q_i = p_i$, for a given probability predictions p_i.
- Cox: $\log(q_i/(1-q_i)) = \beta_0 + \beta_1 \log(p_i/(1-p_i))$, test $\beta_0 = 0$, $\beta_1 = 1$
  - Key limitation: may not have power to detect certain forms of mis-calibration; especially for rich/overparametrized models

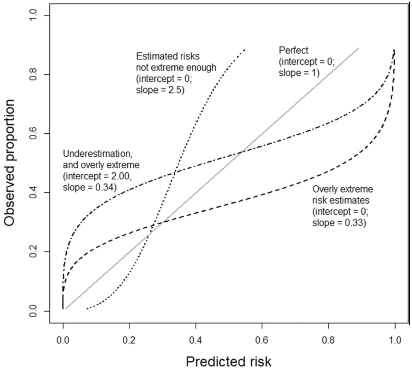


Figure: Van Calster et al, 2015

Testing Calibration

- Miller: after "suitable grouping" of $\{p_i\}$, use chi-squared test for per-group average predicted probability vector
  - Key limitation: how to choose groups?

Testing Calibration

- Miller: after "suitable grouping" of $\{p_i\}$, use chi-squared test for per-group average predicted probability vector
  - Key limitation: how to choose groups?
- Further key questions: which test statistic? optimality?

T-Cal: An optimal test of calibration

- Our contribution: T-Cal: An optimal test of calibration
  - Resolves key limitations, answer key questions
  - Readily applicable to state-of-the-art/overparametrized deep nets

T-Cal: An optimal test of calibration

- Our contribution: T-Cal: An optimal test of calibration
  - Resolves key limitations, answer key questions
  - Readily applicable to state-of-the-art/overparametrized deep nets
- How to detect general forms of mis-calibration?
  - Nonparametric testing model, chi-squared type test

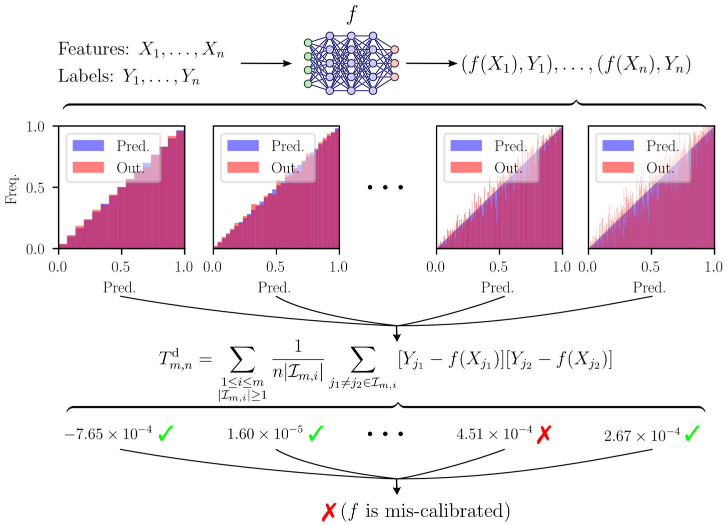
T-Cal: An optimal test of calibration

- Our contribution: T-Cal: An optimal test of calibration
  - Resolves key limitations, answer key questions
  - Readily applicable to state-of-the-art/overparametrized deep nets
- How to detect general forms of mis-calibration?
  - Nonparametric testing model, chi-squared type test
- How to choose groups?
  - Adaptive binning scheme

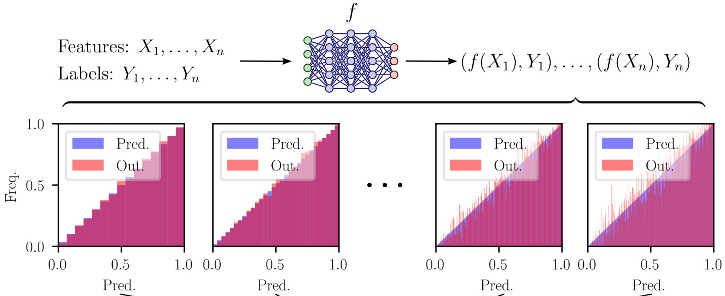
T-Cal: An optimal test of calibration

- Our contribution: T-Cal: An optimal test of calibration
  - Resolves key limitations, answer key questions
  - Readily applicable to state-of-the-art/overparametrized deep nets
- How to detect general forms of mis-calibration?
  - Nonparametric testing model, chi-squared type test
- How to choose groups?
  - Adaptive binning scheme
- Which test statistic? optimality?
  - Debiased plug-in estimator of Empirical Calibration Error (ECE)
  - Minimax optimal over Hölder smooth calibration curves

T-Cal: An optimal test of calibration



T-Cal: An optimal test of calibration



$$T^{\mathrm{d}}_{m,n} = \sum_{\substack{1 \le i \le m \\ |\mathcal{I}_{m,i}| \ge 1}} \frac{1}{n|\mathcal{I}_{m,i}|} \sum_{j_1 \neq j_2 \in \mathcal{I}_{m,i}} [Y_{j_1} - f(X_{j_1})][Y_{j_2} - f(X_{j_2})]$$

-5.95×10^{-5} ✓ -1.48×10^{-4} ✓ $\cdots$ -2.33×10^{-4} ✓ -4.36×10^{-4} ✓

✓ (f is perfectly calibrated)

Overview

Regression/Residual Function

- Recall: A classifier (probability predictor) $f : \mathbb{R}^d \to \Delta_{K-1}$ is calibrated if

$$\mathbb{E}[Y \mid f(X) = \mathrm{z}] = z \quad \text{for any } z \in \Delta_{K-1}.$$

Regression/Residual Function

- Recall: A classifier (probability predictor) $f : \mathbb{R}^d \to \Delta_{K-1}$ is calibrated if

$$\mathbb{E}[Y \mid f(X) = \mathrm{z}] = z \quad \text{for any } z \in \Delta_{K-1}.$$

- The calibration curve (*regression function*) $\mathrm{reg}_f : \Delta_{K-1} \to \Delta_{K-1}$ is

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

Regression/Residual Function

- Recall: A classifier (probability predictor) $f : \mathbb{R}^d \to \Delta_{K-1}$ is calibrated if

$$\mathbb{E}[Y \mid f(X) = \mathrm{z}] = z \quad \text{for any } z \in \Delta_{K-1}.$$

- The calibration curve (*regression function*) $\mathrm{reg}_f : \Delta_{K-1} \to \Delta_{K-1}$ is

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

- The classifier f is calibrated if

$$\mathrm{reg}_f(Z) = Z,$$

almost surely w.r.t. the law of $Z = f(X)$.

Expected Calibration Error

- For any $p \geq 1$, the ℓ_p-ECE is

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

- $\ell_p\text{-ECE}_P(f) = 0$ iff f is calibrated under P

Hypothesis Testing Setup

- Consider distributions over $(f(X), Y) = (Z, Y) \in \Delta_{K-1} \times \mathcal{Y}$

[1] for a Hölder smoothness parameter s and a Hölder constant L

Hypothesis Testing Setup

- Consider distributions over $(f(X), Y) = (Z, Y) \in \Delta_{K-1} \times \mathcal{Y}$
- Goal: Test the *null hypothesis* of calibration against the *alternative* of miscalibration:

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

[1] for a Hölder smoothness parameter s and a Hölder constant L

Hypothesis Testing Setup

- Consider distributions over $(f(X), Y) = (Z, Y) \in \Delta_{K-1} \times \mathcal{Y}$
- Goal: Test the *null hypothesis* of calibration against the *alternative* of miscalibration:

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

- Nearly calibrated data distributions

$$\mathcal{P}_0(\varepsilon_n) := \{P : \ell_p\text{-ECE}_P(f) \leqslant c\varepsilon_n,\ P_Z\text{-a.s.}\}.$$

[1] for a Hölder smoothness parameter s and a Hölder constant L

Hypothesis Testing Setup

- Consider distributions over $(f(X), Y) = (Z, Y) \in \Delta_{K-1} \times \mathcal{Y}$
- Goal: Test the *null hypothesis* of calibration against the *alternative* of miscalibration:

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

- Nearly calibrated data distributions

$$\mathcal{P}_0(\varepsilon_n) := \{P : \ell_p\text{-ECE}_P(f) \leqslant c\varepsilon_n, \; P_Z\text{-a.s.}\}.$$

- Distributions P of (Z, Y) under which the ℓ_p-ECE of f is at least ε_n:

$$\mathcal{P}_1(\varepsilon_n, s) := \{P \in \mathcal{P}_{s,L,K} : \ell_p\text{-ECE}_P(f) \geq C\varepsilon_n\}.$$

$\mathcal{P}_{s,L,K}$: $\mathrm{z} \mapsto [z - \mathrm{reg}_{f,P}(\mathrm{z})]_k$ is (s, L)-Hölder continuous[1] for every $k \in \{1, \dots, K\}$.

[1] for a Hölder smoothness parameter s and a Hölder constant L

Hypothesis Testing Setup

- Consider distributions over $(f(X), Y) = (Z, Y) \in \Delta_{K-1} \times \mathcal{Y}$
- Goal: Test the *null hypothesis* of calibration against the *alternative* of miscalibration:

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

- Nearly calibrated data distributions

$$\mathcal{P}_0(\varepsilon_n) := \{P : \ell_p\text{-ECE}_P(f) \leqslant c\varepsilon_n,\ P_Z\text{-a.s.}\}.$$

- Distributions P of (Z, Y) under which the ℓ_p-ECE of f is at least ε_n:

$$\mathcal{P}_1(\varepsilon_n, s) := \{P \in \mathcal{P}_{s,L,K} : \ell_p\text{-ECE}_P(f) \geq C\varepsilon_n\}.$$

  $\mathcal{P}_{s,L,K}$: $\mathrm{z} \mapsto [z - \mathrm{reg}_{f,P}(\mathrm{z})]_k$ is (s, L)-Hölder continuous[1] for every $k \in \{1, \dots, K\}$.
- Open problem: Can we improve if f has more structure?

[1] for a Hölder smoothness parameter s and a Hölder constant L

Reduction to two-sample testing: Intuition

- If a classifier is calibrated, then its probability predictions match the true class probabilities.

Reduction to two-sample testing: Intuition

- If a classifier is calibrated, then its probability predictions match the true class probabilities.
- Randomly sampling new labels according to the probability predictions yields a sample from the true distribution.

Reduction to two-sample testing: Intuition

- If a classifier is calibrated, then its probability predictions match the true class probabilities.
- Randomly sampling new labels according to the probability predictions yields a sample from the true distribution.
- After sample splitting, we can use classical two-sample tests to check if the two samples are from the same distribution.

Experiment

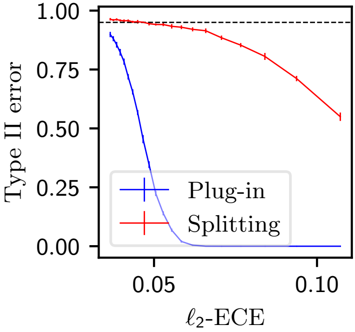


Figure: $s = 0.6$, $\rho = 100$

- Due to sample splitting, effective sample size is smaller than that of T-Cal.

Overview

Plug-in Estimator

- Recall

$$\ell_2\text{-ECE}(f)^2 = \mathbb{E}_{Z\sim P_Z}\left[\|\text{reg}_f(Z) - Z]\|_2^2\right]$$

- Given a partition $\mathcal{B}_m = \{B_1, \ldots, B_{m^{K-1}}\}$ of Δ_{K-1}, with

$$\mathcal{I}_i := \{j \in \{1, \ldots, N\} : Z_j \in B_i\},$$

the plug-in estimator for $\ell_2\text{-ECE}(f)^2$ by piecewise averaging is defined as

$$T^{\text{b}}_{m,n} := \sum_{\substack{i\in[m^{K-1}] \\ |\mathcal{I}_i|\geq 1}} \frac{|\mathcal{I}_i|}{n} \left\| \frac{1}{|\mathcal{I}_i|} \sum_{j\in\mathcal{I}_i} (Y_j - Z_j) \right\|^2 . \tag{1}$$

Bias of the Plug-in Estimator

- Consider $K = 2$, $Z \sim P_Z = \text{Unif}[0, 1]$, $P_0 : P_Z \times \text{Ber}(Z)$ and $P_1 : P_Z \times \text{Ber}(\text{reg}_f(Z))$ depicted below (left).

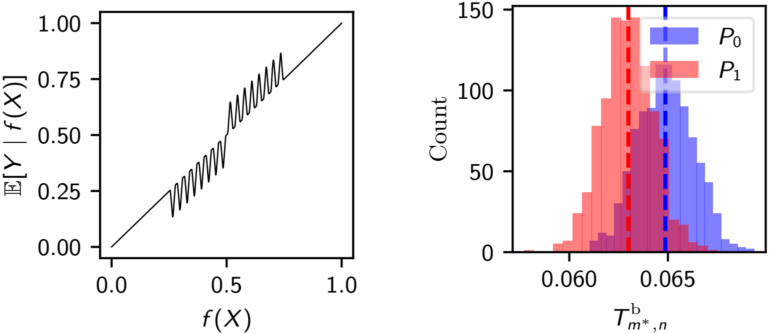


Figure: **Left:** A graph of the calibration curve $z \mapsto \text{reg}_f(z)$ under P_1. **Right:** Histograms of $T^{\mathrm{b}}_{m,n}$ and $T^{\mathrm{d}}_{m,n}$ under P_0 and P_1 are obtained from $1,000$ independent observations.

Debiasing the Plug-in Estimator

- The plug-in estimator is biased, because we are estimating both $\mathbb{E}[Y \mid Z \in B_i]$ and $\mathbb{E}[Z \mid Z \in B_i]$ using the same sample $(Z_i, Y_i), i \in \{1, \ldots, n\}$.

Debiasing the Plug-in Estimator

- The plug-in estimator is biased, because we are estimating both $\mathbb{E}[Y \mid Z \in B_i]$ and $\mathbb{E}[Z \mid Z \in B_i]$ using the same sample $(Z_i, Y_i), i \in \{1, \dots, n\}$.
- We define the *Debiased Plug-in Estimator* (DPE):

$$T^{\mathrm{d}}_{m,n} = \sum_{\substack{i \in [m^{K-1}] \\ |\mathcal{I}_i| \geq 1}} \frac{1}{n|\mathcal{I}_i|} \left[\left\| \sum_{j \in \mathcal{I}_i} (Y_j - Z_j) \right\|^2 - \sum_{j \in \mathcal{I}_i} \|Y_j - Z_j\|^2 \right].$$

Debiasing the Plug-in Estimator

- The plug-in estimator is biased, because we are estimating both $\mathbb{E}[Y \mid Z \in B_i]$ and $\mathbb{E}[Z \mid Z \in B_i]$ using the same sample $(Z_i, Y_i), i \in \{1, \ldots, n\}$.
- We define the *Debiased Plug-in Estimator* (DPE):

$$T^{\mathrm{d}}_{m,n} = \sum_{\substack{i \in [m^{K-1}] \\ |\mathcal{I}_i| \geq 1}} \frac{1}{n|\mathcal{I}_i|} \left[\left\| \sum_{j \in \mathcal{I}_i} (Y_j - Z_j) \right\|^2 - \sum_{j \in \mathcal{I}_i} \| Y_j - Z_j \|^2 \right].$$

- The mean of $T^{\mathrm{d}}_{m,n}$ is not exactly $\ell_2\text{-ECE}(f)^2$ under $P \in \mathcal{P}_1(\varepsilon, s)$, but debiasing makes it comparable to $\ell_2\text{-ECE}(f)^2$.
- Open problem: Can we also use it to improve calibration during training?

T-Cal: Debiased Plug-in Test

- We use $T^{\mathrm{d}}_{m,n}$ as our test statistic.

$$\xi_{m,n}(\alpha) = \xi_{m,n} := \begin{cases} I\left(T^{\mathrm{d}}_{m,n} \geq \sqrt{\frac{2K}{\alpha}} m^{\frac{K-1}{2}} n^{-1}\right) & \text{if } m^{K-1} \leq n, \\ I\left(T^{\mathrm{d}}_{m,n} \geq \sqrt{\frac{2K}{\alpha}} m^{-\frac{K-1}{2}}\right) & \text{if } m^{K-1} > n. \end{cases}$$

T-Cal: Debiased Plug-in Test

- We use $T^{\mathrm{d}}_{m,n}$ as our test statistic.

$$\xi_{m,n}(\alpha) = \xi_{m,n} := \begin{cases} I\left(T^{\mathrm{d}}_{m,n} \geq \sqrt{\frac{2K}{\alpha}} m^{\frac{K-1}{2}} n^{-1}\right) & \text{if } m^{K-1} \leq n, \\ I\left(T^{\mathrm{d}}_{m,n} \geq \sqrt{\frac{2K}{\alpha}} m^{-\frac{K-1}{2}}\right) & \text{if } m^{K-1} > n. \end{cases}$$

- One can choose critical values by bootstrapping (or, consistency resampling) in practice.

Main Theorem I

Theorem (T-Cal: Calibration test via debiased plug-in estimation)

Suppose $p \leq 2$. *For* $m^* = \lfloor n^{2/(4s+K-1)} \rfloor$,

$$\varepsilon_n = n^{-\frac{2s}{4s+K-1}},$$

we have

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

Main Theorem I

Theorem (T-Cal: Calibration test via debiased plug-in estimation)

Suppose $p \leq 2$. For $m^* = \lfloor n^{2/(4s+K-1)} \rfloor$,

$$\varepsilon_n = n^{-\frac{2s}{4s+K-1}},$$

we have

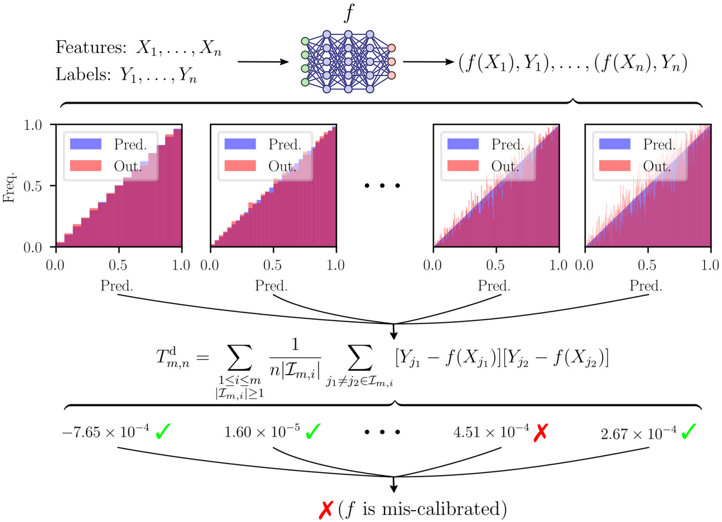
1. **False detection rate control.** *For every P for which f is calibrated, i.e., for $P \in \mathcal{P}_0(\varepsilon_n)$, the probability of falsely claiming mis-calibration is at most α, i.e., $P(\xi_{m^*,n} = 1) \leq \alpha$.*
2. **True detection rate control.** *There exists $c > 0$ depending on (s, L, K, α, β) such that the power (true positive rate) is bounded as $P(\xi_{m^*,n} = 1) \geq 1 - \beta$ for every $P \in \mathcal{P}_1(\varepsilon_n, s)$—i.e., when f is mis-calibrated with an ℓ_p-ECE of at least $C\varepsilon_n$.*

Combined with lower bounds we show, T-Cal is minimax optimal over Hölder smooth calibration curves

Adaptive T-Cal

For a number $B = \lceil \frac{2}{K-1} \log_2(n/\sqrt{\log n}) \rceil$ of tests performed, let

$$\xi_n^{\text{ad}} := \max_{b \in \{1,\ldots,B\}} \xi_{2^b,n}\left(\frac{\alpha}{B}\right).$$



Main Theorem II: Adaptive T-Cal

Theorem (Adaptive T-Cal)

Suppose $p \leq 2$. For

$$\varepsilon_n = (n/\sqrt{\log n})^{-\frac{2s}{4s+K-1}},$$

the adaptive test ξ_n^{ad} enjoys

1. **False detection rate control.** *For every P for which f is calibrated, i.e., for $P \in \mathcal{P}_0(\varepsilon_n)$, the probability of falsely claiming mis-calibration is at most α, i.e., $P\left(\xi_n^{\text{ad}} = 1\right) \leq \alpha$.*

Main Theorem II: Adaptive T-Cal

Theorem (Adaptive T-Cal)

Suppose $p \leq 2$. For

$$\varepsilon_n = (n/\sqrt{\log n})^{-\frac{2s}{4s+K-1}},$$

the adaptive test ξ_n^{ad} enjoys

1. **False detection rate control.** *For every P for which f is calibrated, i.e., for $P \in \mathcal{P}_0(\varepsilon_n)$, the probability of falsely claiming mis-calibration is at most α, i.e., $P\left(\xi_n^{\mathrm{ad}} = 1\right) \leq \alpha$.*
2. **True detection rate control.** *The power (true positive rate) is lower bounded as $P(\xi_n^{\mathrm{ad}} = 1) \geq 1 - \beta$ for every $P \in \mathcal{P}_1(\varepsilon_n, s)$—i.e., when f is mis-calibrated with an ℓ_p-ECE of at least $C'\varepsilon_n$.*

Overview

Synthetic data

- We compare T-Cal with classical calibration tests [2, 9, 4, 10]:
  - Cox's Logistic Score test
  - test based on plug-in $\widehat{\ell_1\text{-ECE}}$

Synthetic data

- We compare T-Cal with classical calibration tests [2, 9, 4, 10]:
  - Cox's Logistic Score test
  - test based on plug-in $\widehat{\ell_1\text{-ECE}}$
- $H_1 : (Z, Y) \sim P_{1,m}$ – reasonable if model class is sufficiently rich/over-parametrized

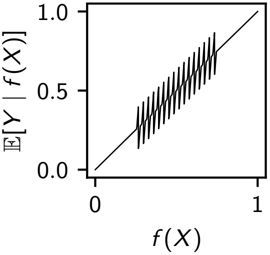


Figure: Calibration curve under $P_{1,m}$

Synthetic data results: T-Cal vs other tests

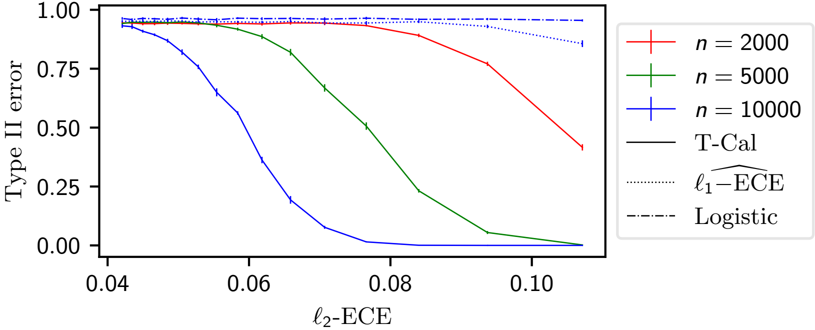


Figure: Comparison of calibration tests: T-Cal (with m^*) is more sample-efficient than other methods (with $n = 10,000$)

Synthetic data ablation: m^*, debiasing, ℓ_2 vs ℓ_1

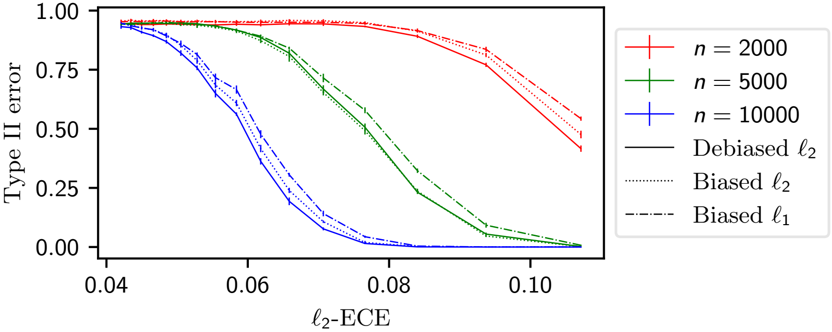


Figure: Type II error comparison for $T^{\mathrm{d}}_{m^*,n}$ (T-Cal), $T^{\mathrm{b}}_{m^*,n}$, and $T^{\ell_1}_{m^*,n}$. Using ℓ_2 is better than ℓ_1, and debiased ℓ_2 (T-Cal) is better than biased ℓ_2. Standard error bars are plotted over 10 repetitions. m^* has largest effect.

Empirical data

- We test adaptive T-Cal on various deep neural networks trained on CIFAR-10/100 and ImageNet.

Empirical data

- We test adaptive T-Cal on various deep neural networks trained on CIFAR-10/100 and ImageNet.
- We test models calibrated by standard post-hoc methods (and uncalibrated ones).

Results on empirical data: CIFAR-10

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

Table 1: The values of the empirical ℓ_1-ECE (Guo et al., 2017) and the testing results, via adaptive T-Cal and multiple binomial testing, of models trained on CIFAR-10.

Figure: Results roughly align with the magnitude of the empirical ECE.

Results on empirical data: CIFAR-100

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

Table 2: The values of the empirical ℓ_1-ECE (Guo et al., 2017) and the testing results, via adaptive T-Cal and multiple binomial testing, of models trained on CIFAR-100.

Figure: Results roughly align with the magnitude of the empirical ECE. However, T-Cal *not* the same as ℓ_1-ECE: see ResNet-56.

Results on empirical data: CIFAR-10, reliability diagrams

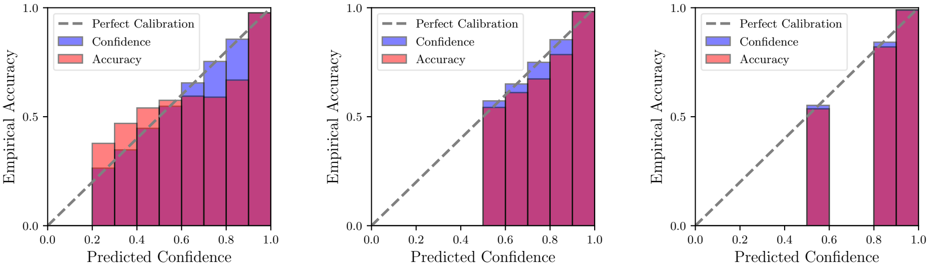


Figure: The reliability diagrams for VGG-19, trained on CIFAR-10, calibrated by Platt scaling (left - reject), polynomial scaling (middle - accept), and histogram binning (right - accept). Bins containing less than 10 data points, where the sample noise dominates, are omitted for clarity.

Results on empirical data: ImageNet

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

Table 3: The values of the empirical ℓ_1-ECE (Guo et al., 2017) and the testing results, via adaptive T-Cal and multiple binomial testing, of models trained on ImageNet.

Overview

Impossibility for continuous mis-calibration curves

- If the mis-calibration curve can oscillate with arbitrarily high frequency, mis-calibration cannot be detected from a finite sample. (Caution when using complex/too overparametrized models!)

Impossibility for continuous mis-calibration curves

- If the mis-calibration curve can oscillate with arbitrarily high frequency, mis-calibration cannot be detected from a finite sample. (Caution when using complex/too overparametrized models!)
- Define minimax type II error for distributions in the alternative with *continuous* mis-calibration curves

$$R_n^{\text{cont}}(\varepsilon) := \inf_{\xi \in \Phi_n(\alpha)} \sup_{P \in \mathcal{P}_1^{\text{cont}}(\varepsilon)} P(\xi = 0).$$

Proposition

Let $\varepsilon_0 = 0.1$. For any level $\alpha \in (0, 1)$, the minimax type II error $R_n^{\text{cont}}(\varepsilon_0)$ for testing the null hypothesis of calibration at level α against the hypothesis $P \in \mathcal{P}_1^{\text{cont}}(\varepsilon_0)$ of continuous mis-calibration curves satisfies

$$R_n^{\text{cont}}(\varepsilon_0) \geq 1 - \alpha$$

for all n.

Hölder continuous calibration curves

- We consider detecting mis-calibration when the mis-calibration curves are Hölder continuous; as usual in nonparametric statistics [5, 8, 3, 6].

Hölder continuous calibration curves

- We consider detecting mis-calibration when the mis-calibration curves are Hölder continuous; as usual in nonparametric statistics [5, 8, 3, 6].
- Rich class of mis-calibration curves, including non-smooth ones.

Lower bound for Hölder continuous mis-calibration curve

- Test the calibration of the K-class probability predictor f assuming (s, L)-Hölder continuity of mis-calibration curves at a level $\alpha \in (0,1)$.
- Minimax type II error

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

- Minimum separation needed for a minimax type II error of at most β

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

Lower bound for Hölder continuous mis-calibration curve

- Test the calibration of the K-class probability predictor f assuming (s, L)-Hölder continuity of mis-calibration curves at a level $\alpha \in (0, 1)$.
- Minimax type II error

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

- Minimum separation needed for a minimax type II error of at most β

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

Theorem

There exists $c_{\text{lower}} > 0$ depending only on $(p, s, L, K, \alpha, \beta)$ such that, for any $p > 0$, the minimum ℓ_p-ECE of f, i.e. $\varepsilon_n^*(s)$, required to have a test with a false positive rate (type I error) at most α and with a true positive rate (power) at least $1 - \beta$ satisfies

$$\varepsilon_n^*(s) \geq c_{\text{lower}} n^{-\frac{2s}{4s+K-1}}$$

for all n.

Context, continued

- Nonparametric rate $n^{-\frac{2s}{4s+K-1}}$ slower than parametric $n^{-1/2}$ rate. Slower as the number of classes K grows.

Context, continued

- Nonparametric rate $n^{-\frac{2s}{4s+K-1}}$ slower than parametric $n^{-1/2}$ rate. Slower as the number of classes K grows.
- What one expects based on nonparametric two-sample goodness-of-fit testing for densities on Δ_{K-1} [6].

Context, continued

- Nonparametric rate $n^{-\frac{2s}{4s+K-1}}$ slower than parametric $n^{-1/2}$ rate. Slower as the number of classes K grows.
- What one expects based on nonparametric two-sample goodness-of-fit testing for densities on Δ_{K-1} [6].
- *T-Cal is minimax optimal*.

Context, continued

- Nonparametric rate $n^{-\frac{2s}{4s+K-1}}$ slower than parametric $n^{-1/2}$ rate. Slower as the number of classes K grows.
- What one expects based on nonparametric two-sample goodness-of-fit testing for densities on Δ_{K-1} [6].
- *T-Cal is minimax optimal*.
- Evaluating multi-class model calibration on a small dataset and a complex model can be challenging.

Summary and takeaways

- Uncertainty quantification important for ML; even more with overparametrization. Calibration is a key problem with rich history.

Summary and takeaways

- Uncertainty quantification important for ML; even more with overparametrization. Calibration is a key problem with rich history.
  1. **The need for statistical significance to claim calibration.**

Summary and takeaways

- Uncertainty quantification important for ML; even more with overparametrization. Calibration is a key problem with rich history.
  1. **The need for statistical significance to claim calibration.**
  2. **Potential suboptimality of popular approaches.**

Summary and takeaways

- Uncertainty quantification important for ML; even more with overparametrization. Calibration is a key problem with rich history.
  1. **The need for statistical significance to claim calibration.**
  2. **Potential suboptimality of popular approaches.**
- T-Cal: adaptive test for calibration of ML models; supported by empirical & theoretical results.
  - Available at `https://github.com/dh7401/Calibration-Test`

Reduction

- For $i \in \{\lfloor n/2 \rfloor + 1, \ldots, n\}$, we generate random variables $\tilde{Y}_i \sim \mathrm{Cat}(Z_i)$.

Reduction

- For $i \in \{\lfloor n/2 \rfloor + 1, \ldots, n\}$, we generate random variables $\tilde{Y}_i \sim \mathrm{Cat}(Z_i)$.
- For each $k \in \{1, \ldots, K\}$, define

$$\mathcal{V}_k := \left\{ Z_i : [Y_i]_k = 1, 1 \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor \right\},$$

$$\mathcal{W}_k := \left\{ Z_i : [\tilde{Y}_i]_k = 1, \left\lfloor \frac{n}{2} \right\rfloor + 1 \leq i \leq n \right\}.$$

Reduction

- For $i \in \{\lfloor n/2 \rfloor + 1, \ldots, n\}$, we generate random variables $\tilde{Y}_i \sim \mathrm{Cat}(Z_i)$.
- For each $k \in \{1, \ldots, K\}$, define

$$\mathcal{V}_k := \left\{ Z_i : [Y_i]_k = 1, 1 \leq i \leq \left\lfloor \frac{n}{2} \right\rfloor \right\},$$

$$\mathcal{W}_k := \left\{ Z_i : [\tilde{Y}_i]_k = 1, \left\lfloor \frac{n}{2} \right\rfloor + 1 \leq i \leq n \right\}.$$

- $\mathcal{V}_k$ and $\mathcal{W}_k$ have densities

$$\pi_k^{\mathcal{V}}(\mathrm{z}) := \frac{[\mathrm{reg}_f(\mathrm{z})]_k}{\int_{\Delta_{K-1}} [\mathrm{reg}_f(\mathrm{z})]_k dP_Z(\mathrm{z})} = \frac{[\mathrm{reg}_f(\mathrm{z})]_k}{\mathbb{E}[Y]_k},$$

$$\pi_k^{\mathcal{W}}(\mathrm{z}) := \frac{[\mathrm{z}]_k}{\int_{\Delta_{K-1}} [\mathrm{z}]_k dP_Z(\mathrm{z})} = \frac{[\mathrm{z}]_k}{\mathbb{E}[Z]_k}$$

with respect to P_Z.

Reduction detail

- Let $\texttt{TS}_{\alpha,\beta}$ be an optimal two-sample goodness-of-fit test (e.g., due to Ingster, Arias-Castro et al., Kim et al., [5, 1, 7]).

Reduction detail

- Let $\mathtt{TS}_{\alpha,\beta}$ be an optimal two-sample goodness-of-fit test (e.g., due to Ingster, Arias-Castro et al., Kim et al., [5, 1, 7]).
- For $k \in \{1, \dots, K\}$,

$$
\begin{aligned}
T_{1,k} &= \frac{1}{n}\sum_{i=1}^{n}[Y_i - Z_i]_k, \\
T_{2,k} &= \frac{1}{n}\sum_{i=1}^{n}[Z_i]_k[Y_i - Z_i]_k, \\
b_k &= I\left(|T_{1,k}| \geq \sqrt{\frac{3K}{\alpha n}}\right) \vee I\left(|T_{2,k}| \geq \sqrt{\frac{3K}{\alpha n}}\right) \vee \mathtt{TS}_{\frac{\alpha}{3K},\frac{\beta}{2}}(\mathcal{V}_k, \mathcal{W}_k).
\end{aligned}
$$

Reduction detail

- Let $\texttt{TS}_{\alpha,\beta}$ be an optimal two-sample goodness-of-fit test (e.g., due to Ingster, Arias-Castro et al., Kim et al., [5, 1, 7]).
- For $k \in \{1, \dots, K\}$,

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

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

$$b_k = I\left(|T_{1,k}| \geq \sqrt{\frac{3K}{\alpha n}}\right) \vee I\left(|T_{2,k}| \geq \sqrt{\frac{3K}{\alpha n}}\right) \vee \texttt{TS}_{\frac{\alpha}{3K},\frac{\beta}{2}}(\mathcal{V}_k, \mathcal{W}_k).$$

- Reject H_0 if $\max\{b_1, \dots, b_K\} = 1$.

Main Result: Known smoothness

Theorem (Optimal calibration test via sample splitting)

Suppose $p \leq 2$ and let ξ_n^{split} be the test described in the previous slide. Assume the Hölder smoothness parameter s is known. We have

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

Main Result: Known smoothness

Theorem (Optimal calibration test via sample splitting)

Suppose $p \leq 2$ and let ξ_n^{split} be the test described in the previous slide. Assume the Hölder smoothness parameter s is known. We have

1. **False detection rate control.** *For every P for which f is calibrated, i.e., for $P \in \mathcal{P}_0(\varepsilon_n)$, the probability of falsely claiming mis-calibration is at most α, i.e., $P(\xi_n^{\text{split}} = 1) \leq \alpha$.*
2. **True detection rate control.** *There exists $c_{\text{split}} > 0$ depending on $(s, L, K, d_c, \alpha, \beta)$ such that the power (true positive rate) is bounded as $P(\xi_n^{\text{split}} = 1) \geq 1 - \beta$ for every $P \in \mathcal{P}_1(\varepsilon, s)$—i.e., when f is mis-calibrated with an ℓ_p-ECE of at least*

$$\varepsilon \geq c_{\text{split}} n^{-\frac{2s}{4s+K-1}}.$$

Main Result: Adapting to smoothness

- Consider an adaptive two-sample goodness-of-fit test $\mathtt{TS}^{\mathrm{ad}}_{\alpha,\beta}$.

Corollary (Adaptive test via sample splitting)

Suppose $p \leq 2$ and let $\xi_n^{\text{ad-s}}$ be the test described abvoe with $\mathtt{TS}$ replaced by an adaptive two-sample test $\mathtt{TS}^{\mathrm{ad}}$. We have

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

Main Result: Adapting to smoothness

- Consider an adaptive two-sample goodness-of-fit test $\mathtt{TS}^{\text{ad}}_{\alpha,\beta}$.

Corollary (Adaptive test via sample splitting)

Suppose $p \le 2$ and let $\xi_n^{\text{ad-s}}$ be the test described abvoe with $\mathtt{TS}$ *replaced by an adaptive two-sample test* $\mathtt{TS}^{\text{ad}}$. *We have*

1. **False detection rate control.** *For every P for which f is calibrated, i.e., for $P \in \mathcal{P}_0(\varepsilon_n)$, the probability of falsely claiming mis-calibration is at most α, i.e., $P(\xi_n^{\text{ad-s}} = 1) \le \alpha$.*
2. **True detection rate control.** *There exists $c_{\text{ad-s}} > 0$ depending on $(s, L, K, d_c, \alpha, \beta)$ such that the power (true positive rate) is bounded as $P(\xi_n^{\text{ad-s}} = 1) \ge 1 - \beta$ for every $P \in \mathcal{P}_1(\varepsilon, s)$—i.e., when f is mis-calibrated with an ℓ_p-ECE of at least*

$$\varepsilon \ge c_{\text{ad-s}}(n/\log\log n)^{-\frac{2s}{4s+K-1}}.$$

References I

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

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

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk.
A distribution-free theory of nonparametric regression, volume 1.
Springer, 2002.

Frank E Harrell.
Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis, volume 3.
Springer, 2015.

References II

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

Yuri Ingster and Irina A Suslina.
Nonparametric goodness-of-fit testing under Gaussian models, volume 169.
Springer Science & Business Media, 2012.

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

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

References III

Robert G Miller.
Statistical prediction by discriminant analysis.
In *Statistical Prediction by Discriminant Analysis*, pages 1–54. Springer, 1962.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön.
Evaluating model calibration in classification.
In *International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.